University of Reading

School of Psychology and Clinical Language Sciences

Multisensory Integration: Does Haptics Improve Tumour Delineation?

Thesis submitted for the degree of Doctor of Philosophy

Julie Bauge Skevik

December 2020

Declaration of Authorship

Declaration: I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Julie Bauge Skevik

Acknowledgements

Firstly, I would like to thank my husband, Connor, who has been by my side through thick and thin, and in general has gone above and beyond to be there for me throughout it all. I could not have done this without you – one day I'll even read *your* thesis.

I also want to say a thousand thank-you's to my family, especially to Pia, Patrick and Amy, for being incredibly patient with me through this ridiculous journey and when I vanish off the face of the earth for months at a time, only to surface for birthdays and holidays, disappearing just as quickly as I appeared. Tusen takk for alt, elsker dere høyere enn himmelen.

I am also fortunate enough to have found a loving family in my in-laws, Byrion, Philip and Gina, who welcomed me with open arms from the start, and have been nothing but sweet and welcoming to me – I deeply appreciate the support through it all.

I also extend my deepest appreciation to my friends, especially to Mike, Lui, and Mark for patiently standing by me through everything from enthusiastic rants to dumb questions, and keeping me marginally sane through the years.

I would also like to thank my supervisors Peter Scarfe and William Harwin, who have helped, guided and supported me along this path, with all its twists and turns. I also want to thank Andrew Glennester for all his help and support over the years.

Lastly, I would like to extend my thanks to the Engineering and Physical Sciences Research Council for funding the studentship – it has been an amazing opportunity.

In memory of my late mother, Randi Bauge Skevik (1952 – 2015).

She would have loved to read this.

Abstract

The ability to use touch in addition to vision when searching for anomalies and differences in texture is well known to be beneficial to human perception in general. The aim of this thesis is to evaluate the potential benefit of using a haptic signal in conjunction with visual images to improve detection and delineation of tumours in medical imaging data. One of the key issues with tumour delineation in the field today is the interclinician variance in delineating tumours for diagnostics and treatment, where even clinicians who have similar sensitivity and precision levels tend to delineate widely different underlying shapes. Through three experiments we investigate whether the ability to touch a medical image improves tumour delineation. In the first experiment, we show that combined visuohaptic cues significantly improves performance for signal detection of a 2D Gaussian embedded in a noisy background. In the second experiment, we found that the relative dissimilarity of different images per modality did not systematically decrease precision in a two-alternative forced choice (2AFC) slant discrimination task, in a spatially coaligned visuohaptic rig. In the third and final experiment we successfully found that observers are significantly better at delineating generated 'tumours' in synthetic 'medical images' when the haptic representation of the image is present compared to drawing on a flat surface, in a spatially coaligned visuohaptic rig.

Preface

"If you can put your five fingers through it, it is a gate, if not a door. Shut your eyes and see."

— James Joyce, *Ulysses*

In this thesis we will be exploring the effect haptic cues have on performance through a series of experiments, covering the topics of simple detection, slant discrimination and ultimately tumour delineation. The first chapter covers the theoretical foundations and the current state of the field, while the second chapter goes into detail about the hardware and software used, and the measures taken to ensure a precise calibration. Chapter 3 is the first experimental chapter, and covers the series of experiments and controls run on visuohaptic detection of 2D Gaussian bumps embedded in varying levels of visual noise. Chapter 4 explores the potential influence that using a different visual to haptic texture might have on performance in a slant-detection task, while Chapter 5 covers the potential benefits of haptic height-mapping on a tumour delineation task, using simulated medical images. An overall discussion and summary is located in Chapter 6, with contributions and implications discussed in Chapter 7.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis will look at addressing the issue of interclinician variance in tumour delineation through the use of combined vision and touch, exploring the potential benefit of adding a haptic signal in the delineation process of tumours in medical imaging data, approaching this question as a detection task from the perspective of perceptual psychology. In order to investigate this, a series of visuohaptic experiments were run with the aim of establishing whether a haptic signal aids in basic signal detection, whether an incongruent haptic signal will detract from task performance in a slant discrimination task, and finally whether adding haptic height-mapping of a simulated medical image improves accuracy compared to outlining on a haptically 'flat' simulated medical image. As this project is rooted in haptics, engineering, psychology and clinical medicine, we aim to bring together concepts and methods from each discipline to aid in resolving the question of whether being able to explore medical images haptically as well as visually improves precision in detecting, and accuracy in delineation of tumours. In this chapter I will start with focussing on the current clinical approach for delineation,

before moving onto the topics of perception, signal detection and psychophysics.

1.1 Detection and delineation of tumours in medical imaging data

A crucial part of a successful cancer treatment is to accurately delineate the tumour so as to get a precise picture of its shape. This gives the radiation oncologist or surgical oncologist responsible for determining a treatment the most informative basis for doing so. The initial task of inspecting a medical image, detecting the presence of a possible tumour, and drawing the boundary between cancerous and healthy tissue is typically done by a clinical radiologist. As cancerous tissue grows out of healthy, regular tissue – such as sarcomas, or soft tissue tumours, from connective tissue – this is not an easy task. The ability to notice cancerous tissue requires a strong grasp of anatomy and familiarity with how healthy tissue in the affected area normally presents. If the delineation is too cautious and broad then too much healthy tissue may be negatively affected and recovery may be halted; while if the delineation is too aggressive and attenuated then parts of the tumour may be left behind, which would greatly increase the likelihood of cancer recurrence (Nelms et al., 2012).

Detection in and of itself is a very common task performed regularly, both consciously and subconsciously. For example, detecting the face of a relative in a crowd or noticing a friend on a train. Detection of a tumour is, however, a more constrained task. The boundaries between healthy and cancerous tissue are often ill-defined, as the origin of a growing tumour is in cells originating from within the human body itself. The task of delineating a tumour can be broken

into the aspects of detecting an anomaly in the tissue, and differentiating the size and shape of the anomaly compared to the surrounding healthy tissue, such as soft tissue tumours contrasted with the surrounding soft-tissue structures such as fatty tissues, blood vessels and lymph nodes.

1.1.1 The current clinical approach

A key aspect of tumour detection in radiology is the ability to detect anomalies in medical imaging data. Trained radiologists will visually inspect the data and make decisions based on locating potentially cancerous tissue embedded in healthy human tissues (Castella et al., 2009; Kompaniez-Dunigan et al., 2015). This is done by going over the medical imaging slice-by-slice and drawing the outlines with either a mouse or, more recently, on a tablet using a pen-shaped stylus. Locating abnormal tissue is effectively a signal detection task, where the trained radiologists attempts to locate a signal, the abnormal tissue, obscured by noise, the surrounding healthy tissue. As human biology is a complex, non-uniform body of varying densities and individual differences, this is a very difficult visual task, with a high level of inter-clinician variability (Elmore et al., 2009; Jager et al., 2016; Njeh, 2008). This delineation task can be viewed as a standard 2-alternative forced choice (2AFC) task, which has four possible outcomes as shown in Table 1.1. The two correct responses, shown in Table 1.1 as green, are correctly outlined tumour ('Hit') and correctly excluded healthy tissue ('Correct rejection'). The two incorrect responses, shown as red, are 'under-delineation' ('Miss'), which misses parts of the tumour, and 'over-delineation' ('False positive'), which is to erroneously include healthy tissue as part of the estimated tumour body. With more practice, the number of incorrect responses (Miss & False-positive) reduces,

which in turn increases the accuracy of the correct outcomes (Hit & Reject). The negative consequences of the incorrect responses are not weighted equally, as missing out on tumourous tissue is in most cases considered to have a more negative impact on treatment and recovery than the inclusion of additional healthy tissue (Nelms et al., 2012).

		Tumour	
		Present	Absent
Response	Yes	*Hit*	*False positive*
	No	*Miss*	*Correct rejection*

Table 1.1: The four alternatives of a 2AFC task with Tumour as the signal. 1) 'Hit', the response is yes when the tumour is present ('True positive'), 2) 'Miss', the response is no when the tumour is present ('True negative'), 3) 'False-positive', the response is yes when the tumour is absent, 4) 'Correct rejection', the response is no when the tumour is absent ('False negative').

In order to look at the details of detection sensitivity in visual tasks, experiments typically use stimuli that are less realistic, opting instead to use simulated stimuli, which allow for precise manipulations of both signal and noise (Abbey, 2013; Castella et al., 2009).

1.1.2 Key Problem to be solved: Variance across clinicians

Currently, the issue of variability between clinicians' delineations is considered to be one of the biggest challenges and the weakest link in the treatment chain (Njeh, 2008). Several studies look at the different aspects of detection and delineation of, tumours, aiming to help identify the sources of, and to reduce the overall

variability. This variability across clinicians can have a range of different causes (Van de Steene et al., 2002). Some might be overly cautious, not wanting to leave any residual cancer cells, while others might be expecting a different overall shape to the tumour and as such be affected by their subconscious bias. Additionally, it is important to take into account that different medical specialities have different outlining methodologies, which affects whether a person has been trained to draw the outlines erring on the outer boundary of the tumour, aiming to hit the exact centre on the perceived boundary, or erring towards the inner boundary of the tumour. Oncologists in general outline a wider contour to avoid missing tumour tissue, while radiologists are more concerned with making false-positives and will outline more aggressively. Regardless of training background, several margins of safety are added on top of the base outline. In general, three margins are added: a 'conservative' smaller margin, a 'safe' medium margin and a 'cautious' large margin, the size of which scale with relative position of cancer site (Horan et al., 2006).

In a study aiming to identify performance-related characteristics of radiologists, Elmore et al. (2009) had a group of 205 radiologists interpret a large sample of 1,036,155 screening mammograms in 531,705 women, of which 4961 had breast cancer. The study found that there was a wide range of variability in the outlined delineations between clinicians, one example of variability was highlighted from a mammography study that had 187 radiologists interpreting mammograms associated with cancer diagnoses, where 119 radiologists interpreted 10 or more images each. The median sensitivity of these 119 participants was 82.2%, with the full range presenting between 40% and 100%, and the interquartile range being 76.5% to 88.2%. The large variability across clinicians was also found to hold true even

for radiologists with similar false-positive rates. The authors conclude that the inclusion of fellowship training might help improve the sensitivity of detection in this task, though the training was also associated with higher false-positive rates.

Also looking to investigate the aspect of observer performance, a study by Castella et al. (2009) explores the influence of signal variability in a detection task. The authors ran an experiment on detection of simulated masses superimposed on both real and synthesised mammographic backgrounds, using a 2AFC vision-only detection task. In the experiment, participants were asked to select which of two images most likely contained the signal, they were given no time limit, feedback was provided after the completion of each trial, and a summary performance measure of percent correct was presented after every 25 trials. The stimuli used were images of synthetic breast masses, which were based off of image analyses of real-world breast lesions, embedded into a background of either a real mammographic image or a simulated one.

The experiment used two separate tasks, Signal-Known-Exactly (SKE) and Signal-Known-Statistically (SKS). In SKE, which is a match-to-sample task, the observers were both trained and tested on a pre-existing pool of signal images with a high-contrast replica displayed alongside the presented images, so they knew exactly what the signal would look like. In the SKS task, the signal was randomly chosen from a pool of different candidate signals, but the observer did not know which signal has been selected, though the size of the mass and whether it was malignant or benign remained the same throughout the task. Most studies on tumour detection variability historically focus on SKE tasks, as it is more easily modellable. SKS tasks are, however, much more closely related to the challenges of real-world radiological tasks where the ground truth tumour is unknown, but is

therefore also inherently more complex to analyse than the more controlled SKE task.

The experiment used a total of 1400 image pairs for a 2AFC signal detection task, using a total of 13 possible background and signal combinations, the order of presentation of these being the same across all observers. However, by not randomising the order of presentation of the background and signal combinations, it is difficult to identify a learning effect as such. The experiment had no time limit, feedback was given after every trial, and positional cues were embedded in both the signal-present and signal-absent images. The signals were either quantified as "malignant" or "benign", with a signal size of 6.5 mm when embedded in real medical background images, and either 6.5 mm or 9.5 mm when embedded in computer-generated clustered lumpy backgrounds (CLB). Additionally, for the CLB backgrounds, "benign" signals in the SKS task, there was also a variation in both shape and size, whereas for the other 12 combinations, size was static across all blocked trials. The main finding of the study was that participants did not perform significantly differently between the SKE and SKS tasks, for "benign" signals, in spite of the uncertainty and variability regarding the shape of the signal. Whereas for the "malignant" masses, there was a significant difference between the CLB images and the realistic medical images, which would suggest that the human observers utilise a different strategy which involves more complex signal and texture properties compared to the presumed strategy of the theoretical model observer.

One of the reasons for variability is the methodology used by the individual clinicians, which is a joint effort of their speciality, their training and personal bias. To that effect, the following two studies look to create guidelines to help reduce

the difference introduced by training and speciality and aiming to counter some of the individual bias. Jager et al. (2016) developed and tested new guidelines for tumour delineation on MRI for throat cancers. Aiming to develop guidelines to get the volumetric gross tumour volume (GTV) closer to the unknown ground truth (GT), they found that delineated GTV on MRI is generally twice as much as the volume of the GT. These guidelines help decrease overestimation of volume, while still maintaining similar tumour coverage. Similarly, Njeh (2008) also looked at validating guidelines for tumour delineation on MRI for specific subtypes of throat cancer; laryngeal (voice box) and hypopharyngeal (gullet) cancer.

However, one of the bottlenecks for delineation is the images that are being delineated; a noisy, blurry medical image cannot produce an accurate delineation, as low-resolution cannot simply be scaled up. The following three studies look into and compare different imaging methods, potential use cases, and the potential for multimodal imaging which allows the radiologist to view different scans of the same area either in parallel or in series.

In their meta-analysis of non-invasive imaging methods, Kinkel et al. (2002, pp. 748–756) compared different methods of detection for stomach cancers. They found that FDG-PET[1] is the most sensitive noninvasive image-modality. Meanwhile, Leclerc et al. (2015) investigated automated tumour delineation based on FDG-PET imaging for head and neck based tumours, which were to be treated by combined chemotherapy and radiotherapy. They aimed to confirm their pilot results across three participating hospital centres, as well as to evaluate the clinical outcome of the automatic delineation on the FDG-PET images using a combination of contrast enhanced CT or MRI and FDG-PET scans prior to treat-

[1]Positron Emission Tomography with ^{18}F-Fluorodeoxyglucose

ment. The results showed that GTV_{PET}[2] contours were significantly smaller and as such required smaller doses of radiotherapy than GTV_{CT}[3] contours in all but one cases, with the preventative-dose volumes based on PET also being significantly smaller in throat cancer cases, reducing over-treatment in organs at risk. As a result, dosimetry measurement of the ionising radiation doses showed significant decrease in cheek and oral cavity mean dose from PET based plans compared to CT based ones.

Shrikhande et al. (2012) have created a comprehensive literature review of different multimodal imaging of pancreatic cancer, aiming to reduce exploration time during surgery and to increase awareness of any aberrant anatomy surrounding the target area. They compared Endoscopic Ultrasonography (EUS), Multi-detector computed tomography (MDCT), Magnetic Resonance Imaging (MRI), Magnetic Resonance Cholangiopancreatography (MRCP), Multidetector CT (MDCT). They found that MDCT and MRI/MRCP have comparable sensitivity and specificity[4] for both diagnosis and for staging of pancreatic cancer, where EUS has the best sensitivity for lesions smaller than 2 cm. The authors concluded that MDCT with angiography or MRI/MRCP would be the best option for first image modality on suspected pancreatic cancer, while EUS would be the best option for suspected but not detected lesions from CT/MRI, MDCT would be best to assess vascular involvement in 'borderline resectable' tumours, and lastly that PET-CT would be benficial when used to help rule out distant metastases.

[2]Ground Truth Volume for FDG-PET scans

[3]Ground Truth Volume for CT scans

[4]Here, Sensitivity is the true-positive ('Hit') rate of pixels labelled as tumour, while Specificity is true-negative ('Correct rejection' rate of non-tumour pixels.

Furthermore, another aspect of tumour delineation that can benefit from improvement is the technology available for exploring these images. One such avenue is the potential to be able to view the images using stereoscopic displays, as outlined by Held and Hui (2011). In their guide they outline the potential benefits for stereoscopic 3D rendering for viewing vascular, mammographic and ophthalmic imaging as well as strong evidence for benefit when used in training, for telesurgery, and in surgical planning for lung-cancer treatment. The increased depth information requires fewer viewpoints than the current structure-from-motion technique used, it improves 3D anatomical understanding for medical students with low visual-spatial skills and provide a more intuitive understanding for relative shapes, sizes and positions that are missing from indirect viewing and 2D displays.

In addition to being able to view the images in 3D, there is also the potential for touching through simulated haptics, as described by Vidholm et al. (2008). In this paper the authors evaluated semi-automatic method for liver segmentation in CT images using haptic feedback for exploration by touch, and stereo graphics for a 3D visual representation of the workspace, giving the observers a fully 3D spatially coaligned rig. The medical images were shown slice by slice, with the segmentation mesh rendered in stereoscopic 3D. The segmentation mesh was initiated as a rough sphere, which could be scaled and positioned by the observer, who then deforms and refines the segmentation mesh to fit their perceived outline. To test the accuracy of the segment, a soft-edged 'fuzzy' ground truth was created by averaging across manual delineations of the tumours made on 2D surfaces, resulting in a mean precision of 96.9% for their seeded fast-marching method, with a decrease in interaction time from 5-18 minutes on a manual delineation

to a mean interaction time of 93 seconds with the seeded fast-marching method. These findings are promising for the potential use of haptic signal when working with medical images.

1.1.3 Tumour detection as a perceptual detection task

While the visual inspection of medical images is a common, well-established practice, there are still a lot of questions surrounding the underlying process used by the human sensory system to perceive and categorise image features (Abbey, 2013, p. 1). One of the fundamental aspects of optimising image acquisition and image processing is related to reducing the signal-to-noise ratio available to the clinician, which requires understanding the noise magnitudes and frequencies encountered in medical images, and how they are processed perceptually by human observers. In a series of 2AFC experiments and papers, Abbey and Eckstein first developed a framework to obtain and analyse classification images (Abbey & Eckstein, 2002), then by using these classification images they found that human observers differ significantly from the ideal observer in both a detection task and a contrast discrimination task (Abbey & Eckstein, 2006). Where the ideal observer uses a common non-linear mechanism for all three tasks, implying the same classification image in all three cases, the human observers used significantly different classification images in each of the three tasks. Expanding onto these findings, they found that human observers modify their perceptual template relative to the correlated noise textures (such as low-pass noise and high-pass noise), allowing them to focus on the more informative aspect of the image (such as looking at higher spatial frequencies in the low-pass experiment) (Abbey & Eckstein, 2007). In their 2009 paper, they found that human observers also adapt their perceptual

template to the different levels of magnitude of image noise (Abbey & Eckstein, 2009).

These noise adaptations as performed by the visual system have also been found to occur when viewing medical images, in a study by Kompaniez-Dunigan et al. (2015). They found that adaptation to images that were textually noisy in the same manner of the target image improved reaction time to locate a hidden signal embedded in dense or fatty images, suggesting that grouping the images by density would allow clinicians to be more attuned to the overall noisy background and decrease the required adaptation time.

1.1.4 Haptic exploration of medical images

While there is an abundance of information available through the synergistic use of different medical imaging modalities, there still remains the question of how best to access and take advantage of this information. Currently viewing the different scans in parallel is done by either switching between which of the modalities is being displayed to compare the corresponding anatomical features, or by using a slider to adjust relative opacity and overlap between the two imaging modalities. The amount of information that is available through these methods is limited by the nature of the presentation method; while the medical scans are rich 3D datasets of the body, they are compartmentalised and reduced to a stack of 2D images which are viewed on standard 2D computer monitors in monochrome grey-scale or artificially colour-mapped to highlight regions and features.

Through the use of new and novel technologies known as 'sensory substitution devices' we are able to access and integrate information in different sensory modalities than the ones conventionally used to perceive the respective sensory

signals. Sensory substitution devices can help translate sensory information including but not limited to: using audition (Alfaro et al., 2015; Bologna et al., 2008; Hamilton-Fletcher & Ward, 2013; Hamilton-Fletcher et al., 2016; Maidenbaum, Abboud, et al., 2014; Osinski & Hjelme, 2018); haptic vibration (Carcedo et al., 2016; Delazio et al., 2017; Schwerdt et al., 2009; Tapson et al., 2008) to perceive colour on a fully aligned spectrum; or even using haptics to explore simplified graphical images and recognise facial features extracted from photographs (Lim et al., 2019).

Some information is lost, however, when artificially conveying information through these methods, not unlike cochlear implants which circumvent auditory signals and transfer them directly to a neural signal, some level of detail will be lost (Richardson et al., 2019). In order to render photographs haptically, the different image features must be artificially adjusted to allow for higher contrast and differentiability between features (Lareau & Lang, 2012). There are also several other uncertain factors, including how the difference between vision and touch behave in terms of statistical properties of textures (Kuroki et al., 2019), and whether the addition of haptic information simultaneously explored alongside vision would be beneficial or detrimental. For example, the addition of augmented reality (AR) can be detrimental to task performance in surgery (Dixon et al., 2013), and whether the simultaneous presentation of signals could be trained (Bertram & Stafford, 2016). Additionally, even sighted individuals can easily recognise objects from touch alone, thanks to the well-established object features such as texture, compliance, shape and volume (Chit & Yap, 2012; Klatzky et al., 1987; Lacey & Sathian, 2014). This brings us to the old philosophical question of Molyneaux's Problem: whether someone born-blind who gains sight at a later

point in life is able to identify an object by sight alone. This problem has recently been solved via breakthroughs in medical science (Held et al., 2011), and the answer is a simple 'no, they cannot'. However, they do quickly learn to combine the senses shortly after gaining vision.

While the use of height-based haptic depth rendering has been shown as very beneficial for intuitively understanding topographic maps (Evreinova et al., 2012), there is a distinct difference between the scale of a mountain range and the level of signal present in a medical imaging scan. This is especially considering that medical image scans are not themselves height maps, but instead density maps created by the various tissues of the body's ability to absorb the relevant scanning technology such as x-rays or magnetic resonance. Rendering a density map as a height map could be considered a minor adaptation, or it could be considered a completely novel mapping between two arbitrary signals.

However, even arbitrary signals can be learnt to be integrated, as suggested by Ernst (2007). In their study, Ernst (2007) explored the ability of observers to integrate arbitrary, unlinked sensory signals after a period of training. The study consisted of an experiment where twelve trained observers were asked to identify which of three trials was different in a three-interval forced-choice (3IFC) oddity task, where the stimulus was a 25 by 25 mm flat square presented at 50 cm distance to the observer on a spatially coaligned visuohaptic setup, with the three conditions of haptic-only, vision-only, and combined visuohaptic signal. The haptic signal was surface stiffness, while the visual signal was surface luminance. They first recorded the single-cue just-noticeable difference (JND) thresholds prior to training, which allowed for having normalised stimulus space on a per-observer basis, in JND units. Having normalised the stimulus reliability, data was collected

for the two-cue stimuli (being both haptic stiffness and visual luminance), along the two respective directions of the individually JND-normalised space, again using the 3IFC odd-one-out task. Following this, the observers went through a training phase where half were trained to correlate higher visual luminance with higher haptic stiffness, while the other half were trained to correlate lower visual luminance with higher haptic stiffness. This was done by only presenting stimuli along the "congruent" direction of the respective groups, where the strength of the stimuli were equally distributed along the possible range of presented signal strengths, and providing feedback after each trial. After the training phase, observers were again tested on both "congruent" and "incongruent" visual and haptic signals, and finally the single-cue JND units were measured per observer to control for overall improvement associated with simply having more experience with the task.

The data was fit to cumulative Gaussian (CG) psychometric functions (PF). A repeated-measures ANOVA was run comparing the performance in the single-cue tasks performed at the beginning and end of the experiment, which found no significant effect of having experience with the task, and no significant difference was found between "congruent" and "incongruent" signals in the pre-training block. A two-factor within-subject ANOVA showed no significant main effect for pretest nor post-test, nor congruent and incongruent, but did show a significant interaction effect between pre-training and post-training, and congruence versus incongruence. Whereas pre-training had no difference between congruence and incongruence, the post-training block shows significant difference in perceptual thresholds. This shows that the observers had learnt to utilise the redundant information between an object's surface stiffness and its luminance, the two arbi-

trary signals in question. While the effect of the study is still small, the training time was only a single 60-minute session for two completely arbitrary sensory signals, and as such was not expected to be large. This study indicates that, even if the information from the medical imaging scan is displayed as a visual density map and a haptic height map, it could be possible to learn to associate and benefit from these signals, regardless of whether or not they are considered by the sensory system to be arbitrary.

However, it is worth noting that, for arbitrary signals to be maximally beneficial, it would be sensible to normalise the reliabilities in a manner similar to, or simply following, the JND methodology of estimating exactly how much the sensory signal can change before the observer can reliably detect a difference, on a per-person basis. For medical images specifically, this would likely be through the use of a 3IFC odd-one-out task on statistically matched synthetic medical images, similar to the procedure used in Ernst (2007).

1.2 Introduction to Perception

Having established the potential of visual perception metrics for tumour detection tasks, we are next going to delve into perception as a field. A large part of tumour detection comes down to a person visually exploring an image and attempting to perceive a difference, an anomaly in the source image. A lot of the current experimental procedures compare overall performance as a percentage correct as well as by using self-reported measures, neither of which lend themselves well to modelling trends or to investigating the effects of variable manipulation. However, through the use of psychophysics and its subsequent methods of measuring and

modelling human perception we are able to investigate an observer's precision and accuracy in detection and delineation tasks, and compare numerically the effect of adding information such as haptics.

The three main methods used in psychophysics to measure and quantify human perception are 1) Magnitude estimation, where a subject rates an aspect of the stimuli on a scale; 2) Match-to-sample, where the subject adjusts the stimulus to match a ground-truth sample, and 3) Detection and Discrimination, where the subject is asked to detect small differences between several stimuli, or whether a signal is present or absent. A detection task is a specific subcase of the discrimination task, where the observer compares the target signal to a signal that is always absent, rather than comparing between signals of differing strengths (Prins & Kingdom, 2016).

1.2.1 Perceptual cues

The way we, as humans, perceive and explore the world is through our various senses, such as vision, hearing and touch. Through these senses we are able to pick up information that combined with both previous experience and other senses allows us to create a percept of the world around us. The information that we acquire is made up of a collection of 'cues', which can be thought of as "any sensory information that gives rise to a sensory estimate" (Ernst & Bülthoff, 2004, p. 168). Much research has gone into breaking down the various underlying processes utilised by the brain for perceptual tasks. Vision, for example, is one of the most well-developed of the human senses, and like most primates, comprises a substantial part of the cerebral cortex (Zeki, 1993). Vision is used to perceive a lot of the properties of objects, for example the shape and orientation of an object,

relative distance between several objects, as well as their relative difference in size to one another (Howard, Rogers, et al., 1995). These property estimations are made from a whole host of visual cues such as the varying shading of the object, the disparity between the signals the two eyes receive (Lovell et al., 2012), and the presumed perspective of the shape (Purves et al., 2013), to name a few. For the perception of touch, or 'haptics', examples of sensory cues include temperature, pressure, proprioception and physical texture. It is possible to touch a shiny, transparent surface and identify if it is likely to be glass or plastic (Gueorguiev et al., 2016) or whether an object is metallic or plastic (Chen & Chuang, 2014).

1.2.2 Cues in Detection and Discrimination tasks

The act of searching visually is an everyday task; as one might search for a familiar face in a crowd, or a four leaf clover in a field. The difficulty of the task depends on several factors such as how distinguishable the target is from the background. If for example your friend is tall, they will be easier to spot in a crowd than a person of short or average height. However, for the clover to be comparably distinguishable it would have to be several times taller, as the background and overall scene is visually very different. In a complex naturalistic scene, the human observer subconsciously takes in the shading, motion and sounds, as well as the disparity provided by their binocular vision. In order to be able to narrow down on the specific aspects of these sensory processes, one feature is typically selected and a simplified, non-realistic stimulus is created to explore how this feature is identified in isolation. For example, for looking at how human observers perceive medical images, one would simplify the texture from the noisy, real-world medical scan into a simplified stimulus of a basic tumour-like shape or Gaussian bump

in a white-noise background (Abbey & Eckstein, 2002; Kompaniez-Dunigan et al., 2015). By starting with very simplistic, controlled stimuli it is possible to create a baseline for measuring performance, and adding onto this by making the experiment more complicated and adding in more features, like using more complicated 'tumour' shapes or by using real medical image as a background, building up to a full-on high detail representation of a realistic scene.

The difference between signal discrimination and detection lies within the task itself. In a discrimination task, the observer aims to determine which of several signals is 'stronger' than the others, with all the possible alternatives containing a signal of predetermined strength. While a detection task is a subclass of the discrimination task where the non-target alternatives have a signal strength of zero, the subtle difference in task goal alters the underlying mechanisms used by the observer. While both detection and discrimination are used to measure the precision of an observer, a detection task can be used to measure the minimum required signal strength (the 'absolute threshold', often shortened to just 'threshold') for the task, whereas a discrimination task instead compares between several non-zero signals and is instead used to measure the just-noticeable-difference (JND, also sometimes referred to as the 'difference threshold') and bias at the point of subjective equality (PSE).

In psychophysics, one typically presents the stimulus at a range of strengths, either from very strong to very weak (or absent, as is the case in a detection task), or on a scale of strongly one parameter change to the opposite parameter change (i.e. slanted left or right, or positioned above/below). These are done as N-alternative forced choice, so even if the observer cannot detect the signal, or the parameter change is neutral (frontoparallel, in the centre), the observer has

to select which of the alternatives they perceive to be the most appropriate on a per-trial basis. By repeating this process over a large number of trials, it is possible to estimate at what level the observer can reliably perceive the stimuli, by using parameters from the psychometric function fits. However, as the process used by the brain to perceive information from the different senses (i.e. vision, touch) is not the same for each respective sense, it is important to separate out and investigate these processes a bit more in detail on a single-sense basis.

For example, a study by Klatzky et al. (1987) found that haptic perception is better than visual when it comes to quickly gathering a large amount of information about the material properties of an object, such as its hardness and texture, while visual perception is more salient when it comes to object shape, contour and to some extent size. In this study the authors performed two experiments in sorting objects, the first requiring that the objects be sorted by shape, size, texture and hardness, and the second that they be sorted by similarity. From their findings they conclude from their results that haptics uses its own object encoding pathway compared to vision.

There is also a difference in how the haptic information is received. In their paper, Lederman and Klatzky (2009) contrast passive versus active haptic exploration of rough surfaces, theorising that fingertip estimation of rough surfaces uses the cutaneous feedback of how much of the fingertip is touching an edge. The authors conclude that observers are better at estimating roughness when actively exploring compared to passively being touched by an object. The paper also notes that, while both sighted and blind observers have been shown to have a decline in fingertip spatial acuity with age, other studies that used active exploration of braille-like dot patterns have shown the expected haptic acuity decline for sighted

people while blind observers retain their acuity to a much higher extent. This benefit is also not limited to the braille-reading finger.

As people have access to a multitude of senses, they seldom rely on the sole input of a single sense, instead the brain greatly benefits from combining the information from several. Exactly how people combine and integrate sensory information is currently a very active area of study, with a lot of research going into modelling and describing the mathematical models of cue combination.

1.3 Cue Combination and Models of Fusion

There are multiple ways in which a human observer can combine information (Jones, 2016), with different models of accuracy, ideal observers and at what point of the signal processing noise is introduced. The main basis of most of these models is that having access to multiple cues give a better performance than just a single cue, where "better" is the ability to detect a signal at or around threshold, and precision in discriminating a supra-threshold stimulus.

All senses are potential sources of information to the brain, some of which may contain information about the same object. The brain can, in most circumstances, combine this information to create a stronger, more reliable idea of the various properties of the object in question, known also as 'fusion' (Ernst & Bülthoff, 2004). There are a few different distinctions between terms: sensory combination is used for non-redundant signals, where each signal has unique characteristics and may well be in different units, different coordinate systems or even be concerning complementary aspects of the same perceptual property – while sensory integration on the other hand, requires there to be redundancy between the signals, so

they are in the same units, same coordinate system and about the same aspects.

1.3.1 Modality appropriate hypothesis

One of the oldest cue combination theories in the field is known as the Modality Appropriate Hypothesis, also known as 'sensory dominance'. While some of the earliest studies into visuohaptic cue combination found that vision strongly dominated the task. One such example is the study by Rock and Victor (1964) who presented an object that was optically distorted from its physical shape, and had participants either draw the perceived object or match to sample. Their findings strongly favoured the visual representation. McDonnell and Duffett (1972) aimed to investigate the visuohaptic cue conflict with added control groups, breaking up the participant pool into three groups. One group of combined visuohaptic exploration, one of vision-only exploration, and one of haptic-only. After the initial exploration, observers in all three groups were asked to select a comparison block of equal size to the one they had been investigating. By comparing the means of the separate groups, the authors found that the observers had made their judgements conforming either to their visual impression, or to the haptic one – indicating that vision does not always dominate in the task. Welch and Warren (1980) thoroughly and critically reviewed the cue conflict literature out at the time, and suggested a new integration model that labels the intersensory bias as an attempt of the perceptual system to integrate information from discordant cues. This would imply that the sensory system attempts to discern which cue is more reliable, and assigns a higher weighting to this cue than to the less-reliable one, whether automatically or consciously.

Where these modality-specific reliabilites have great variance, it would be con-

sistent with one cue dominating over the other. However, Guest and Spence (2003) found that task-irrelevant cues can influence perception, where given a haptic roughness signal and a visual one, the haptic roughness perception could influence the visual perception of roughness, but the visual roughness perception did not in return influence the haptic perception.

These findings were matched by Bresciani et al. (2006), who investigated the integration of sequence of events for visual and tactile events. They presented their subjects with a series of visual flashes and haptic taps simultaneously, and asked them to count either the flashes or the taps. They also found that haptic influences touch more than touch influences haptic. Interestingly, they found that the variance was lower when both modalities were presented compared with the task-relevant modality alone. They conclude that the integration of sensory information is automatic across modalities, even in the case of task-irrelevant information.

1.3.2 Weak fusion

The weak fusion model is a sensory integration model where the sensory cues are thought to be separate and modular sources of information that contribute their own estimate of a property (such as slant or depth), prior to combining into a single, combined estimate (Clark & Yuille, 2013; Landy et al., 1995).

This model has an additional benefit that it is easy to test empirically. There are however two primary issues with weak fusion, firstly that it cannot take into account any qualitative differences between the sensory cues, and secondly that the averaging of the cues assumes an equal weighting of all inputs. The first concern is rooted in the fact that cues are often given in disparate units of measurement,

where for example identifying the shape of objects in a visual scene using both shading and stereo vision would include both the relative strength of illumination gradient as one cue metric, while the distances between objects perceived using binocular disparity as a cue would be a completely different and incomparable metric, as binocular disparity and visual luminance are unrelated sensory cues of disparate sensory units. The second issue is the assumption that all cues are equally reliable and should contribute equally to the fused percept, even though it is well established that cues can often have different levels of reliability, depending on the information available in the scene (Landy et al., 1995).

In short, the cues are themselves independent sources, and when combined in the weak fusion model they are assumed to have equal reliability and would inform the final estimate equally. However, some signals are more reliable than others (Backus & Banks, 1999), so this is not an optimal model. A more elaborate model of fusion should, instead of assuming an equal weighting if cues, adapt the weightings in the combined estimate to the variance in cue reliabilities, and adjust the contributions of the respective cues.

1.3.3 Strong fusion

This brings us to the strong fusion model. In the strong fusion model, cues are thought to interact before the estimation occurs, and they are inseparable sources of information. The cues themselves are not thought to be separate or modular, nor that they have individual estimates that exist separately to the final estimate. While the strong fusion model has the potential of reweighting cues dynamically and that the cues themselves do not require to have common sensory units of measurement, it also means that any separation of cues at the perceptual source

level is considered to be an artificially derived reconstruction (Landy et al., 1995).

The negative aspect of the strong fusion model is its inherent lack of constraints regarding the combination rules, making it difficult to test and model experimentally. As there are no formal rules surrounding the combination itself the number of interactions between the cues can be limitless. This complicates finding the method of which the sensory system creates its final percept, as the interactions themselves could be arbitrarily complex. The majority of support for this model is found in the absence of support for other fusion models, such as the work done by Rosas et al. (2007).

1.3.4 Modified weak fusion

If one places the weak fusion and strong fusion models on a continuum of sensory integration, where on one end is the weak fusion model where it is assumed that all signals are of equal importance and in comparable units as considered by the sensory system, and on the other end is the strong fusion model where the cues are inherently inseparable in terms of percepts but the cues themselves are not neccessarily weighted equally – there will be points between the two that could allow for the brain to weight the inputs relative to their reliability, as well as having interaction between the cues, though some limits would be placed on these interactions. Aiming to address this, Landy et al. (1995) constructed a modified version of the weak fusion model, naming it modified weak fusion (MWF).

The MWF model differs from the basic weak fusion model in that it allows cues to be translated, or 'promoted', into a common set of units, which then allows arbitrary linked cues to be combined. Like in the strong fusion model, the MWF model has dynamic weighting of the cues, where the weight given to a cue

is relative to its reliability compared to the other cues, which is done prior to the linear summation into a unified percept. The basic mathematical description for MWF is given in Equation 1.1, first as the general case for n cues and for the specific case of $n = 2$ cues. However, a few assumptions must be met for these equations to apply. The signals, such as the likelihood and the priors (Hillis et al., 2004), must be Gaussian distributions, they must be independent and they must be combined in a linear fashion (Oruç et al., 2003).

$$\widehat{S_T} = \sum_{i}^{n} w_i \hat{S}_i = w_1 \widehat{S}_1 + w_2 \widehat{S}_2 \tag{1.1}$$

Where the total Signal $\widehat{S_T}$ comprises the sum of the weighted estimates of Signal 1, $\widehat{S}_1$, and Signal 2, $\widehat{S}_2$, and the respective weightings in 1.2 and 1.3. An example illustration is shown in Figure 1.1.

If we constrain the distributions to be independent and Gaussian, with a sum of weight equal to 1 (Maloney & Landy, 1989), we can express the individual weights w_i as the relative reliability over the sum of all reliabilities, shown in Equation 1.2.

$$w_1 = \frac{r_1}{r_1 + r_2}, w_2 = \frac{r_2}{r_1 + r_2} \tag{1.2}$$

The individual reliabilities r_i are further expressed as a function of the variance of their respective cues, σ_i^2.

$$r_1 = \frac{1}{\sigma_1^2}, r_2 = \frac{1}{\sigma_2^2}, \tag{1.3}$$

By plugging the values for w_i from Equation 1.2 into Equation 1.1 and multi-

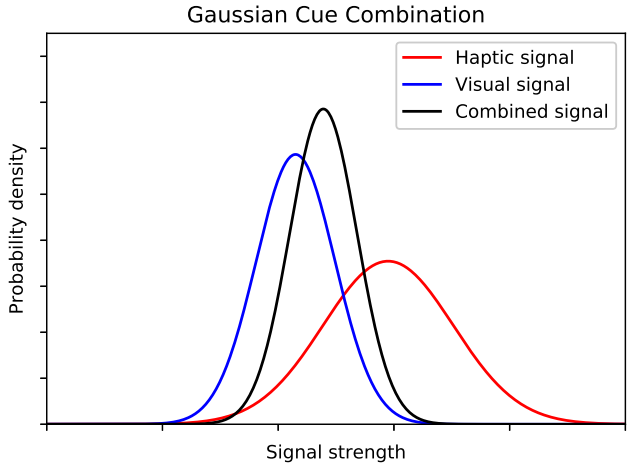


Figure 1.1: Example image of basic two-cue Gaussian cue combination. Where the blue line represents vision, the red line haptics, the black line is the calculated weighted average of the two signals. As vision has higher reliability (lower standard deviation), it is weighted more than the haptic signal which is less reliable, illustrated by the centre of the combined distribution being closer to vision than to haptics. However, the combined distribution has a lower standard deviation than either vision or haptics, as it is more reliable than either signal in isolation.

plying by $(r_V + r_H)$ on both sides, we get Equation 1.4.

$$\widehat{S_C} = \frac{r_V}{r_V + r_H}\widehat{S_V} + \frac{r_H}{r_V + r_H}\widehat{S_H} \tag{1.4}$$

$$(r_V + r_H)\widehat{S_C} = r_V\widehat{S_V} + r_H\widehat{S_H}$$

By then substituting the values for r_i from (1.3), we get (1.5).

$$\begin{aligned}(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2})\widehat{S_C} &= \frac{1}{\sigma_V^2}\widehat{S_V} + \frac{1}{\sigma_H^2}\widehat{S_H} \\ (\frac{\sigma_V^2 + \sigma_H^2}{\sigma_V^2\sigma_H^2})\widehat{S_C} &= \frac{1}{\sigma_V^2}\widehat{S_V} + \frac{1}{\sigma_H^2}\widehat{S_H}\end{aligned} \tag{1.5}$$

Which allows us to isolate the σ_c variance on the left-hand side, shown in 1.6.

$$\frac{1}{\sigma_C^2} = \frac{\sigma_V^2 + \sigma_H^2}{\sigma_V^2 \sigma_H^2}$$
$$\sigma_C^2 = \frac{\sigma_V^2 \sigma_H^2}{\sigma_V^2 + \sigma_H^2} \tag{1.6}$$

Which shows that the variance of the combined Gaussian will always be smaller than the smallest individual variance of each respective cue (Oruç et al., 2003). While there are a number of models for combining cues, MWF is the most widely researched and the supported model for sensory cue combination (Knill & Richards, 1996; Mamassian et al., 2002; Rohde et al., 2016). However, the Weighted Averaging model used by the MWF framework is also one specific example of the more general Bayesian Framework for fusion.

1.3.5 Bayesian cue combination

Bayes Theorem is a statistical model that describes the probability of an event occurring, based on prior knowledge of factors related to the event, contrasted with most frequentist models which do not take prior probability into account. For example, the likelihood of having cancer if a positive result occurred on a cancer test is proportional to the likelihood of a cancer test giving a positive result when cancer is present, multiplied by the likelihood of having cancer, divided by the likelihood of the test giving a positive result regardless. This is expressed in Equation 1.7.

$$P(A|B) \propto \frac{P(B|A)P(A)}{P(B)} \tag{1.7}$$

The probability of A occurring (having cancer), given B (positive test results) is the probability of P(Positive test | having cancer), times the probability of cancer P(A), over the sum of all positives P(B) (which includes both 'hit' and 'false positive'), P(B|A)*P(A) = P(Positive test | not having cancer)*P(not having cancer).

A different example would be the probability of an object being a given size in the real world A = Size, given B = a Visual cue. $P(A)$ is then the probability of it being that size in the world, also known as the 'prior' probability. It is independent of any current sensory data. $P(B)$ is the likelihood of the visual input, here it is considered a normalising constant and can be dropped (Mamassian et al., 2002). $P(B|A)$ is the likelihood of that visual input occurring if the property is in fact that size, physically, in the world. By dropping the normalising constant $P(B)$ we obtain Equation 1.8.

$$P(A|B) \propto P(B|A)P(A) \tag{1.8}$$

$P(A)$ is the prior probability of a property in the world, here the probability of an object being a certain size. If this is flat (i.e. uninformative), it too can be dropped and we end up with equation 1.9.

$$P(A|B) \propto P(B|A) \tag{1.9}$$

Which informs us that the probability of the object being a specific size given the visual cue is dependent on the probability of the visual cue itself if the object is of that physical size. For two sensory cues with conditionally independent distributions, it is written as 1.10.

$$P(A|B,C) \propto P(B|A)P(C|A) \tag{1.10}$$

Where C represents the input of a second cue, in this case haptics. Here, the probability of the object being a certain size (A), given inputs from both vision (B) and touch (C), depends on both the probability of the visual cue given the size ($P(B \mid A)$), and the probability of the haptic info given the size ($P(C \mid A)$). This also assumes the different cues (B) and (C) are conditionally independent, and that any sensory noise associated with each respective cue is independent from one another. This model, as shown in Equation 1.9 and 1.10, is known as the Maximum Likelihood Estimate (MLE) model.

1.3.6 Maximum Likelihood Estimate

As mentioned, MLE is a subset of Bayesian Inference, where the normalising constant $P(B)$ is dropped, the prior probability $P(A)$ is considered 'flat' and therefore uninformative, and can be dropped. Additionally, the modality noise of the cues is assumed to be conditionally independent and of a Gaussian nature. In this specific case of the probability $P(B|A)$ and the probability $P(C|A)$ both being Gaussian, the Maximum-Likelihood Estimation is also equivalent to weighted averaging, where $\widehat{S}$ is the property being estimated and w is the weight given to that estimate. In this case, Equation 1.11 is equivalent to Equation 1.1, where the Combined estimate $\widehat{S_C}$, in this example being the size of the object, equals the sum of the Weighted Visual estimate $w_V\widehat{S_V}$ and the Weighted Haptic estimate $w_H\widehat{S_H}$. As we discussed in MWF, the weighting of the cues is proportional to the reliability of said cue, the more reliable a cue is, the higher its weight in the linear

summation. This is statistically optimal as the Combined variance is lower than the lowest variance of either of the single cues in isolation (Ernst & Banks, 2002).

$$\widehat{S_C} = w_V \widehat{S_V} + w_H \widehat{S_H} \tag{1.11}$$

The ability to benefit from having multiple cues is present both when the different cues are from the same modality (e.g. disparity and occlusion are both vision), or from different modalities (vision and touch). Maximal benefit occurs when the cues have matched reliabilities[5]. If reliabilities are not equivalent, the magnitude of improvement decreases. As MLE is also a part of the MWF framework, it is an empirically testable model, and has been shown to work both within modality, such as visual disparity and texture (Knill & Saunders, 2003), or across modalities such as vision and touch (Ernst & Banks, 2002).

Within modality

Currently there is a lot of evidence for the optimal integration of within-modality cues, such as Knill and Saunders (2003), Hillis et al. (2004) and Machilsen and Wagemans (2011). In their paper, Knill and Saunders (2003) test their predictions of humans optimally integrating stereo information and texture information when judging surface slants presented visually in 3D. In their study, subjects were asked to select which of two visually textured surfaces they perceived to be more slanted, in a 2AFC experimental paradigm. The authors report findings to be largely consistent with an optimal combination of stereo cues and texture cues, with the variance of the combined cues being lower than the variance of either single cue in isolation. The findings of Hillis et al. (2004) supports the results of the

[5]Under the Modified Weak Fusion framework.

study by Knill and Saunders (2003). In their experiment participants were asked to select which of two sequentially presented visually textured slanted surfaces they perceived to be more slanted than the other. The results of the experiment found that the variance of the combined visuohaptic modality was lower than the variance of either of the single-modality cues in isolation, as represented by Equation 1.6.

Machilsen and Wagemans (2011) explored the integration of visual cues, by investigating contour cues and surface information for the detection of shapes. Their study comprised of two experiments, both of which used a 2AFC detection task with arrays of Gabor elements to explore the integration of contour cues and surface information, contrasting between the probability summation model and the additive summation model. The probability summation model postulates that the performance improvement of several cues compared to single cues is not due to the integration of the cues, but simply the combined likelihood of detecting any of the single cues individually. Additive summation on the other hand, is a mathematical model which suggests integrating information from the different cues. In the study, participants were asked to identify which of two sequentially presented visually noisy squares contained a hidden visual shape. In the first experiment, the authors found that participants use both cues to detect the shapes, which indicates a clear double-cue benefit. In the second experiment, which focussed on a subset of the jitter levels for the detection thresholds, the results indicate that the performance of the observers was better than predicted by the probability summation model, while no evidence was found to be against the performance following an integration model, to which the authors conclude that the observers may well be combining the information of the surface texture

and the contour in a statistically optimal fashion.

While fewer studies have been carried out on within-haptic cue combination, Drewing and Ernst (2006) looked at the effect of separate force-curve and position-curve within haptic signal. In their study they ran two experiments, both of which asked observers to select which of two arches that were presented visuohaptically they perceived to be more 'convex'. The arch-pairs were presented sequentially in a randomised order following the procedure of a 2 interval forced choice (IFC) discrimination paradigm. In the first experiment, observers were tested on a range of 9 different arch heights matched with 13 different 'control' arches (for 117 unique pairs), while in the second experiment they were tested on only the 'shallow' and 'high' extremes of the arches. The authors found that by using active exploration of a virtual haptic arch, where both the position of exploration and the force used to explore were independent, that participants followed a linear weighted combination model where the cue weighting changed dynamically in a manner proportional to the manipulated reliabilities.

Across modality

Previous research has shown that people perform with a higher level of precision when discriminating features of objects when using both vision and haptics, compared to using either modality in isolation (Ernst & Banks, 2002; Helbig & Ernst, 2007). In Ernst and Banks (2002), the authors ran a study where they investigated the effect of an added haptic cue on combined performance, tested across a set range of visual reliabilities. In their spatially coaligned 2AFC size-discrimination experiment participants were asked to identify which of two blocks were 'taller' than the other, by means of either vision, haptics or by using combined visual-

haptic signals. By manipulating the level of noise in the visual modality while keeping the haptic signal constant across all visual noise levels, the authors found that people performed in a statistically optimal fashion when the reliabilities of the cues were matched. Additionally, the results show that the higher the visual noise (the lower the visual reliability) the more weighting was afforded the haptic signal, which suggests a dynamic adaptation for vision and touch. The results of this study were confirmed by Helbig and Ernst (2007), who also did a 2AFC experiment where observers were asked to estimate whether the target signal, an ellipse, appeared horizontally or vertically elongated, again using vision-only, haptic-only or combined visual-haptic signals in a cue-conflict paradigm. The results of the experiment indicate that redundant information from vision and touch were combined in line with the predictions of the MLE model, which supports the findings of Ernst and Banks (2002). Helbig and Ernst (2007) additionally found that modality-specific distractors did not interfere with the process of multisensory integration.

Although object detection has been studied in both the visual and haptic domains respectively, the research on using the combined visuohaptic cues have heavily focussed on object discrimination. One study that looked at this cross-modal signal detection is Louw et al. (2000), who explored haptic detection thresholds of Gaussian profiles embedded in otherwise flat strips of material. They ran a 2AFC experiment where subjects compared a curved and flat strip presented successively in a random temporal order. The authors found that for Gaussian width (σ) greater than 1.0 mm, the signal detection threshold presents as a linear function of the width σ on a double-logarithmic scale, which holds true for both convex and concave stimuli.

Another study looking at cross-modal detection, Plaisier et al. (2010) investigated whether vision helps with haptic signal detection in an environment that was both haptically and visually noisy. The participants were asked to locate a target dot using a haptic device. The target dot was presented haptically but not visually, as the target dot was surrounded by 'noise' – additional non-target dots that were simultaneously presented both haptically and visually. The authors found that while the visual information was not very informative, it still aided in improved the overall spatial representation of the scene. This finding can be extrapolated to indicate that it could still be beneficial to show the position of a probe in teleoperational systems, even when the overall visual information in the scene is poor.

However, some contradictory findings exist regarding full optimality in cue combination, especially regarding suboptimal performance in slant perception (Plaisier et al., 2014; Rosas et al., 2005; Rosas et al., 2007), and for reaching and grasping (Adams, 2019), to name a few.

The study by Plaisier et al. (2014) investigated how exploration mode might affect the perceived orientation of a surface, comparing between temporally parallel and sequential exploration modes. In their experiment the participants were asked to explore a slanted surface using vision, haptics or a combination of the two, using either the parallel or the serial exploration mode. The parallel exploration mode had two static points of the surface presented simultaneously, while the serial exploration mode had only one point of contact used to dynamically explore the surface. In the first experiment the same exploration mode was used for both cues while for the second experiment the different exploration modes were used for the different cues. The authors conclude that using different exploration

modes for information acquisition in the respective cues caused a suboptimal integration of cues, disrupting the effect of sensory integration that occurred when the exploration mode was the same for both cues.

1.4 Sensory correspondence

The different models mentioned in the previous section all discuss the potential underlying method of how the sensory system integrates the different sensory signals, however, cue integration is only beneficial and should only occur when the signals are related and have a common source. The basic cue combination models do not address whether the signals should be integrated, nor how the sensory system determines which signals are related and should be integrated, and which signals are not and as such should not be. This is known as the 'sensory correspondence problem'. The sensory system is being constantly bombarded with information from the different sensory modalities, and if it were to attempt to combine every single of these sensory inputs that occur at any given point in time, it would be a combinatorial explosion of potentially linked signals, placing a prohibitively costly computational load on the brain.

It is also worth noting that, even for sensory signals that are known to be from the same source, there are often naturally occurring discrepancies, such as magnitude for haptic and visual textures, where the visual texture may look coarser than it feels haptically, or duration for visual and auditory signals, where an auditory beep will appear to last for a longer duration than an equally long flash of light (Marks, 2014). It is possible that this is due to some percepts being more developmentally critical in early evolution, and as such have their own 'wiring' in

the brain, which may be imperfect in overlap between the senses (Marks, 2014).

While signals are commonly considered strongly correlated when they are presented either spatially or temporally co-inciding (Holmes & Spence, 2005; Spence, 2013), the existence of tool-use in humans, suggest that the sensory anomaly introduced by the spatial offset between signals is something which the brain can adapt to. While tool-use is an ability commonly observed in humans, other primates, and some birds, the exact mechanisms behind this is a very active topic of research (Holmes et al., 2004, 2007; Holmes, Sanabria, et al., 2007; Maravita et al., 2002; Takahashi et al., 2009). As tools introduce a spatial offset between the visual signal and haptic signal, a mechanism must in the human sensory system which allows for solving the correspondence problem in order to adapt to the use of tools. Here we discuss two of the more common models for solving the sensory correspondence problem.

In a study by Kang and Kim (2018), the authors explore the multimodal discrimination of curvatures using a combination of vision and haptics in a 2AFC signal discrimination paradigm. In their experiment participants were asked to explore a curved texture using a 3D rendered haptic signal, a monocularly presented shaded visual signal or using combined visuohaptic signal – the observers would then select which of the two temporally presented alternatives they perceived to be more curved. Using these responses, the authors calculated the JND values for the three conditions, and found that, for JND units, the weight placed on haptics increases as the relative reliability of vision decreases, in line with the findings of Ernst and Banks (2002).

1.4.1 Coupling priors

The first of the suggested models is the 'coupling priors' model, proposed by Ernst (2006). The 'coupling priors' model is based on the concept that redundant information between sensory signals is used to improve the sensory estimate of an object, and that previous experiences create mapping between sensory signals, which is stored and used to assist in future estimations. If the mapping between two sensory signals remains truly constant, then the sensory system has no need to retain the individual estimates and can instead fuse the signals completely. In this case, the coupling between the signals would be very strong. If the signals are completely independent, the coupling between them is non-existent, and the sensory system calculates the estimates in isolation. However, one would often estimate something in-between, where the pre-existing mapping between the sensory signals is present but not truly constant, the coupling between the signals is considered to be strong, but not as strong as for fused percepts. One example of a strong coupling between signals is visual disparity between the eyes and certain distortions to texture signals which are both strongly associated with slant. Because the coupling between these signals is so strong, it is heavily relied upon in perceptual illusions such as the 'Ames Window'[6].

In the paper by Ernst (2007), which is covered in detail in Section §1.1.4, it is suggested that there exists a statistically linked component between the neural signals given by the sensory inputs, occurring in the brain. When interacting and exploring an object both visually and haptically, there are neural signals that appear in the brain which are linked to the respective sensory modalities. These

[6]The 'Ames Window' or 'Ames Trapezoid' is a forced-perspective illusion of a slanted window, first described by Ames Jr in their 1951 paper *"Visual perception and the rotating trapezoidal window."*

neural signals have a naturally occurring statistical relationship, which could be used by the brain to form an object property concept which we might call, for example, the "size" of the object (Ernst, 2007). In Chapter 4, this model is discussed in relation to a study by Adams et al. (2016).

1.4.2 Causal inference

Another potential model for solving the correspondence problem is the 'causal inference' (CI) model, proposed by Körding and Tenenbaum (2007). In the CI model, an observer is thought to identify multiple signals to be inherently linked due to having a common source, such as the visual representation and proprioceptive awareness of ones own hand, or the sensation of touching a hard surface and hearing a tapping sound.

The CI model is a generalisation of the weighted linear summation model, which was discussed in Section §1.3.4, where in the CI model the weighting of the respective cues is further affected by the probability of the signals having arisen from the same source. This is shown in Equation 1.12, where $\widehat{S}$ is the perceived signal, V is vision, H is haptic, x is the single-cue estimate, μ is the mean of the prior and σ is the variance of the cue. The top row shows the weighting of the cues when the sources are separate (C = 2), the bottom row when the both the signals are from the same source (C = 1).

$$\begin{aligned}
\widehat{S}_{V,C=2} &= \frac{\sigma_P^2 x_V + \sigma_V^2 \mu_P}{\sigma_V^2 + \sigma_P^2} \\
\widehat{S}_{H,C=2} &= \frac{\sigma_P^2 x_H + \sigma_H^2 \mu_P}{\sigma_H^2 + \sigma_P^2} \\
\widehat{S}_{V,C=1} = \widehat{S}_{H,C=1} &= \frac{\sigma_H^2 \sigma_P^2 x_V + \sigma_V^2 \sigma_P^2 x_H + \sigma_V^2 \sigma_H^2 \mu_P}{\sigma_H^2 \sigma_P^2 + \sigma_V^2 \sigma_P^2 + \sigma_V^2 \sigma_H^2}
\end{aligned} \tag{1.12}$$

If the probability of the cues originating from a common source is at the logical extremes of $P = 0$ or $P = 1$, then the CI model is, as shown above, the same as the basic weighted linear summation. In the event of P being a number between 0 and 1, the cues will instead be combined non-linearly.

There are several studies which have found that their results are well matched by the causal inference model (Hong et al., 2021; Rohe & Noppeney, 2015; Shams & Beierholm, 2010), though one of particular interest is a study by Takahashi et al. (2009). In their study, the authors aimed to investigate the sensory correspondence problem in regards to the visual and haptic signals present in tool-use. The study consisted of three interlinked 2AFC (two-alternative forced-choice) size-discrimination experiments aimed at comparing the detrimental effect of spatial offset between visual and haptic signals. In the first experiment, a spatial offset of ± 0, ± 50, or ± 100 mm was present between the visual representation of the hand and the haptic representation. In the second experiment, the spatial offset was the same as in the first one, but a rigid-link 'tool' was rendered visually alongside the representation of the hand so that the tool-tip end effector was rendered as touching the object in line with the offset. In their third and final experiment, they had an additional offset of the same range as in experiment 1 was added onto the spatially offset but visually rendered tool of experiment 2, but with the

length of the tool set as a static 50 mm; which is to say, the offset between the 'seen' and 'felt' was in the range of 50 mm ±0, ±50, or ±100 mm, with the initial 50 mm rendered visually as the size of the tool. Additionally, in order to control for and compare the performance of both available sensory cues, the JND units of the distance between the two planes were measured on a per-participant level, where the JNDs are defined as the standard deviation σ of the CG PF fit, when using the maximum-likelihood criterion. The estimated maximum improvement is predicted to be when the variances of the two sensory signals are matched, according to the principles in Equation 1.6. These JNDs were then used to estimate the 'ideal' improvement of size discrimination as expected under the statistically-optimal cue integration model outlined in Equation 1.6.

The results of the first experiment showed that, when the spatial offset was non-existent, observers performed near what would be expected from optimal cue integration, while larger spatial offsets are significantly worse, but still perform better than either single-cue conditions alone. This finding is as expected, based on previous studies on visuotactile spatial offset such as Gepshtein et al. (2005), who also found that humans combine signals optimally when signals came from the same spatial location. However, for the second experiment, the presence of the visual tool greatly improved performance in all levels of spatial offsets, though the improvement was not as large when the tool-length was ±100 mm, likely due to the 'unnaturally' large offset. However, the overall results strongly indicate that humans integrate the spatially offset visual and haptic signals optimally when perceived to be using a simple tool. The final experiment, which introduced spatial distortion in addition to the use of the simple tool, the effects of the spatial offset replicated the findings of their first experiment, where at larger offsets the

observers performed much worse than at no additional offset. The results of these three experiments are consistent with the hypothesis that the sensory system can integrate spatially offset visual and haptic signals by taking into account the geometry and dynamics of simple tools, when it is considered relevant to the task.

1.4.3 Sensory correspondence and medical imaging

As discussed in Section §1.1.4 and §1.4, it is fully possible to add sensory information otherwise absent to great benefit, such as training the use of vibrotactile wearables to allow blind individuals to perceive colour through torso-worn belt with haptic feedback and a glove with colour sensors (Maidenbaum, Arbel, et al., 2014). However, as these senses are not previously mapped, long periods of training is often required. If the sensory inputs are not sufficiently integrated, they will be either ignored by the sensory system, or they will add a detriment to performance, especially if feedback is given in the exploratory modality (Ho & Spence, 2014).

For the signals to be beneficially combined, the sensory system needs to be able to identify that the signals should correspond before it makes attempts at integrating the information from the signals. Typically, this starts with having the signals spatially and or temporally aligned (Holmes & Spence, 2005; Spence, 2011, 2013), but the signals still need to be perceivable as coming from the same source, which can be difficult for any difference between the signals. Even when using the same imaging modality as the source in both of the sensory modalities, there is the issue of naturally occurring sensory discrepancies such as texture roughness and size estimates (Marks, 2014).

Sensory dimensions do not always align perfectly, and as such it is possible

that a texture that is visually different in higher order statistical features may still be haptically metameric. In a study by Kuroki et al. (2019), it was shown that, while both vision and touch use lower order statistical features (such as the mean and standard deviation of the luminance of the pixels in the image) when categorising realistic textures, haptics was not found to share the same sensitivity as vision regarding changes in the texture's higher order statistical features. In a series of 2AFC experiments the authors found that, as expected, tactile perception was sensitive to low-order statistical differences, though unexpectedly, not higher-order statistical differences. The paper itself is discussed in more detail in Section §4.1.3, Chapter 4. While it would be possible to have textures within the medical images which were visually distinct, those textures may not be haptically distinguishable. Whether this is a positive or negative aspect depends on whether it would introduce textural discrepancy between the sensory modalities, such as the difference between an area rendered from CT compared to from MRI where the difference in density mapping could introduce incongruence, or whether the difference in available information between the scans is lost due to haptic texture perception's inability to distinguish these features. Which is to say, it needs to be accounted for and planned around. The different recorded densities may also lend themselves more, or less, to being displayed as height maps rather than density maps.

A difference also exists in how the imaging modalities themselves affect the information in the image; CT scans have highly accurate spatial *information* but poor level of detail for soft tissues, whereas MRI has a superb level of soft-tissue detail as well as great spatial *resolution*, but in turn it suffers from spatial distortion due to the magnetic field, making it difficult to align with a CT scan.

Of course one can attempt to adjust the distorted MR scan, as is commonly done (Seibert et al., 2016), but as the distortion correction that comes with some MR machines only works on 2D slices, there is something lost since the distortions are across the entire 3D object, and even when corrected for, the distortion itself exists in the raw data.

The suggested approach of using haptic normal-mapping displaying a synthetic height-map of the imaging modalities implies that the addition of a haptic signal will improve performance, while the approach of using different imaging modalities as the image sources for the sensory signals implies that there can be up to a certain amount of discrepancy allowed between the signals before the sensory correspondence breaks down and integration stops. This requires the sensory integration system to adapt a density map to a height map, to view a subsection of a 3D person in 2D while being aware of other anatomical features not present in the current image, to adapt to tool-use in a novel exploration task, and to combine these beneficially for an improved delineation of a tumour.

1.5 Overview of subsequent chapters

As the overall aim of this thesis is to investigate whether the addition of a haptic signal improves detection of tumours in medical imaging, three main experiments were set to each answer a different aspect of the query. The first experiment looks at the foundation of the research question, general cue combination of vision and haptics for signal detection in a spatially misaligned rig. Having established this, the second experiment investigates the effect of increasing incongruence between spatially coaligned visual and haptic signals, for the purpose of establishing

whether it would be possible to use different medical imaging modality sources for the respective sensory cues. The third and final experiment explores the effect haptic image rendering has on delineating a simulated 'tumour' embedded in a generated 'medical image', compared to drawing on a haptically flat surface, in a spatially coaligned rig. Together they answer the questions of whether a spatially misaligned set-up can still be beneficial, whether using different-but-similar signal sources negatively impacts performance, and whether haptically textured images can improve delineation over non-haptically textured images.

The second chapter of the thesis concerns the hardware and setup used and how this was calibrated and tested. Following this, the three experimental chapters, before finally a concluding Discussion chapter which discusses the overall contents and findings of the thesis as a whole. Additional supplementary figures such as photographs of kit and individual results for all experiments can be found in Appendix A and Appendix B, respectively.

Chapter 2 explains in detail all the hardware devices used, the two different experimental rigs used and the calibration procedure for spatially coaligned vision and touch. It also contains the different pieces of software used to run experiments, to create stimulus, analyse results and plot figures.

Chapter 3 explores the use of haptics and vision to locate a Gaussian blob embedded in a visually noisy background in a 2AFC signal detection paradigm, what effect haptic training has and whether the incidental auditory cue of the haptic device was utilised to a degree that significantly affected the results.

In Chapter 4 we investigate whether adding an incongruent visual texture to a haptic texture plane disrupts precision in a 2AFC slant discrimination task.

Chapter 5 looks into the usage of the haptic device as a tool and the height

rendering of synthetic medical images has an effect on accuracy and variability in delineation of generated 'tumours'.

The general discussion, Chapter 6 discusses overall results and conclusions of the research, while the final chapter, Chapter 7 discusses the overall contributions and implications of this thesis.

Appendix A contains photographs of the experimental rigs and the equipment used to calibrate the coaligned rig, while Appendix B contains the individual observers' plots for all experiments.

Chapter 2

Setup, calibration and general methods

In this chapter I will be going over the different pieces of hardware used throughout this thesis, where all devices are used for every experiment unless otherwise specified. Additionally, I will introduce the two different experimental rigs used for spatially misaligned and spatially coaligned visuohaptics, respectively, and detail the calibration procedure used for the coaligned rig. Lastly I will go over the main pieces of software used to run experiments, create stimuli, analyse results and produce plots. All relevant pieces of kit and software are also referenced in the individual experimental chapters where relevant.

Visual and haptic stimuli were rendered in OpenHaptics using C++ and a custom modified version of the open source CHAI3D frame work (Conti et al., 2003), where all visual stimuli were rendered with parallel axis asymmetric stereo frustums to give perspective correct stereo projection. To limit unwanted auditory cues, all participants were outfitted with ear plugs throughout testing (Moldex

7800 Spark Plugs). Additionally, prior to all experiments, the inter-ocular distance (IOD) was measured for each participant, and the individual IOD used to create the stereo projections on a per-participant basis.

2.1 Hardware and riggings

In the experiments run for this project we used two separate experimental rigs – one where vision and touch were spatially misaligned and one where they were spatially coaligned. Both rigs, and by extension all experiments, used the same hardware; a gamma-corrected 3D monitor and a stylus-tipped haptic device.

2.1.1 Hardware

In all experiments we present stereoscopic 3D visual stimuli using a VIEWPixx /3D monitor (Vpixx Technologies). The VIEWPixx /3D has a resolution of 1920x1080 on a 24" monitor, giving a DPI of 91.79 pixels per inch. Additionally, it has a refresh rate of up to 120 Hz, which when combined with the use of Nvidia 3D Vision shutter glasses allows temporally interleaved stereoscopic imaging to be presented to each eye at a frequency of 60 Hz. The shutter speed of the glasses is synchronised with the refresh rate of the monitor using blueline stereo and an Infrared (IR) emitter connected directly to the VIEWPixx /3D monitor. In order to ensure a linear relationship between the requested and displayed luminance of the visual stimuli, the monitor was gamma-corrected to a γ=1, using a Datacolor Spyder5Elite monitor calibrator, shown in Figure 2.1b.

The haptic device used throughout the project is the 3D Systems Touch X (formerly Geomagic Touch X, Geomagic Inc.), with 6 degrees of freedom (DOF)

positional tracking, 3 DOF haptic feedback, and a positional resolution of approximately 0.023 mm (greater than 1100 DPI), shown in Figure 2.1a. It has an available haptic workspace of 160 W * 120 H * 120 D mm, and a maximum force of 7.9 Newtons. It comes with full calibration software and out-of-the-box compatibility with OpenHaptics and subsequently the CHAI3D framework.

(a) Haptic device (b) Spyder calibration

Figure 2.1: (a) The haptic device in question, formerly the 'Geomagic Touch X', now simply the 'Touch X'. (b) The calibration device in question, Spyder5, which allows for fine-tuning of colour balance in monitors.

2.1.2 Spatially misaligned rig

The first experiment and the following controls found in Chapter 3 were all done on a pre-built, spatially misaligned rig. While the maximum benefit of an added haptic signal occurs when vision and touch are spatially coaligned precisely, the focus of this thesis is aimed towards the potential real-world use case of haptic feedback. As such, the spatially misaligned rig is a closer match to the realistic

configuration of a haptic device in a medical setting, which allows us to explore the benefit of haptic addition in this non-ideal physical setup shown in Figure 2.2. A photograph of the rig can be found in Appendix A, Figure A.1. Observers were placed in a chin rest 60 cm from the monitor screen, with their eye height adjusted to match the midpoint of the monitor screen. The haptic device was positioned between the screen and directly in front of the observer, the device being situated at 31.5cm from the centre of its base to the base of the chin rest. While the presented visual and haptic stimuli were misaligned, the scaling of the haptic device had a one-to-one movement mapping from the physical device to the virtual cursor.

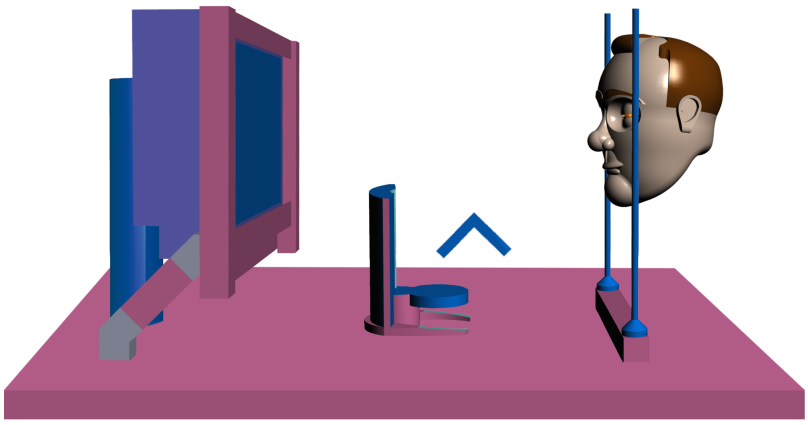


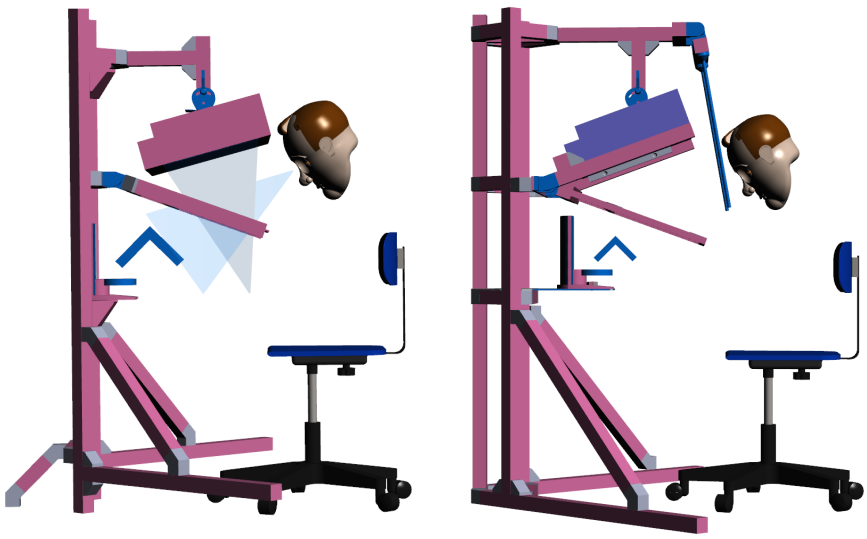
Figure 2.2: Diagram of spatially misaligned experimental rig. The participant is placed in the chin rest, regulated to keep all participants' eye height consistent. The haptic device is placed at a comfortable distance from the participant, at 31.5cm from the base of the chin rest.

2.1.3 Spatially coaligned rig

The spatially coaligned experimental rig is used in Experiment 2 and 3, and shown in Figure 2.3. A photograph of the rig can be found in Appendix A, Figure A.2. Observers were placed in a chin rest, with eye height was adjusted to match the midpoint of the reflected monitor screen. The haptic device was positioned directly in front of the observer under a front-surface silvered mirror, for spatially coaligned vision and touch. While the chin rest is not at a direct 90° angle to the viewing angle, the adjustable chin bar position was adjusted on a per-participant basis to keep their eyes in the appropriate position and the angle of the head as close to 90° as the chin bar position would allow. The exact measurements for this was calculated on a per-participant basis using the distance between the chin and eyes as measured on a different chin rest, this distance was then used to calculate the required height and depth of the chin bar trigonometrically.

2.1.4 Calibration

For the spatially coaligned experimental rig to be accurately representing the visual and haptic stimuli in a spatially coaligned manner, the physical properties of all components needed to be measured and the different coordinate frames used by the software, and by the external motion tracking system Vicon, needed to be fit and transformed into a single shared coordinate frame, quantifying any positional distortions near the edges of the haptic space. In order to ensure overall accuracy of the virtual representation compared to the physical components, this was done by first using Vicon to record the physical room-coordinates of the mirror and monitor, then by mapping the visual coordinates from the simulated Chai space



(a) Original coaligned setup (b) Final coaligned setup

Figure 2.3: The progression of the coaligned rig design. (a) shows the initial 3D design of the coaligned visual-haptic rig, showing the visual reflection of the visual signal overlapping with the haptic space. Here the monitor was suspended in-air and the chin rest not yet added. In (b), additional supports were added to protect the monitor and increase stability overall, as well as including a chin rest for controlling participant head position.

to the physical Vicon coordinates, before collecting the haptic coordinates in both simulated Chai space and physical Vicon space simultaneously. These coordinates were used to map the transformation from Chai coordinates to Vicon coordinates, before finally calculating a transform between the visual and haptic coordinates in Chai.

Vicon coordinates

In order for Vicon to accurately locate and track objects, the majority of the five Vicon cameras must be able to detect the markers used. Immediately this raises an issue, as a large part of the area of the rigging that requires careful calibration is a

partially enclosed space with limited visibility from above. In order to accurately collect several points on the coaligned rig, we require a vector that can convey the physical positions obscured to the Vicon system, for the purpose of which a 'vector stick' was created, illustrated in Figure 2.4. For a photograph of the vector stick see Appendix A, Figure A.3, A.4 and A.5.

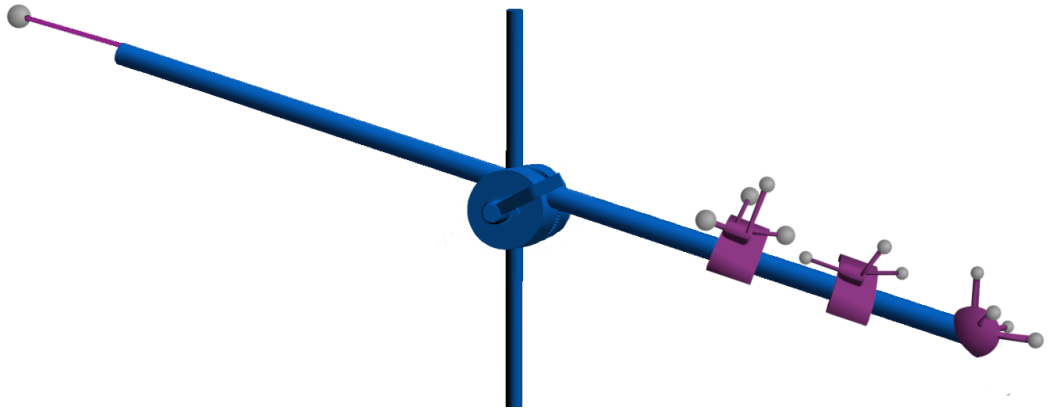


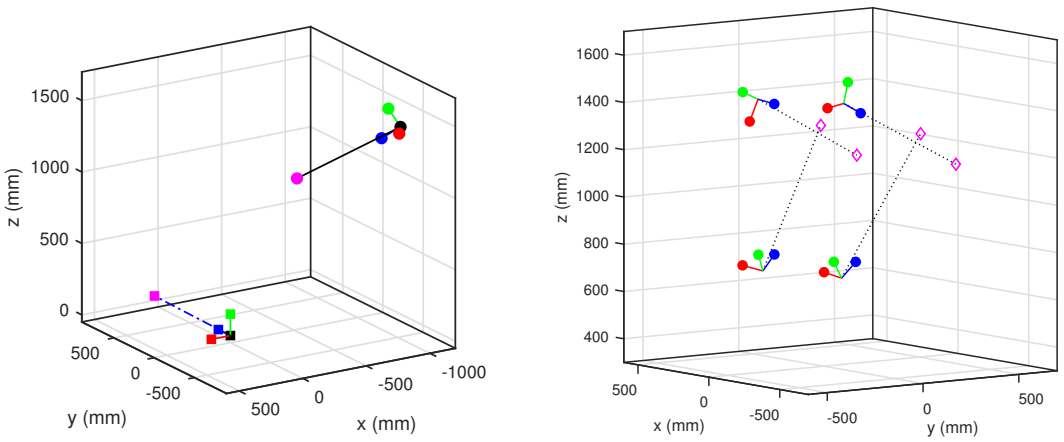
Figure 2.4: Vector stick. On the bottom right are three objects that together form the grouped 'Caterpillar' as tracked by Vicon, and to the top left is the precision point unlabelled marker. After calibrating the length and relative position of the vector stick the top-left marker is removed, leaving the tip of the rod to be used to mark the locations of points being measured.

A Vicon 'object' was placed at one end of a long stick which was mounted on a tripod, while at the other end of the stick was placed a thin metal rod onto which a single unlabelled marker was affixed. This would form the basis of the vector. The Vicon Object is tracked using translation and rotation, where the rotation uses Quaternions – which are a representation of rotation of either a vector or coordinate frame, depending on the viewpoint. The exact position and rotation of the object is recorded by the Vicon tracking software. Having polled the position of the single marker and the 'Caterpillar' object simultaneously, the position of the vector stick's tip can be extrapolated for any given point in space, based on the

position and rotation of the Vicon object at its base. By using this method, the vector stick acts like a vector transformation, allowing for the accurate and precise polling of specific points on the coaligned rig normally obscured from the Vicon cameras. As the unlabelled marker was positioned where the vector stick would be in contact with the surface of the screen and the mirror, the relative position of these points in space could be recorded and calculated to millimetre precision. After first running an initial calibration the Vicon tracking system itself, a quick Matlab routine was run to measure the length of the vector stick, by method of tracking the Vicon 'Caterpillar' object and the relative position of the tip itself compared to the 'Caterpillar', which were then used to calculate the normalised vector distance between the two, as shown in Figure 2.5a. Once this has been done, the marker was removed from the tip of the vector stick and a number of measurements were taken of the various segments of the rig, as shown in Figure 2.5b which shows the collection of the corners of the monitor. All Quaternion calculations were performed using the Robotics toolbox for Matlab (Corke, 2017).

Fitting the points

A basic representation of the points used in the fitting procedure is shown in Figure 2.6a, where the physical monitor (blue plane) reflects off the front-silvered mirror (orange plane), to the reflected monitor position (teal plane). Above the physical monitor is the edge of the frame for the monitor and the additional distance of the attached chinrest, from which the observer angle (black) projects downward. The ideal gaze vector (magenta line) starts from the centre of the reflected monitor, and extends towards the observer angle-vector, where the point of intersection between the vectors is highlighted as a red diamond with a green border. In the initial



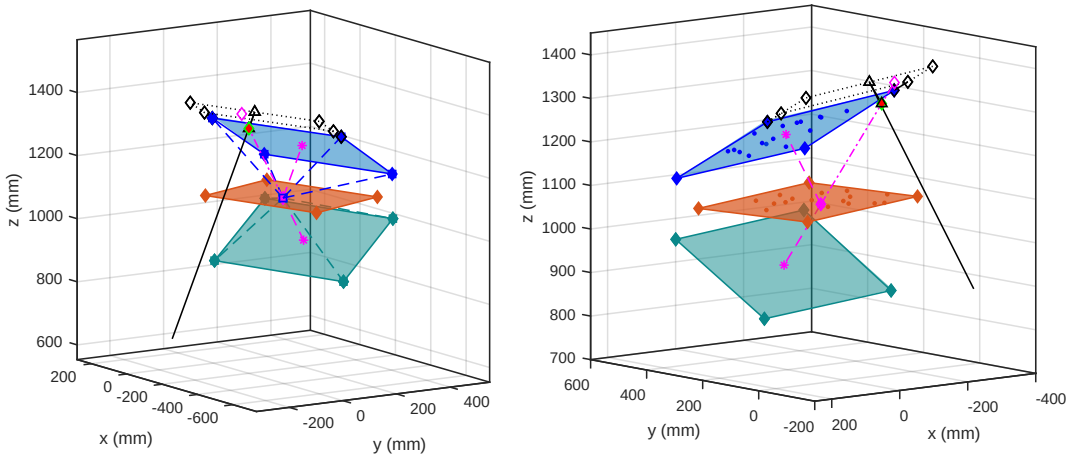
(a) vector stick relative transform

(b) Example point collection

Figure 2.5: (a) shows the calculated relative positions in space, where the top-right shows the actual recorded position of the 'Caterpillar' object in relation to the reflective tip (magenta sphere). By inversely transforming this to get the relative position of the tip in space, the extrapolated tip is shown as the magenta square in the bottom-left, with the 'Caterpillar' Vicon object as the origin. (b) shows an example of how the vector stick is moved around to perform the point collection, where the each of the four coordinate clusters (RGB standard of basis, spheres) points to a specific, extrapolated point in space as represented by the dotted line leading to the magenta diamonds.

calibration procedures, the four corners of the mirror and monitor were collected first and the normalised vector distance between the points was compared to the known physical dimensions of the objects in real-world coordinates, doing a total of six comparisons – the first four being between the four corners (1-2, 2-3, 3-4, 4-1), and the final two of the two diagonals (1-3, 2-4) of the plane. These additional diagonal comparisons allows for the verification of the angles between the collected points being approximately 90°, as the four corner comparisons themselves could still be correct for a parallelogram rather than a regular rectangle. Once all six comparisons were within 5 mm tolerance of the physical dimensions, we proceeded to collect an additional 15 points used to fit the surface normals to the screen and the mirror, where a minimisation function was used to fit the surface normals to

the series of points. These additional points are illustrated in Figure 2.6b, where the additional points are marked on the surfaces of the mirror and the physical monitor, respectively.



(a) Core points and reflection lines (b) Additional surface points

Figure 2.6: (a) shows the core points of the 3D representation of the rig as points recorded from the vector stick in Vicon space, complete with reflection lines for how the monitor (blue) reflects off the mirror (orange) to create the virtual screen (teal). The X, Y and Z axes show the relative position in the room as described in Vicon coordinates, in mm relative to room origin. while (b) shows all the points collected on the surfaces of the mirror and monitor, respectively, which were then used to calculate the surface normals of the mirror and the monitor.

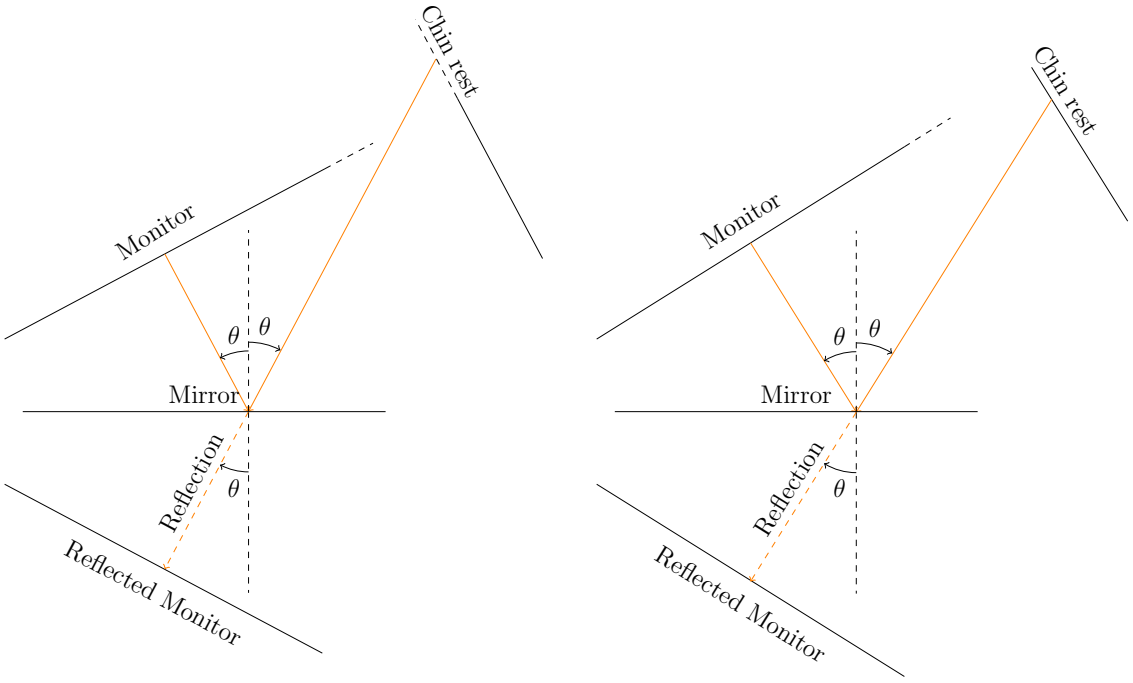
The surface normals of the mirror and monitor were then used to calculate their respective incident angles, using a combination of linear algebra, Pythagorean theorem and trigonometry. Following this, the incident angle and the normal of the reflected monitor were used to calculated the required viewing angle of an observer. By following along the line of the reflected normal, while working with the known distance between the monitor screen to the monitor edge and the distance between the chin rest and the monitor, the ideal position of the eyes along the viewing normal was calculated trigonometrically. The position of the

eyes along the gaze line was added to the distance metric between the reflection point of the mirror and the centre of the reflected monitor to find the focal distance used for rendering things visually that align perfectly with physical space, for the spatially coaligned visuohaptic signals.

Adjusting the angle

For the initial calibration, each of the four corners of the monitor and mirror were collected and the distance between the eight points were calculated in Matlab using a vector translation script. As the dimensions of the monitor and mirror were known, the physical dimensions were used as a ground truth to compare against. Once the script produced the expected fit of the points, the vector stick was modified for increased sturdiness and improved accuracy, and new and improved measurements were taken. These measurements found that the initial angle between the mirror and monitor would produce an unfeasible position for the observers' eyes at the correct angle, as shown in Figure 2.7a.

Here, the view of an observer would have been obscured by the case of the monitor or unless they were positioned physically within its casing. To correct this, the angle between the monitor and mirror was increased in order to lower the required position of the eyes, as shown in Figure 2.7b where the chin rest and observer is comfortably below the top edge of the monitor. After readjusting and testing the angle, a series of new measurements were taken and controlled for, and the position of the eyes was calculated and visually verified using a laser-emitting digital level on a tripod, a photo of which is available in Appendix A §A.6. These measurements provided the reference points of the cyclopean eye, and the resulting focal distance was calculated trigonometrically allowing the setup of



(a) Initial 28 degree angle

(b) Improved 32 degree angle

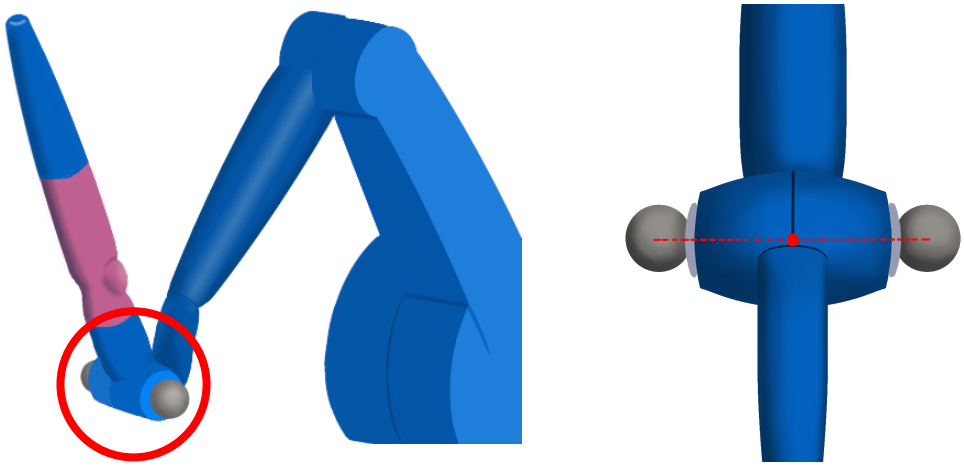
Figure 2.7: Initial and final angles of the spatially coaligned rig, where the ideal eye position in (a) would place the head of the observer partially inside the monitor case, which is corrected for in (b)

the asymmetric frustum used to render the separate views per eye in stereo vision.

Haptic coordinates

Using the same Vicon calibration and physical position of the coaligned rigging, a series of positions were collected for the haptic device as well, using both Vicon markers as well as the CHAI3D relative positioning of the haptic point to calculate the required transformation matrix between the coordinate systems to allow for a fully coaligned visuohaptic experience. The calibration of haptic space to real-world space was done by placing two Vicon markers on the 'barrel' of the haptic device (shown in Figure 2.8), importing the coordinates of these markers

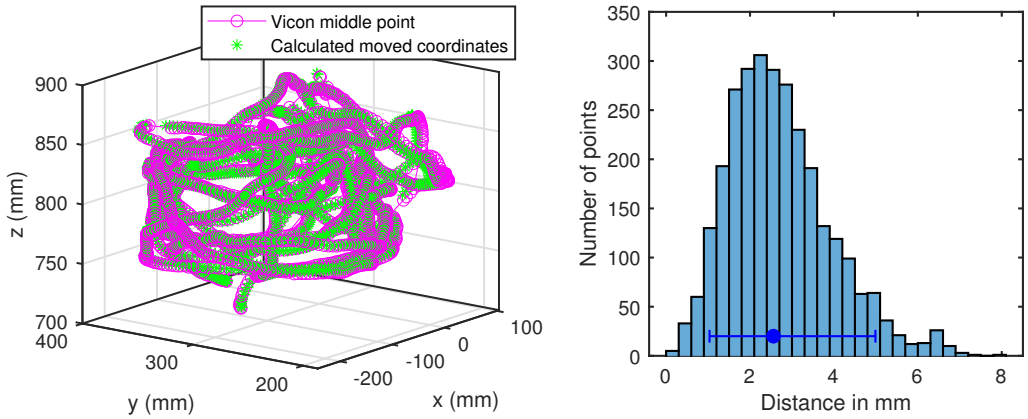
from Vicon into Chai while concurrently saving out the relative haptic coordinates from Chai. From these coordinates, any frames where both Vicon markers were not present were discarded from both the Vicon set and the haptic data set. The remaining frames were used to first calculate a theoretical middle point between the two Vicon markers, indicating the centre of the haptic device barrel and by extension the theoretical haptic point. Having matched the frames of both coordinate sets, it was possible to create a rigid transform of rotation and translation to ensure a high-precision spatially coaligned fit, as shown in Figure 2.9a. For this we used the 'fminsearch' function, which uses a root-mean-square (RMS) optimisation routine over the 'fminsearch' routine in MATLAB, to find the best-fit transformation between the two data sets. This transformation of the haptic coordinates compared to the calculated Vicon middle point gave an $e_{RMS} = 2.54 \pm (0.05)$ mm, and is shown in the histogram in Figure 2.9b.



(a) Haptic device barrel (b) Vicon markers

Figure 2.8: (a) Illustration of the positioning of the Vicon reflective markers on the effective 'tip' of the haptic device, which is roughly shaped like a barrel. (b) The two Vicon reflective markers are placed on the centre of each of the sides of the round 'barrel' structure of the haptic device, allowing us to calculate the centre point between the two tracked markers, which coincides with the virtual haptic interaction point.

The calibration of the rig was done by first using a half-silvered mirror where the front-surface mirror would later be placed. This allowed us to perform a visual check of the calibration by comparing the reflected monitor image with the real-world position of the haptic device. Figure 2.10 shows the coaligned rig complete with the transformed haptic coordinates in the Vicon room-coordinate frame.



(a) Calibration Traces (b) Trace Accuracy

Figure 2.9: (a) The 3D representation of the scaled haptic space from Chai3D (green) compared with the real-space of Vicon (magenta), where the Chai3D points were originally in metres and centred around the haptic device, but were then scaled to mm and their origin point shifted in space relative to the Vicon mapping of the position of the physical haptic device. (b) The histogram of errors for the 'fminsearch' transformation fit between the haptic coordinates and Vicon coordinates across the entire haptic workspace, with error bars showing 95% confidence interval.

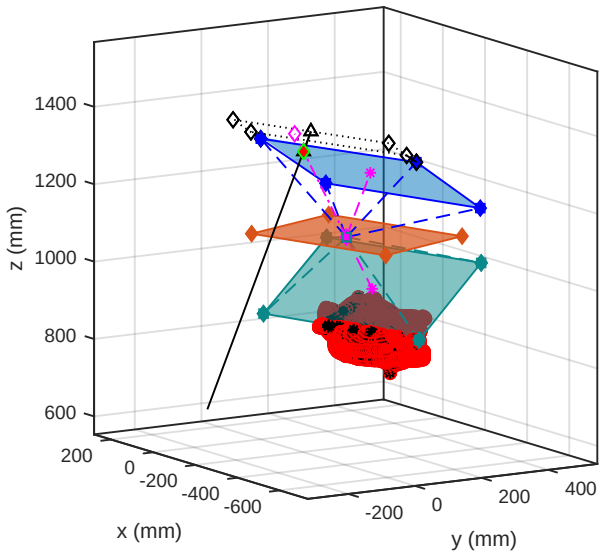


Figure 2.10: Illustration of the coaligned haptic and visual coordinates in the same physical coordinate frame. After having found the transform between the visual coordinates for the monitor, and the physical coordinates from Vicon, the haptic coordinates (red and black cloud) could be transformed into the monitor's visual coordinate frame, presented here in the Vicon recorded physical room coordinates.

2.2 Software

2.2.1 About Chai

CHAI3D is an open-source C++ framework for haptic rendering using OpenGL (Conti et al., 2003). The version of CHAI3D used throughout this project is 3.1.1. While the framework itself offered a great deal of the required functionalities, there were some additional modifications done to fully tailor it to the project. While a newer version of Chai, version 3.2 was released during the span of this project, there were issues with how the modified height-mapping algorithm changed direction of the normal maps depending on the exploration direction of the haptic device. Namely, on a plane with a 2D Gaussian 'bump', in version 3.1.1 there would be a smooth gradient from all directions, where the centre of the Gaussian was the rendered 'top'. The same image in version 3.2 would on the right-to-left exploration render the Gaussian as the 'top', on a top-to-bottom exploration the Gaussian was rendered as a 'bottom'. If switching directions mid exploration, as required by our experiments, it would rapidly switch from rendering it as a 'bump' to a 'dip' – which was considered unusable for our purposes and we kept with version 3.1.1.

Modifying the Chai framework

The default CHAI3D source code stereo frustum did not allow objects to be rendered in the plane of the monitor, nor at negative parallax, so a small modification was made to the camera function code to allow this. In the original code, the left eye offset vector[1] was multiplied by -1.0, while no modifications were made to

[1]The offset vector is the cross product of the up-vector and the look vector.

the vector for the right eye. The modification subsisted of multiplying the right eye offset vector by -1.0, and the left-eye offset vector by 1.0. This allowed for geometrically correct stereo projection, with zero parallax occurring on the plane of the monitor, with the ability of rendering objects as both in front of and behind the plane of the monitor, depth-wise.

After calibrating to coaligned visual and haptic space using Vicon and Chai, there was a notable offset of the depth-position (Z-axis) of the haptic device. Comparing the physical position to the rendered position revealed a 10 cm offset within the libraries of the haptic device itself. The source code has for a different device from the same manufacturer added a negative 10 cm depth positioning to the rendered position, and copying this method allowed the rendered position of the haptic device to line up perfectly with its physical position in space. This was confirmed using the physical size versus the visually rendered size of the affixed Vicon-markers, which were of a known size and comparable due to the partial transparency of the half-silvered mirror.

Lastly, in order to display one texture visually and render a different texture haptically, a quick modification had to be made to the stock CHAI3D code. This added an optional second texture to objects, which could be used to render a haptic texture that was separate from the visual texture. This texture modification was used in Experiment 1 (Chapter 3) and Experiment 2 (Chapter 4).

2.2.2 Stimulus creation and data handling

All of the in-house stimulus generation and the different function fits, boundary fits, texture similarity comparisons and model comparisons were done using MATLAB (*MATLAB version 9.4.0.813654 (R2018a)*, 2018). The psychometric

function fits and model comparisons were all done with the Palamedes toolbox for Matlab (Prins & Kingdom, 2016), while the realistic medical images used in Chapter 5 were made using the textureSynth toolbox (Portilla & Simoncelli, 2000). For more details on specifics for each experiment, see the methods section of the respective experimental chapters.

2.2.3 Statistical analyses and figures

The statistical analyses were run in JASP, a state-of-the-art open-source statistical software program (JASP Team, 2019). Most of the data-containing figures in the thesis were made in Python 2.7, using the matplotlib (Hunter, 2007) and the Seaborn data visualisation packages (Waskom et al., 2018). All of the geometrically correct diagrams, such as the slant illustrative ones, the angle of the rigging ones and the n-dimensional vectors were made in LaTeX using the TikZ package (Tantau, 2013). The remaining figures are either screenshots and photographs, realistic textures from the Edinburgh PerTex texture database (Halley, 2012), or generated in Matlab.

Chapter 3

Experiment 1

3.1 Introduction

As discussed in Chapter 1, the task of locating abnormal tissue is effectively a signal detection task where trained radiologists attempt to locate a signal, in this case the abnormal tissue, obscured by noise, here being the surrounding healthy tissue. One of the main complications in the field is the inability to know the ground truth – as tumours are organic in nature, and the boundaries between healthy tissue and tumours are between indistinct at best and indistinguishable at worst. Since it is easier to manipulate aspects of the stimuli and make inferences on how these manipulations affect perception when the stimuli are more simplistic, simulated stimuli were chosen for investigating whether the additional haptic signal is beneficial to the observer. As the core aspect of the research question is looking at basic signal detection we condense the issue down to a very basic stimulus; a 2D Gaussian embedded in a noisy background, similar to the works of Abbey and Eckstein (2009). By simplifying the stimulus from a realistic,

highly detailed tumour into a basic 2D Gaussian we are able to explore the novel task with highly controlled variables, while the noisy background is reduced from healthy tissue to Gaussian white noise.

In this experiment we explore whether adding haptics increases the accuracy of people's performance for detection tasks using visual stimuli similar to the stimuli used in Abbey and Eckstein (2009), a 2D Gaussian embedded in white noise. In order to test this, we collect psychometric functions for two different modalities, being haptics and vision, both separately as single cues and when both are available simultaneously as combined visuohaptic. For both vision-only and combined visuohaptic, the visual performance is manipulated using four different visual noise levels, which provides us with a range of visual performances to which we match the haptic performance level (Ernst & Banks, 2002; Rohde et al., 2016; Takahashi et al., 2009). This matching allows us to compare individual participants' single-cue condition with their respective combined-cue condition, and investigate the specific improvements across a short range of different visual noise additions. As the haptic-only condition was to be used as a single, constant comparator, no noise was added to the haptically rendered stimuli.

3.1.1 Stimuli

Observers viewed two image patches of Gaussian noise (361*361 pixels) and had to decide which contained a 2D Gaussian hidden signal. The stimuli were generated in Matlab using the formula given in Equation 3.1, to create an image size of 361*361 pixels, equivalent to $9.998cm^2$ on the monitor screen.

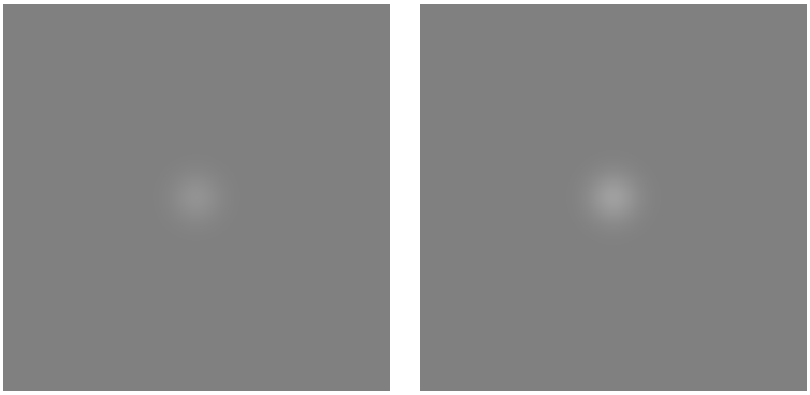
$$G(x, y) = c * e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{3.1}$$

Where σ is the standard deviation of the Gaussian, here set to be 3/64 of the width and height of the image, in order to keep the proportions the same to that used by Abbey and Eckstein (2009).

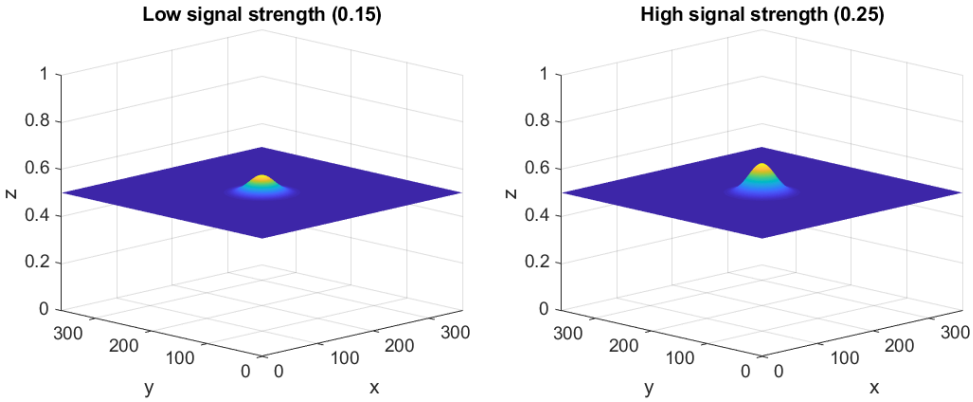
In order to map out a psychometric function the signal strength of the stimulus was varied to measure detection thresholds. The variable c is used to create a range of signal strengths, being image luminance values set in the unit interval, i.e. between a minimum of 0 (black) and maximum of 1 (white). The resulting Gaussian signal was then further scaled to change the background from black to grey, by halving the signal and adding the desired base level, using the following: $0.5 * G + 0.5$, where G is the strength of the Gaussian signal. There were 11 different signal strengths within each set, linearly spaced values between the highest signal strength and the lowest, which was set to be 1/128, the lowest detectable signal strength for an image with a bit-depth of 8 bits, as limited by the rendering framework CHAI3D (Conti et al., 2003). The resulting 2D Gaussian height differences are illustrated in Figure 3.1.

3.1.2 Cue conditions

The task consisted of testing three separate cue conditions: vision-only, haptic-only and combined visuohaptic. In the combined visuohaptic condition the signal strength of the Gaussian was the same for both the visual signal and the haptic signal. In order to keep the cues consistent, both the visual and haptic signals were created to the same signal strength and standard deviation of the 2D Gaussian. To ensure a reasonable match of the relative reliability between haptics and vision we modified the haptic scale factor, a variable in the CHAI3D framework that sets the relative size of the haptic steps, on a per-participant level where needed. This



(a) 2D Gaussian, weak (b) 2D Gaussian, strong



(c) 3D Gaussian, weak (d) 3D Gaussian, strong

Figure 3.1: Example Gaussian signal strength difference. (a) shows the 2D representation of a weak Gaussian, (b) shows the 2D representation of a strong Gaussian. (c) shows the 3D representation of a weak Gaussian, while (d) shows the 3D representation of a strong Gaussian. The x-axis shows the width of the image in pixels, the y-axis shows the relative strength of the signal on a scale from 0 (black) to 1 (white), 0.5-1 after scaling for a grey background. To the left, (a) and (c) have a signal strength of c=0.875. To the right, (b) and (d) have a signal strength of c=0.95. The top, (a) and (b) shows the signal images as would be presented in the experiment, while the bottom row, (c) and (d), illustrates the relative height of the Gaussian signals after scaling for signal strength.

ensured the magnitude of the haptic and visual signal were in the same range. After creating the Gaussian on a neutral background, as used for rendering the haptic stimuli, the visual stimuli were further processed to add Gaussian White

Noise with a standard deviation σ of 0.05 to 0.20, as shown in Figure 3.2. For each signal-containing image, a matching signal-absent neutral image was also produced. The noise was added using Matlab's 'imnoise' function, which adds a Gaussian white noise, for an average of 0 and a variance of σ^2. The haptic stimuli were produced using the same algorithm as the visual ones, and were exported prior to the addition of noise. Also, in order to maximise the available haptic resolution as limited by the rendering framework, they were created to double the image size of the visual images, then shrunk to present at the same physical dimensions as the visual stimuli.

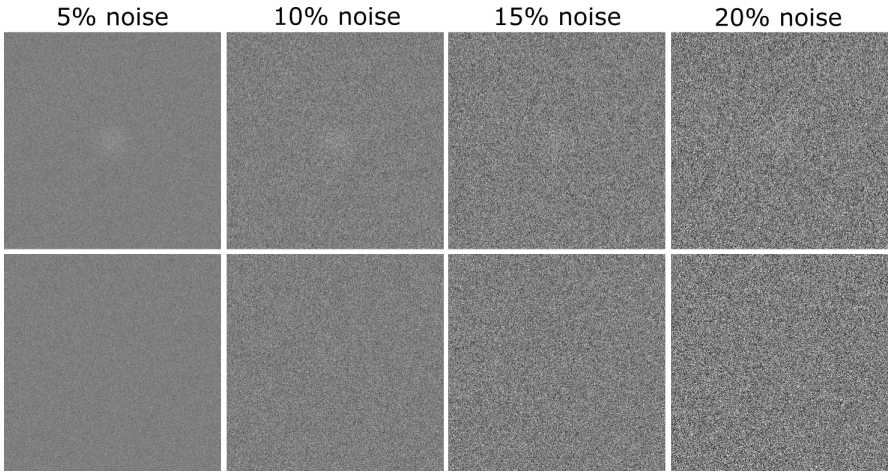
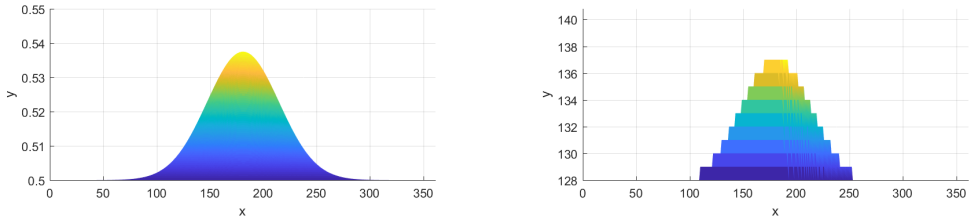


Figure 3.2: Signal present/absent pairs for the visual stimuli noise levels, with a signal strength of c=0.1. A subset of participants completed an additional noise level of 0% Visual noise, but this noise level was dropped due to 0% noise having a prohibitively high accuracy level compared to the other noise levels in all 5 tested participants, where the 0% noise level accuracy was magnitudes higher than the remaining four noise levels.

All the stimuli were produced in Matlab and exported as PNG image files, which were then loaded into the programme, and rendered haptically using OpenHaptics and visually using OpenGL, using the CHAI3D framework. As CHAI3D only supports PNG images with a resolution up to a maximum of 8 bits, the hap-

tic resolution of the stimuli is limited to 256 discrete levels. CHAI3D proceeds to use the luminance of the individual pixels in the image to encode the 3D haptic height levels. A representation of the resulting signal can be seen in Figure 3.3. As the haptic stimulus is explored by a stylus, and rendered at an 8 bit depth, it tends to feel more like a gentle grating than a bump per se, as described in Unger et al. (2011).



(a) Smooth Gaussian

(b) 8-bit Pixelated Gaussian

Figure 3.3: Calculated versus haptically rendered Gaussian signals. (a) shows the initial, smooth, mathematically calculated Gaussian signal as created in Matlab, while (b) shows the discrete, haptically rendered Gaussian signal CHAI3D presents, recreated from the luminosity levels in the 8-bit PNG (a) was exported to. The y-axis shows the relative height of the Gaussian, while the x-axis shows the image width in pixels, while the colours are indicative of Gaussian strength (y-axis) as rendered by Matlab. In (a) the y-axis is on a scale from 0 being black, to 1 being white. In (b) it is in pixel luminance, where 0 is black and 256 is white.

3.1.3 Procedure and task

The physical setup used is the spatially misaligned visuohaptic rig as described in detail in Chapter 2 §2.1.2. The primary task the participant will be doing is to explore two planes, which are presented side-by-side, as shown in Figure 3.4. One of these planes contains the hidden signal. The participant will then select which plane they believe the signal is in, in a simple 2AFC task. For each signal strength there are 40 iterations, and for each psychometric function there are 11 different

signal strengths, giving a total of 440 data points per psychometric function. There were a total of nine psychometric functions collected per participant: one for the noise-less haptic-only condition, and four each of the vision-only condition and of the combined visuohaptic condition, being one per visual noise level.

As the visual and haptic sensitivities differ between individuals, we need an appropriate difficulty range to measure a good psychometric function. For a 2AFC signal detection task, this ranges between chance (0.5) and always correct (1). Before starting data collection, participants did a shortened version of the main signal detection task which used a binary search algorithm to adjust the levels of signal strength relative to the observers performance. This gauging task was performed per noise level, per participant, allowing us to set an appropriate range of stimulus strengths on a per-individual basis.

Once the difficulty ranges have been established, the main part of the experiment began. The stimulus presentation order was modelled after that of Ernst and Banks (2002), where single cue and combined cue stimuli are presented in a counterbalanced order to reduce order effects, as shown in Table 3.1. First, the participants collect half of the single-cue functions, in an order randomised per individual. At the beginning of each new block a single signal-containing image, at the maximum signal strength for that block, was presented to allow the participants to familiarise themselves with the specific stimulus they were searching for. This was not time limited, and once they were ready to proceed, they were asked to select the plane by touching the haptic device to it and clicking the button on the device. This procedure is used throughout the experiment to select the perceived signal-containing plane. In Figure 3.4, it is the left-most plane. Whether the signal containing plane was presented on the right hand side or left hand side,

as well as the order in which the signal strengths were presented, was randomised on a per trial basis.

After collecting all the initial single-cue functions, the haptic performance was compared, and adjustments made to the scale factor if needed and the participant retested for the haptic-only cue. If the participant performed much better in haptic-only compared to their visual performance, the haptic scale factor was reduced, while if the visual performance greatly exceeded the haptic-only performance, the scale factor was increased. The default scale factor was 0.8, with the minimum and maximum used by individual participants were 0.56 and 1.2, respectively. Any previous haptic-only data were excluded from analyses. Once set, the haptic scale factor remained constant throughout all relevant conditions for the respective participants. Then they collect all the combined-cue functions, in a randomised-per-individual order. Finally, the remainder of the single cue functions are collected, again in a randomised-per-individual order.

3.2 Experiment 1

3.2.1 Methods and Procedure

Participants

A total of 10 observers participated in this experiment, of which 7 were naïve to the purposes of the experiment. All observers were right hand dominant, and had normal or corrected to normal vision, and all but one of the participants had a stereoacuity of 60 arcsec or better, as measured with the Laméris Ootech TNO Stereo Vision test. One participant (S8) could not do the stereo test at all, giving

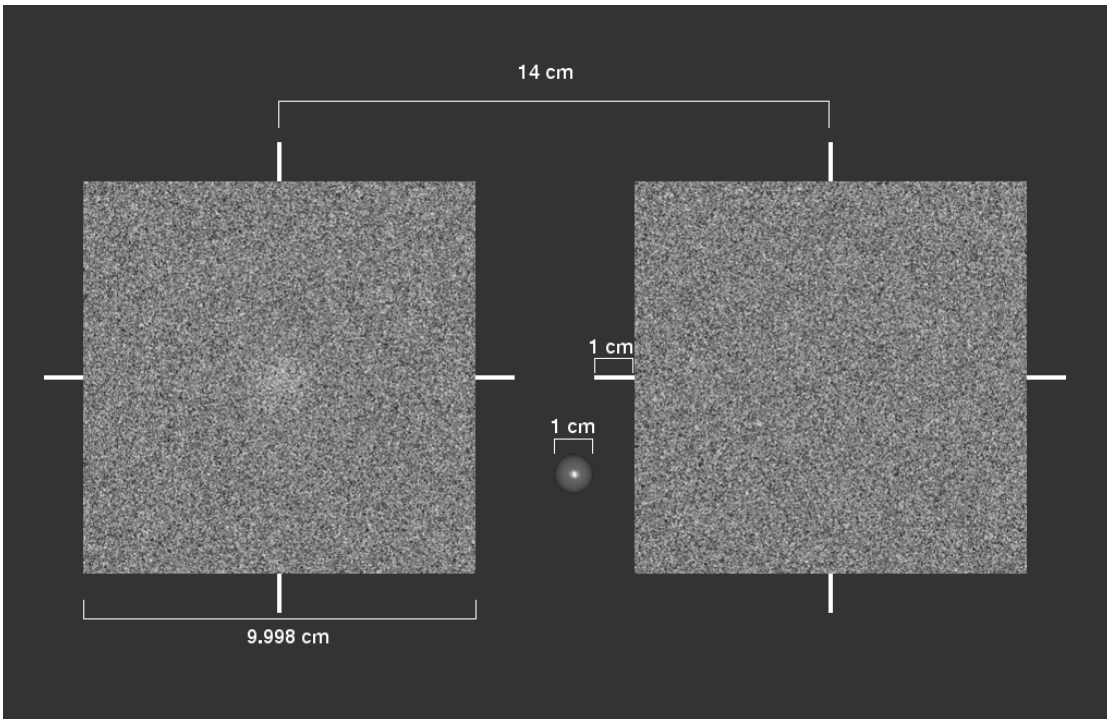


Figure 3.4: Screenshot of the experimental screen, presented with 20% visual noise. The planes are presented side by side, equidistant from the centre of the monitor. The sphere is the visual representation of the haptic device, and the Gaussian signal is present in the left image.

them a theoretical score of 480 arcsec. They did not report any issues with stereo and could see 3D appropriately in the task and in several different demos. As such we included them. When to reject participants on the basis of stereo tests is considered an open question. Heron and Lages (2012) suggest that as stereo performance is task dependent, it is important to screen and report the stereopsis results, but not necessarily use it as an exclusion criteria for an experiment, as while an observer might fail a stereoacuity test, they might fully well be able to perform the task with the actual experimental stimuli.

The first five participants completed a total of 22 blocks, comprising two haptic blocks, two vision only blocks per noise level and two blocks of combined

visuohaptic for each visual noise level. Each block contained 20 iterations of 11 different stimulus levels, resulting in 440 total trials per individual condition, 4840 total. However, the 0% noise condition gave a prohibitively high-precision function, with the size of the slope being magnitudes higher than the other noise conditions for all 5 tested observers, and any signal present was met with a 90% correct performance, so this condition was subsequently dropped for the remaining participants. The remaining five observers completed a total of 18 blocks, comprising two haptic blocks, two vision only blocks per noise level and two block of combined for each visual noise level. Each block contained 40 iterations of 11 different stimulus levels, resulting in 440 total trials per psychometric function, 3960 total.

The visual blocks were all carried out in a single session, where each block consisted of 20 iterations per stimulus level, while the haptic only condition was broken down into 4 smaller blocks of 5 iterations to prevent fatigue as haptic exploration takes longer. The combined-cue condition was subdivided into 2 smaller blocks of 10 iterations, but both blocks were grouped, effectively creating a superset of 4 subblocks of 10 iterations, rather than the two sets of 2 subblocks of 10 iterations. Each of these sub-blocks took between 6 and 20 minutes to complete, with individual variation. Participants were encouraged to take breaks when needed, and each test session ran on average between 60 and 120 minutes. As the haptic device occasionally emitted a scraping sound when exploring certain textures, all participants were outfitted with ear plugs throughout testing (Moldex 7800 Spark Plugs) to limit the effect of spurious auditory cues from the haptic device.

First group, single cue	Second group combined cues	Third group, combined cues	Fourth group, single cue
Vision-only, 5% noise	*Combined, 5% visual noise*	*Combined, 5% visual noise*	*Vision-only, 5% noise*
Vision-only, 10% noise	*Combined, 10% visual noise*	*Combined, 10% visual noise*	*Vision-only, 10% noise*
Vision-only, 15% noise	*Combined, 15% visual noise*	*Combined, 15% visual noise*	*Vision-only, 15% noise*
Vision-only, 20% noise	*Combined, 20% visual noise*	*Combined, 20% visual noise*	*Vision-only, 20% noise*
Haptic only			*Haptic only*

Table 3.1: Block group order of Experiment 1.0, excluding the 0% noise condition. Each cell represents a block of 20 trials per stimulus level, where vision-only was always collected in single sets of 20, while due to time and fatigue associated with haptic exploration, the haptic-only condition was collected in four sets of 5 per cell, and combined cues was collected in two sets of 10 per cell. During testing, the order of the presented noise conditions were randomised on a per-person basis.

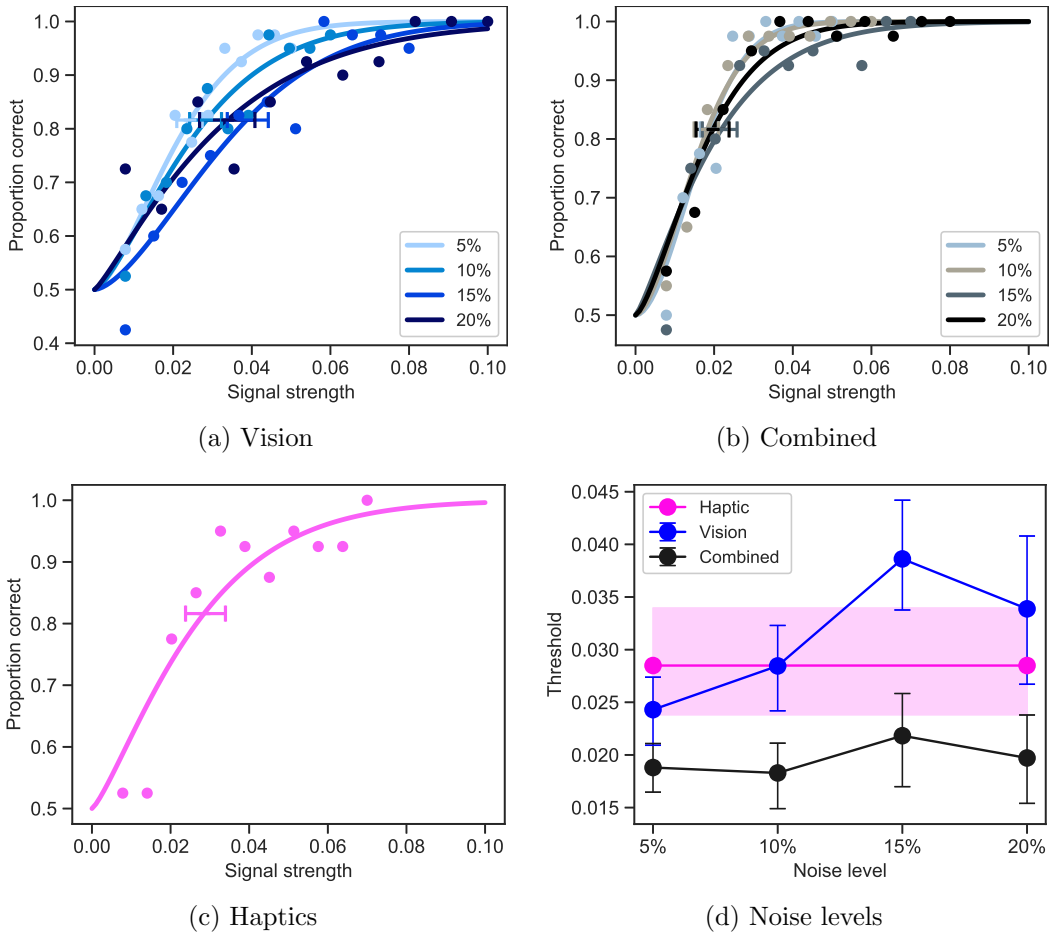
3.2.2 Results

Weibull functions were fitted to observers' data with a maximum likelihood criterion using the Palamedes toolbox (Prins & Kingdom, 2016). The threshold of the function, defined as the 81.61% inflection point of the psychometric function, was estimated and a bootstrapped 95% confidence interval was calculated around this value. Figure 3.5 shows the dataset of a single participant, plotting the measure of performance in the three respective conditions. Figure 3.5d plots the thresholds of the Weibull fitted psychometric functions on the y-axis, against the respective noise levels on the x-axis. The haptic performance is uniform across all condi-

tions due to its lack of visual signal and thereby lack of visual noise. Example functions fits for Observer S1 are shown in Figure 3.5 (a) through (c). As can be seen in Figure 3.5a, the addition of noise to the stimuli successfully degraded observers' performance in the vision-only condition. The accuracy of haptic performance was successfully manipulated on an individual basis to be comparable to the mid-range visual noise level condition. Mean results in the same format as 3.5d are shown in Figure 3.6. As seen in Figure 3.5d and 3.6, the thresholds were consistently lower for the combined-cue condition compared to either of the other single cues. From this we can see that by adding Gaussian white noise to the visual stimuli, the 81.61% performance point was successfully shifted to vary the participants' level of accuracy, allowing us to compare the relative performance of haptics to vision and combined visuohaptic.

A two-way repeated measures ANOVA was run on the threshold values of the Weibull psychometric fits, comparing the performance of vision only and combined vision and haptics, for the four respective noise levels. A Mauchly's test indicated that the assumption of sphericity was not violated for Noise, $W = 0.266, p = 0.071 > 0.05$ nor for the interaction of Condition* Noise, $W = 0.425, p = 0.254 > 0.05$. The ANOVA used the within-subjects factor of Condition (Vision, Combined) and visual noise levels (5%, 10%, 15% and 20% noise), results shown in Table 3.3 with descriptives shown in Table 3.2. The analysis revealed a significant main effect of Condition $F(1, 9) = 37.799, p < .001, \epsilon^2 = 0.410$, a significant main effect of Noise $F(3, 27) = 24.94, p < .001, \epsilon^2 = 0.319$, and a significant interaction effect between Condition and Noise $F(3, 27) = 19.67, p < .001, \epsilon^2 = 0.104$.

In order to investigate the interaction, the data was split into the vision-only and combined conditions, on which another repeated measures ANOVA was run



(a) Vision (b) Combined

(c) Haptics (d) Noise levels

Figure 3.5: Dataset of a single participant S1, plotting the Weibull fits of proportion of correct responses over increasing signal strength, in the three respective conditions: (a) Vision-only, (b) Combined visuohaptic and (c) Haptic-only. (d) Thresholds of the Weibull fitted psychometric functions on the y-axis, plotted against the respective noise levels on the x-axis. The haptic performance is uniform across all conditions due to its lack of visual signal and resulting absence of noise. All error bars show the 95% confidence intervals.

on both conditions respectively. For the vision-only condition, there was a significant effect of Noise level on the threshold, $F(3, 27) = 31.81, p = 5.21e - 09, \epsilon^2 = 0.470$. The same was shown for the combined condition, a significant effect of Noise level on the threshold, $F(3, 27) = 8.72, p = 3.30e - 04, \epsilon^2 = 0.151$. Furthermore, a set of paired-samples t-tests were conducted to compare the threshold

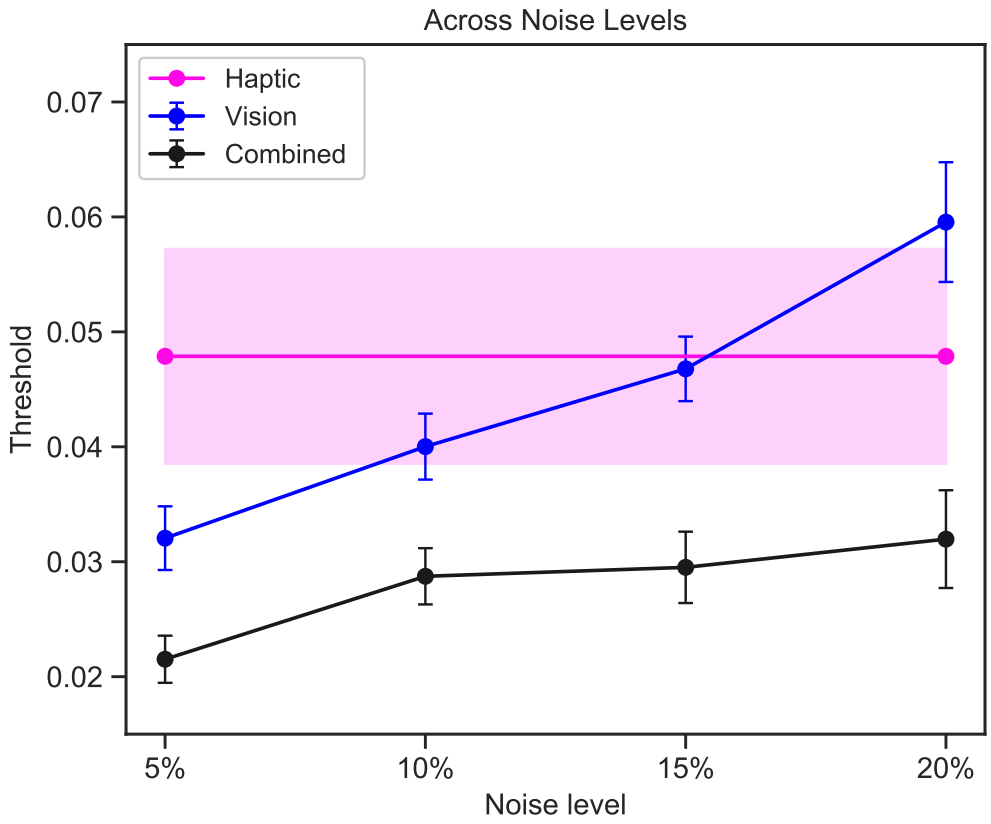


Figure 3.6: Mean performance thresholds per noise level as calculated across all 10 participants, with error bars showing the standard error of the mean.

levels in vision and combined visuohaptic conditions, for each of the four noise levels respectively. There was a significant difference for all noise levels, with the t-value and mean difference increasing as the noise level of the comparisons increases. (5% $t_5(9) = 4.18, p_5 = 0.002$, 10% $t_{10}(9) = 4.84, p_{10} < 0.001$, 15% $t_{15}(9) = 5.06, p_{15} < 0.001$ and 20% $t_{20}(9) = 7.03, p_{20} < 0.001$). As expected, we found that as the noise increased so did the thresholds, though this increase was found to be larger in the vision-only modality than for the combined condition. In line with predictions, we also found that combined thresholds are significantly lower than for vision-only, as seen in Table 3.2 and in Figure 3.7.

Descriptives

Condition	Noise	Mean	SD	N
Vision	5	.032	.009	10
	10	.040	.009	10
	15	.047	.009	10
	20	.060	.016	10
Combined	5	.022	.006	10
	10	.029	.008	10
	15	.030	.010	10
	20	.032	.013	10

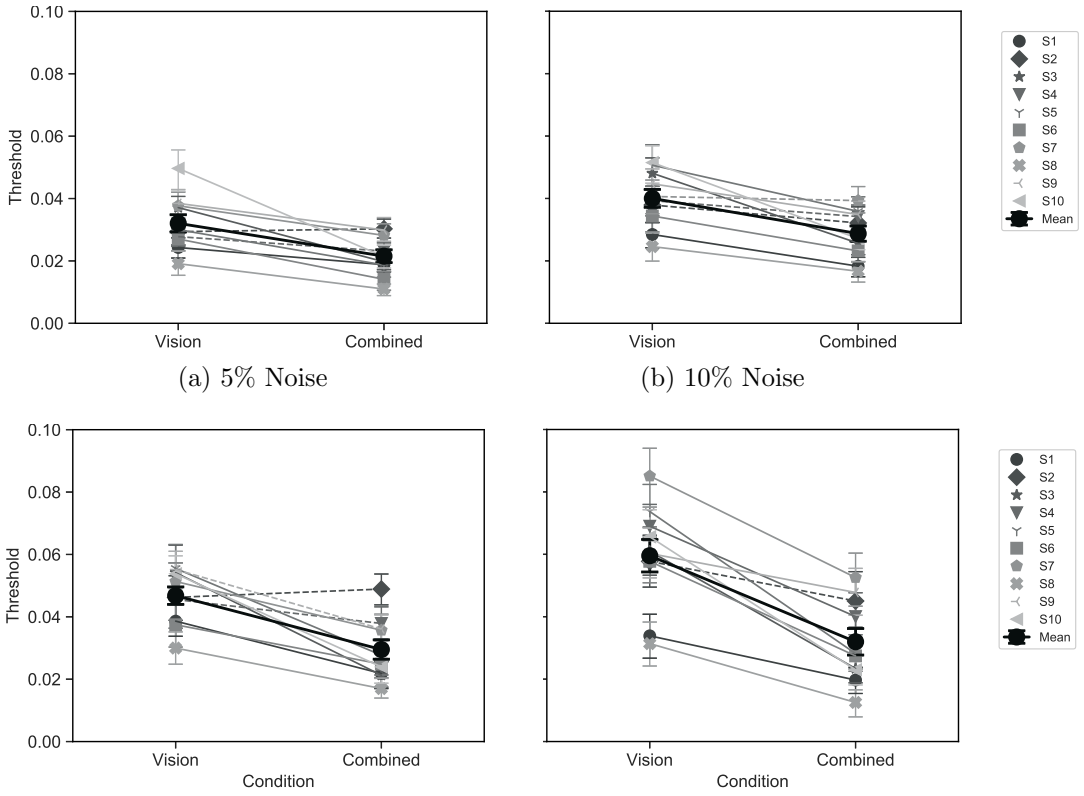
Table 3.2: Descriptives of visual thresholds and combined thresholds per noise level.

Within Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Condition	0.006	1	0.006	37.80	< .001*
Residual	0.001	9	1.470e -4		
Noise	0.004	3	0.001	24.94	< .001*
Residual	0.001	27	5.002e -5		
Condition * Noise	9.303e -4	3	3.101e -4	19.67	< .001*
Residual	4.256e -4	27	1.576e -5		

Table 3.3: Repeated Measures ANOVA shows a main effect of Condition and Noise, as well as an interaction between Condition and Noise.

This strongly indicates that as the visual noise increases, the task gets significantly more difficult to do, while with the addition of haptics to combined cues the increase in thresholds is not as sharp as for vision-only. Not only does haptic improve detection across all individual noise levels, but as expected from the literature, as the reliability of vision decreases towards the reliability of haptics, the benefit of haptics in the combined condition increases.



(a) 5% Noise (b) 10% Noise

(c) 15% Noise (d) 20% Noise

Figure 3.7: Comparative performance on a per-individual level, across all four noise conditions, where a solid line indicates a significant difference, and a dashed line indicates an insignificant difference. The individual error bars are the 95% confidence intervals of that respective threshold, while the error bars on the mean are of the standard error of the mean. (a) 5% Visual noise, where 8 out of 10 participants had significantly lower thresholds for the combined condition compared to the Visual-only condition. (b) 10% Visual noise, where 8 out of 10 participants had significantly lower thresholds for combined. (c) 15% Visual noise, where 7 out of 10 participants had significantly lower thresholds for combined. (d) 20% Visual noise, where 9 out of 10 participants were significantly better. As shown overall, as the noise increases so do the thresholds, and the improvement of combined over vision-only. Observer S8 is the only observer with a stereoacuity greater than 60 arcmin.

3.2.3 Model Comparison

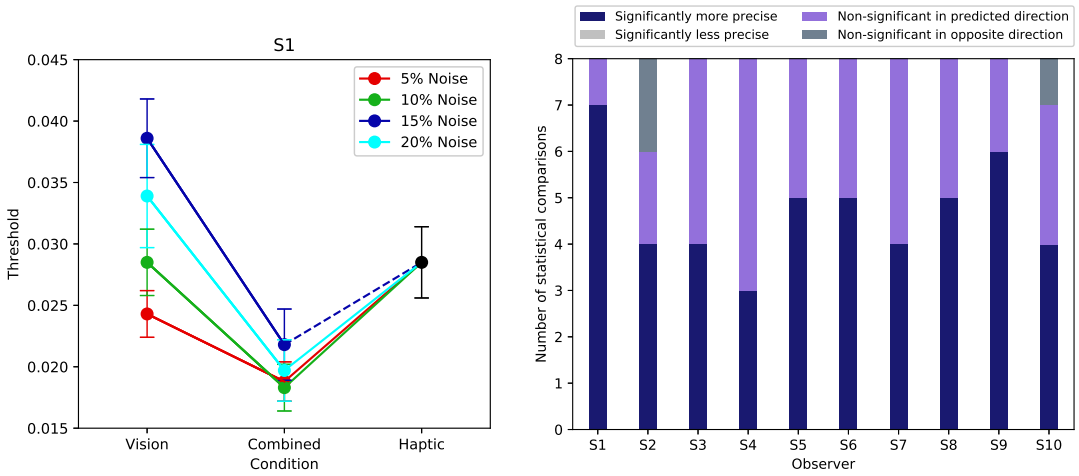
While ANOVAs and other least-squared related statistical methods are more commonly used for formal hypothesis probability, they do not work as neatly for

psychophysical parameters that instead lean more towards maximum-likelihood criterion and Bayesian statistics (Prins et al., 2018).

As we collected a large amount of data per participant across a series of conditions, there is enough data to test our hypotheses on an individual level as well as on a group level. The benefit of being able to perform individual analyses per participant is the perceptual differences found on a per-individual level can be obscured by averaging when analysing only on a group-level.

Model comparison specifically looks at the likelihood of whether two different sets of data are better fit to a single psychometric function (PF) – known as the 'lesser model' or 1PF model – or whether the parameters are too different to fit the same psychometric function – known as the 'fuller model' or 2PF model. In the 'lesser model' fit, one or both of the threshold and slope parameters are fixed, constrained to be equal across all comparisons, whereas in the 'fuller model' the thresholds and slopes are unconstrained and can take any of the valid values during the fitting procedure. The likelihood ratio test between the 'lesser model' and 'fuller model' is used to calculate a statistical 'p-value', where the null-hypothesis is that the 'lesser model' is the better fit, with the 'fuller model' as the better fit being the alternative hypothesis. For this experiment we contrast the thresholds of visual performance and combined performance, leaving the slopes free to ensure the best fit. An example of this can be seen in Figure 3.7, which shows the comparison between the vision-only and combined cue thresholds of each participant, per noise level. As shown in the figure, the data belonging to observer S8, the only observer who measured a stereoacuity of greater than 60 arcmin, was well-fit by the model and fully in line with expected results.

A model comparison was performed in Matlab using the Palamedes toolbox



(a) Significance (b) Barchart

Figure 3.8: Bootstrapped likelihood test results. (a) Bootstrapped likelihood test comparing fuller and lesser model for Observer S1, with error bars showing the 95% confidence interval. Here, 7 out of 8 total comparisons were statistically significant with a $pTLR < 0.05$. This analysis was performed on all 10 participants. (b) Bar chart of performance per participant, where each point on the y-axis represents one of the possible bootstrapped model comparison tests shown in(a) – the majority of the total comparisons showed significantly improvement, where only 3 comparisons showed a non-significant reduction in performance, and there were no comparisons resulting in significantly reduced performance. The full individual results in the form of the plot in (a) can be found in Appendix B, Figure B.1.

(Prins & Kingdom, 2016). This was run on matching the thresholds, but not the slopes, for vision-only and combined visuohaptic on a per-noise-level, as well as between haptic-only and combined visuohaptic – as shown in Figure 3.8a – checking whether there was a significant difference between vision-only and combined visuohaptic, as well as haptic-only and combined visuohaptic, for each respective noise level. The results are plotted in Figure 3.8a, where a solid line indicates a significant difference and a dotted line a lack of significance. This was done for all 10 participants, with the results as shown in Figure 3.8b. The majority of the 8 possible comparisons per individual show a significant overall improvement

in performance for the combined condition compared to either cue in isolation, supporting our hypothesis on an individual level as well.

3.2.4 Discussion

Time as a factor

Experiment 1.0 showed a significant improvement of combined cues compared to vision-only both across noise levels and across participants, as shown in Figure 3.8 (a) and (b) respectively. However, participants used between 1.2 and 5.5 times as long on the combined, due to the additional, slower, haptic exploration. For individual mean times, see Table 3.4. A possible problem arises from this: could the improvement be simply due to spending longer on the task itself?

Total Exploration Time

	0% Noise	5% Noise		10% Noise		15% Noise		20% Noise	
	H	V	C	V	C	V	C	V	C
S1	3271	676	2093	583	2381	513	2304	561	2305
S2	3847	575	1807	683	1689	693	1786	692	1631
S3	2304	714	2113	680	2368	882	1706	847	1753
S4	7980	751	1578	738	2418	843	2150	637	1913
S5	2169	690	2105	780	2131	678	2702	674	2684
S6	5951	858	1314	537	1455	1084	1475	722	2221
S7	2591	582	2067	663	1565	504	1698	516	2356
S8	4645	858	3464	936	5194	971	1830	773	2258
S9	2733	735	1749	613	1690	553	1950	718	1551
S10	2426	840	1716	1054	1295	737	1746	675	1636
Mean	3792 ± 1900	728 ± 103	2001 ± 578	727 ± 106	2219 ± 1124	746 ± 197	1935 ± 359	681 ± 96	2031 ± 385

Table 3.4: Exploration time of vision-only and combined visuohaptic conditions, in Seconds.

Potential auditory confound

While efforts had been made to eliminate auditory sounds as a cue through the use of ear plugs, there were reports of some participants being able to hear a sort of 'grating' sound emitted from the haptic device when exploring the stimulus haptically, even through their foam ear plugs. The emergence of this sound is considered to be a potential confounding auditory cue, if a link exists between the volume, frequency, tone or other auditory sensory dimension with the statistical features of the visual stimuli (Evans & Treisman, 2010; Marks, 1974; Marks et al., 2003; Marks et al., 1987; Misselhorn et al., 2016). If the sound emitted from exploring the haptic surface is related to the rendered signal strength of the haptic surface rather than simply being uniformly present for all haptic exploration, this can create a potentially linked auditory cue which could emerge in the data in place of the 'haptic' signal data, as the haptic signal is calculated based on the visual luminance of the pixels.

In theory, if the confounding auditory signal were stronger or easier to infer to be from the same source as the visual signal, it would be sufficient to the task and could effectively cue veto in favour of the auditory cue; and no information would be obtained from the haptic signal itself. If this were the case, the data labelled as 'haptic-only' would in fact be 'audio-only', and a similar cue combination effect would have occurred had the haptic device given no haptic feedback while auditory exploration sounds were played to the participant during exploration of the stimulus. While it is also possible for there to have been a cue combination of the three potential sensory cues, with information being perceived from both the confounding auditory cue as well as the intentional haptic cue, this was not the aim of the experiment and as such that would need to be controlled for. If

auditory exploration sounds were to greatly improve tumour detection and delineation that would be an interesting result, but that question is beyond the scope of this thesis.

3.2.5 Pilot controls for Time and Audio

As mentioned in the previous section, two distinct limitations in the procedure of Experiment 1.0 were discovered. Firstly, the unconstrained viewing time in all conditions meant that participants spent less time overall exploring the stimulus visually compared to their exploration time when using haptics or when using both. Secondly, some participants reported occasionally being able to hear a sound emitted from the motors of the haptic device while exploring, through the provided foam ear plugs. In order to investigate these concerns we ran two pilot control studies, one for the effect of time and one for the effect of auditory masking noise.

Experiment 1.0.1, Controlling for time

Participants and methods

The first piloted control experiment was run in order to investigate the effect of visual exploration time on performance. For this study, four of the participants from Experiment 1.0 were tested on two separate noise conditions, 5% and 20%, the lowest and highest noise levels used in Experiment 1.0. They underwent one block of free-exploration time, vision-only task which mirrored the one done in the original experiment, and one block of matched-time vision-only task that displayed the stimulus for a set duration, which was matched, on an individual level, to the median time of the observer's own original combined visuohaptic cues

condition from Experiment 1.0, for each of the two noise levels respectively. All other aspects of the experiment, such as the task, stimulus strengths, physical setup and participant positioning were otherwise the same as in Experiment 1.0.

Results

First, the data was fit to Weibull functions as previously described, and the threshold of the psychometric functions were found and then compared between conditions. This compares the level of signal required for the observers to reliably perform above chance. To compare performances, a repeated-measures ANOVA was run on the thresholds comparing Experiment 1.0 and the control experiment, 1.0.1. These compared the conditions of free-exploration vision for both experiments, with combined for Experiment 1.0 and matched-time for the control, 1.0.1. This was done for both low noise level and high noise level, as shown in Table 3.5.

This test found a significant main effect of condition and noise, and an interaction effect of noise, condition and experiment. As there was a three-way interaction, another ANOVA was run on the thresholds comparing Experiment 1.0 and Control, 1.0.1, for the conditions of vision and combined/timed for only the higher noise level, with the outputs shown in Table 3.6. Again, an effect was found on condition, as well as the interaction of experiment and condition. To follow up on this, a paired t-test was run on condition. This resulted in a significant effect for Experiment 1.0 ($t_{Experiment\ 1}(3) = 4.499, p < 0.025$), but not for the control ($t_{Control}(3) = 1.181, p = 0.323$). These findings indicate that, while the thresholds for combined visuohaptic were significantly lower than vision-only in Experiment 1.0, the thresholds for the median-time matched vision-only in the control experiment, 1.0.1, were not significantly lower. However, with a sample

size of four participants, the ANOVA is likely to be underpowered, so a model comparison was also run.

Within Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Experiment	1.040e -5	1	1.040e -5	0.601	0.495
Residual	5.196e -5	3	1.732e -5		
Condition	4.062e -4	1	4.062e -4	20.301	0.020*
Residual	6.003e -5	3	2.001e -5		
Noise	0.002	1	0.002	40.559	0.008*
Residual	1.617e -4	3	5.389e -5		
Exp * Condition	1.271e -4	1	1.271e -4	8.517	0.062
Residual	4.476e -5	3	1.492e -5		
Exp * Noise	7.236e -6	1	7.236e -6	0.166	0.711
Residual	1.304e -4	3	4.348e -5		
Condition * Noise	1.328e -4	1	1.328e -4	5.073	0.110
Residual	7.854e -5	3	2.618e -5		
Exp * Condition * Noise	7.996e -5	1	7.996e -5	11.711	0.042*
Residual	2.048e -5	3	6.827e -6		

Table 3.5: Repeated measures ANOVA shows a main effect of condition and noise, and an interaction effect of noise, condition and experiment.

Within Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Experiment	1.750e -5	1	1.750e -5	0.311	0.616
Residual	1.687e -4	3	5.624e -5		
Condition	5.018e -4	1	5.018e -4	11.893	0.041*
Residual	1.266e -4	3	4.219e -5		
Exp * Condition	2.043e -4	1	2.043e -4	13.967	0.033*
Residual	4.388e -5	3	1.463e -5		

Table 3.6: Repeated measures ANOVA shows a main effect of condition, and an interaction effect of condition and experiment.

Model comparison and Discussion

As previously, we are comparing the fit of the psychometric functions of the two

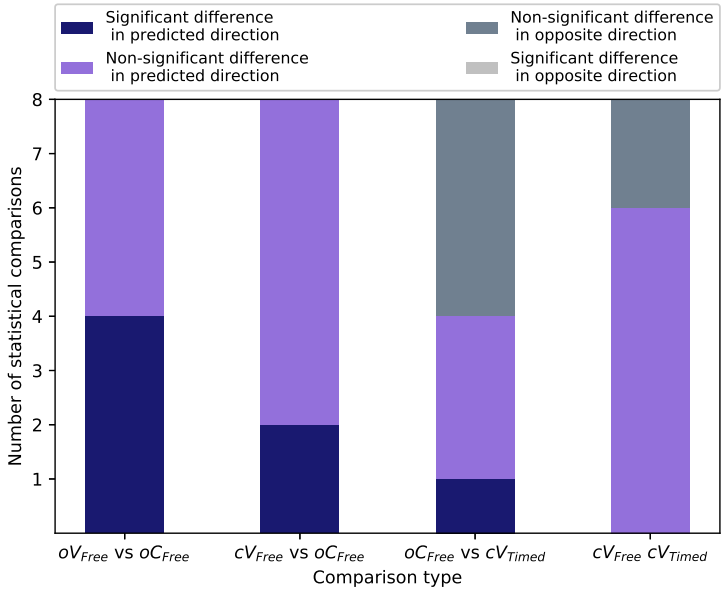


Figure 3.9: Comparisons between (1) Original free-exploration vision and original combined visuohaptic (original results for Exp 1.0), (2) New free-exploration vision and original combined visuohaptic, (3) Matched time and original combined visuohaptic, (4) New free-exploration vision and matched time. If exploration time was the only contributing factor, we would expect the results in (3) to be non-significant, while the results in (4) would be significant improvement in the predicted direction. However, with the small sample size this was underpowered and more data required.

conditions, free-exploration vision and matched-time vision. We run a model comparison which allows us to contrast the likelihood of the psychometric function fits of each condition being significantly distinct from one another ('fuller model') or whether they are likely to stem from the same psychometric function fit ('lesser model'). Out of the total 8 possible comparisons in the model comparison, we look at how many of these fits that fit the fuller model in the predicted direction (significantly 'better'), how many fit the lesser model in the predicted and non-predicted directions (not-significantly 'better' and not-significantly 'worse', respectively), and finally how many of these fit the fuller model in the non-predicted direction (significantly 'worse'), as shown in Figure 3.9.

Unfortunately, the results of this control were inconclusive, as shown in Figure 3.9. Where (1) shows the original effect from Experiment 1.0, improvement between the original free-exploration vision and combined visuohaptic, (2) is a cross-comparison effect between the new free-exploration vision with the the original combined visuohaptic data from the Experiment 1.0. As these do not differ much, the learning effect is considered to be negligible. (3) shows the comparison between the original combined from Experiment 1.0, compared with the matched time condition in the control experiment. The prediction is that combined would perform at a higher level of accuracy than the matched time condition, but the results indicate only a trend in the predicted direction. Lastly, (4) shows the effect in the control, comparing the new free-exploration vision and matched time collected. Had time been the only factor, one would expect (4) to be identical to (1) and (2), and for (3) to be completely insignificant. However, had timing not been a factor at all, one would expect (3) in Figure 3.9 to have a greater number of significant comparisons in the predicted direction, while (4) in the same figure would be completely insignificant.

Experiment 1.0.2, Controlling for audio

Participants and methods

The second piloted control experiment was run in order to investigate the potential influence of noise emitted by the motors of the device as a separate cue. This was done by adding an auditory masking noise on haptic-only tasks. For this study, four of the participants from Experiment 1.0 were tested on the haptic-only modality from Experiment 1.0. They underwent one block of haptic-only with no additional noise masking, mirroring the one done in the original experiment, and

one block of the haptic-only modality but with the addition of a headset playing an auditory white noise to fully mask any sounds emitted from the device. All other aspects of the experiment, such as the task, stimulus strengths, physical setup and participant positioning were otherwise the same as in Experiment 1.0.

Results and discussion

As previously, the data was fit to Weibull PF functions in the Palamedes toolbox, in order to extract the thresholds of the psychometric function fits. These again are used to ascertain the level of signal required for the observers to reliably perform above chance. To compare performances between the two conditions, a paired-sample t-test was run on the thresholds of the performance of the replicated haptic condition which had no auditory masking, and the repeated haptic condition that had an additional auditory masking. No significant effect was found ($t(3) = 0.065, p = 0.953$). As the results of the two pilot controls were underpowered, a full-scale control experiment (Experiment 1.1) was devised and implemented to attempt to answer both of these.

3.3 Control Experiment 1.1, Controlling for Time and Audio

3.3.1 Setup and Methods

For the first full control experiment, 10 naïve participants were tested on the 20% noise condition, the condition from Experiment 1.0 that would have the most to gain from haptic feedback. This is the same number of observers as in Experiment

1.0. The participants underwent 4 repetitions of a non-randomised 3-block set comprising 3 separate conditions, as shown in Table 3.7. The first block was free-exploration time vision-only, followed by one block of free-exploration combined visuohaptic, and lastly one block of matched-time vision-only where the duration of the stimulus presentation was matched to the median time of the previous combined condition. One data set was dropped due to the participants performance dropping significantly in the second session, resulting in their complete inability to perform the task, despite having good initial results in the preliminary tests. The participant's overall performance ended up at almost chance levels, even at very high signal strength where the participant themselves could identify the Gaussian with the naked eye, but for some reason not during the experimental trials.

In order to rule out the auditory sounds of the haptic device, a pair of Bluetooth headphones playing a Brownian noise were used in addition to the standard foam ear plugs. This drowned out any stray sound made by the motors of the haptic device in exploration. In all 3 conditions the sound was played for the duration of each individual trial. With exception of these modifications, all other aspects of the experiment, such as the task, stimulus creation, physical setup and participant positioning were otherwise the same as in Experiment 1.0.

The aim for this experiment was to replicate the findings from the original Experiment 1.0, where the performance of the combined visuohaptic condition is better than for the free-exploration vision-only. It was also hypothesised that the matched-time vision-only condition would perform worse than the combined visuohaptic condition, though it might perform better than free-exploration vision-only.

	First group, no time limit, single cue	Second group no time limit, combined cues	Third group, matched time, single cue
Four repetitions	*Vision-only Free exploration 20% noise*	*Combined Free exploration 20% visual noise*	*Vision-only Matched time 20% noise*

Table 3.7: Block group order of Experiment 1.1. Each group cell represents a block of 10 iterations. In order to match the time spent exploring in the combined condition, the matched-time condition always used the median exploration time of the combined block immediately preceding it. 10 iterations was selected as the highest number of iterations for the slowest condition, combined cues, as while Experiment 1.0 could break up the collection blocks into smaller sets to avoid observer fatigue, this experiment looks at comparing the combined cues with matched time vision-only. Instead, the block order is repeated a total of four times to collect the required amount of data.

3.3.2 Results

The data was fit to Weibull psychometric functions using the Palamedes toolbox, and the threshold of those functions were compared between the three conditions by using model comparisons, as shown in Figure 3.10. No significant differences were found between any of the three conditions. While we expected to find the thresholds for combined visuohaptic condition to be lower than the thresholds of the free exploration vision and of the matched-time vision conditions, none of the overall results show significant differences between any of the three conditions. This was confirmed by a repeated measures ANOVA, shown in Table 3.8 which resulted in $p = 0.970$.

3.3.3 Discussion

In this control experiment, Experiment 1.1, we set out to investigate if the improvement found in Experiment 1.0 was due to the increased time spent exploring

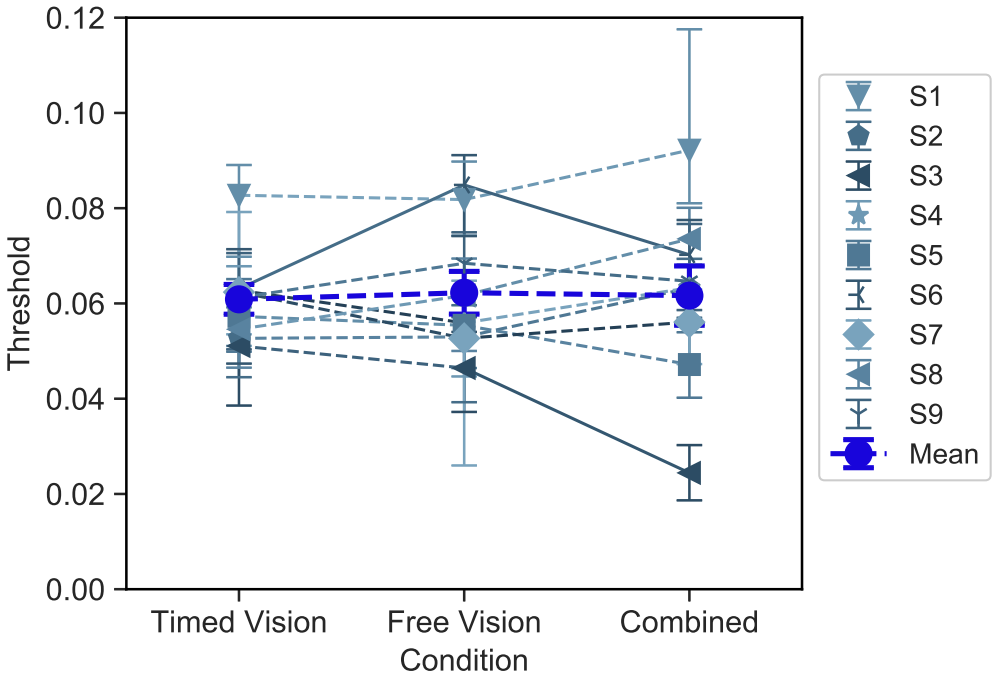


Figure 3.10: Performance for Experiment 1.1. Solid lines indicate a statistically significant difference between the thresholds, while a dashed line shows a non-significant difference. Individual result error bars show the 95% confidence interval while the error bars on the mean show the standard error of the mean. Unexpectedly, the initial effect was not replicated for either combined or matched-time vision conditions, as the majority of the observers did not perform significantly better, or worse, in any of the three conditions. For individual plots of this experiment, see Appendix B, Figure B.2.

Within Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Condition	5.860e -6	2	2.930e -6	0.030	0.970
Residual	0.002	16	9.752e -5		

Table 3.8: Repeated Measures ANOVA shows no effect of Condition.

the stimulus both visually and haptically. However, the findings of this experiment show no difference between free-exploration vision-only condition, combined visuohaptic condition or matched-time vision-only condition, as shown in Figure 3.10. As our expectations were to find the combined visuohaptic condition

perform better than the free-exploration vision-only condition, and matched-time vision-only condition performing somewhere between the free-exploration vision-only and combined visuohaptic conditions, the lack of difference in any of the conditions is very unexpected. We would expect the original improvement in combined to appear, with some improvement possibly occurring in the matched-time vision-only condition as well, but this is not the case. Comparing the results from Experiment 1.0 and Experiment 1.1, shown in Figure 3.11, there is a comparable performance for the free-exploration vision-only condition, whereas the combined visuohaptic condition shows no improvement.

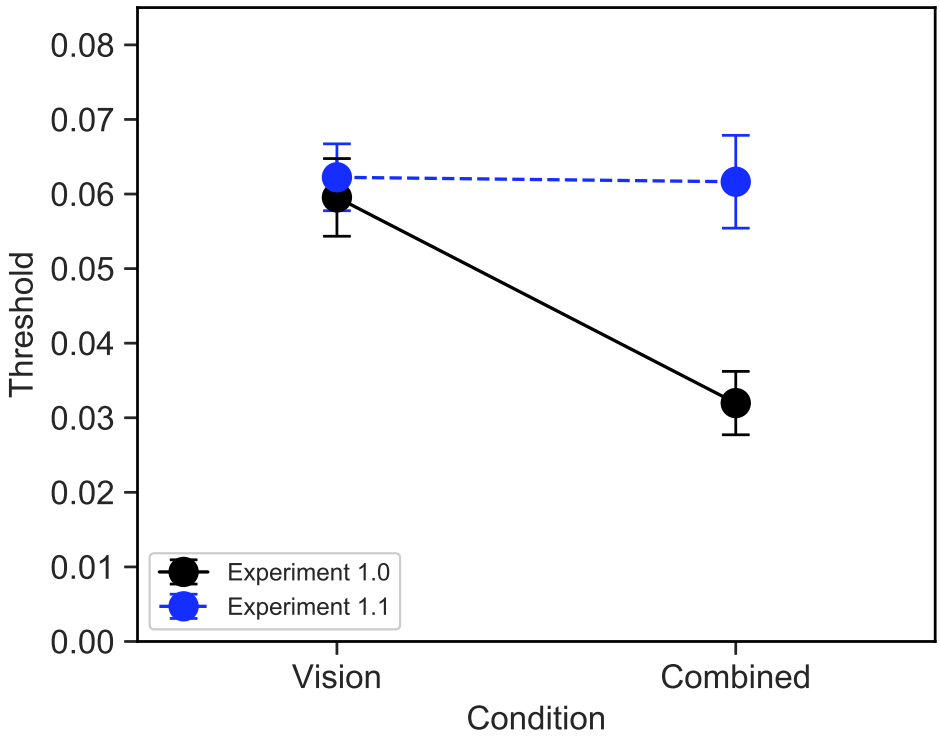


Figure 3.11: Mean performance contrasting Experiment 1.0 with Experiment 1.1, error bars showing the standard error of the mean. This plot illustrates that the initial thresholds of free vision did not vary significantly between the two participant groups, while the performance in combined condition varied greatly.

Experiment 1.1 lack of result

The lack of result in Experiment 1.1 was initially unexpected, however a more in-depth review of the experimental setup revealed several plausible explanations. Had the initial experimental design been more rigorously reviewed to ensure all potential variables were in fact controlled for, it would have been unlikely for the experiment to have a lack of result. While it is possible that the difference in performances were due to differences in the individuals in the respective experimental test groups, it is more likely that the now-removed haptic-only condition originally present in Experiment 1.0 acted as a training or familiarisation task, that the haptic-only signal was poorly matched to the participants in Experiment 1.1, or that the initial improvement in Experiment 1.0 was due to an errant auditory cue which was covered up by the Brownian auditory masking.

This could have been avoided through the collection of additional data, such as had the data for the single cue haptic-only condition been available, it would be possible to model and measure whether the haptic-only signal was at an appropriate signal strength for the observer, as well as being possible to make estimations of what values the threshold of the combined visuohaptic condition would be expected to have based on the results of both of the single cue modalities. However, in the initial experimental design it was presumed that the data for the single cue haptic-only condition would be neither important nor relevant to the experiment's hypothesis, and as such it was erroneously excluded from the experiment.

The timing data shown in Table 3.9 clearly show that the time spent exploring the stimuli was comparable to the time spent in Experiment 1.0, and as shown in Figure 3.11, the free-exploration vision-only performance was also highly comparable to the first experiment. This means that the participants in Experiment

1.1 spent, on average, 2.5 and 1.7 times longer exploring in the vision-only and combined visuohaptic conditions respectively, while having performance at 1.0 and 0.5 times the respective rates. Additionally, the time spent in the matched-time vision-only condition is 4.9 times higher than the free-exploration vision-only condition of Experiment 1.0, and 1.9 times higher than the free-exploration vision-only condition of Experiment 1.1, while having 1.0 and 0.98 times the performance. This clearly shows that while the exploration time was comparable between Experiments 1.0 and 1.1, the relative performance of the combined visuohaptic condition was not, allowing us to still rule out the effect of exploration time as the sole contributor to the initial improvement found in Experiment 1.0.

Total Exploration Time

	20% Noise		
	C	M-V	F-V
S1	1452	1467	925
S2	2671	2457	1411
S3	8250	7690	2848
S4	1746	1982	1307
S5	5455	5089	2061
S6	2814	2713	2039
S7	3696	3644	2101
S8	3451	3428	1676
S9	1722	1736	1341
Mean	3473 ± 2180	3356 ± 1974	1746 ± 578

Table 3.9: Exploration time of vision-only and combined visuohaptic conditions for Experiment 1.1, in seconds.

3.4 Control Experiment 1.2, Haptic as Training

3.4.1 Introduction

In order to disambiguate the potential causes for the lack of results in Experiment 1.1, a final control experiment was run. Experiment 1.2 replicated the experimental design of Experiment 1.0 for the 20% visual noise level, aiming to explore whether the initial improvement was due to the errant sounds emitted by the haptic device, or due to having a haptic-only modality prior to combined visuohaptic condition. To eliminate exploration noise as a cue, all participants were equipped with the Brownian auditory masking used in Experiment 1.1. While it is unlikely that the difference is due to individual differences in the test groups, a new group of 10 naïve participants were recruited. In repeating the procedure for Experiment 1.0, half of the vision-only data was collected first, then the haptic-only performance was compared and matched to the vision-only performance before the haptic-only data was collected. Following this, the entirety of the combined visuohaptic data was collected, and lastly the remainder of the vision-only data. As the actual performance of the haptic-only condition was not considered related to the improvement the combined visuohaptic condition, only the initial half of the haptic data was collected for investigating the potential training effect. This was done in the block order shown in Table 3.10, which is the same as the order in Table 3.1 for the 20% visual noise group specifically.

3.4.2 Methods

We collected 10 naïve participants on three different conditions: vision-only, haptic-only, and combined visuohaptic. The blocking order of the conditions was

selected to be the same as used in Experiment 1.0, both to replicate the methodology and to balance out any training effect associated with task repetition, and is illustrated in Table 3.10. All participants were outfitted with ear plugs (Moldex 7800 Spark Plugs) and a pair of headphones emitting the Brownian masking noise, which ran for the duration of each trial in order to mask any potential sound from the haptic device during exploration. The masking noise was played in all three conditions. As the effect of cue combination is the strongest when the relative reliabilities of the cues are matched, the observer's individual haptic-only performance was compared to the accuracy of their vision-only, and if needed the haptic scale factor was adjusted on a per-individual basis. This was done for the same reasons as in Experiment 1.0, so that the haptic-only and vision-only performances were of the same magnitude. As the haptic-only condition was absent from all previous controls, so were the haptic scale factor adjustment. These adjustments were made prior to collecting the data for the haptic-only and combined visuohaptic conditions.

3.4.3 Results and discussion

As done previously, Weibull psychometric functions were fit to the data using the Palamedes Toolbox in Matlab. From these psychometric functions we obtained the observers' accuracy (threshold) and precision (slope). Firstly a model comparison was run in the Palamedes toolbox per participant, and found that 6/10 participants were significantly better described by the fuller model, with 4/10 being better described by the lesser model. The results of the model comparison is shown in Figure 3.12. Additionally, in order to assess any difference in accuracy between the vision-only condition and the combined visuohaptic condition a

First group, single cue	Second group combined cues	Third group, combined cues	Fourth group, single cue
Vision-only, 20% noise	*Combined, 20% visual noise*	*Combined, 20% visual noise*	*Vision-only, 20% noise*
Haptic only			

Table 3.10: Block group order for Experiment 1.2. Each group cell represents a block of 20 iterations. The haptic-only condition was only used to 'train' the observers in using the haptic signal and to ensure the reliability of the haptic signal was comparable to the visual signal, and therefore the second half of haptic was not collected. As in Experiment 1.0, the vision-only cells were single blocks of 20 iterations, while haptic-only used four sets of 5 iterations per group, and combined cues used two sets of 10 iterations per group.

paired t-test was run on the threshold of the vision-only and combed visuohaptic conditions and found to be statistically significant ($t(9) = 2.839, p = 0.019$).

As we have two data-sets of the same size, with the same number of trials per observer, a within-between repeated-measures ANOVA was run on the two data-sets, shown in Table 3.11 and Table 3.12. These results show that the act of using haptic-only without the aid of a visual signal greatly improves the performance of combined visuohaptic signal detection compared to having no haptic signal (vision-only), and to having no haptic training prior to the combined condition. Note however that for Experiment 1.1, participant 9 was dropped due to inability to perform task in spite of preliminaries.

Table 3.13 which shows the mean exploration times for Experiment 1.2 shows that people spent, again, comparative time in the visual-only and combined visuohaptic conditions, and Figure 3.13 shows that observers performed comparatively in the visual-only condition in all three experiments, whereas Experiment 1.0 and

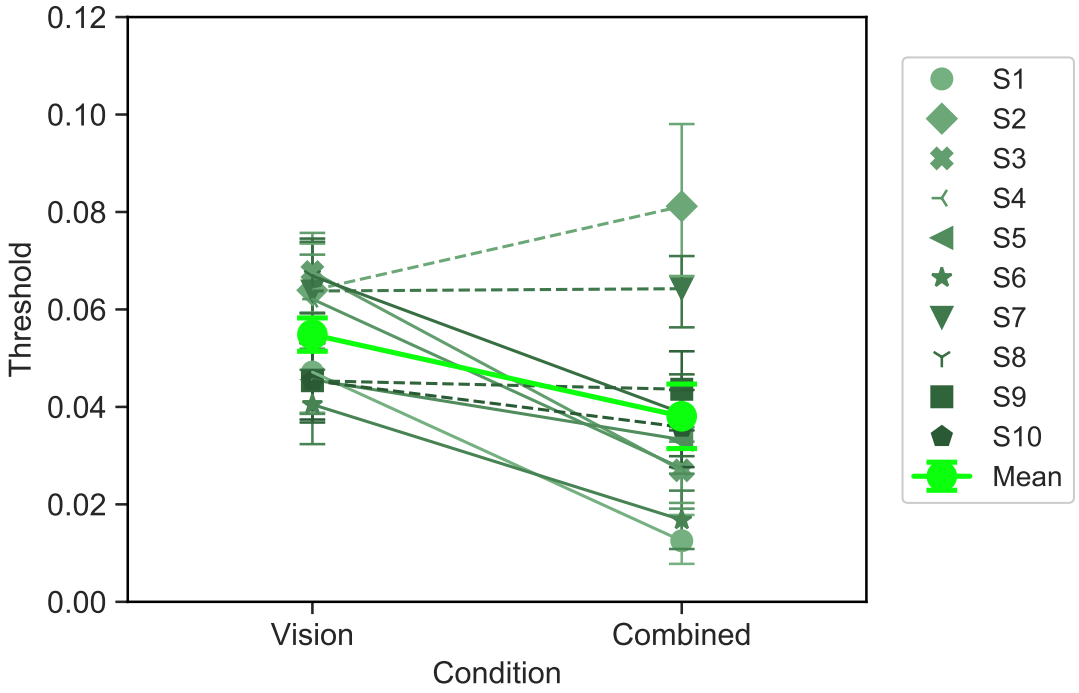


Figure 3.12: Performance for Experiment 1.2. Solid lines show a statistically significant difference between the conditions, while a dashed line show a non-significant difference. The individual error bars are the 95% confidence intervals, while the error bars on the mean are of the standard error of the mean. As the plot illustrates, the effect of the combined condition has returned, as a result of reintroducing haptic-only as a training phase. For individual plots of this experiment, see Appendix B, Figure B.3.

Within Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Condition	5.880e -4	1	5.880e -4	4.337	0.053
Condition*Training	7.488e -4	1	7.488e -4	5.524	0.031*
Residual	0.002	17	1.356e -4		

Table 3.11: Repeated measures ANOVA shows an interaction effect between condition and training, with the main effect of condition being just shy of significance.

1.2 had a comparable effect of combined visuohaptic while Experiment 1.1 did not.

These results strongly indicate that the original effect found in Experiment 1.0 is not due to subtle auditory cues from the haptic device, nor the length of

Between Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Training	0.003	1	0.003	5.831	0.027*
Residual	0.008	17	4.644e -4		

Table 3.12: Repeated measures ANOVA shows a main effect of training.

Total Exploration Time

	20% Noise	
	V	C
S1	1178	3180
S2	705	2181
S3	1368	2829
S4	3005	7396
S5	1112	2694
S6	922	2361
S7	1407	2887
S8	1307	2402
S9	790	2277
S10	888	1745
Mean	1145 ± 658	2382 ± 1626

Table 3.13: Exploration time of vision-only and combined visuohaptic conditions of Experiment 1.2, in seconds.

time spent viewing the visual stimulus. However, the results also indicate that for the addition of haptics to have a measurable effect a training phase may well be required for the observers to successfully recruit the somewhat novel method of using a stylus for perceiving a haptic cue.

3.5 Overall results

Comparing and contrasting the results from all three experiments, we find that combined vision and touch improves participants' ability at detection compared

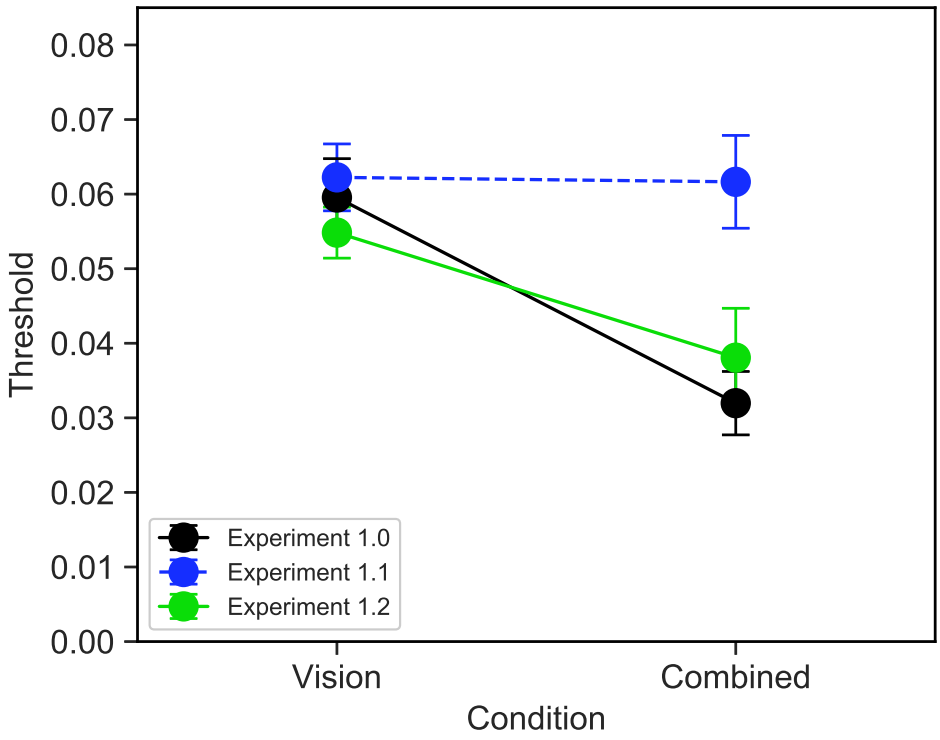


Figure 3.13: Mean performance of all three full experiments, error bars showing the standard error of the mean. Experiment 1.0 and 1.2 both have haptic-only prior to combined condition, and show a significant improvement in performance. While both Experiment 1.1 and 1.2 have auditory masking, Experiment 1.2 still shows significance, thereby ruling out audio as the main source of the effect.

with either vision or touch in isolation. While the combined condition is slower than vision-only, a forced increase in visual exploration time did not significantly improve performance over free-exploration time in Experiment 1.1. From this we can conclude that the improvement found in Experiment 1.0 is not due to the increase in time spent. However, there was at this point the need for a second and final control, replicating the exact procedure of the first experiment but with the addition of the auditory masking Brownian noise. The results from Experiment 1.2 show a significant improvement in detection again for the combined condition,

allowing us to rule out the errant sounds of the haptic device in motion as a significant source of information. This does however lead to the conclusion of having haptic-only as a training task is key to benefiting in combined visuohaptic tasks.

3.6 Discussion

In this experiment we set out to investigate whether or not the addition of a haptic cue would improve the performance in detecting a hidden Gaussian bump in a 2AFC detection task. In Experiment 1.0 we found that, in line with our hypothesis and previous research, the addition of a haptic cue significantly improves performance and accuracy in the visuohaptic condition compared to either cue in isolation (Figure 3.6). However, a concern was that the highest-performing condition was much slower than the other two, increasing the available exploration time which could translate to an improvement in detection, rather than the improvement being due to a cue combination effect from the addition of a haptic signal. Additionally, as the haptic device moves across the simulated surface, some participants reported hearing a scraping sound emitted by the motors of the haptic device. This could potentially be used as an unintended auditory cue and needed to be controlled for. These two queries were further explored in Experiment 1.1, where the effect of an auditory cue was controlled for by adding in an additional auditory Brownian noise which masked the errant sounds of the haptic device, while the time element was examined by having participants view a visual-only stimulus for the median duration of time they themselves spent on the immediately preceding combined condition. This allowed us to investigate

the effect of matched-time vision compared to both free-exploration vision and combined visuohaptic.

The results from Experiment 1.1 did not find any significant improvement in either condition (Figure 3.10), which was expected for the matched-time visual-only condition but unexpected for the combined condition. We found that performance in visual-only does not improve when forced to view the stimulus for the duration spent in the combined condition, ruling out the increase in time as the sole reason for the initial effect. However, no further improvement was found for the combined condition, raising the question of whether the improvement was due to the spurious auditory cue of the haptic device motors or whether haptic-only acts as a training phase. This incidental removal of the haptic-only condition used in Experiment 1.0, half of which took place before the combined condition, could have inadvertently removed a vital training aspect for a novel perceptual cue. These concerns were further explored in the next control experiment. For the final experiment, Experiment 1.2, we replicated the procedure of Experiment 1.0 for the highest level of visual noise (20%), with the addition of the headphones playing auditory Brownian noise to mask the errant sounds of the haptic device.

In Experiment 1.2, we expected to replicate the results found in Experiment 1.0, where combined visuohaptic performed significantly better than either cue in isolation. Our findings confirm the results of Experiment 1.0 (Figure 3.12). We found that the readdition of haptic-only condition prior to starting the combined condition restores the improvement for combined visuohaptics. This reaffirms the findings of Experiment 1.0, and disproves the concern that additional auditory cues were the cause of the improved performance. The results of the three experiments 1.0, 1.1 and 1.2 lead us to the potential importance of training for haptic

cues, as illustrated by Figure 3.13. Interestingly, a study with similar design to this found no such effect. The study by Kang and Kim (2018) as mentioned in Chapter 1, had a very similar experimental design of a 2AFC visuohaptic signal discrimination of a rendered Gaussian bump, using the same haptic device model, the Touch X. While cue combination benefit is known to be at its highest when the reliabilities of the cues are matched, Kang and Kim (2018) did not attempt to match their haptic signal reliability to match that of their visual signal. Additionally, the visual and haptic signals were presented perpendicular to one another – where the visuals were presented top-down on a standard monitor, while the haptic signal was presented top-down parallel to the ground.

3.6.1 Training

What we have discovered over the course of these three experiments is just how crucial the haptic training phase can be to successfully completing the task, whether or not that training is intentional. The question remains of exactly what might be 'learnt' through the use of the haptic-only condition as a potential training or familiarisation task.

It is worth noting that the haptic device used in these experiments is a stylus-based device, where the observer holds the handle of the device similar to how one would hold a pen. This differs from several of the classic multisensory visuohaptic tasks which predominantly use thimble-based devices such as PHANToMs used in Ernst (2007), Ernst and Banks (2002), Gepshtein et al. (2005), Helbig and Ernst (2008), Hillis et al. (2002), Plaisier et al. (2010), Plaisier et al. (2014), and Takahashi et al. (2009), Takahashi and Watt (2014), to name but a few. Whereas a thimble-based haptic interaction design mimics the natural grip movements of

the hand and fingers, using a stylus-based device is more akin to using a tool, though tool-use is a skill to which humans are well adapted – though the exact mechanisms behind how the sensory system successfully links the information from the two sources is still highly debated.

One of the studies that goes into detail about solving the correspondence problem during tool-use is Takahashi et al. (2009). In their study, they ran three experiments with different levels of spatial separation between the visual and haptic representations of an object, in order to establish whether the perceived use of a tool can mitigate the negative effects associated with large spatial offsets between the visual and haptic signals. The results of their experiments show that an equivalent distortion between vision and touch is no longer detrimental when a rigid tool is presented as linking the visual and haptic signals, and also that when an additional offset is introduced to the tool-use condition, the resulting detriment matches the detriment found when introducing the spatial offset in the no-tool condition, strongly indicating that humans learn to adapt to the geometry of the tool in use. For a more in-depth description of this paper, please see Section §1.4 in Chapter 1.

Another aspect could be the lack of active tool use for a haptic-only signal in Experiment 1.1. In Experiment 1.0 and 1.2, the observers each collected four blocks of haptic-only data prior to collecting the data for the combined condition, spending around an hour on average performing the haptic-only task. This absence of intentional haptic-only experience could be depriving the participants of a sense of agency and familiarity with active tool-use of the haptic device as a novel tool. As shown in a study by Maravita et al. (2002), the skills associated with active tool use develops further with increased experience using the tool. This

theory is also supported by several other studies, which have found that skilful tool use and 'warming up' with the tool which enables the sensory system to extend the perceptual awareness to the tool-tip (Holmes et al., 2004, 2007; Holmes, Sanabria, et al., 2007). In the study by Holmes et al. (2004), the authors aimed to investigate whether tool-use extended the peripersonal space of the observer, projects the peripersonal space to the tool-tip, that it forms new visual receptive fields (VRF) around the tool-tip as an addition to the VRF of the hand, or lastly that tool-use has no effect on VRF. Through a series of three experiments they found that the perceptual space did not get extended, as observers did not integrate signals that were presented along the shaft of the tool, but rather suggesting that the tool-tips themselves were perceptually highlighted. These findings are validated by later studies (Holmes, Calvert, et al., 2007; Holmes, Sanabria, et al., 2007), where the authors compared different spatiotemporal congruency effects and both visual and auditory distractors. Lastly, in order to further test their hypothesis that simple tools caused an enhanced processing of visual stimuli at the functional end of the tool, a functional MRI study was run by Holmes et al. (2008). Continuing on the visuohaptic tool integration experiments with congruent and incongruent distractors, they ran the experiment with healthy human participants situated within MRI scans, in order to obtain quantifiable brain activity (blood-oxygen-level-dependent signals, often abbreviated 'BOLD') in the occipital cortex. The results show significant support for their hypothesis, and the magnitude of visuohaptic interactions in the behavioural responses of the observers were significantly predicted by the BOLD signals in the occipital cortex.

Comparing these findings with our own, there is a high chance that the absence of dedicated haptic-only exploration in Experiment 1.1 could have unintentionally

removed critical familiarisation time and active, skilful tool-use that were afforded to the participants in both Experiment 1.0 and 1.2.

The initial theory of the haptic signal rendering was that it would be a simple, intuitive height-map, comparable to ones used in topographic mapping. Previous research has shown that, when exploring mountain ranges that are being presented both visually and haptically, trained observers perform significantly better and are more precise with the additional haptic signal compared to using vision alone (Evreinova et al., 2012). While this shows that haptic height mapping can beneficially be learnt as a signal, the haptic signal as rendered in our experiment was a luminance-based normal-mapping of the underlying image, which was limited to the bit-depth of an 8-bit PNG. Contrasting this resolution to the sub-pixel spatial resolution of the haptic device itself, the signal as presented haptically could easily have been more akin to perceptual roughness and grating than low-intensity height map. If this were the case, the haptic signal would not be a simple height map, but instead an arbitrary signal which is unlikely to have been perceptually linked with the visual luminance of the visual stimuli.

However, arbitrary signals can also be learnt to be integrated, as discussed in detail in Section §1.1.4, Chapter 1. The study by Ernst (2007) found that it is possibly to learn to integrate fully arbitrary sensory signals, to which no prior perceptual linking exists, such as between visual brightness and haptic surface stiffness. By collecting the single-cue JND thresholds of both vision and touch before and after training, Ernst established that the observers learnt to integrate the two signals, comparing the effect of 'congruent' and 'incongruent' sensory signal pairs both before and after training. Whereas there was no distinction between 'congruent' and 'incongruent' signals prior to training, there was a significant dif-

ference in JND thresholds between the conditions after training occurred.

It is possible that the haptic-only condition offered more explicit familiarisation with the haptic device as a novel tool, or that the haptic signal was rendered as a haptic roughness texture or a miniscule height map, or that even the potentially unrelated, 'arbitrary' visual and haptic signals were trained to form a coupling; any of these hypotheses could be true. Unfortunately, as the signals were not matched to JNDs on a per-observer level, there was no single-cue data collected for haptic-only in Experiment 1.1, and no post-combined condition haptic-only data was collected in Experiment 1.2, no further conclusions can be drawn without running further experiments.

Future studies carrying on the findings of these experiments could include rendering the haptic signal using a 3D surface mesh of the underlying signal, which would help compare whether the initial haptic signal was treated like a height map or as something more akin to a coarseness metric. Another alteration would be to match the signal strengths in JND space on a per-observer basis, as well as collecting all the single-cue data both before and after any familiarisation phase of the experiment, which would have helped with modelling the specific 'learning effect' associated with prolonged exposure to the task as well as disambiguating what exactly might have been 'learnt'.

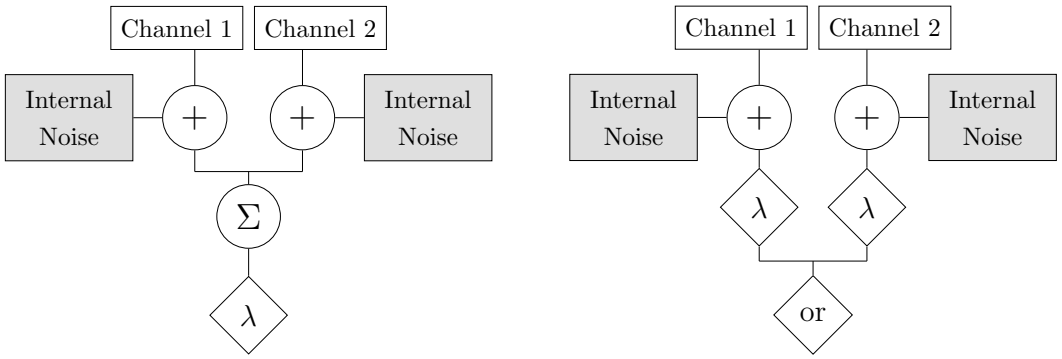
3.6.2 Comparing cue combination models

One final query is whether the cue combination effect is a true cue integration effect, where additive summation (AS) would have the brain integrate the information from both cues – a requirement for any of the linear combination models discussed in Chapter 1 and illustrated in Figure 3.14a, or whether probability

summation (PS) is likely, where the likelihood of the brain locating the signal is higher just due to the increased channels of information available, as illustrated in Figure 3.14b. If PS is the source of the improvement, it would suggest that the brain does not *combine* the visual and haptic cues, but rather that it has a higher chance of detecting the signal as there are multiple cues available. To compare whether Probability Summation (PS) or Additive Summation (AS) presents a better fit to the data, we ran the data through a model comparison using the Palamedes toolbox (Prins & Kingdom, 2016), which compares the fits for PS and AS, using the Akaike's Information Criterion (AIC) which is an alternative to p-value in comparing model fits, as described in Kingdom et al. (2015). When comparing the difference between the AIC fit for the AS model and the PS model, a negative ΔAIC implies that the AS model is a better fit, while a positive ΔAIC implies that the PS model is a better fit. The comparisons were run on all 10 participants for all four noise levels respectively. For this dataset, there were 21 occurrences of AS is better fit (negative values), and 19 occurrences of PS is the better fit (positive values). As such, we found no evidence that one model fits better. For exact numbers, see Table 3.14.

	ΔAIC									
	1	2	3	4	5	6	7	8	9	10
5%	1.05	-0.56	-10.82	-0.91	-3.78	-1.94	-0.14	-2.71	-11.83	-0.37
10%	-3.38	1.25	-7.96	0.91	5.98	-0.23	4.13	2.93	7.16	-5.07
15%	14.34	1.63	-5.20	7.85	3.43	5.58	-6.17	-5.66	-2.73	8.53
20%	1.49	10.62	-5.79	13.85	-7.63	6.90	1.67	17.23	-2.44	-0.12

Table 3.14: Table of ΔAIC values per participant over noise levels, where negative values, coloured red, indicate that additive summation is a better fit, while positive values, coloured blue, indicate that probability summation is a better fit.



(a) Additive Summation (b) Probability Summation

Figure 3.14: The two contrasted sensory summation models, as per Jones (2016). (a) Additive summation model, where the individual sensory channels are linearly combined, the sum of which is used to calculate the internal selection criterion λ The MWF model is a specific subcase of the Additive summation model, where the Σ represents Equation 1. (b) Probability summation model, where the sensory channels are independently evaluated and the internal selection criterion λ is calculated per individual channel before the sensory system selects a boolean or for which of the channels hits the internal selection criterion λ.

Maximum likelihood estimation model fit

One of the cue combination models mentioned in the introduction is the Maximum likelihood estimation (MLE) model, which is a specific instance of the additive summation model (Jones, 2016), where an ideal observer always combines the sensory cues in a statistically optimal fashion. Incidentally, it also follows the

same model and equation as current through resistors in parallel.

However, it is worth noting that the MLE calculations are traditionally performed on JND thresholds, which are based on cumulative Gaussian psychometric functions. When taking the JND threshold of cumulative Gaussian functions, assuming independent Gaussian distributions, then the higher order terms in the Bayesian probability equation vanish and give an exact equation where the signal weights are inversely proportional to the variance of the Gaussian distributions (Yuille & Bülthoff, 1993). As our data is not collected for JNDs, and uses the threshold values of Weibull psychometric function fits which are not modelled as pure Gaussian distributions, there is no equivalent or directly transferable equation to test and compare the optimal cue combination as described by the MLE model. However, whether our data would have been better modelled using cumulative Gaussians or following the JND paradigm, it was at the time not considered the appropriate choice for the experimental protocol. The underlying equation of the MLE model, given in Equation 3.2, is still used in a wide range of other applications such as calculating the current through resistors in parallel, the basic MLE model equation was applied to our data as collected, and is statistically not different from the optimal combined cues, as shown in Figure 3.15.

The standard MLE model equation (Equation 3.2) was applied to the respective thresholds of the collected Weibull psychometric functions to calculate the theoretical combined visuohaptic performance level, on a group-level average as well as on a per-individual basis. These calculated combined thresholds were compared to the actual collected combined thresholds using a repeated measures ANOVA with within-subjects factor of type (collected, calculated) and between-subjects factor of visual noise levels (5%, 10%, 15% and 20% noise). The ANOVA

revealed that there was a significant effect of noise ($F(3, 27) = 10.846, p < .001$), but no main effect of type ($F(1, 9) = 2.757, p = 0.131$) and no interaction between type and noise ($F(3, 27) = 2.406, p = 0.089$). This is also shown in Figure 3.15. This indicates that the results of Experiment 1.0 are well fitted to the estimated cue combination described by the MLE model, strongly suggesting that the sensory cues are being integrated in accordance to the MLE optimal cue combination model. Individual prediction graphs per observer are shown in Figure 3.16.

$$\alpha_{VH}^2 = \frac{\alpha_V^2 \alpha_H^2}{\alpha_V^2 + \alpha_H^2} \tag{3.2}$$

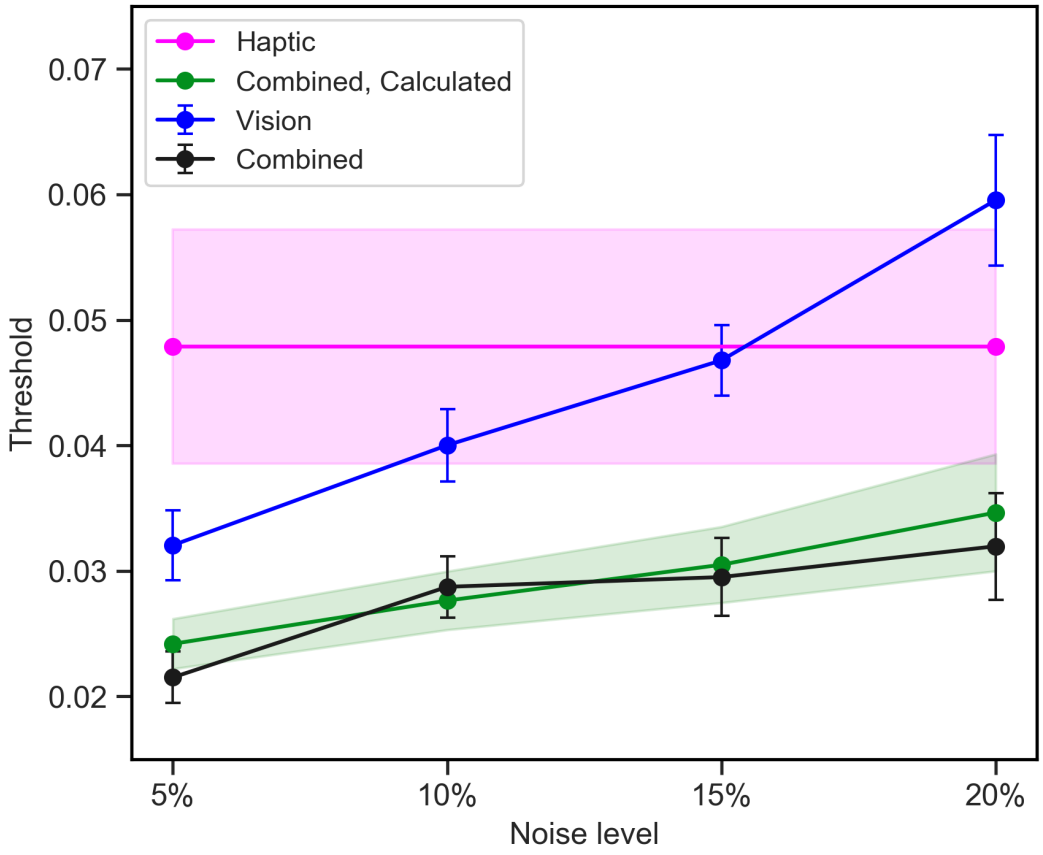


Figure 3.15: Mean MLE predictions in the same form as Figure 3.6. The black, blue and magenta lines are all collected data, with error bars showing the standard error of the mean, while the green line and error region is the region of optimal combination of the single-cue thresholds as collected.

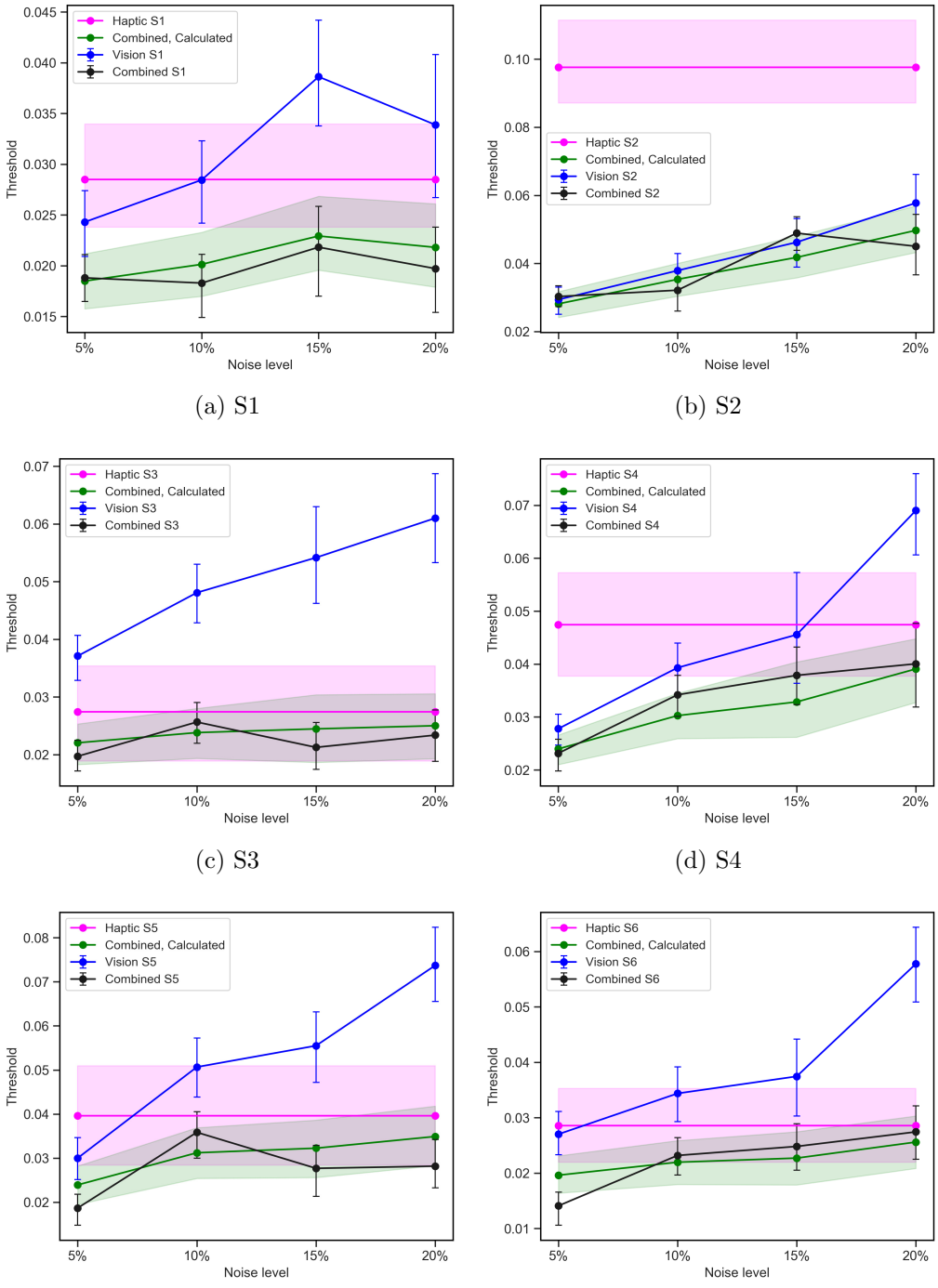
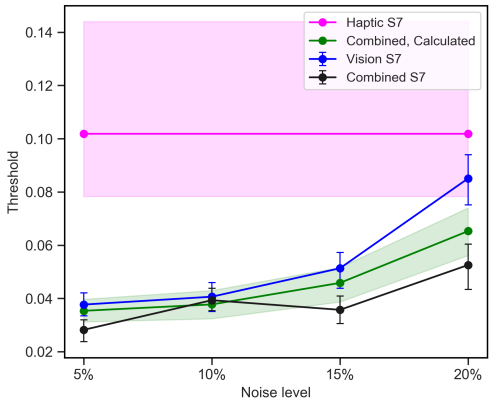
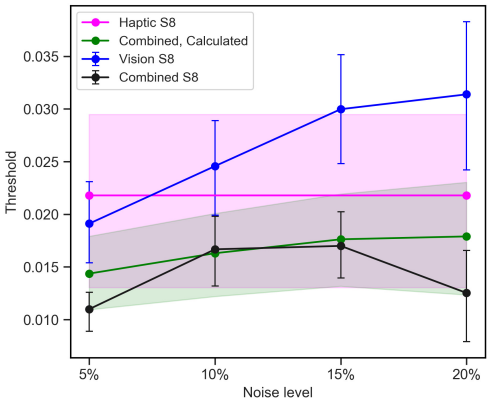


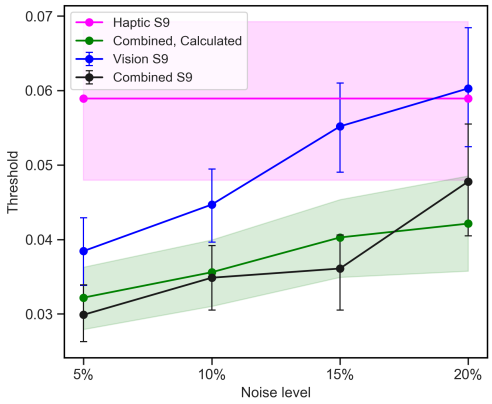
Figure 3.16: Experiment 1.0 individual MLE predictions, with unconstrained y-axis for within-participant comparison purposes. For the between-participant comparisons with constrained y-axes, please see Section §B.1.4, Appendix B



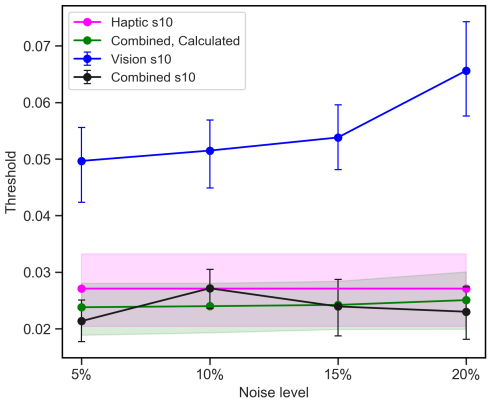
(g) S7



(h) S8



(i) S9



(j) S10

Figure 3.16: Experiment 1.0 individual MLE predictions, with unconstrained y-axis for within-participant comparison purposes. For the between-participant comparisons with constrained y-axes, please see Section §B.1.4, Appendix B

3.6.3 Implications of findings

As stated in the introduction, while multisensory cue combination has been widely studied, most of the body of research focusses on discrimination tasks rather than detection tasks, and most of these are performed on spatially coaligned experimental rigs. Our findings verify that the cue combination benefit also occurs for signal detection specifically, and also has a significant benefit in a spatially misaligned setup similar to rigs more feasible to use in a medical clinician setting.

3.7 Summary

In this chapter we aimed to explore whether the addition of a haptic cue would improve detection of a hidden Gaussian bump in a 2AFC detection task on a spatially misaligned experimental setup. In Experiment 1.0 we expected to find, in line with previous research, that the addition of a haptic cue would improve performance in the visuohaptic condition compared to either cue in isolation and our findings are well aligned to our initial theory. We found that people significantly improved their accuracy when both vision and haptic were available simultaneously, even though the visual and haptic signals were spatially misaligned. There was however a concern that there was a discernable difference in exploration time between the three conditions, where the highest-performing combined visuohaptic condition was much slower than the other two conditions, which could possibly have lead to an improvement in detection from exploration time alone, as well as a concern that a faint scraping noise emitted from the haptic device during exploration could be unintentionally contributing as an errant auditory cue. Experiment 1.1 used auditory masking and compared the performance between

untimed vision-only, untimed combined visuohaptic and matched time vision-only where durations were matched per participant between the combined visuohaptic condition and matched-time vision-only, using the untimed vision-only as a control. We predicted that participants would again have improved performance in the combined visuohaptic condition, with little to no improvement predicted in the matched time vision-only compared to the untimed vision-only condition.

The results show that, contrary to the predictions and findings of Experiment 1.0, there was no significant improvement in either of the combined visuohaptic condition or the matched time condition, despite exploration time being comparable between the two experiments. While this result ruled out exploration time as the sole contributor to the improvement, two new theories arose. Either the auditory masking removed the improvement, or the auditory cue was the beneficial addition or that the haptic-only condition unintentionally doubled as a form of training for haptic detection. The final experiment, Experiment 1.2, replicated the basic order of Experiment 1.0 with additional auditory masking, aiming to disambiguate whether the effect was due to an errant auditory cue or the haptic cue after training. The results of the final experiment matches the findings of Experiment 1.0, indicating that the improvement in signal detection was not due to the auditory cue, nor the exploration time, but through the addition of a haptic signal. From this series of experiments we can conclude that even for spatially misaligned vision and touch, the addition of haptics improves performance in 2AFC signal detection tasks for Gaussian bumps.

Chapter 4

Experiment 2

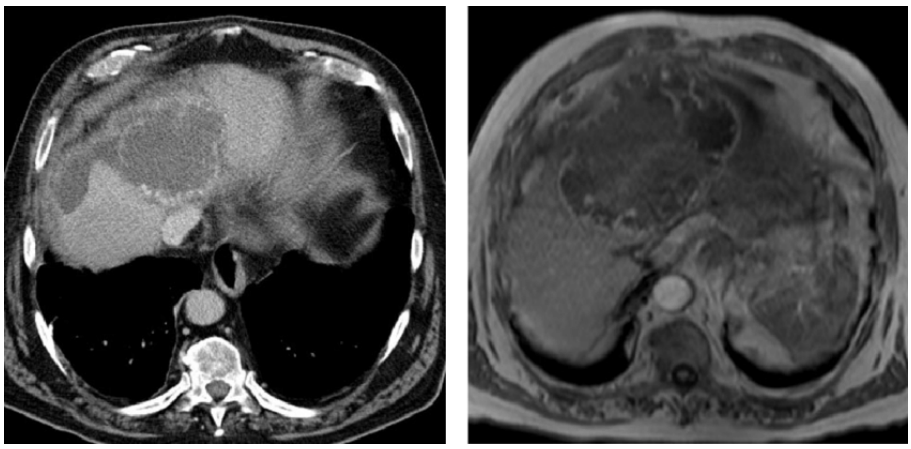
4.1 Introduction

4.1.1 Medical imaging modalities

In the field of medical imaging there exist a range of different technologies used to image the human body, such as computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) to mention a few. Each of these types have benefits and drawbacks, such as CT having higher accuracy than MRI for lung assessments while MRI has a much higher soft-tissue resolution, or the cost difference between the more expensive PET/MRI and the more affordable PET/CT (Shrikhande et al., 2012; Spick et al., 2016). With these trade-offs as considerations, it is important to know which imaging modality is the most appropriate for the specific region under assessment. It is also possible to have multimodal imaging, such as PET/CT, which is either done asynchronously, acquiring images at different times and fusing them through digital image manipulation techniques; or synchronously, simultaneously acquiring and automatically

fusing them (Martí-Bonmatí et al., 2010).

An example of the different modality outputs can be seen in Figure 4.1, where Figure 4.1a shows a CT-scan, and Figure 4.1b shows an MRI scan of a liver showing hypervascularized hepatic (source: Kohler et al., 2018, Figure 1). With the current level of technology, the primary ways of viewing the different scans is to either switch between which modality they are viewing, use a blended image of the two modalities as often done for PET/CT, or by viewing the images simultaneously side by side (Kinkel et al., 2002; Leclerc et al., 2015).



(a) CT scan (b) MRI scan

Figure 4.1: Contrasting image results from a CT-scan (a) and (b) an MRI. As can be seen, the different imaging modalities offer different focus points and levels of detail, source: Figure 1 in Kohler et al. (2018).

In their literature review, Spick et al. (2016) found that the combined PET/MRI protocol is a feasible imaging protocol for all types of cancer. However, they were not able to establish a clear diagnostic advantage when the PET/MRI protocol was used for mainly an anatomic framework. Additionally, they surmise it would be difficult to demonstrate this as the nearest comparator PET/CT already has high accuracy for a large number of cancers. Of high interest is the fact that

multiparametric[1] PET/MRI (as opposed to standard PET/MRI) may have additional advantages for better allocation of bone metastases and disease involvement in the prostatic areas. On the other hand, PET/CT is still superior in terms of lung assessment and still holds great relevance for many types of cancer. They conclude that PET/MRI will be more beneficial for cancers that are routinely imaged with MRI when an addition of PET scans (with various probes) can provide added informational value. However, they recommend that PET/MRI, still more expensive than PET/CT, should not be used simply as a replacement at this point in time (Spick et al., 2016). However, it still has use cases where it will be recommended, such as studying tumour perfusions[2] at baseline, and in response to therapy – this has the potential to provide useful insights into both the delivery and effectiveness of drug-treatment.

One interesting subfield of cancer diagnostics and treatment is the field of Radiomics, which comprises such features as size, shape, descriptors of image intensity distribution and relationship between voxels, to name a few. Yip and Aerts (2016) compared radiomic features, which can be unstable between imaging scans acquired within weeks or even minutes of each other, due to minor breathing and shifting of patients during imaging. They concluded that PET/CT provides greater diagnostic accuracy than either PET and CT on their own. It is also a high-performing modality for determining positive malignancy, sites of disease, primary tumour detection when unknown, staging disease, estimating prognosis, identifying residual disease and confirming sites of recurrence, planning for radiotherapy, predicting early response after treatment starts, and lastly, objectifying

[1]MRI with additional imaging and contrasting
[2]The passing of blood through tumourous tissue.

the efficacy of treatment. MRI has excellent tissue contrast and multidimensional functional, structural and morphological information, compared to a CT scan. PET/MRI has some drawbacks, such as MRI's ability to interfere with PET as the magnetic field and radio frequency can interfere with the electronics of the PET, while the PET might interfere with the magnetic field and radio frequency of the MR imaging. PET imaging has accurate attenuation correction which is less direct with MRI than CT, as MRI has information on proton density while attenuation is proportional to electron density. MRI-based PET attenuation correction has the issue that attenuation is not directly correlated with the signal measured by MRI.

4.1.2 Imaging modalities and sensory correspondence

As discussed more thoroughly in the literature review, being able to take in stimuli from multiple different cues at once can improve precision in psychophysical tasks compared to only having a single cue available (Ernst & Banks, 2002; Gepshtein et al., 2005; Helbig & Ernst, 2007; Hillis et al., 2002). This holds true for a number of different sensory combinations, such as having multiple cues of the same modality (such as texture and shading from vision) or from different modalities (vision and touch, vision and sound). Based upon this research it would be predicted that having several different cues available at once would improve object detection in medical images as well. However, due to the nature of these medical images it is difficult to visually combine the different imaging modalities without losing some of the important aspects like spatial resolution, or introducing undesirable artefacts such as aliasing and noise (Yadav & Yadav, 2020). While Yadav and Yadav (2020) discuss and contrast some of the methods used, another way to

view these images simultaneously would be to externalise the second cue to a different modality, such as haptics. Being able to explore an image both visually and haptically would greatly increase the information bandwidth available, which would overcome the issue of how to simultaneously present two separate images at once, and could lead to an increase in precision. That is, assuming that the use of different medical imaging modalities in the different sensory modalities would be automatically integrated by the sensory system, and that the natural incongruences of the images created by the different scanning technologies would not be sufficient to cause the sensory correspondence between the signals to break down.

While both MRI and CT scans are rich 3D density maps of the human body, with the same overall anatomical landmarks, they capture different levels of detail and relative contrasts depending on how the different tissue types of the body absorb and reflect the scanning signal. In a CT scan, which uses x-rays, there is a large difference in contrast between bone tissue, soft tissue, the vascular system, and 'empty' space, as bones readily absorb the radiation, while soft tissue like muscle and some denser organs absorb some, and 'empty space' doesn't absorb it at all, giving a range of densities from 'white' being bones or metal, soft tissue and liquids being different levels of grey, while air is shown as 'black' (Broder, 2011). As liquids and fat are both less dense than soft tissue, one typically injects a contrast liquid into the veins when wanting to look at the vascular system in detail, which cause them to react much more strongly to the radiation of the scan, giving a clearer and more detailed scan of the circulatory system (Broder, 2011). For an MRI scan, which uses magnetic resonance, there is a much larger level of detail between the different soft tissue structures, due to the difference

in resonance of the different fluid characteristics of a given tissue. As neither air nor bones themselves have fluid characteristics, both of these appear as 'black', signal-absent areas on an MRI scan, while adipose tissue and soft tissue show up in different levels of greyscale. MRI scans also have a spatial distortion associated with the expansion of the magnetic field (Seibert et al., 2016), and while this can be corrected to some extent, the act of doing so may extinguish details in the raw data or introduce unwanted artifacts.

As such, even if one were to have a platonic ideal pair of images from the two scanning modalies, where the only difference is the absorption and reflection of the scanning signal itself, there will still be a not-insignificant level of incongruence between the two, in terms of bone tissue, adipose tissue, and soft tissue, and what the density value means within the respective modalities and image regions. However, due to real-world practicality issues, it is impossible to perform these scans simultaneously. Other errors are introduced through difficult-to-control-for issues such as patients not lying perfectly still throughout the scan, changes and positional shifts of the internal anatomy associated with normal breathing, or even'chaotic' movement of the organs themselves (Papiez & Langer, 2006). Even when controlling for and correcting spatial distortion and relative density mapping, it is therefore impossible to obtain the platonic ideal, perfectly matched image scans.

Lastly, there is the naturally occurring incongruences between visual and haptic texture perception (Kuroki et al., 2019; Marks, 2014), which may cause difficulties in integrating the signals even when the signal sources are from the same imaging modality, discussed in more detail in Chapter 1. This experiment aims to look at just how different can the visual and haptic textures be and still be

considered to have come from the same source, and as such, be integrated by the sensory system. By having sensory signals which are spatially and temporally congruent, with congruent texture features, we are aiming to show that an increase in perceptual dissimilarity breaks down this integration, within smaller levels of dissimilarity. Specifically, by looking at perceptually 'similar' textures, seeing if a larger dissimilarity rating between these negatively impacts performance, where dissimilarity is measured in Euclidean distance in an 8-dimensional perceptual space. If a lower-level of dissimilarity does not significantly reduce slant discrimination ability, it has a positive indication to whether it would be feasible to render the two sensory modalities using the different imaging modalities.

In this experiment, we are expecting that an increase in dissimilarity and incongruence between the visually and haptically rendered textures will cause a gradual decrease in slant discrimination precision, as modelled by cumulative Gaussian psychometric function slopes.

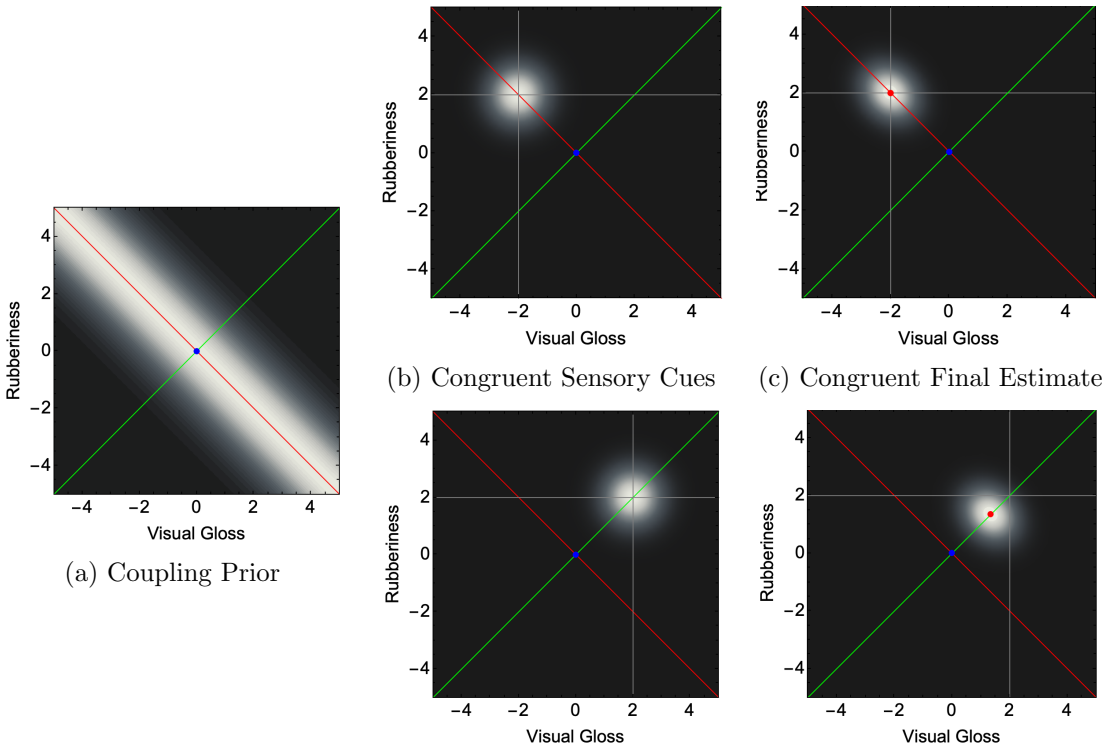
Effect of signal incongruence

Several studies have shown that intentionally introducing incongruence causes the expected perceptual disruption of task performance. In a study by Adams et al. (2016), the authors used a 3IFC odd-one-out paradigm to investigate how visual gloss interacts with haptic perception of the surface of a computer generated potato-like lumpy object, aiming to confirm whether the sensory incongruence of haptic roughness would affect the perceived level of visual gloss. The study consisted of two experiments with stimuli that intentionally violated perceptual predictions, achieved by manipulating three variables independently: visual gloss, haptic rubberiness and haptic friction. Their overall results showed that the intro-

duced incongruence increased thresholds, which translates to lowered discrimination performance by the observers. The first experiment had two subconditions: the congruent Type A, where haptic rubberiness and friction were inversely proportional to gloss and decreased as the visual gloss increased (red diagonal, highlighted, Figure 4.2a); and the incongruent Type B, where the haptic rubberiness and friction were directly proportional with gloss and increased as visual gloss increased (green diagonal, Figure 4.2a). In a real-world scenario, one typically would associate high-gloss surfaces to be hard and smooth, while matte surfaces are more associated with softer, more pliable object, as shown in Figure 4.2.

The predictions of the proposed model is that the probability distribution of the final sensory estimate will trend towards the expected coupling, where the expected coupling is that as gloss goes up, rubberiness goes down, and vice versa (Figure 4.2a). In a congruent trial (Figure 4.2b and Figure 4.2c), the presented stimulus (red dot) is following the expected prior information, which gives an estimate with accurately perceived gloss compared to the standard (blue dot). In incongruent trials the cues are presented in a manner that does not follow prior information (Figure 4.2e and Figure 4.2d), where the perceived gloss of the stimulus (red dot) is dragged towards the expected coupling, compared to the standard (blue dot). The authors found that the Type A condition had a higher performance in the form of lower psychometric thresholds, compared to the results of the Type B condition. In their second experiment, they manipulated the three variables independently and found that gloss was negatively correlated with friction, but unrelated to compliance. Friction was negatively correlated with gloss, but also strongly negatively correlated with compliance.

Compliance, however, was found to be highly related to the perceived higher



(a) Coupling Prior

(b) Congruent Sensory Cues

(c) Congruent Final Estimate

(d) Incongruent Sensory Cues

(e) Incongruent Final Estimate

Figure 4.2: (a) shows the expected link along the red diagonal, where high rubberiness is associated with low visual gloss. In (b) the cues are congruent and lie along the coupling prior between rubberiness and visual gloss which, as shown in (c) the perceived estimate of rubberiness and visual gloss of the target stimulus (red dot) remains the same as was presented in comparison to the standard (blue dot). In (d) the sensory cues are not following the expected prior information, instead presenting a high level of rubberiness with a high level of visual gloss. (e) shows the final estimate when the integrated cues go against the coupling prior; the final estimate shows the target stimulus (red dot) perceived as less glossy and less rubbery than were presented in (d) compared to the standard (blue dot), shifting the estimates towards the coupling prior.

friction, but unrelated to perceived gloss. These findings strongly indicate that as visual gloss increases the perception of friction decreases, and as friction increases the perception of gloss decreases. The visual glossiness was unrelated to compliance, but as compliance increased the perceived friction also increased. In short, a glossy surface is perceived as smoother than it necessarily is, but no perception

changes are made to how soft and malleable the surface might be. On the other hand, if a surface is very pliable it is perceived as having higher friction than it might actually have, and a higher friction surface might seem more pliable than it objectively is.

These findings strongly indicate the perception of visual gloss and haptic friction are inversely proportional (as visual gloss increases the perceived friction decreases and vice versa). In short, a glossy surface is perceived as smoother than a matte surface with the same haptic friction, but no perceptual changes are made to how soft and malleable the surface is perceived as. And conversely, a surface that is presented as highly pliable and soft is perceived as having higher friction than a harder, less pliable surface is, even though they have the exact same friction level. This firmly suggests that different cues influence each other based on prior experience, and that the strength of the coupling between cues varies depending on the respective cues. However, it is unknown to what extent this could affect performance when using similar-but-different sources, such as two superimposed medical images in presented in different modalities such as vision and touch.

In addition to incongruent signals, signals that are unrelated but similar and presented simultaneously can be detrimental to performance. In a study by Bresciani et al. (2006) on the effect of unrelated but similar signals presented simultaneously. In the first session the task was to count the number of flashes and the taps were task irrelevant, while in session 2 the task was to count the number of taps and the flashes were task irrelevant – in both sessions the number of flashes and taps could differ by ± 1. They found that the two sensory modalities would mutually bias one another and the perceptual estimates were being influenced by the task-irrelevant modality. They also showed that the influence of the more

reliable modality on the less reliable one was significantly stronger than the other way around.

However, even though unexpected and unrelated signals can be detrimental overall, that does not necessarily bleed over to the arbitrary signal types. Ernst (2007), which is discussed in more detail in Section §1.1.4, explored the possibility of observers learning to combine two arbitrary signals, compared to the norm of simply different signals of the same property. In their experiment, observers were trained using stimuli that are usually completely unrelated: haptic stiffness and visual brightness. In order to measure the influence of learning on precision, they compared the performance in a 3 interval forced-choice (IFC) odd-one-out detection task from before and after the observers underwent training. They found that, in line with their prediction, successful integration makes discrimination performance worse for incongruent stimuli, and the more certain the observers are about the newly learnt mapping, the stronger the influence is expected to be on performance. Again, this shows great promise to the potential of having observers learn to integrate the information of having different medical images presented for the visual and haptic modalities, respectively.

4.1.3 Textures

In this experiment we are interested in the potential for using medical images from different sources, in vision and touch. From a computer vision standpoint, medical images can in essence be viewed as a collection of different textures. By using known, discretely comparable textures for vision and touch we are aiming to make a judgement as to whether the usage of different image sources is viable in visuohaptic image exploration.

Texture categorisation

Before we can delve into similarity between textures we must first define what it means when two textures are considered to be 'similar'. When looking at real-world textures, they are typically categorised and grouped by their statistical properties, which include luminance levels, standard deviation and skew of the pixel distributions of the image. These properties are divided into two statistical categories; lower-level and higher-level statistical properties. Lower-level properties include luminance and standard deviation of the image's pixel distributions, while higher-level properties include the kurtosis, overall shape of the distribution, and the skew of the luminance distributions of the pixels in the image. When textures are grouped for similarity purposes, this is either done perceptually by human observers, or objectively by grouping images by their respective statistical features (Clarke et al., 2012; Clarke et al., 2011; Kuroki et al., 2019; Motoyoshi et al., 2007).

Texture similarity

As our experiments build around the human perception of visual and haptic signals, we are more interested in the perceptual side of similarity. Clarke et al. (2011) created a large dataset of naturalistic textures using perceptual similarity data collected from human observers, and compared similarity ratings given by several different sorting algorithms to the similarity ratings done by perceptual data. In their experiments they found that, while all the algorithms performed near-ceiling at the classification task where it was tasked to group together images of textures by their 'similarity', the machine performance did not correlate well with that of human perceptual ratings for the same task.

In a related paper, Clarke et al. (2012) looked more into the relative similarity and dissimilarity of these textures as features of an 8-dimensional Euclidean space based on the human-sourced perceptual estimations from previous experiments. The resulting PerTex dataset is a large dataset of 334 naturalistic textures, both represented as a height map and visually rendered with a uniform gloss and illumination angle. Due to its large number of textures and accompanying perceptual ratings, the PerTex database and the accompanying similarity matrix was chosen to be the source of the stimuli used for Experiment 2, with the perception-based 8-dimensional features used as a basis for the different dissimilarity levels used in the texture selection for this experiment, which is explained in more detail in the Texture selection subsection (§4.2.2). The exact method used in the selection procedure is explained in detail in the Methods section (§4.2).

Haptic texture perception

The sense of touch is highly developed in humans. We are able to touch a shiny, transparent surface and identify if it is likely to be glass or plastic, depending on whether it feels cool to the touch or whether it bends when pressure is applied. Similarly, it is possible to identify whether an object is metallic or plastic by the relative friction of the surface material and how heavy it is when picked up (Okamoto et al., 2012). A recent study by Kuroki et al. (2019) has shown that while both vision and touch uses lower order statistical features (such as the mean and standard deviation of the luminance of the pixels in the image) when categorising realistic textures, they found that haptics does not appear to share the same sensitivity as vision regarding changes in the texture's higher order statistical features. In a series of experiments the authors aimed to explore

human accuracy in tactile discrimination of 3D printed texture-pairs by using a 2AFC match-to-sample task, which used three 3D printed textures. In the task, the goal texture was the same as one of the two presented textures, rotated 180°. The observers were asked to scan these textures either through passive or active exploration. The authors measured three sets of five textures, where the first two sets were spatially bandpass random noise pattern, with several Gabor components embedded. The manipulated variables for the textures were in set one difference in the Centre Frequency ('coarseness' of presented textures), for set two, the filter bandwidth was manipulated, while the third set instead used five natural visual textures. In the third set, the printed textures were depth-matched across the average and variance of the depth across the images, while leaving the rest of the statistical features intact. Through their series of experiments they found that tactile perception is sensitive to low-order statistical differences, though the tactile perception was not as high as their visual perception of greyscale images.

Where a texture might be easily differentiated visually, it might be considered perceptually identical, also known as a 'metamer', from a haptic perspective. A perceptual metamer is when two objects appear identical, under specific circumstances. For example, a shallow texture and a deep texture when viewed from above with strong illumination would appear identical, while shifting the position of the viewer or of the light source would reveal the difference between the objects (Backus, 2002; Hillis et al., 2002; Ho et al., 2006).

4.1.4 Designing the task

In order to ensure a potential effect of the haptic cue, we need an experimental task that is well understood from a perceptual viewpoint and allows us to explore

the effect of using increasing levels of difference between the visual and haptic stimuli. Additionally, the task needs to have an appropriate training phase to allow observers to familiarise themselves with the task, as well as to verify their ability to discriminate between the stimuli at maximum presented textural dissimilarity. The current method of tumour delineation on a monitor or tablet is in effect a type of frontoparallel planar stimulus, to which a slant-discrimination task could be well suited.

Slant discrimination is a well researched and thoroughly understood perceptual task, with several studies done on the visual perception of slant such as the effects of different textures on visual slant percention (Rosas et al., 2004), intra-modality combination of stereo and visual textures (Hillis et al., 2004; Knill & Saunders, 2003), as well as inter-modality combination of visuohaptic signals for both congruent and incongruent visual and haptic signals (Hillis et al., 2002; Rosas et al., 2005).

Much research has gone into the visual perception of slant, especially surrounding the effect of texture types. In their paper, Rosas et al. (2004) performed two experiments looking at the effect of individual textures on visual slant perception. In both experiments observers performed a 2IFC slant discrimination task where they were asked which of two physically slanted textures they perceived as being more slanted, using four different slant levels and eight different visual textures. In their first experiment, the slant was applied on the pitch rotation, so the textures were at low-slant near vertical and at high-slant near horizontal to the observer, while in the second experiment the plane was rotated 90°, slanting at the yaw rotation. The authors found that both the texture type used and the degree of presented slant had a significant effect on precision, where different textures had

a greater effect at lower slant (26°) compared to higher slant levels (66°), where there was little difference in performance between the two experiments.

Both Knill and Saunders (2003) and Hillis et al. (2004) explored cue combination of slanted textures presented with both slanted textures and stereo disparity in 2IFC slant discrimination tasks, the results of both supporting each other in showing that the cues were combined in a manner consistent with the MLE optimal cue combination model.

Several studies also exist on the effect of haptic cues on the perceived surface slant, such as by Rosas et al. (2005) who performed two 2IFC slant discrimination experiments, the first of which used congruent signals where the presented slant of the textured visual cue and the untextured, 'flat' haptic cues were the same, and the second of which used incongruent signals where small perturbations were added to the slant of the respective cues in order to investigate the underlying mechanics of the cue combination. Contrary to most studies, the authors here conclude that the weighted averaging was a poor fit to the data, indicating a lack of statistically optimal combination. Their findings are similar to ones found by Hillis et al. (2002), who used an odd-one-out 3AFC task to compare the information combination between disparity, visual texture and haptic slant. Hillis et al. (2002) found that within-modality vision did adhere to the MLE optimal cue combination model, but did not find the same benefit with between-modality vision and haptics, citing a probable difference in the coupling priors as one commonly has experience in looking at one object while touching an unrelated one.

As our experiment is looking at the effect of incongruent visuohaptic textures on precision, the body of existing literature on slant and texture provides a good indication that a slant discrimination task would be well suited for our experi-

mental design. As such, a slanted planar stimulus was selected as a basis for a 2AFC slant-discrimination task for the experiment.

Training and cue recruitment

From Experiment 1 we found that in order to benefit from the haptic signal, training the haptic-only modality separately was a key element in successful cue integration. This finding is backed up by the findings of Ernst (2007), as discussed earlier in section §1.1.4. With this in mind, we added two different training tasks, one on haptic texture recognition and one on basic slant discrimination. The primary training task was selected to be a simple match-to-sample, where observers were asked to match one of two haptic-only explorable textures to a vision-only presented target texture, serving the function of both aiding participants with associating the feel of the haptic texture to its visual representation before the main task, and allowed us to potentially screen participants who might be unable to reliably differentiate between the haptic textures presented. Further details of the main and training task is located in the Methods section (§4.2). The second training task was added in order to ensure the range of slants was appropriate to fit psychometric functions per participant. We know from previous studies that cue recruitment for cues that are novel and arbitrary can be trained, as found by Haijiang et al. (2006) who, as mentioned in Chapter 3 looked at vision-only cue recruitment through the use of Pavlovian conditioning. They successfully trained participants to perceive the bistable stimuli in a specific direction depending on where on the screen it was presented. However, another study by van Dam and Ernst (2010) found that pre-exposure to the bistable stimulus can bias the perception of the ambiguous stimulus, preventing the effect found by Haijiang et al.

(2006). van Dam and Ernst conclude that even small variations in an experimental paradigm could have large effects for learning of perceptual biases for ambiguous stimuli. From these studies as well as the results from the previous experiment in Chapter 3, we know there is scope to train people to use the haptic signal – whether it is novel or simply reinforcing a prior coupling – though it is important to keep the experimental procedure consistent, as even small variations could introduce larger biases as found by van Dam and Ernst (2010).

Training

In Chapter 3 we found that in order to benefit from the haptic signal, training the haptic separately was a key element in successful cue integration. This finding is backed up by the findings of Ernst (2007), as discussed earlier. With this in mind, we added two different training tasks, one on haptic texture recognition and one on basic slant discrimination, to ensure the range of slants was appropriate per participant.

Hypothesis

In this study we aim to investigate the effect of incongruent-but-similar visual and haptic signals on precision in a slant discrimination task, emulating a mismatch that would occur if using visual and haptic stimulus from different medical imaging modalities which can in essence be viewed as a collection of different textures. We will be looking at how relative difference between visual and haptic textures affects precision in a simple 2AFC planar slant discrimination task. By having four separate conditions with a baseline of the same texture used for both vision and touch, we linearly increase the difference between the textures through the

use of an additional 3 visual textures which are selected to be of known perceptual difference to the base visual texture, while leaving the haptic texture unchanged throughout all conditions. We are aiming to investigate whether the addition of the texture conflict interferes with the precision of the observers' judgement of slant, decreasing their precision as the level of incongruency increases. If this conflict has an effect, we would expect that to be apparent in decreased precision of slant-discrimination as the difference between the textures increases, as modelled using cumulative Gaussian psychometric function fits over a range of 11 linearly spaced slant angles.

4.2 Methods

4.2.1 Participants and setup

In this experiment we collected data from a total of 10 observers, of which 8 were naïve to the purposes of the experiment. 8 of the observers were right hand dominant, while the remaining two were left handed, and all participants had normal or corrected to normal vision and a stereoacuity of 60 arcsec or better, as measured with the Laméris Ootech TNO Stereo Vision test. The physical setup used in this experiment is the spatially coaligned visuohaptic rig, which is described in detail in Chapter 2 §2.1.3.

4.2.2 Stimuli

The stimuli were naturalistic texture surfaces taken from the Edinburgh PerTex database of naturalistic rendered surface textures (Clarke et al., 2011), selected

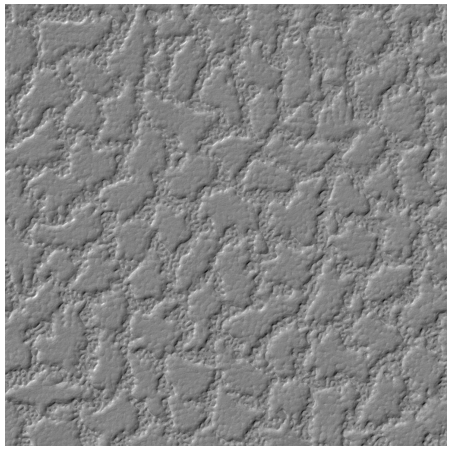
to be of a known, numerically quantified similarity distance to one another. One texture was selected to be the 'base texture' of a set, where the haptic and visual textures were both selected to this initial texture, which has a theoretical Euclidean distance of $(0 * \Delta_{Diff})$ this pairing is hereby termed Δ_0. Each subsequent texture was selected to be visually different at the aforementioned predefined distance interval created from the Euclidean distance between the textures, as defined in an 8-dimensional feature space – the mathematical details of the selection procedure follows in the next section. These texture-pairs are referred to as Δ_1 $(1 * \Delta_{Diff})$, Δ_2 $(2 * \Delta_{Diff})$ and Δ_3 $(3 * \Delta_{Diff})$, respectively.

Some example texture-pairings are shown in Figure 4.3. In Δ_0 trials, the visual and haptic textures were from the same set – for example Figure 4.3a and 4.18b are both of texture 315 – while in Δ_1, Δ_2 and Δ_3 the visual texture was from a different set than the haptic – for example Figure 4.3c and 4.18b, where the visuals is texture 26 but the haptic stimulus is texture 315.

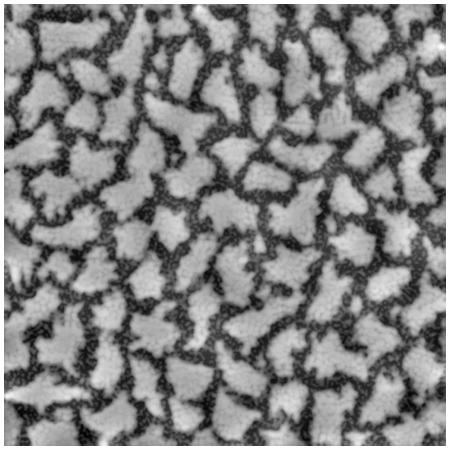
Texture selection

One of the fundamental questions for the selection process of the textures is how does one define dissimilarity on a perceptual scale? There exist many different ways of comparing and deciding on what makes images and textures similar, but very little has been done on quantifying the inverse; a lack of data quantifying a feature is not the same as absence of said feature. None of the conventional measures of similarity (structural similarity, luminance, geometry, amplitude) can individually be directly inverted to measure dissimilarity, as a directly inverted version of an image would still be perceived as similar by human observers.

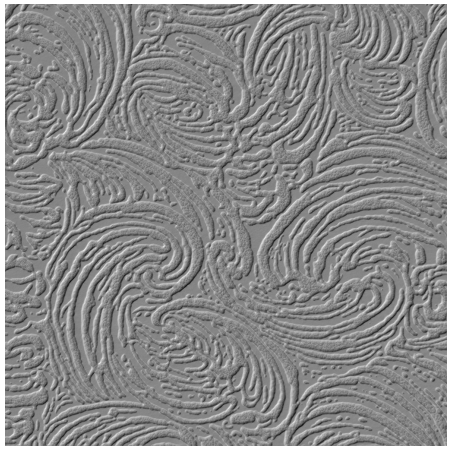
One example of such is the potential issue with structural similarity index



(a) Rendered texture 315



(b) Height-mapped texture 315



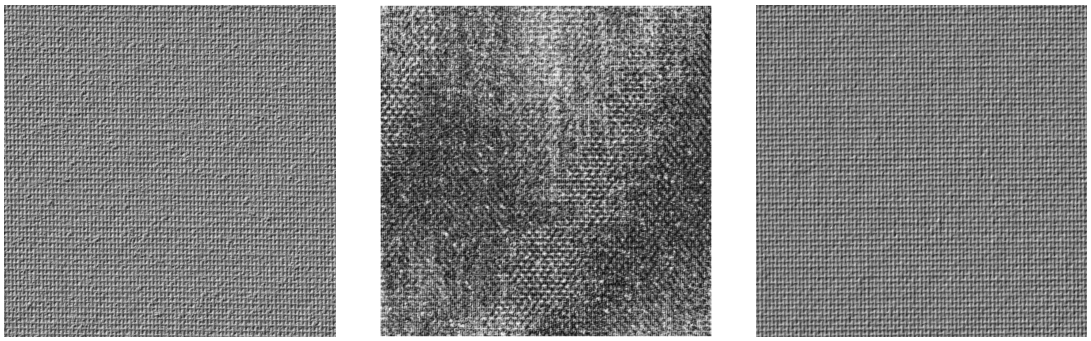
(c) Rendered texture 26

(d) Height-mapped texture 26

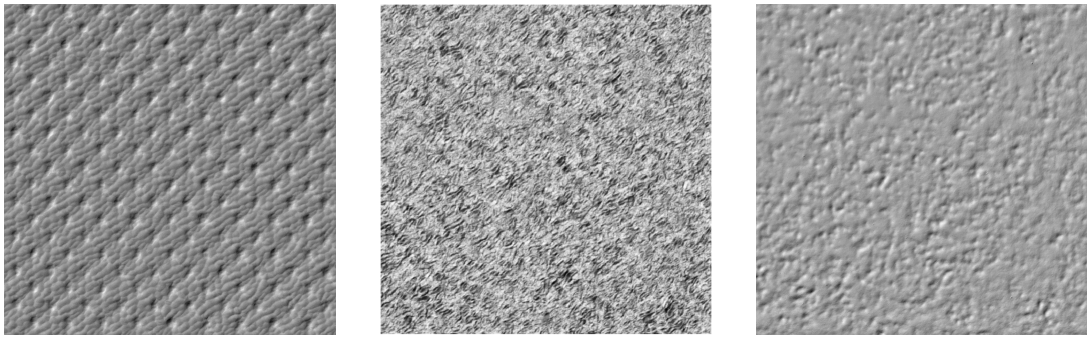
Figure 4.3: The two texture formats from the PerTex dataset. (a) and (c) show the visually rendered textures, with the same illumination angle, albedo and normalised luminance value. (b) and (d) show the height-mapped versions of the same textures, where the luminance of the pixels goes from black (low) to white (high). These height-maps were used in the haptic rendering of the textures.

metric (SSIM), a method that compares the structural qualities of two images. In certain circumstances, SSIM is able to quantify two nigh-identical images as wildly different, as shown in Figure 4.4. The pair in the top row (4.4a and 4.4c) are according to SSIM the least similar texture pair of the entire PerTex database of 334 textures, with an SSIM value of -0.1605 – which on the SSIM scale of 0 to

1 is mathematically impossible – and a PerTex similarity matrix rating of 0.6429. However, the pair in the second row of images (4.4d, 4.4f) are, according to SSIM, the most similar pair in the database, with an SSIM value of 0.2957 and a PerTex similarity matrix rating of 0.1667. From a purely perceptual viewpoint, this is an incorrect judgement.



(a) Rendered texture 107 (b) SSIM pair map (c) Rendered texture 195



(d) Rendered texture 90 (e) SSIM pair map (f) Rendered texture 324

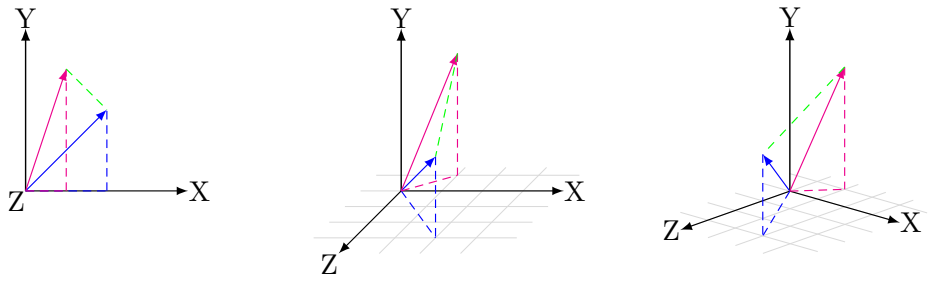
Figure 4.4: One problem with the SSIM similarity function. (a) and (c) show the least similar texture pairs in the database, according to SSIM with a value of -0.1605 (which on the permissible SSIM scale of 0 to 1 is mathematically speaking impossible). In the PerTex similarity matrix (PSM) rating, they have a score of 0.6429. (d) and (f) show the most similar texture pairs in the database, with an SSIM value of 0.2957, but a PSM of 0.1667. (b) and (e) show the perceived differences between the respective textures per pair in the SSIM pair mappings.

A different approach to similarity is Euclidean distance between features. Euclidean distance in space is considered a complicated but robust method of defining

and sorting similarity. While the further reaches of dissimilarity are still vaguely defined, it is possible to enumerate the difference of n-dimensional textures by way of applying linear algebra and comparing the relative positioning between points in n-dimensional space. By treating each quantifiable feature as a separate dimension to compare with, it is possible to create an n-dimensional point per texture, and compare how similar the textures are overall by calculating the length of the vector between the points, using the formula shown in Equation 4.1. A texture might be similar in some dimensions but vastly different in others, increasing the overall length of the vector, as shown in Figure 4.5. This is a basic example of $n+1$ dimensional growth. Figure 4.5a shows the position of two points in a 2D, XY coordinate system. The green dashed line shows the distance between the two. However, when adding in a third dimension of Z, as shown in Figure 4.5b and 4.5c, the distance between the points is shown to be significantly larger than it appeared for the two-dimensional plane, as the new dimension adds in more information, making the position more accurate. While our example covers a $2 + 1$ dimensional growth, the principle holds true for an increased number of dimensions. The more feature dimensions are known for a set of images, the more accurately their position and distance can be calculated (Clarke et al., 2012).

$$\begin{aligned} \mathrm{d}(\mathbf{q}, \mathbf{p}) = \|\mathbf{q} - \mathbf{p}\| &= \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + ... + (q_n - p_n)^2} \\ &= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \end{aligned} \tag{4.1}$$

For this experiment, we calculated and compared the Euclidean distance between the real-world textures from the PerTex dataset (Halley, 2012), using their 8-dimensional coordinates for all the textures' relative positioning in space, which



(a) 2D view of distance (b) 3D view of distance (c) Rotated 3D view of distance

Figure 4.5: Effect of added dimensions on distance between points. (a) shows the distance between two points from a two-dimensional, XY coordinate system. (b) shows the same points from a three-dimensional, XYZ coordinate system. When the Z-dimension was flattened, the perceived distance between the points was reduced, as it did not take into account the distance between them on the Z-axis. (c) shows the same points plotted from a different viewpoint.

is based on an ISOMAP[3] of the perceptual similarity grouping the textures have been rated with. This allowed us to look at the various distances between all the textures, and we were able to select three equally spaced distance values (named 'Deltas', hereby referred to using the Greek symbol Δ), around which upper and lower bounding bands were placed, so the selected textures would be constrained within a reasonable distance respective to their respective Δ measurement. The algorithm used to select both the Δs and the Δ-bands is as follows. The mean of the overall distances between all respective textures and the mean of the standard deviation of the same distances are used to calculate three main values: low Difference, medium Difference and high Difference, as shown in Figure 4.7 ($Diff_L$, $Diff_M$, $Diff_H$ in Equation 4.2). These three initial values were chosen to get a

[3]An ISOMAP is a technique of nonlinear dimensionality reduction that preserves isometric distances by generating features during the scaling transformation from larger to smaller metric space.

wide region of the distance distribution. Note that, as the distance matrix "Dist" is a 334x334 matrix, the mean of the means and mean of the standard deviation is taken.

$$\begin{aligned} Diff_L &= \overline{Mean_{Dist}} - 2 * \overline{\sigma_{Dist}} \\ Diff_M &= \overline{Mean_{Dist}} \\ Diff_H &= \overline{Mean_{Dist}} + 2 * \overline{\sigma_{Dist}} \end{aligned} \tag{4.2}$$

Where $Diff_M$ is the mean of the overall distances and $Diff_{L,H}$ are ± 2 mean standard deviations of the distance ($\overline{\sigma_{Dist}}$) from the mean distance, respectively. In addition to the baseline distance of 0 (representing an image being compared to itself), this produced three thresholds used for the low, medium, and high distances, respectively. The low and medium thresholds were then adjusted so as to produce three equally sized bands of size Δ_{Diff} as described in Equation 4.3.

$$\Delta_{Diff} = \overline{\{Diff_L - 0, Diff_M - Diff_L, Diff_H - Diff_M\}} \tag{4.3}$$

Which is equivalent to $Diff_H/3$, which at $3\Delta_{Diff}$ would put us at the highest possible acceptable value. As we need regions around the Δs to select textures from, the final base value Δ_{Base} was selected to be the middle point between the lowest acceptable value $Diff_L$ and the newly calculated Δ_{Diff}, which gives a selection region of ± 0.1.

$$\Delta_{Base} = \frac{Diff_L + \Delta_{Diff}}{2} \tag{4.4}$$

Which for an integer $i \in \{0, 1, 2, 3\}$ gives us the Δs as described in Equation

4.5.

$$\Delta_i = \begin{cases} 0 & \text{if } i = 0 \\ i * \Delta_{Base} \pm 0.1 & \text{if } i > 0 \end{cases} \tag{4.5}$$

However, as the experiment uses both vision and haptic stimuli, one issue arose around some of the more uniform and featureless textures. These could potentially be considered 'flatter' by the observer, where the cue would be mistaken for 'weaker' and less informative and, as haptic exploration using tools incurs a reduction in perceived reliability (Takahashi et al., 2009), the other sensory channel be given a much higher perceptual weighting. In that eventuality, the data on which the analyses are run would be misrepresented, and as the experiment uses a series of similar textures, there will be a difference in the exact texture composition between visual and haptic textures in 75% of the trials. To combat this, these 'flat' textures were filtered out by sorting through the images and calculating a minimum standard deviation (STD) threshold using Matlab. This 'flatness' threshold was calculated as the mean of the STDs of all 334 textures, minus one STD of all the textures' STDs. The formula (and the calculations) used for this is shown in Equations 4.6 to 4.8, where σ is the STD and c is the intensity, or luminance, of the pixels in the greyscale images. Any texture with a σ lower than 24.46 was deemed 'flat' and excluded from the selection pool. This method successfully excluded 51 unsuitable textures. An example of a 'flat' texture is shown in Figure 4.6.

$$
\begin{aligned}
m_{\sigma_c} &= \frac{1}{N}\sum_{i=1}^{n}\sigma_{c_i} \\
&= \overline{\sigma_c} \\
&= 29.7076
\end{aligned}
\tag{4.6}
$$

$$
\begin{aligned}
\sigma_{\sigma_c} &= \sqrt{\frac{1}{N-1}\sum_{i=1}^{n}(\sigma_{c_i} - \overline{\sigma_c})^2} \\
&= 5.4487
\end{aligned}
\tag{4.7}
$$

$$
Thr = m_{\sigma_c} - \sigma_{\sigma_c} = 24.4589 \tag{4.8}
$$

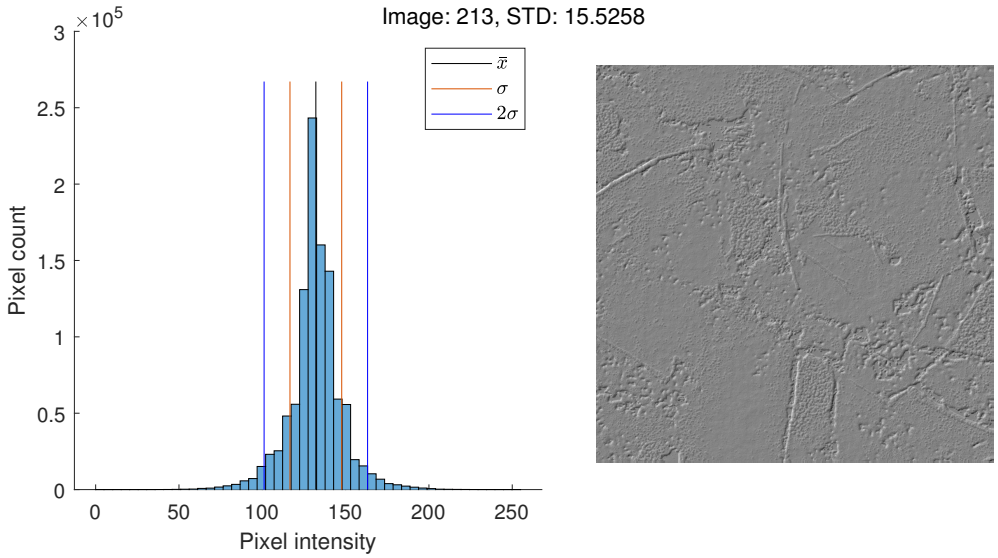


Figure 4.6: Some textures were removed from the potential stimulus pool for having a very small standard deviation of pixel luminance σ, meaning the majority of the pixels were very close in luminance and the textures were considered to be 'flat'. The histogram, left, shows the relative pixel value in the texture, right. The x-axis plots pixel intensity from 0, black, to 255, white. The y-axis plots the number of pixels total in the image that fall within the respective intensity levels.

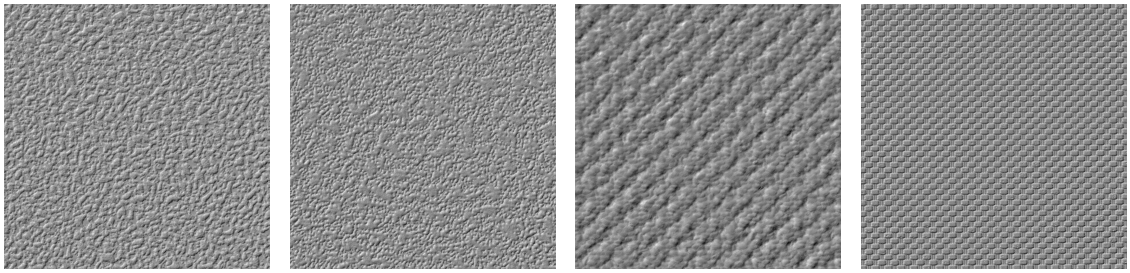
After removing the 51 'flat' textures from the selection pool, some of the remaining textures did not have any viable pairings in some Δ-bands, so an additional 11 textures were further excluded due to lack of viable matches in any of the required ranges. This left us with 272 potential textures to select from, having at least 1 possible texture pair in each Δ band. 21 unique sets were generated, each comprising four evenly spaced textures. Each participant was randomly allocated one of these pre-generated sets.

For the experiment, the four conditions tested were Δ_0, which featured the same visual and haptic texture; Δ_1, where the haptic texture has a Euclidean distance between 0.72 and 0.92 from the visual texture; Δ_2, 1.52 and 1.72; and Δ_3, 2.32 and 2.52. An example texture set is shown in Figure 4.7, with Figure 4.7e showing the number of viable matches to the core texture in Figure 4.7a. The experiment was run on a tilted angle between $\pm$ 4° and $\pm$ 12°, depending on the participant. These angles were spaced evenly between 11 points. Each of these gradients were run 20 times, giving each data point of the psychometric functions shown in Figure 4.13 140 trials. Low difference values are 0.8201 ± 0.1, medium difference are 1.6401 ± 0.1, and for high difference 2.4602 ± 0.1.

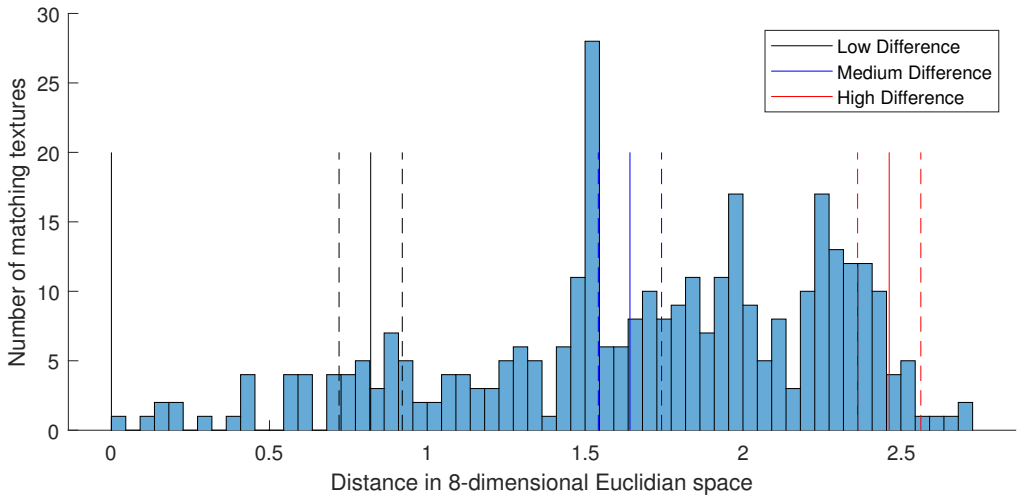
Stimulus presentation

For this experimental set-up the visual and haptic signals were spatially coaligned, where the slanted plane was presented visually using stereoscopically rendered depth in addition to the perspective projection of the visual texture, and haptically using a luminance-based height mapping of the textures.

As slanted geometric shapes can often be estimated by perspective and texture edges, an aperture was included to avoid additional visual cues. The initial



(a) Base texture, Δ_0 (b) Very similar, Δ_1 (c) Dissimilar, Δ_2 (d) Very dissimilar, Δ_3



(e) Significance

Figure 4.7: (e) Histogram showing the number of possible texture matches (y-axis) by relative distance to the base texture in 8D Euclidean space (x-axis), where the dashed lines indicate the limits of the Δ-bands. Here, the base texture 80 (a) has a Euclidean distance of 0.8083 to texture 317 (b), 1.5499 to texture 94 (c) and 2.5278 to texture 46 (d).

piloting of the experiment used a square $10 * 10$ cm aperture around the centre of the image (Figure 4.8a), but the pilot data showed that the visual system far outperformed that of haptics for slant discrimination. In order to ensure that the reliabilities of vision and touch were matched, the reliability of the visual stimulus was intentionally reduced by using a narrower aperture of 1.7 cm wide by 10 cm tall, the dimensions of which are in line with the findings of Burge et al.

(2010). This aperture is illustrated in Figure 4.8b. In their paper, Burge et al. (2010) tried a series of different size apertures to match reliabilities of visual and haptic stimuli for a slant discrimination task. From their estimates of matched reliabilities for their stimuli, the aperture most applicable to our experimental design was 1.6 cm wide for 9.2 cm tall, which is the same ratio used in our 1.7 cm wide and 10.0 cm tall aperture. While using a single constant aperture across all participants and textures assumes a similar visual performance across these, since we are not measuring or comparing discrimination thresholds between the textures we opted for adopting the rough match found by Burge et al. (2010). The alternative would be to collect over 20 psychometric functions per participant prior to starting data collection which was, at the time, determined to be beyond the scope of this experiment.

One of the design considerations of the aperture was the effect of the width of the aperture on the overall viewing geometry, and what the maximum usable slant angles were that would still be obscured by the edges of the aperture, as illustrated in Figure 4.9. For a basic monocular viewpoint this required the angle between the aperture edge and the viewing point (θ_{M1}) to be equal to, or less than, the angle between the edge of the aperture and the edge of the stimulus (θ_{M2}), shown on the left hand side in Figure 4.9a. For a binocular viewpoint this needed to be measured per eye, where the angle between, here, the left edge of the aperture and the right eye (θ_{B1}) must be equal to, or less than, the angle between the left edge of the aperture and the left stimulus angle (θ_{B2}), shown on the right in Figure 4.9a. The same is true for the right aperture angle and left eye. For the pilot aperture of 10 cm^2, as shown on the left of Figure 4.9b, this was at 7°. However, for the new aperture, as shown on the right in Figure 4.9b, this is no

longer a concern.

A further consideration regarding a more narrow aperture is the fusibility of the visual texture, where there is a possibility of the negative parallax being equal to the fusible area, which would prevent fusion from occurring. However, the modified aperture was both verified via piloting, as well as calculated to be within tolerance for the expected inter-ocular distances of participants IOD ≤ 7.0 cm. The fusible area for the wide aperture is shown on the left in Figure 4.9c, while the right-hand side of Figure 4.9c shows the area fusible for the narrow aperture. The effect of the experimental aperture in use is shown in Figure 4.10, comparing the displayed angles of no aperture and with aperture. In order to aid the 3D perception and fusibility of the otherwise flat aperture, and to facilitate the aperture presenting as a 'flat' surface, a randomised series of squares were placed across the surface of the aperture to help indicate depth and placement. The squares were 5 mm in each direction and each of the 39 placement slots for the squares were placed 5 mm from the next, starting at 5 mm from the edges of the aperture and avoiding the centre slot of the aperture itself, removing 5 squares per row for 21 rows. The placement algorithm used a randomised 'coin-toss' to decide whether or not to place a square, averaging 19.5 squares per full row and 17 squares per row for the 21 rows coinciding with the slot. The squares were randomised and placed at the start of each trial, so to avoid familiarisation and spurious visual cues.

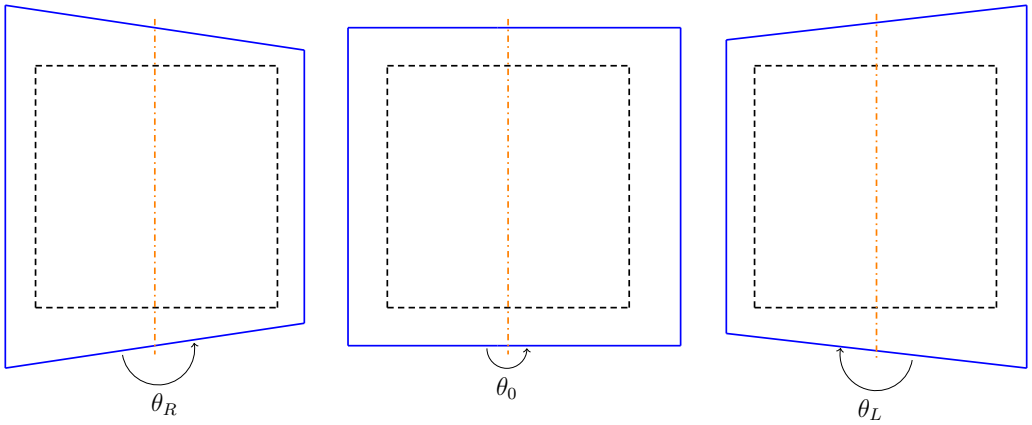
While utmost care has been put in place to ensure the stereoscopic 3D presentation of depth, the experiment itself could be performed under monocular viewing conditions, as the participants also rely on the visual perspective cue to perceive the slant alongside the haptic cue. However, as previously mentioned, it

is well established that spatially coaligned cues are more likely to be integrated and for that the additional stereo vision is required.

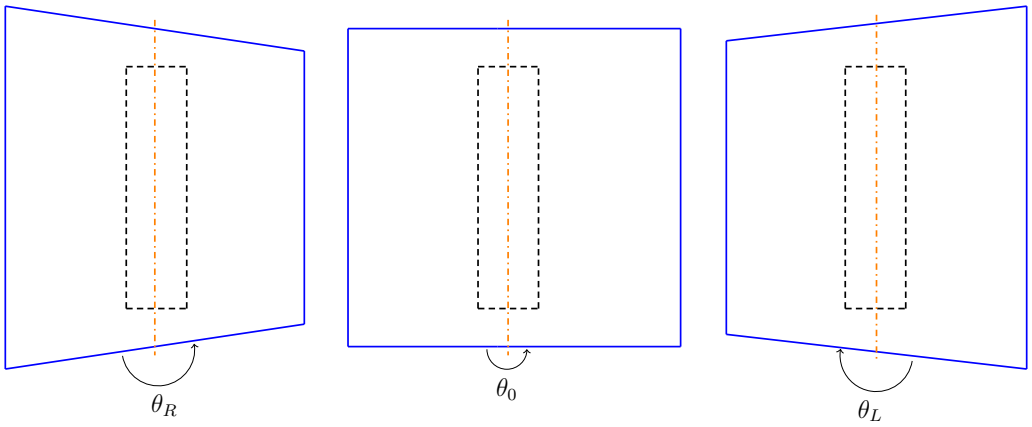
4.2.3 Task/Procedure

Training

As the previous experiment in Chapter 3 found, having a haptic-only training task to recruit haptic perception with the device is a key point in order to benefit from the combined visuohaptic stimuli. While in this experiment we are looking at the negative effect of an incongruent signal, a prior link between the cues is required for the dissimilarity between the cues to have an effect. Furthermore, it is essential that the participants are able to differentiate between the maximum-difference textures on a purely haptic level for the similarity manipulations to have an effect. To ensure that all of the participants can differentiate between these Δ_3 difference textures, a short match-to-sample training task was designed, where the participant is presented with 3 planes in front of them, shown in Figure 4.11. One of the planes contains a visual texture but no haptic information, while the other two planes contain no visual texture and are instead presented as flat, grey surfaces with their own respective haptic texture, as illustrated in Figure 4.11a. For each training trial, one of the two haptic textures will match the target visual texture, and the other will be Δ_3 different. After the participant has selected the texture they believe to be the match, the correct answer is displayed as shown in Figure 4.11b. This training both allowed the participants to familiarise themselves with textures and the use of the device, train themselves to understand the haptic-only aspects of the textural cues, all while allowing us to gauge that they were

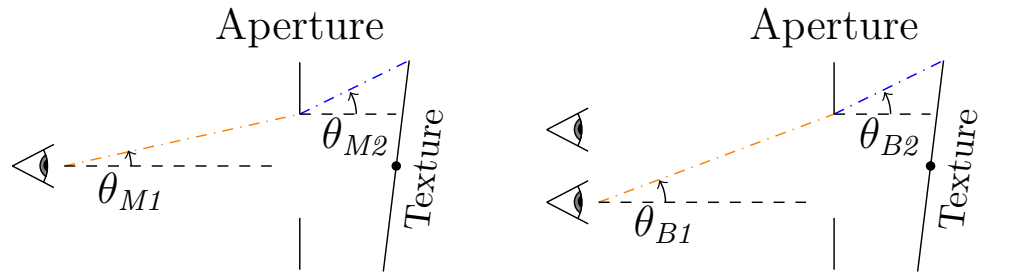


(a) Initial, wide aperture

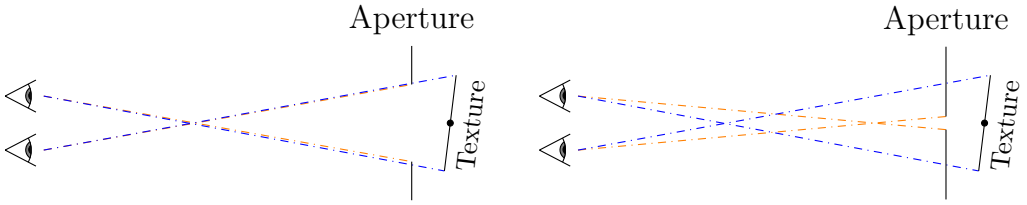


(b) Final, narrow aperture

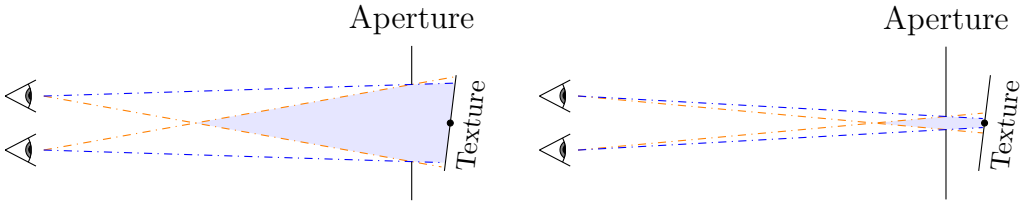
*Figure 4.8: Aperture design considerations, illustrative diagram. (a) The initial design of the rotation and aperture used a 10*10 cm square aperture for the visual stimulus. (b) The final design of the slant and aperture used with a width of 1.7 cm, height of 10 cm, where the size ratios were chosen in accordance with the findings of Burge et al. (2010). While a single constant aperture does assume a similar visual performance across all participants and textures, the alternative being to collect over 20 psychometric functions per participant prior to data collection is beyond the scope of this experiment.*



(a) Exaggerated slant constraints



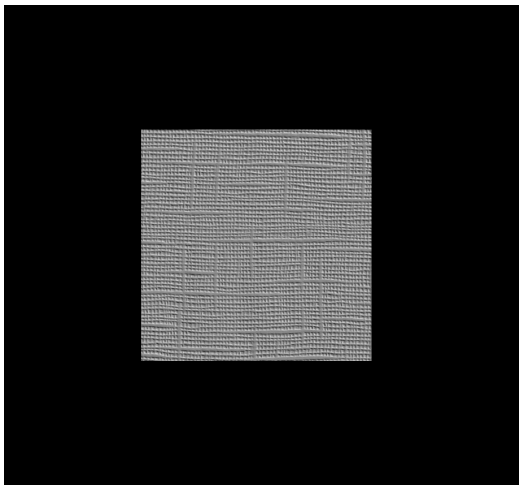
(b) Binocular slant limits



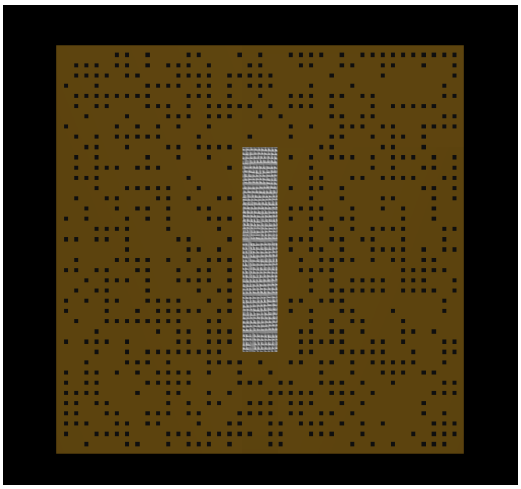
(c) Fusable area

Figure 4.9: The final and initial designs of the slant and aperture used. (a) Stylised angle constraint differences between monocular (left) and binocular (right) vision. In order to ensure the person cannot see beyond the edge of the textured surface, θ_1 needs to be larger than θ_2. For monocular, θ_{M1} and θ_{M2} are at a sufficiently large difference to hide the far edge, while for binocular view θ_{B1} and θ_{B2} are approaching the maximum limit before the edge of the texture is can be seen through the aperture slit. (b) and (c) show, on the left, the initial aperture of width 10 cm and height 10 cm; and on the right, the final aperture of width 1.7 cm and height 10 cm. Both sets are over the 12.5 cm^2 stimulus, at a viewing distance of 47.7 cm, to scale. Measurements were chosen in accordance with the findings of Burge et al. (2010) who had a stimulus of 11 cm^2, with an aperture of width 1.2 to 6.0 cm and height 9.2 cm. The shaded area in (c) highlights the area of visual space that is fusible in binocular vision.

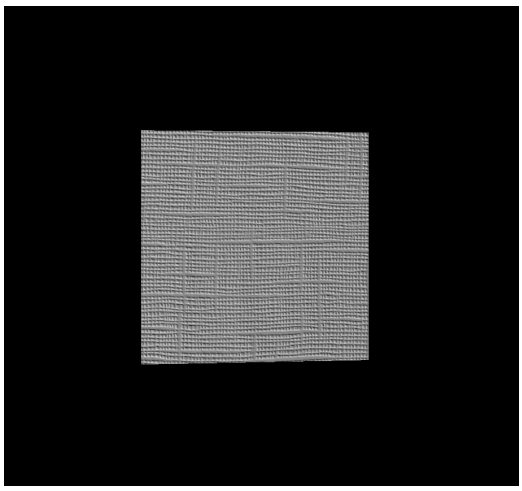
in fact capable of reliably differentiating between the Δ_3 different textures on a haptic level. After four blocks of 10 sets, if the participant hit a correct rate of



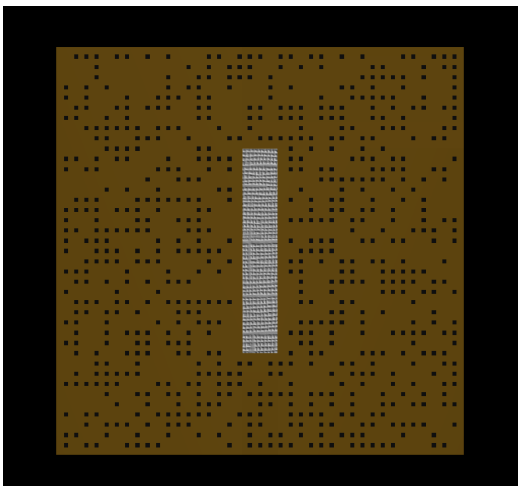
(a) Frontoparallel stimuli, no aperture



(b) Frontoparallel stimuli, with aperture



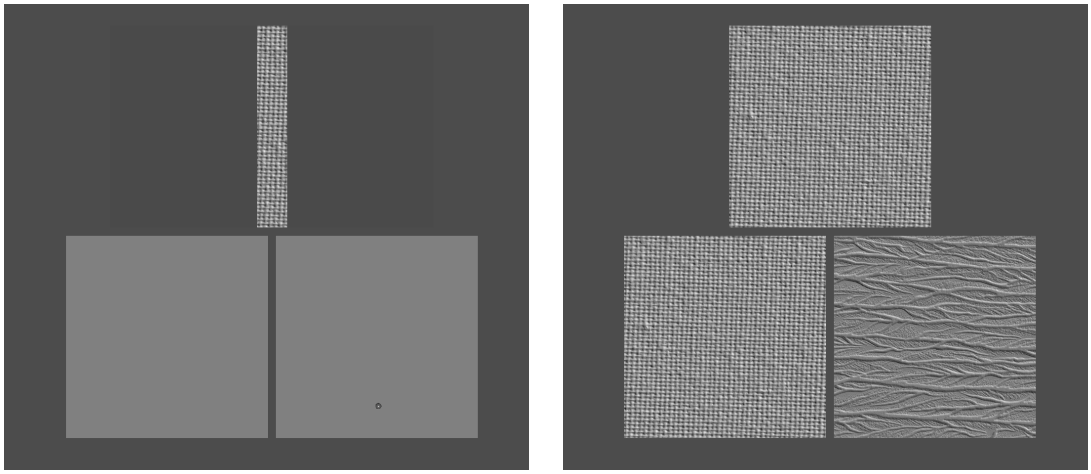
(c) 7° stimuli, no aperture



(d) 7° stimuli, with aperture

Figure 4.10: Screenshots of the second training task and main experiment. (a) and (b) The task starts with a 'menu' screen where the texture is frontoparallel, allowing participants to familiarise themselves with the texture and what frontoparallel looks and feels like. Once they are happy to start, they select anywhere on the texture and the experiment begins. (c) and (d) The stimulus is set to a slant value randomly selected between 11 different slant values spaced linearly between the minimum and maximum angle per participant. As seen in (a), perspective lines are formed at the top and bottom of the texture. In order to combat this, the aperture in (d) and (b) is added on. The small dark squares are randomly generated per trial, and helps establish the aperture as a flat surface, unrelated to the texture. This aids the participants in fusing the texture into a 3D percept.

at least 70%, they were allowed to move onto the experiment. The individual performance of the participants is shown in Figure 4.12.



(a) Training task, prior to selection (b) Training task, post selection

Figure 4.11: Screenshots of the first training task. (a) shows what the participant sees at the start of each training block. One of two textures is randomly selected to be the target texture. The participant explores the two grey squares haptically, before making selecting the texture they believe matches the texture shown. (b) After the selection has been made, all textures are shown visually and the participant has the opportunity to compare the haptic texture with the visual texture present before continuing onto the next trial.

A secondary training phase was used to familiarise the participants with the experimental task itself, and how the slanted stimulus would be presented. It was a shorter version of the main task, using a texture set of Δ_0 that the participant would not encounter in the main experiment. This allowed us to gauge an appropriate maximum-angle to ensure a near 100% performance on a per-participant basis. As the maximum angles are the foundation of the stimulus range presented to each participant in a method-of-constants, it is important to ensure the range is appropriate per individual, for a good psychometric fit to occur (Wichmann & Hill, 2001).

The main experiment

In order to map out a psychometric function of the participant's sensitivity to the right/left slant, the degree of slant was varied at 11 different values spaced linearly between $[-angle_{max}, angle_{max}]$, centred around 0°. At 0° the texture was presented as physically frontoparallel to the observer. The order of the slants presented was randomised on a per-block basis, with each of these slant values repeated for 10 iterations per block, and 2 blocks collected for each of the respective four visual textures. Meaning each participant completed 880 trials total, with every texture-pair having 20 trials per slant level presented.

On each trial, the participant explores a 3D slanted plane with rendered realistic texture, using both vision and touch. The participant then selects which direction they experienced the plane being horizontally slanted to, by touching the tool against whichever side of the plane they experienced as being further

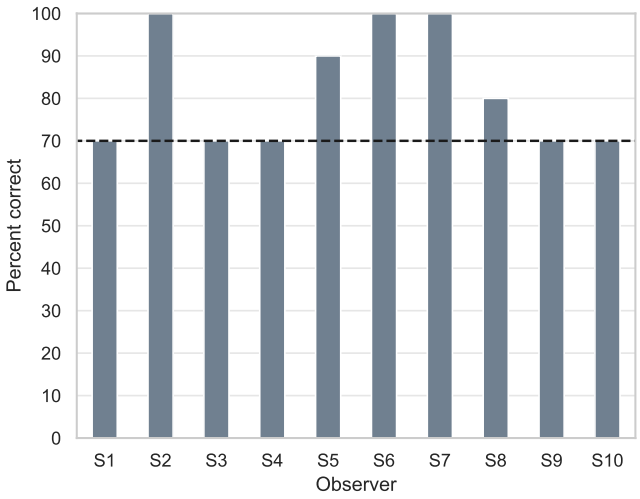


Figure 4.12: Performance of the first training phase, differentiating between two Δ_3 different haptic textures.

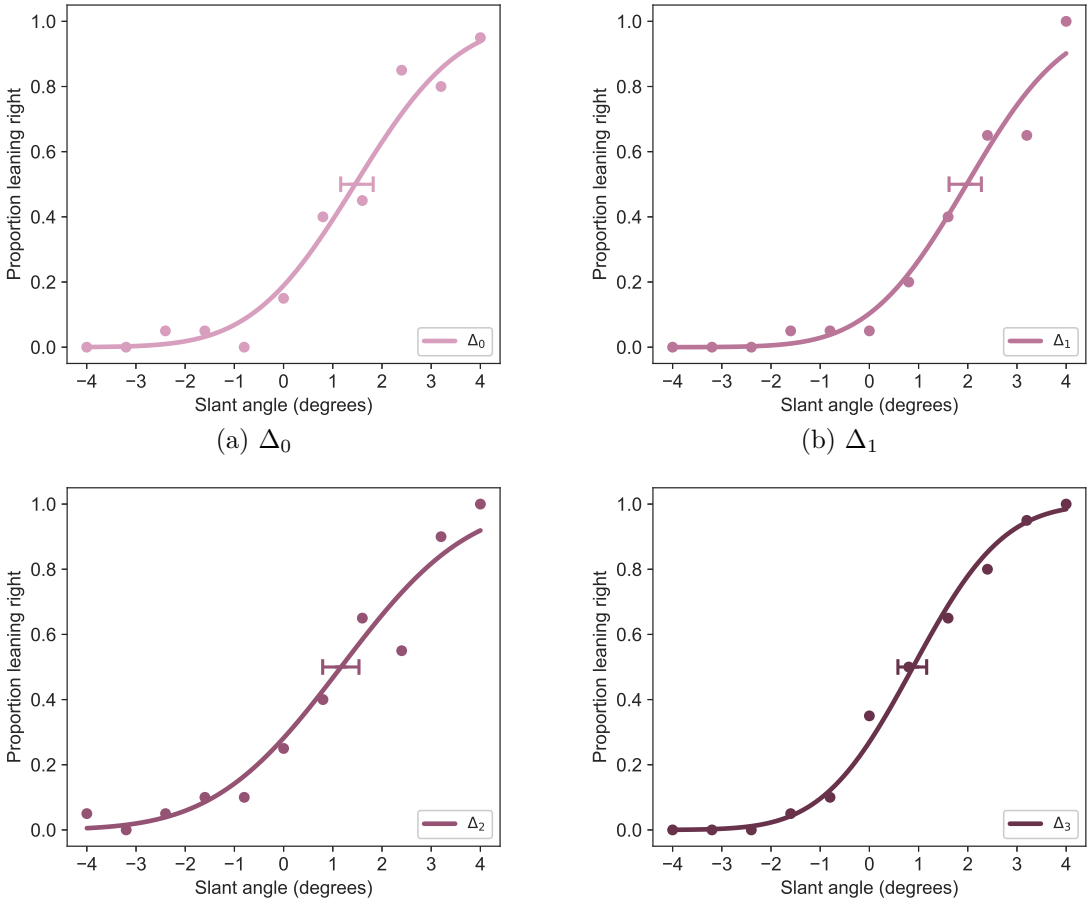
away from them and pressing a button located on the haptic device. This causes the texture to disappear, and the participant starts their next trial in their own time by pressing the button again.

4.3 Results

4.3.1 Psychometric function fits

The main question this experiment aims to answer is whether an increased difference Δ in similarity between visual and haptic stimuli lowers precision across increasing Δ-levels, for a planar slant discrimination task. In order to address this, cumulative Gaussian psychometric functions (PF) were fitted to the participants' data using the Palamedes toolbox (Prins & Kingdom, 2016) in Matlab, as shown in Figure 4.13. Each point represents 20 trials at each respective slant level. From these fits, the slope value was extracted and compared. In a cumulative Gaussian PF, the slope is given as $\beta = \frac{1}{\sigma}$ of the underlying estimator.

Figure 4.13 shows the example data set of a single participant, plotting the measure of performance for the four different Δ-levels. The specific slope values for Figure 4.13 are shown in Table 4.1. Figure 4.14a shows the same four functions overlaid, with Figure 4.14b plotting the slopes of the PF on the y-axis, against the respective Δ-levels on the x-axis, where the error bars are showing the 95% confidence interval obtained through parametric bootstrapping in Matlab, performed 400 times per individual function. As shown in Figure 4.15, the mean slope values does not significantly increase or decrease as Δ increases, indicating an overall indifference to the relative distances between visual and haptic stimulus for these distance levels.



(a) Δ_0 (b) Δ_1

(c) Δ_2 (d) Δ_3

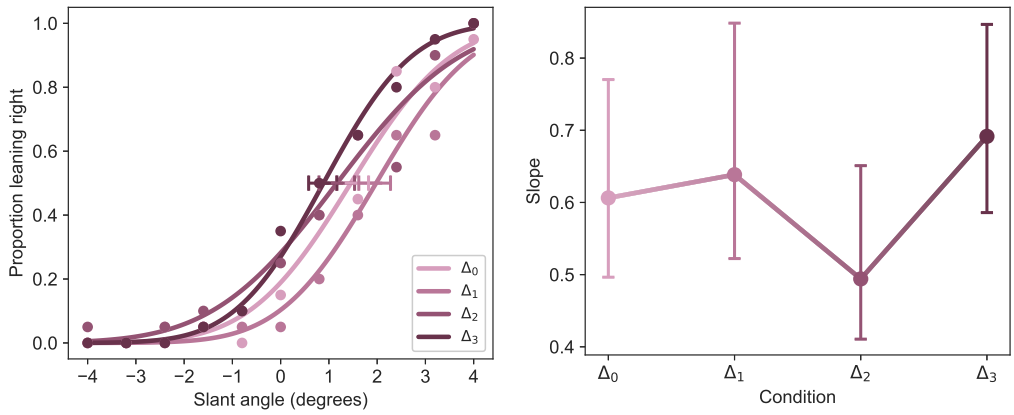
Figure 4.13: Comparative cumulative Gaussian PF fits from the example data set. On the y-axis is proportion of trials the participant perceived the stimulus to be slanted to the right, while the x-axis shows the stimulus slant angle in degrees, and the error bars are of the 95% bootstrapped confidence interval of the PSE.

	Slopes			
	Δ_0	Δ_1	Δ_2	Δ_3
S2	0.6063	0.6385	0.4941	0.6916

Table 4.1: Individual slope performances from example dataset.

4.3.2 Overall results

For us to investigate the predicted detrimental effect of increased dissimilarity of texture cues in a multisensory slant discrimination task, we compare the difference



(a) Psychometric function fits (b) Comparative slopes by condition

Figure 4.14: (a) Comparative psychometric function fits from the example data set. Y-axis is proportion of trials where stimulus perceived to be slanted to the right, where x-axis shows stimulus slant in degrees. (b) Slope performance per Δ between modalities, where the y-axis shows the value of the slope while the x-axis is for the relative distance between the modalities, and error bars show the 95% bootstrapped confidence intervals of the slope. As shown, performance does not significantly differ as the Euclidean distance between the visual and haptic stimuli increases. For the individual functions per observer, please see Appendix B, Section §B.2.5.

in precision for slant discrimination a function of increased Euclidean distance between haptic and visual textures. This is done by extracting the slope values of the PF fits, which were then compared across all participants per Δ level. A repeated measures ANOVA was run on the slopes of the functions and the Δ-levels, where no significant differences were found between the different Δ-levels, $F(3, 27) = 0.496, p = 0.688$ (see Table 4.2). Overall, these results do not strongly indicate that an increased Euclidean distance between visual and haptic textures negatively affect performance.

4.3.3 Model comparison

The effect of increasing Δ-levels could follow a negative linear model (even decrease in precision), a 'flat' linear model (no change in precision), a positive linear

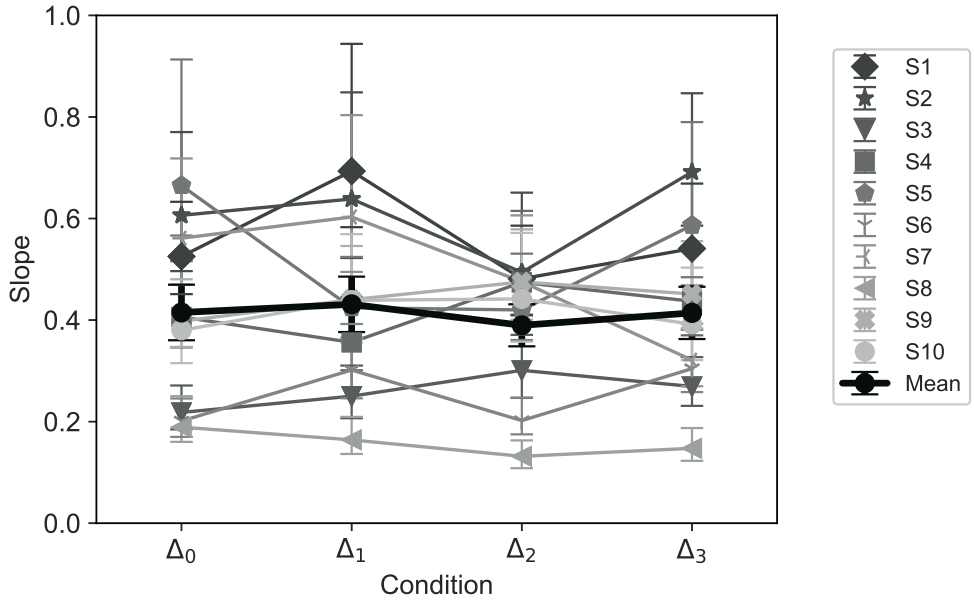


Figure 4.15: Slope performance per Δ between modalities, for all participants. The y-axis indicates the value of the slope while the x-axis indicates the relative distance between the modalities. Individual error bars show the 95% confidence intervals of the slope. The black line shows the mean across all participants, with error bars showing the standard error of the mean. As shown across participants, the overall performance does not significantly differ as Euclidean distance increases. This strongly indicates that the texture features of the individual textures are being weighted more than the similarity between visual and haptic textures for this task. For a detailed look at individual results, see Appendix B, Figure B.5.

Within Subjects Effect

	Sum of Squares	df	Mean Square	F	p
Δ	0.009	3	0.003	0.496	0.688
Residual	0.158	27	0.006		

Table 4.2: Repeated Measures ANOVA shows no main effect of distance Δ on slopes, $p = 0.688$.

model (improved precision), or a quadratic model (initial decrease in precision, followed by an improvement in Δ_3). For an example of these different models, see Figure 4.16. In order to investigate the effect on a per-person basis, we generated the most-fitting polynomial equations per participant based on their individual

performance, and ran a model comparison on a per-individual level using the Palamedes toolbox, with an individual example fitting shown in Figure 4.17, with full individual fits in Appendix B, Figure B.6. Running model comparison analyses on the relative performances of the texture-pairs on a per-participant basis makes it possible to conclude that the inherent difference in similarity between visual and haptic textures does not reliably improve, or reliably impair, the precision of planar slant discrimination.

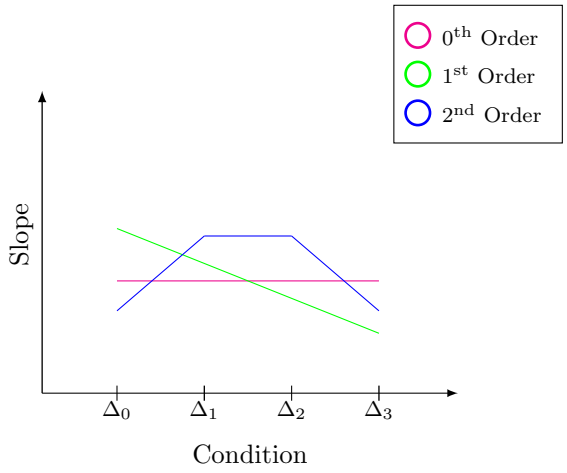


Figure 4.16: Examples of fits to different degrees of polynomials. The magenta line is a 0^{th} degree polynomial, a straight line with a constant $f(x) = a$. The green line is a 1^{st} degree polynomial, $f(x) = -ax + b$. The blue line is a 2^{nd} degree polynomial, $f(x) = -ax^2 + bx + c$. Both the 1^{st} and 2^{nd} degree polynomials could also be fit in the opposite direction, creating an increasing line for the 1^{st} degree fit and a horse shoe shape for the 2^{nd} degree fit.

4.3.4 Post-hoc tests

As we had a strong theory that the difference in textures would degrade performance, we ran some post-hoc tests on other subsequent theories. First, we ran polynomial model fit comparisons – using the model fit function provided in the Palamedes toolbox for Matlab (Prins & Kingdom, 2016) – between 0^{th} and

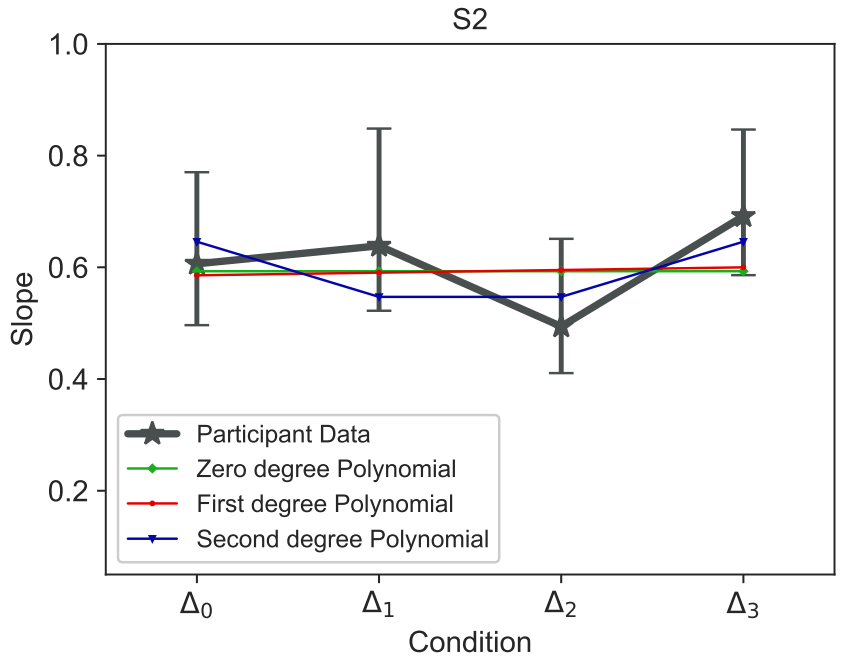


Figure 4.17: Model comparison of Δ-level effect, contrasting polynomial changes to precision, for one representative dataset. For all of the individual observers' polynomial model comparison plots see Appendix B, Figure B.6. In this dataset the 0^{th} degree and 1^{st} degree polynomial fits are overlapping, and the overall data does not fit any of the polynomial models in particular.

1^{st} and 0^{th} and 2^{nd} degree polynomial fits, for all participants, results for which are shown on the left hand side of Table 4.3, Δ_{0-3}. This model comparison ran 1000 bootstrapped simulations per polynomial model, per participant. The output value, $pTLR$ describes the proportion of the bootstrapped likelihood ratios that were smaller for the fuller model (equivalent to 'alternative hypothesis' of ANOVAs) compared to the likelihood ratio of the lesser model (equivalent to the 'null hypothesis'), where the lesser model is usually the observer's collected data – or in the event of two different models such as our polynomial fits, where the lesser model is the 0^{th} degree polynomial and the 1^{st} and 2^{nd} degree polynomials, respectively, are the fuller model – and a $pTLR < 0.05$ is considered statistically

significant. For the test between 0^{th} and 1^{st} polynomial, there was no significant difference between the models, except for S7 whose fits converged fewer than 50% of the time. For the 0^{th} and 2^{nd} degree polynomial, there were only two significant differences (S5, S7). For the remaining 8 participants, there was no significant difference between the models. As the training phase required participants to be able to differentiate textures at Δ_3 on a purely haptic level at 70% precision or higher, we ran the polynomial model fit on the Δ_0 and Δ_3 conditions, the two extremes of our experiment, results shown on the right hand side of Table 4.3, Δ_0&Δ_3. For only the extremes, we found that 8 of 10 participants did not significantly differ from 0^{th} degree, while 2 participants had a pTLR < 0.05 (S6, S7). Lastly, we ran a paired-sample t-test on the slopes for Δ_0 and Δ_3, $p = 0.979$, $t = 0.027$, $df = 9$. This shows no statistical difference between the Δ-levels. As these findings are robust at both an individual level and the group level, we conclude that, for textures of these similarity ratings, having a different signal for visual and haptic stimuli does not degrade precision in a slant-discrimination task.

As we had a strong theory that the difference in textures would degrade performance, we ran some post-hoc tests on other subsequent theories. First, we ran polynomial model fit comparisons – using the model fit function provided in the Palamedes toolbox for Matlab (Prins & Kingdom, 2016) – on zero and first degree polynomials (straight line and sloping line) for all participants, results for which are shown in Table 4.3. This model comparison ran 1000 bootstrapped simulations per polynomial model, per participant. The output value, $pTLR$ describes the proportion of the bootstrapped likelihood ratios that were smaller for the fuller model (equivalent to 'alternative hypothesis' of ANOVAs) compared to the likelihood ratio of the lesser model (equivalent to the 'null hypothesis'), where

the lesser model is usually the observer's collected data – or in the event of two different models such as our polynomial fits, whether the lesser model is the 0th degree polynomial and the 1st and 2nd degree polynomials, respectively, are the fuller model – where the lesser model is the observer's data and the fuller model is the linear regression values – where a $pTLR < 0.05$ is considered statistically significant.

For all but one participant (S7), the 0th degree polynomial fit was considered accurate for too many of the simulated functions for the 1st degree to be considered the significantly better fit (pTLR < 0.05). Further to this, as the training phase required participants to be able to differentiate textures at Δ_3 on a purely haptic level at 70% precision or higher, we ran the polynomial model fit on the Δ_0 and Δ_3 conditions, the two extremes of our experiment. For only the extremes, we found that 8 of 10 participants did not significantly differ from 0th degree, while 2 participants had a pTLR < 0.05 (S6, S7). Lastly, we ran a paired-sample t-test on the slopes for Δ_0 and Δ_3, p = 0.979, t = 0.027, df = 9. This shows no statistical difference between the Δ-levels. As these findings are robust at both an individual level and the group level, we conclude that, for textures of these similarity gradings, having a different signal for visual and haptic stimuli does not degrade precision in a slant-discrimination task.

In the polynomial model fit procedure, some of the sloping regressions were in the opposite direction than was predicted, and while these were not significant, we decided as a result to run a linear regression on the data points, both on an individual level and on the group as a whole, the results shown in Table 4.4 and in Appendix B, Figure B.8. On the standard formula for a linear equation, $f(x) = ax + b$, there was no emerging trend between participants, where

6 participants had a slight trend in the positive direction, while 4 participants had a slight trend in the negative direction. For the mean, the coefficient a was -0.004, which, when considering the scale of the different slopes, is considered to be negligible and 'flat', the results shown in Table 4.4 and in Appendix B, Figure B.8. As shown, four participants showed a negative regression (expected direction, decreasing performance), while six showed a positive regression (unexpected, improving performance). On a group level, the slope of the linear regression was 'flat' ($a = -0.004$, see Appendix B, Figure B.8k).

In the polynomial model fit procedure, some of the sloping regressions were in the opposite direction than was predicted, and while these were not significant, we decided as a result to run a linear regression on the data points, both on an individual level and on the group as a whole, the results shown in Table 4.4 and in Appendix B, Figure B.8. As shown, four participants showed a negative regression (expected direction, decreasing performance), while six showed a positive regression (unexpected, improving performance). On a group level, the slope of the linear regression was 'flat' ($a = -0.004$, see Appendix B, Figure B.8k).

4.4 Discussion

In this experiment we aimed to investigate whether an increase in dissimilarity between a visual and haptic texture degrades performance in a slant-discrimination task using realistic textures for both vision and touch – emulating a potential use-case for viewing medical images from different imaging modalities simultaneously. The dissimilarity was quantified using higher-order linear algebra applied to an existing similarity matrix consisting of 8-dimensional texture features. From our

	pTLR		
	Δ_{0-3}		Δ_0&Δ_3
	1st	2nd	1st
S1	0.633	0.710	0.848
S2	0.895	0.163	0.448
S3	0.120	0.282	0.188
S4	0.310	0.673	0.630
S5	0.710	0.006*	0.580
S6	0.078	0.930	0.010*
S7	0.003*	0.010*	0.003*
S8	0.093	0.333	0.155
S9	0.330	0.486	0.412
S10	0.858	0.290	0.873

Table 4.3: Model comparison to different polynomials, where a $pTLR < 0.05$ is considered statistically significant (*) and the alternative $pTLR \geq 0.05$ is better fit to the 0^{th} degree polynomial. For Δ_{0-3} the comparison is on all four data points per observer, and compares between the 0^{th} and 1^{st} degree polynomial function or between the 0^{th} and 2^{nd} degree polynomial functions, respectively. For Δ_0 & Δ_3, as it is only the two edge points the model comparison is only between the 0^{th} degree and the 1^{st} degree polynomials.

	f(x) = ax + b	
	a	b
S1	-0.017	0.585
S2	0.011	0.591
S3	0.021	0.229
S4	0.021	0.387
S5	-0.024	0.560
S6	0.021	0.221
S7	-0.085	0.617
S8	-0.016	0.182
S9	0.020	0.412
S10	0.004	0.407
Mean	-0.004	0.419

Table 4.4: Linear regression analysis of increasing distance Δ, where the coefficient a and the constant b make up a linear equation of the form $f(x) = ax + b$ and describe the line best fit to the four Δ-points. The majority of the coefficients a non-significantly different from a flat regression of $a = 0.0$, with the exception of S7 who has a decreasing linear function with $a = -0.085$.

hypothesis that an increase in dissimilarity would degrade performance, we predicted that an increase in Euclidean distance Δ would cause a decrease in precision proportional to the increase in dissimilarity. However, our results unexpectedly showed no consistent change of slope values compared to Δ-levels (Figure 4.15, $p = 0.688$), neither as a decrease (lower precision, worse performance) nor as an increase (higher precision, improved performance). This was the case on both the individual and group level, for the polynomial model comparisons as well as

the linear regression models. While initially unexpected, this may be a positive result for the purpose of the experiment – if there is no strong negative effect on the slant discrimination performance, it is fully possible that using different types of medical scans as the source for the different respective modalities can be beneficially integrated by the user for multisensory exploration tasks. However, a 'null' result is not necessarily indicative that there is no effect, just that no effect was found.

4.4.1 Null hypothesis

The null hypothesis is impossible to prove directly (Harms & Lakens, 2018), as there is a clear distinction between proof of absence and absence of proof – especially for studies with few participants or smaller amounts of data. Common methods for bolstering the confidence and statistical power of a result, regardless of whether one expects to support the null hypothesis or not, include having large sample sizes – this is usually done with a combination of high number of participants, as well as large amounts of data collected per observer. Other methods include having control conditions, where one expects to find no effect of alterations as there are no alterations present, or catch-trials, where perturbations and other adjustments are suspended without the participant being aware prior to the trial. Whereas control conditions give us a direct control to compare the results to, catch-trials are often used to investigate adaptation, such as suddenly removing air resistance in a reaching task, causing the participant to wildly overshoot their reach.

Whereas catch-trials specifically are not applicable to the design of this experiment, in retrospect we would have greatly benefited from including the single-

cue data for each of the five single-cue conditions per observer. Had the pre-experimental JND thresholds for the four visual textures and the haptic texture been collected, it would have been possible to confirm whether or not integration did occur at Δ_0, where there was no intended discrepancy, as well as be able to estimate the expected effect of the increased textural discrepancy between the two sensory signals. Without the confirmation that integration did occur when the signals were using the same core texture, there is no baseline to show whether performance was at a level where it would be possible to get worse and whether the task as presented to the observers was too difficult. Since there was no metric to show that performance was not at 'floor' level and there was room for the performance to get worse, it is very difficult to estimate whether an increase in dissimilarity between the texture pairs had a detrimental effect, unless it had happened to be a very strong effect.

Since the task required the surface slant to be congruent between the two sensory modalities, any adjustments to match JNDs would have had to be through adjusting the width of the aperture on a per-person, per-texture basis. This would have been done by collecting the single-cue JND thresholds per texture for a range of different aperture widths, in order to find the aperture width that would match the perceptual JND differences per individual, per texture. With the current experimental design, we do not know the extent to which the aperture at the predesignated width affected the relative reliabilities of the visual textures, nor whether the reliabilities of the textures themselves were comparable to one another – let alone whether the reliabilities were comparable to the haptic-only texture. While effort and care was put into researching and selecting the 'reasonable' width of the aperture, which was based on the work of Burge et al. (2010),

it would have been more scientifically sound to have gone through and manually tested and selected aperture widths on a per-observer, per-texture basis. With the experimental design as run, we can conclude that the increased dissimilarity did not reliably reduce performance for the given aperture width, but as there was no confirmation about cue integration occurring in the first place, no concrete conclusions can be made about whether the cue integration broke down with increased discrepancy.

For more 'soft' metrics of effect of Δs, we can look at the individual performances compared to what the respective textures looked like (examples shown in Figure 4.18 and Figure 4.19), and compare some of the statistical features such as the SSIM values. This was rudamentarily done for all pairs in the Δ-levels, as well as between the visual texture sets and the haptic height maps they were based on. There was no emerging trend or pattern in any of these comparisons. However, a previous study by Rosas et al. (2004) shows that the type of texture shown on the slanted surface will also have an impact on performance in slant-from-texture estimates.

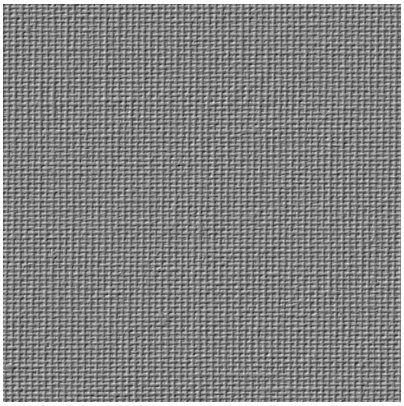
Looking at Figure 4.18, one could have assumed that texture 144 and 114 would have both had a high slope value due to the geometric linearity, or at least performed similarly, which Figure 4.18f shows that texture 114 had significantly worse performance. Whereas looking at Figure 4.19, one might have assumed similar performance between texture 240 and 320, with a higher performance for texture 47, again based on the striations in the surface geometry. As shown in Figure 4.19f, this is not the case, and texture 240 and 47 perform significantly worse than texture 320. For the rest of the individual comparison plots in the same form as Figure 4.18, please see Section §B.2.5, Appendix 1.

Another potential confound to the visual cues is the relative illumination of the rendered objects, where the angle of illumination has a strong impact on the perceived roughness of a texture (Ho et al., 2006). While all the textures in the PerTex database are rendered using shape-from-shading, under Lambertian conditions, and being matched in terms of illumination direction (45°), scaled surface normals and albedo values (Halley, 2012), the illumination of the aperture and rendered tool-point are not. The illumination of the rendered scene itself was set to be directly forward at 90°, which could have caused a subtle visual incongruence which may well have reduced an observer's ability to integrate the cues. To combat this and to introduce genuine stereoscopic visual texture cues as well as conclusive haptic depth mapping, one could have rendered these images as 3D surface meshes of the underlying images, which would have made it simple to match illumination parameters between the signal, the aperture, and the haptic tool point.

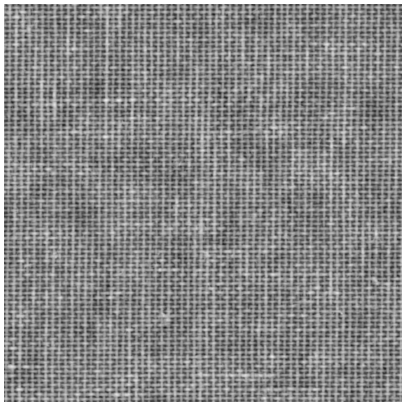
Additionally, this would have ensured that the haptic height mapping would have been truly rendered in quantifiable distance metrics, whereas the experiment itself used the standard Chai3D method of rendering the haptic signal as a luminance-based normal-map of the pixels in an 8-bit PNG, which has the potential of coming across as a roughness signal rather than a depth signal. It is also worth noting that both the training task as well as the main experimental task had the visual pointer vanish when the observer went into contact with the haptic texture. This could have introduced a lack of visual feedback to benefit from tool-use, especially in untrained observers with little to no prior experience using haptic devices.

Lastly, the Euclidean distances between the textures is calculated from the

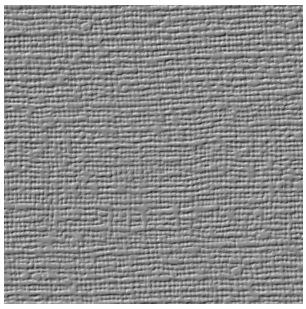
base set Δ_0. However, the textures are defined in 8-dimensional perceptual space and as such, the distance between the visual textures in Δ_1, Δ_2, and Δ_3 are likely to be [and mathematically checked afterwards] different distances between each other than between themselves and Δ_0. Whereas the Euclidean distance between Δ_0 and Δ_2 is twice that of Δ_0 and Δ_1, the distance between Δ_1 and Δ_2 can be anywhere from much shorter to much longer. The 8 dimensions of the provided perceptual matrix are not currently linked to any specific texture features, and so it is possible that any of those 8 dimensions are more task-relevant than others, some of the dimensions may not be discriminable for the haptic system, and some may even be differently related in the haptic system compared to the visual system.



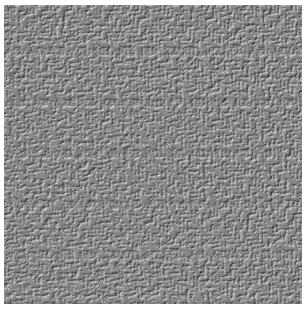
(a) Δ_0, Rendered texture 144



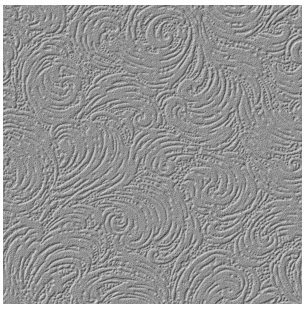
(b) Δ_0, Height-mapped texture 144



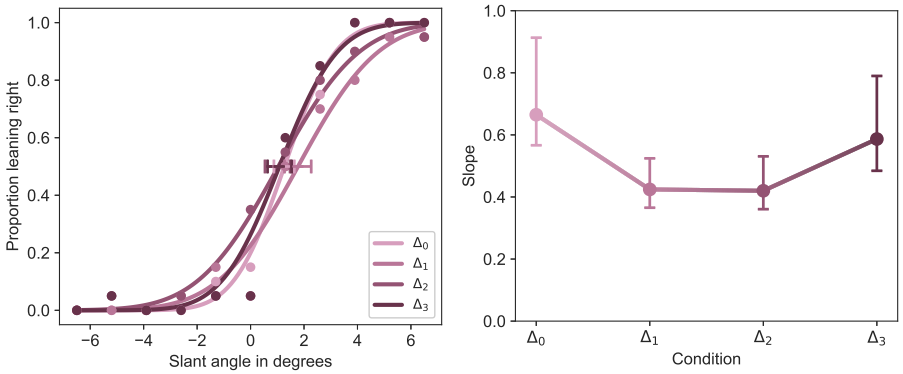
(c) Δ_1, Rendered texture 114



(d) Δ_2, Rendered texture 113

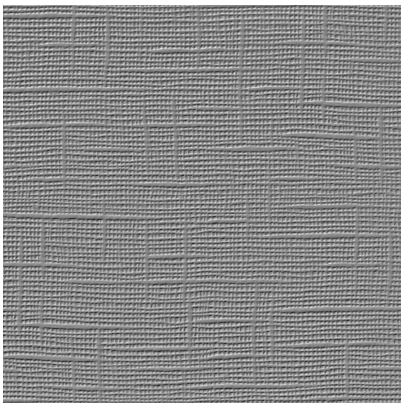


(e) Δ_3, Rendered texture 212

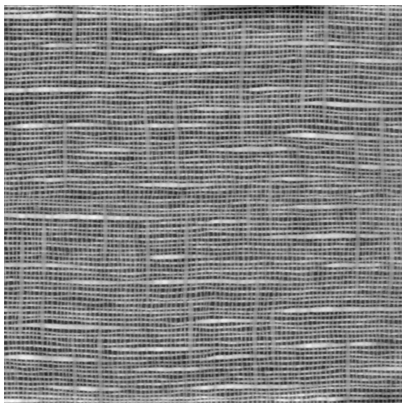


(f) Performance S5

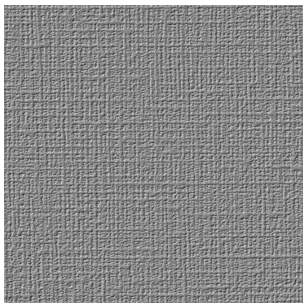
Figure 4.18: For observer S5, texture pairs Δ_0 and Δ_1 showed high geometric regularity, while they performed better in Δ_0 and Δ_3, where Δ_3 had a notable circular geometry. They had a slight bias to the right.



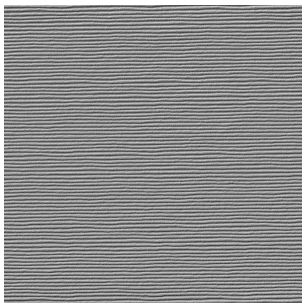
(a) Δ_0, Rendered texture 240



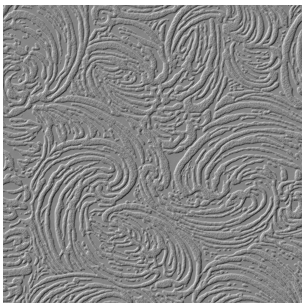
(b) Δ_0, Height-mapped texture 240



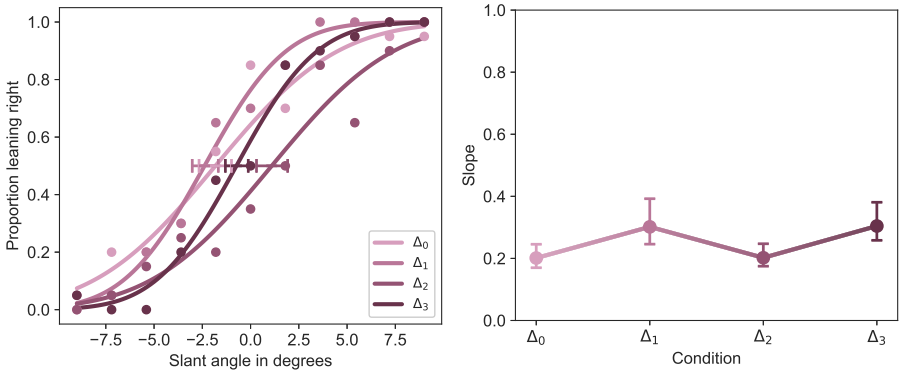
(c) Δ_1, Rendered texture 320



(d) Δ_2, Rendered texture 47



(e) Δ_3, Rendered texture 26



(f) Performance S6

Figure 4.19: For observer S6, texture pairs Δ_0 and Δ_1 had general geometrical regularity in both the horizontal and vertical regions, while Δ_2 had a strong horizontal features. Given the task, one would have expected an improved performance in texture pairs Δ_0 and Δ_2, while the results show a slight improvement in Δ_1 and Δ_3. They had a slight bias to the left.

Haptic perception

The intended magnitude of the discrepancy between the visual and haptic signals was based on the assumption that the sensory dimensions related to the textures were comparable in Euclidean space, which upon further review is a naïve assumption to make; the magnitude of textures can be discrepant between vision and touch even under normal day-to-day circumstances (Marks, 2014). One important aspect is that, while the difference in higher order and lower order statistical features is one of the dominant features in visual perceptual similarity, haptic perceptual similarity is much less sensitive to the higher order statistical features – as such the perceptual similarity rating might not be consistent between visual and haptic perception similarity, as suggested in Kuroki et al. (2019).

Assuming the metrics as defined in one perceptual space to be the same as the metrics in another is a flawed assumption to make. Commonly, this is avoided by finding the just-noticable-difference (JND) threshold in the sensory modality per observer, as done for example by Ernst and Banks (2002). By carefully characterising the perceptual units per sensory cue, per observer, they avoid the assumption that the size would be the same in the visual and haptic modality, which is a potentially unknown relationship between the metrics used by vision and touch, when not utilising a paradigm such as matching JNDs.

As haptic perception of texture features seems to be less sensitive to higher order statistical (HOS) image properties compared to lower order statistical (LOS) properties, visual perception of texture features performs equally well on both sets of statistical properties. When using these statistical properties as features in vector-space, it is possible that the haptic perception is not affected by specific features in the hypothetical N^{th} domain and therefore the distances between the

textures in haptic space would be $n - 1$ dimensions, which one could logically assume to represent a larger or smaller distance than the visual space, as explained more thoroughly in the Methods section (§4.2.2), and as such the similarity – or dissimilarity – between two textures in the visual modality might not be the same as in the haptic modality. However, the 8-dimensional feature space used to characterise the PerTex database was based on human perceptual ranking, and while it is possible they correspond in some way to more conventional measures of statistical features there are no inherent links between the two at this point.

To investigate this further, one could eradicate the HOS and only select similarity based on the LOS features of images to increase the likelihood of the sensory cues overlapping homologously. A different approach could be to run the similarity experiments done by Clarke et al. (2011) on haptic textures in conjunction with visual ones, and attempt to find an alignment between the two similarity spaces. Both of these options are considered to be beyond the scope of this experiment.

In short, there are a number of potential reasons as to why the performance of the task did not reliably differ with increased dissimilarity, several of which would have been simple to control for with more prudent experimental considerations. However, as it stands, for these texture sets, for these levels of perceptual discriminability, where observers can reliably differentiate between textures of Δ_3 Euclidean distance using haptic-only signals, an increase in dissimilarity up to the level of Δ_3 between visual and haptic signals does not reliably deteriorate performance in a slant-discrimination task.

4.4.2 Relevance for Medical imaging

With the likely difference in statistical features used for texture perception in visual perception compared to haptic perception, it is fully possible that the standard cue combination benefit would extend to using different medical images: having one image modality for vision and another for haptics could improve overall precision when viewing medical data, especially if the image sets were synchronously obtained as discussed in the introduction regarding the paper by Martí-Bonmatí et al. (2010). As vision is able to detect more changes in higher-frequency statistical features than haptics (Kuroki et al., 2019), the appropriate modalities would have to be selected depending on the required region. For example, in a brain scan an MRI has higher contrasts overall compared to a CT scan, so one approach would be to explore the CT scan visually with the MRI superimposed haptically. However, when looking at a bonier region such as the pelvis (or when specifically investigating bone-related features), the CT scan offers a higher contrast image with more lower-statistical frequencies compared to an MRI, which excels in the differentiation of tissue density for the various internal organs. This opens up the potential of using different imaging modalities simultaneously without concern for the negative impact of using different sources if the similarity between the features of the presented images is within tolerance.

4.5 Summary

In this chapter we introduced some of the different imaging modalities used by the medical industry, the background of causal inference and the importance of having a perceived common source. We further explained the importance of cue

recruitment and training, how textures are categorised and different ways of estimating textural similarities. The overall question we set out to answer is whether increasing dissimilarity between a visual and a haptic texture would have a negative effect on precision in a 2AFC slant-discrimination task. By using naturalistic textures and selecting three discrete levels of dissimilarity, we were able to poll four levels of visuohaptic incongruence per participant, and contrast these across the group. For the training phase, we ensured that all participants could discriminate haptically between textures of a Δ_3 distance at an accuracy rate of 70% or higher. The precision of slant discrimination was compared using a model comparison between the individual observers' data and 0^{th}, 1^{st} and 2^{nd} degree polynomials with the hypothesis that precision would decrease as the Euclidean distance as described by Δ-levels increased, following a decreasing 1^{st} degree polynomial shape. Contrary to our predictions, the results showed no consistent effect on overall performance. There was no significant overall decrease in precision, nor was there any consistent increase. There were some individual differences where 2 observers were better fit to a 2^{nd} degree polynomial shape, while the remaining 8 observers were not significantly better fit to either 1^{st} or 2^{nd} degree compared to the straight line described by the 0^{th} degree polynomial. A post-hoc regression analysis additionally showed there was no emerging trend in increase or decrease in precision across individual observers, with the linear coefficient of the mean being comparable to that of a 0^{th} degree polynomial – a 'flat' line ($a = -0.004$). Overall, the results clearly show no consistent effect, either positive or negative. In other words, the increase in Euclidean distance between visual and haptic texture does not significantly reduce slant discrimination performance for real-world textures in a combined visuohaptic 2AFC discrimination task. This is a positive

overall result for the purpose of this thesis, as it indicates that slight differences in source image for visual and haptic stimuli do not inherently reduce performance, opening up the possibility of using data from different types of medical scans for different sensory cues used in a multisensory exploration task.

Chapter 5

Experiment 3

5.1 Introduction

As has already been highlighted in Chapter 1, one of the most crucial aspects of cancer treatment is the correct delineation of tumourous tissue. This is however also the weakest link of the diagnostics chain, with each clinician delineating tumours using internal criteria that are influenced by where they trained, what medical field they specialised in and their own perceptual biases. These unique, individual criteria are considered to be one of the main contributors to high inter-clinician variability in a task with a low margin for error (Njeh, 2008; Nowee et al., 2019). Trained radiologists will visually inspect the data, sometimes even looking at several different imaging modalities, and make decisions based on locating potentially cancerous tissue embedded in healthy human tissues (Castella et al., 2009; Cooper et al., 2007; Kompaniez-Dunigan et al., 2015). As previously mentioned in Chapter 1, the diagnostic task of locating abnormal tissue can be simplified to a signal detection task, where the radiologist locates the ab-

normal tissue (the 'signal') hidden within the healthy tissue (the 'noise'). This is a very difficult task, with a high level of inter-clinician variability in the final delineations. This variability issue is a problem related to both overdelineation, defined as the inclusion of healthy tissue, and underdelineation, defined as the unintentional exclusion of tumour, as shown in Table 5.1.

		Tumour	
		Present	Absent
Response	Yes	*Correct inclusion*	*Overdelineation*
	No	*Underdelineation*	*Correct exclusion*

Table 5.1: The four potential outcomes of a tumour delineation task. 1) 'Correct inclusion', outlining 'tumourous' tissue as 'tumour', 2) 'Overdelineation', outlining 'non-tumourous' tissue as 'tumour', 3) 'Underdelineation', not outlining 'tumorous' tissue as 'tumour', 4) 'Correct exclusion', not outlining 'non-tumourous' tissue as 'tumour'.

In the case of overdelineation, the effect of excessive tissue inclusion is at best lower recovery time and an increase in required dosage of treatment, and at worst it can cause the unnecessary removal of tissue that's imperative to leading a fully functional human life, such as brain tissue. For underdelineation, the effect of inadequate tumour inclusion is at best extending the required treatment time by underestimating the amount of treatment needed, and at worst misses out on critical levels of cancerous tissue, greatly increasing the chance of recurrence. These two aspects are two of the key considerations that clinical radiologists must weigh and balance when diagnosing and treating patients in real-world clinical situations.

The current methods of comparing performance predominantly use real images

of tumours, and have a large group of clinicians delineating individually, and these delineations are compared with a 'master outline' which is drawn by an expert radiologist. This is however not a perfect method, where issues include a lack of certainty, as it is currently not possible to find the exact ground truth (GT) shape of the tumour itself without surgically removing it directly after imaging, which is unlikely to be feasible in the vast majority of cases (Ertl-Wagner et al., 2009). There is also the issue of confounding, where if the images have a high level of realism and complexity, it is harder to manipulate the stimulus in a controlled manner and to make inferences of what features and aspects of the stimuli – such as signal strength and size – are more informative, and how manipulating these affect the perception of the stimulus as a whole. Lastly, magnitude of scale – it would be difficult to source a sufficiently large amount of real medical images that are perceptually similar and have a comparable information density.

As we do not know the ground truth in real cancer images, this complicates the synthesis and analysis processes available – it would be difficult to embed a synthetic tumour in a medical image in a realistic way. If a simulated tumour were to be embedded in a real medical image, that would require in-depth analyses of tumour shapes, sizes and placement relevant to the specific cancer sites of the body in order to keep everything at a comparable level of realism – for example, pancreatic cancer tumours and bone cancer growths are incomparable in shape, size and frequency. What we can do instead is to generate our own images and tumours, and embedding these using a randomised positional algorithm. The benefits of using simulated images are that we know the ground truth, we don't have to be concerned about the tumours overlapping with organ boundaries and there is no inherent biases of the anatomical location of the cancer, allowing us to

generalise the overall results. Additionally, as we can control the different statistical features of the images we can easily generate large sets of similar-looking but unique images per individual observer, circumventing the need to source medical images and obtaining permission and anonymising. We can also generate these to the required resolution and aspect ratio as the experiment requires. The drawbacks however being that as these are analytically generated, any findings are not guaranteed to carry over to a real-world physical application without further work, which will be discussed in greater detail in Chapter 6.

The addition of haptics has, as discussed in Chapter 1 and demonstrated in Chapter 3, been shown to improve observers' perceptual judgement. However, for a tumour delineation task precision alone is not the most suitable metric; accuracy is. As the high-resolution modern haptic device is a relatively novel technology for the task of tumour delineation, and the proposed method of height-mapping a haptic signal using a normal-map of the relative luminance of the pixels of the source image is novel as well, we first need to see if the method has an effect on performance, whether this effect is positive or negative, and if this effect is significant or not. Additionally, we want to explore whether the method affects variability and what effect it might have on volumes, such as amount of tumour excluded and amount of healthy tissue included.

5.1.1 Hypotheses

In the previous two experiments, we first established that the addition of a haptic signal can improve an observer's detection of a simple signal a 2AFC detection task[1], and secondly that using a pair of stimuli that are perceptually similar but

[1]When haptic-only training occurs

non-identical for separate visual and haptic modalities is not inherently detrimental to an observer's precision of slant discrimination of a planar stimulus in a 2AFC discrimination task. For the final experiment we aim to investigate whether the addition of a haptic signal can aid observers in the location and delineation of tumours in medical images. In this experiment we aim to identify whether the inclusion of a haptic signal in a standard tumour delineation task has an effect on accuracy, both as a distance-to-ground truth (GT) measure and as volumetric measures such as the volume of tumour included and volume of healthy tissue included. Based on previous research we expect to find that the addition of a haptic signal height mapped to the luminance of the pixels would have a positive effect by way of improving accuracy. This would be expected to present itself as a reduction in distance-to-GT, an increase of included 'tumour', a decrease of included 'healthy tissue' or a combination of several of the three. If there is a reduction in distance-to-GT, this would be due to either of the latter two, or a combination thereof. If distance-to-GT stays the same or decreases, that implies that the volume of included 'tumour' must be significantly improved for it to be a positive result – if one were simply to overdelineate everything this would also increase the distance-to-GT and would not be considered to be an overall positive effect. In essence, we theorise that observers will be more accurate at delineating a 'tumour' when drawing on and exploring the texture with additional haptic height-mapped feedback compared to drawing on a flat planar surface.

5.2 Methods

5.2.1 Participants and setup

For this final experiment, a total of 10 observers participated, 8 of which were naïve to the purposes of the experiment. All observers were right hand dominant, had normal or corrected to normal vision and all of the participants had a stereoacuity of 60 arcsec or better, as measured with the Laméris Ootech TNO Stereo Vision test. The physical setup used is the spatially coaligned visuohaptic rig as described in detail in Chapter 2 §2.1.3.

The participants were asked to outline a tumour to the best of their ability, first starting with the training phase, containing 10 trials of no haptic height-mapping (NH) and 10 trials of with haptic height-mapping (WH), where feedback was provided to the participant after each trial was submitted. After the completion of the training, they were performing the task 20 times with NH, and 20 times WH. Each of these outlines were compared to the ground truth (GT), from which we extracted two main measures, distance-to-GT and volume of tumour or healthy tissue delineated. Distance-to-GT is the main measure used today in the medical field, and consists of finding the X-Y coordinates of the individual points that make up the traced outline and performing a point-by-point comparison of the distance from a given point on the traced outline to the nearest point of the outline of the GT, which does not discriminate between negative and positive distances. We also have measures of volume – how much volume was delineated, how much of this was 'tumour', how much of the 'tumour' was missed, and how much 'healthy tissue' was accidentally included.

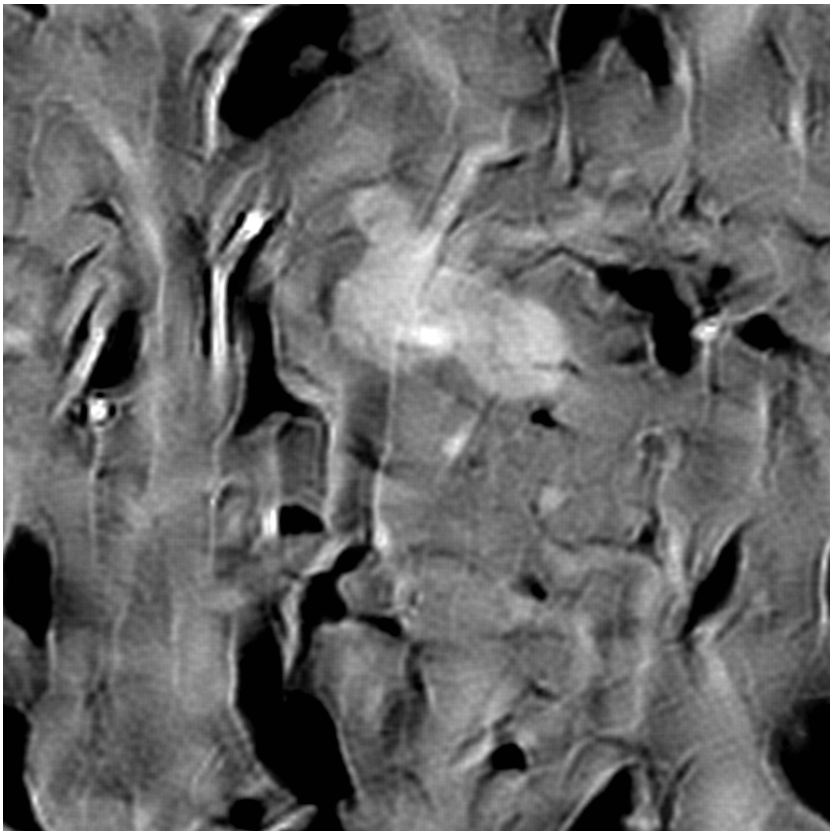
5.2.2 Stimuli

The stimulus used in this experiment is a simulated 'tumour', created by 7 randomly positioned, partially overlapping circles of various sizes. These 'tumours' are then blurred at the edges and embedded in simulated medical images. The simulated image is a texture synthesised from frequencies found in real-world medical images. The simulated tumour is randomly rotated and placed within the medical image, restricted to not overlap too much with pure black ('empty') areas to avoid creating too simple a task. An example of the full stimulus generation steps can be seen in Figure 5.1, while the steps for generating the 'tumour' cluster itself steps is shown in Figure 5.2. A full set comprised 40 of these images per participant, where the first 20 were used for training and the last 20 for testing.

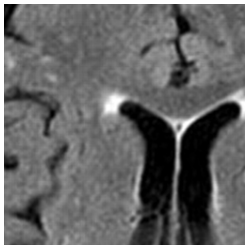
Medical image synthesis

The first part of the presented stimuli is the synthetically generated medical image background. As previously mentioned, the background needs to be variable enough to realistically match real medical images, but without obvious organ boundaries which complicate tumour placement. Additionally we need 20 unique-but-consistent backgrounds per individual observer, 200 total. In order to do this, we chose to generate the backgrounds using the textureSynth Matlab package by Portilla and Simoncelli (2000), which works by inputting a source image and defining how many times to run the synthesis loop, how many orientations and scales to sample, and the size of the surrounding spatial neighbourhood to consider. For our 'medical image' backgrounds, we selected a source image was a subsection of a brain scan from an original image sourced from Debette and Markus (2010),

Figure 1 – the subsection itself shown in our Figure 5.1b. This subsection was chosen as it had a lack of the obvious organ boundaries, had no sharp edges and did not result in obvious banding in the synthesised images. For the remaining parameters the final values were: image size of 768 pixels, synthesised over 25 iterations, 6 scales, 6 orientations and a spatial neighbourhood of 13, resulting in images such as Figure 5.1d.

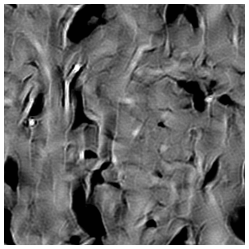


(a) Embedded Tumour Presented Stimuli



(b) Source image

(c) Simulated Tumour



(d) Simulated Medical Image

Figure 5.1: The process of simulating medical image and embedding the tumour. (a) shows the completed trial stimulus, complete with an embedded tumour in the simulated background. (b) is the subsection of the brain image used as a base for simulating the stimulus backgrounds. (c) is the tumour created in MATLAB by overlapping a number of randomly-sized circles at random angles and distances to centre. (d) is the simulated texture as output by the textureSynth software.

Tumour synthesis

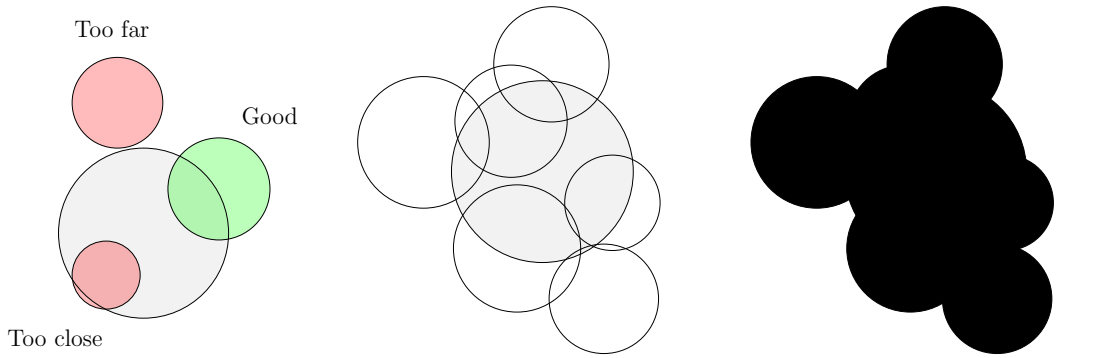
For the 'tumour' cluster itself, the initial generating procedure is to first generate a 'base' circle ($i = 1$), and 6 surrounding circles ($i = 2..7$). For each of these circles i, we generate a radius r_i and centre coordinates x_i and y_i, each chosen randomly from a range of permitted real numbers:

$$r_i \in \begin{cases} [0.2, 0.3] & \text{if } i = 1 \\ [0.1, 0.2] & \text{if } i > 1 \end{cases} \tag{5.1}$$

$$x_i, y_i \in \begin{cases} [0, 0.25] & \text{if } i = 1 \\ [-(r_i + r_1), (r_i + r_1)] & \text{if } i > 1 \end{cases} \tag{5.2}$$

Where the dimensions of the image are 1.8 by 1.8 and the centre of the image is at 0.0. However, the placement of a circle i is recalculated if its generated centre coordinates (x_i, y_i for $i > 1$) would result in the circle not overlapping with the 'base' circle ($\sqrt{(x_i - x_1)^2 + (y_i - y_1)^2} \geq (r_i + r_1)$), or if the distance between x_i and y_i, and a given circle's x_j and y_j, respectively, is less than 0.35 ($\sqrt{(x_i^2 + x_j^2)} < 0.35$ and $\sqrt{(y_i^2 + y_j^2)} < 0.35$ for $1 \leq j \leq i$). The theoretical minimum width or height after scaling, which is part of the placement process and described in the next section, is 100 pixels if all subsequent circles lie within the minimum and maximum axis-values of the smallest possible 'base' circle – which is to say, in a straight line perpendicular to the axis – while the theoretical maximum width or height occurs with the largest 'base' circle and two of the largest subsequent circles, one on each side of the 'base' circle, at 348 pixels. For an example of the metrics, see Figure 5.2a. Figure 5.2b shows the positions of the individual overlapping circles that form the basis of the shape, while the final

used shape is shown in Figure 5.2c.



(a) Overlap rules (b) Outline of circles (c) Combined tumour

Figure 5.2: Overlapping rules and example process for the creation of the synthetic 'tumour' cluster. (a) shows example overlapping circles, where the two red circles ('Too far' and 'Too close') would require repositioning, while the green circle (Good) would be accepted. (b) shows an example outline of 7 accepted circles of varying radii and positions, while (c) show the filled in version of the 'tumour' cluster, as embedded in the simulated medical images.

Tumour placement

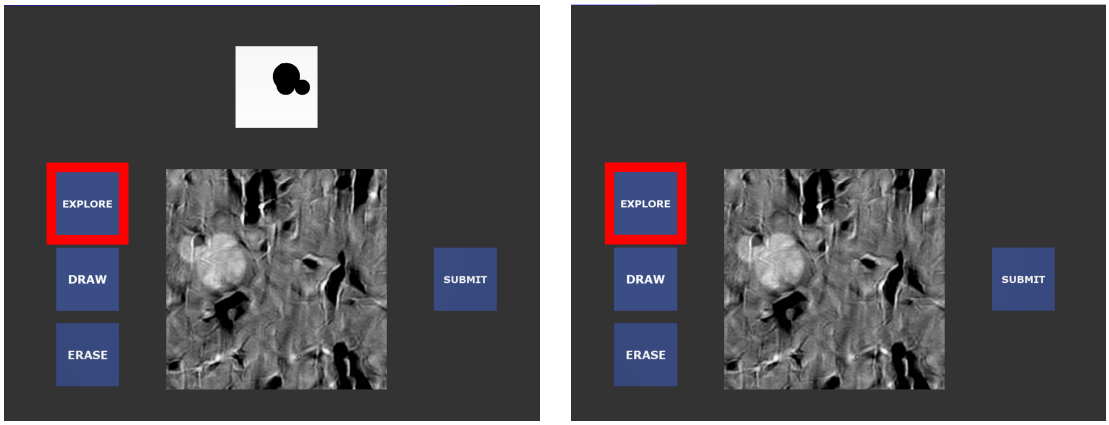
For the embedding of the 'tumour', the cluster image (see Figure 5.2c) was first scaled to a 448 pixel square, which is subsequently rotated randomly in the range $[0°, 360°]$ before being put through a Gaussian blur to soften the edges, using Matlab's 'imgaussfilt' function with a $\sigma = 15$ pixels. Once this preprocessing has been completed, the cluster is positioned randomly within the medical image with the constraints that 1) it cannot be touching the edge, 2) it cannot overlap too much with a pure-black ('empty') area. For the first constraint, the minimum and maximum coordinates of both the x and y position of the rotated 'tumour' is used to calculate the acceptable image bounds, adding an additional randomised displacement value between $[0, imSize - maxVal]$, where imSize is the size of the image in pixels (768 px) and maxVal is the last occurrence of a 'tumour'-present

pixel in the respective x and y directions. This displacement value is then added to the existing bounds $[minVal, maxVal]$, where minVal is the first occurrence of a 'tumour'-present pixel in the respective x and y directions – leaving the final position bounds as a random value between $[minVal, imSize]$, which is calculated individually for both the x and the y variables. For the second constraint, a temporary embedding of the 'tumour' is performed, and the luminance value of the specific area defined by the bounds in condition 1 are compared. As the strength of the 'tumour' in pixel luminance is 72 and black is 0, any luminance in the specified area equal to 72 is tallied up. If the overlap area is equal to or greater than 1% of the total specified area, the fit is rejected and the 'tumour' shifted randomly on the x and y axes respectively by a value in the range of $[-15, 15]$ pixels, if the new maximum value does not violate the boundaries set in condition 1. If, after 10 tries, the position still violates condition 2, the 'tumour' is rotated again and placed anew starting with condition 1, for a maximum of 10 retries where at retry number 5 the source image is rotated 180 degrees.

5.2.3 Training

All participants first underwent a training phase, which both familiarised them with the task as well as what shape of tumours to expect. The training task consisted of a match-to-sample delineation task, where the participant was presented with an image of what shape the tumour is, though no information is given on what orientation the tumour is or what placement it has in the simulated medical image, as shown in Figure 5.3a. This is the same overall task as the main part of the experiment (Figure 5.3b), with the addition of the 'source' image of the tumour to aid in learning the different types of shape. Each participant underwent

20 training trials; 10 trials with the additional haptic height mapping and 10 trials without the height mapping – in which case they were drawing on a simulated flat surface, similar to drawing on a graphics tablet.



(a) Training (b) Main experiment

Figure 5.3: (a) Screenshot of training phase. The participant is instructed to find a shape matching that of the sample tumour shown above the main image. The embedded tumour is 448/1294 of the original size (approximately 7/20), and can be of a different rotation than the source image – as shown in (a), where the tumour in the stimulus is approximately 180 degrees from the source image shown above it. As the participant cycles through the pen modes, the red selection frames moves to the appropriate square. Once ready, the participant selects 'Submit' twice to move on to the next image. The blue panels are approximately (presented as exactly) 5.0 cm across (but 5 cm deeper), while the source image is 6.5 cm across and the stimulus is 17.5 cm across (but 5 cm deeper). The basic tumour is not rotated, simply generated. (b) Screenshot of main task. The overall function is the same as in the training task, without the addition of the source tumour as an aid. As the participant cycles through the pen modes, the red selection frames moves to the appropriate square. Once ready, the participant selects 'Submit' twice to move on to the next image.

5.2.4 Task

The task consisted of testing two separate conditions, the first of which was using the Touch X haptic device without the additional haptic height-mapped feedback,

akin to drawing on a tablet. The second condition used the haptic device and included the haptic height-rendering of the luminance of the pixels, giving a haptic 3D relief of each image. These two conditions were interspaced in blocks, where each block contains 10 trials. On each trial, the participant explored the image and attempted to draw an outline of the embedded mass. The order of the blocks were randomly selected and counterbalanced per participant, so that half the images were presented first in one condition and the other half of the images were presented first in the second condition, and participants were informed prior to each block which condition would be presented.

For the two training sets, there was no repetition of tumour or background, and none of the backgrounds or tumours used in training were later used the experiment. In the main experiment, there were 20 embedded tumour images. The first 10 were presented in one condition (for example, 'no haptic height mapping' (NH1)), the order of which were randomly selected and counterbalanced per participant. These 10 images were then delineated again in the second condition (for example, 'with haptic height mapping' (WH1)). The last 10 images were then first presented in the second condition (e.g, WH2) , and repeated in the first condition again (e.g, NH2). As any differences in 'tumour' shape and placement, as well as background texture, can have a large impact on the final delineation, participants were asked to delineate the same 'tumours' once per condition. This was done so that we could perform statistical analyses specifically on the effect that the WH condition has on delineation compared to the standard NH condition.

5.3 Results

In this experiment we aim to investigate the effect of added haptic height-mapping using luminance based normal-maps of the greyscale simulated medical image. An example of the drawn outline for each of the two conditions is shown in Figure 5.4 which shows the NH (no-haptic) and WH (with-haptic) outlines superimposed on the same source 'tumour' image. In each trial the observer only delineated one tumour, these have been combined in the figure for illustrative purposes. Before the boundary analyses could be performed for either of the conditions, the drawn outlines first had to be 'cleaned up' in Matlab by removing any line markings not attached to the main outline body, the outlines were then smoothed by Matlab's 'boundary' function algorithm which fits an outline to any 2D or 3D grouping of points – fitting to the exterior of the delineation, disregarding line thickness. The border of the 'tumours' had been blurred using Gaussian filtering, so the luminance value chosen to find the boundary of the 'tumour' was any pixel luminance greater than 5/256. A matching outline was drawn from the GT in the same manner, all of these are illustrated in Figure 5.5. This fitting procedure was performed on both NH and WH conditions so as to avoid differences over conditions.

These outlines were then compared using Matlab's 'dsearchn' function which computes a convex hull on the data and finds the nearest points in Euclidean distance using Delaunay triangulation, our comparison is shown in Figure 5.6 which compares the fitting of the outlined 'tumour' to the ground truth outline. The relative precision is defined to be the difference in outlined volume of tumour versus healthy tissue, where the volume was calculated using Matlab's 'area' function on

the areas described by the respective outlines. Under-selecting and over-selecting both count as poor performance. Figure 5.6a and 5.6b show the traced outline of one participant compared to the outline of the GT while Figure 5.6c shows the distances of the respective conditions with overlaid histograms of distances. The results were then compared per participant between conditions, and across all participants per condition.

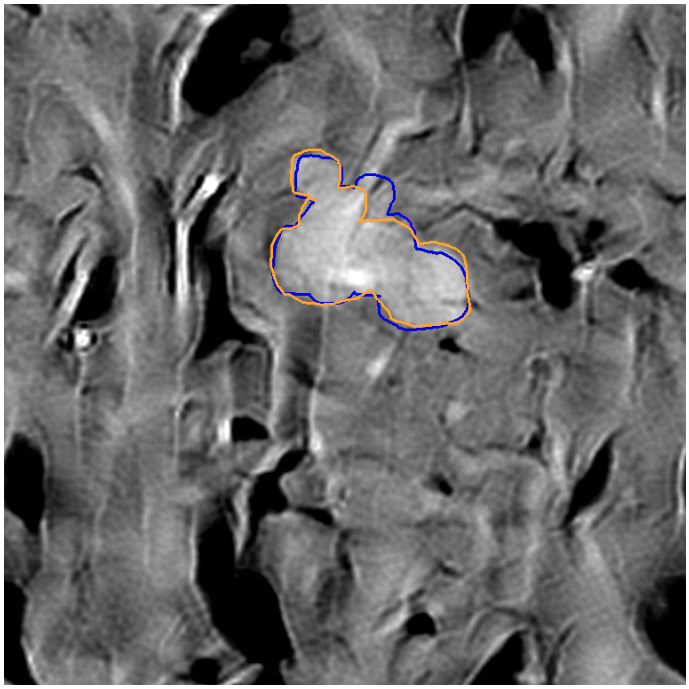
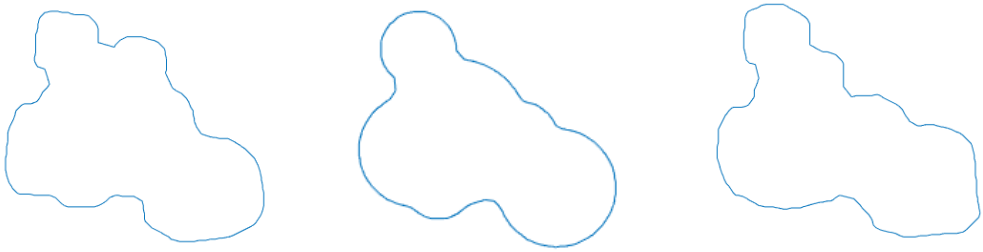


Figure 5.4: The drawn outlines of the perceived tumour overlaid onto the trial image, where blue is the NH (no haptics) condition and orange is WH (with haptics) condition. The orange outline was coloured in post-processing to highlight the difference; both conditions used the same shade of outline during the experiment.

In both individual and group analyses, a paired-samples t-test run on the distance-to-GT per condition shows that the addition of haptic topography and height-mapping based on the relative luminance of the pixels improves performance when compared to a haptically flat stimulus for the delineation of tumours

(a) NH outline (b) Ground Truth (c) WH outline

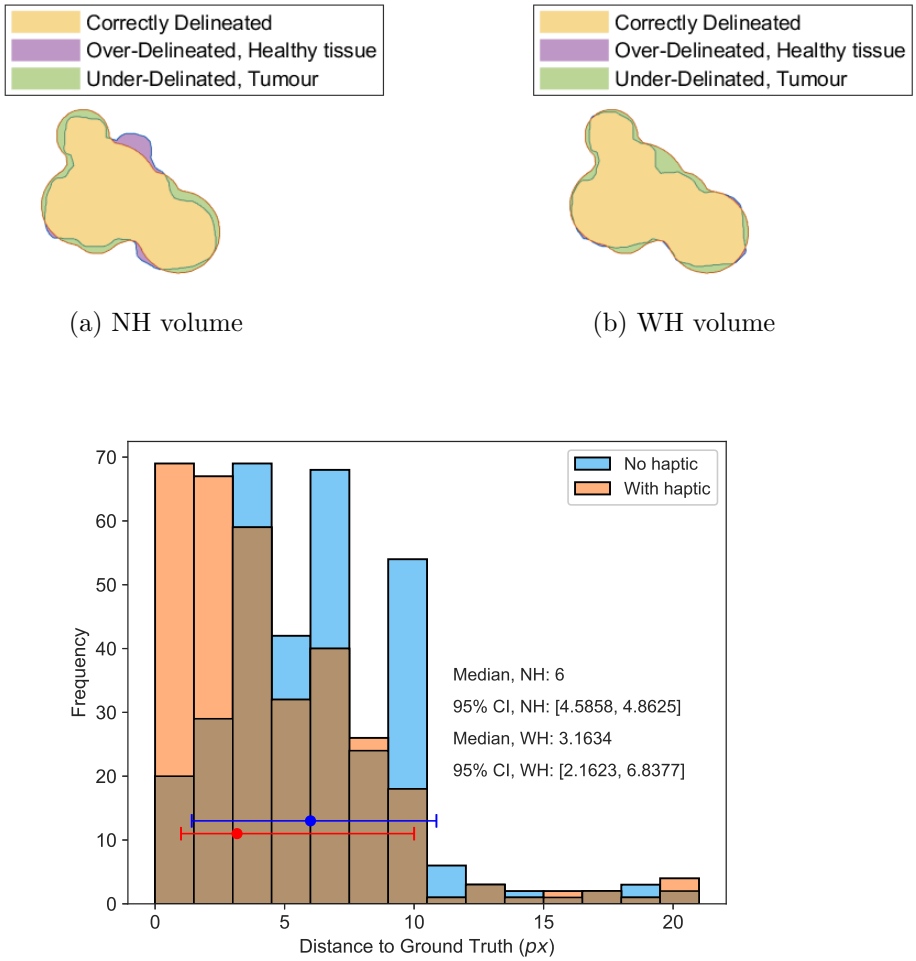


(d) Cleaned up NH (e) Cleaned up GT (f) Cleaned up WH

Figure 5.5: The output outline of (a) drawn by participant on the non-haptic (NH) surface, (b) the outline of the GT tumour, (c) drawn by participant on the with-haptic (WH) surface. (d) shows the cleaned up boundary of NH, (e) of the GT, and (f) of the WH condition.

in these tasks.

Additionally, using the distances between the outlines as obtained by the Delaunay triangulation, the median distance to the ground truth in pixels was calculated for each of the 20 images per person (see Figure 5.6), from which a paired-samples t-test was run on the distance-to-ground truth, showing a significant decrease in distance ($t(9) = 6.123, p < 0.001$) for haptic height mapping WH ($M = 3.74$ pixels, $SD = 1.044$), compared to no haptic height mapping NH ($M = 4.79$ pixels, $SD = 1.413$), as shown in Figure 5.7, with full results shown in Table 5.2.



(a) NH volume

(b) WH volume

(c) Distance histogram

Figure 5.6: Comparison and overlap between the ground truth and delineated tumour. (a) shows delineation of NH compared to GT, (b) shows delineation of WH compared to GT. (c) displays the histogram of distances for NH-to-GT (blue) and WH-to-GT (red), with error bars showing the 95% confidence interval.

These analyses allow us to investigate and compare the volume of the 'tumour' that was correctly outlined, the volume that was missed, and the volume of 'healthy tissue' that was erroneously outlined as part of the tumour, and compare the volume measurements between conditions and participants, as illustrated in

Distance			
	t	df	p
S1	1.269	19	0.220
S2	2.805	19	0.011*
S3	6.140	19	<.001*
S4	3.800	19	0.001*
S5	3.089	19	0.006*
S6	4.536	19	<.001*
S7	2.739	19	0.013*
S8	4.965	19	<.001*
S9	0.982	19	0.338
S10	3.558	19	0.002*
Mean	6.123	9	<.001*
Median	5.587	9	<.001*

Table 5.2: The values of the t-test per participant for the Delaunay contours, and for the mean per participant, as well as the median per participant – per condition. The results were confirmed with Wilcoxon signed-rank tests when normality was in doubt, tables shown in Appendix B, Table B.1 and assumptions in Table B.2.

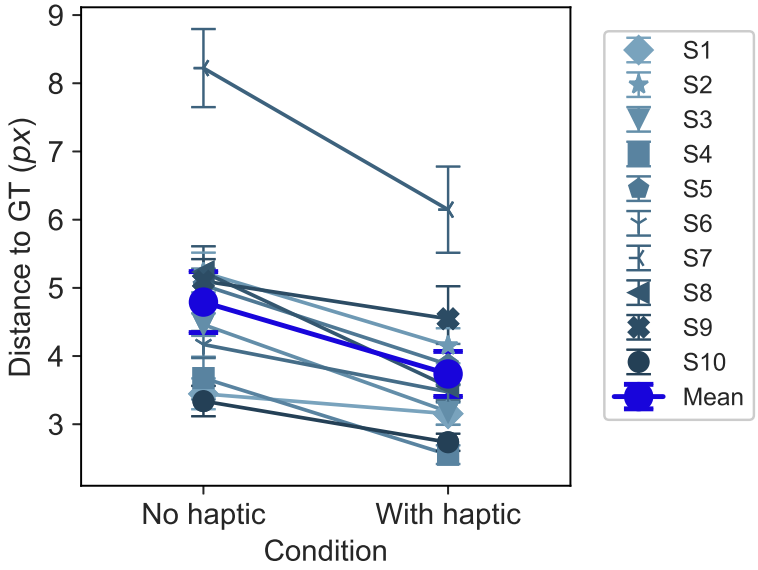
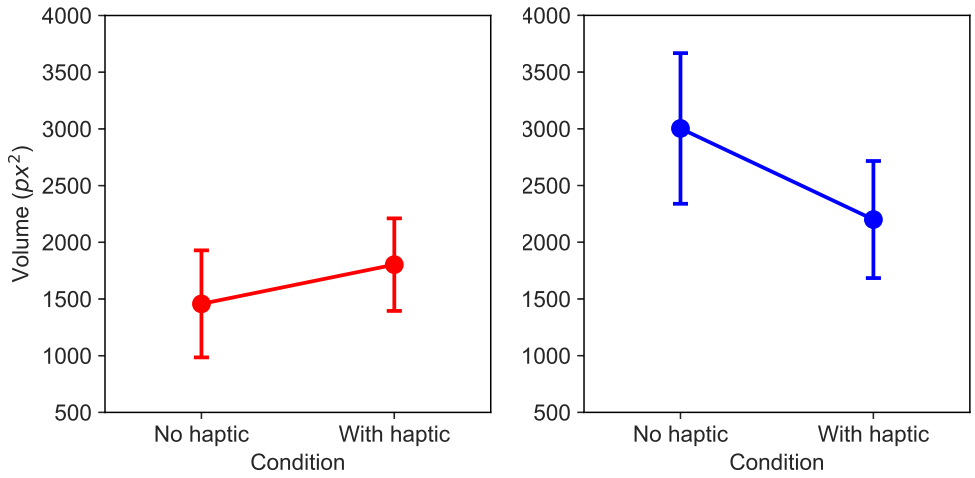


Figure 5.7: Comparative performance of distance-to-GT in pixels, per condition, with error bars showing the standard error of the mean. For a more detailed look at the respective individual plots, see Appendix B, Figure B.19.

Figure 5.8. From Figure 5.8a we can see that the haptic condition delineated a non-significant increase of the 'healthy tissue' than the non-haptic condition

($t(9) = -2.620, p = .028$), while from Figure 5.8b we can see that the haptic also underdelineated significantly less ($t(9) = 8.004, p < .001$). In other words, it correctly delineated significantly more of the 'tumour', leading to less 'cancerous tissue' being left behind. A table with the full results of the t-test, Shapiro-Wilk normality test and and Wilcoxon signed-rank can be found in Appendix B, Table B.3.



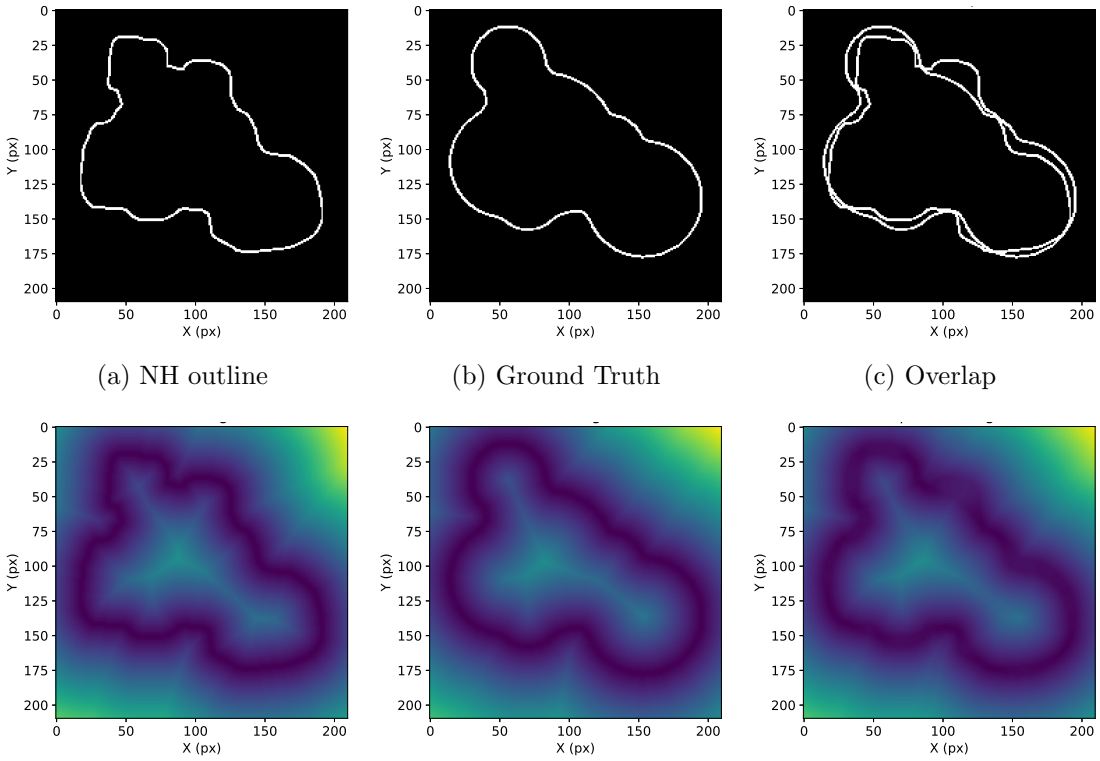
(a) Overdelineation (b) Underdelineation

Figure 5.8: Comparative difference in delineation errors per condition, with error bars showing the standard deviation. (a) shows a small increase ($t(9) = -2.620, p = .028$ in the volume of 'healthy tissue' being erroneously included (overdelineation in Table 5.1), while (b) shows a significant decrease ($t(9) = 8.004, p < .001$) in 'tumour' volume erroneously excluded (underdelineation in Table 5.1), which is equivalent to a significantly larger volume of the 'tumour' correctly delineated. A table with the full results for the volume metrics t-tests, Shapiro-Wilk normality tests and and Wilcoxon signed-rank tests can be found in Appendix B, Table B.3.

5.3.1 Fitting data using a medical algorithm

There are several different ways currently in use to fit the data of delineations done by clinicians, depending on what protocol each individual hospital goes by. Some hospitals use a specialised piece of software called SAS (Vadakkumpadan & Sethi, 2018), other hardware linked software such as Siemen's e.soft (Tateishi et al., 2009) while others use a binary dilation algorithm in Python, and manually perform statistical analyses on the resulting metrics – such as Christie hospital in Manchester. We have been supplied with one of these fitting algorithms as used by our collaborator at Christie hospital in Manchester Dr McWilliam – allowing us to compare our fitting procedure and validate our findings. In the provided medical algorithm, the first step is to locate the outlines by adding a boundary around a masked volume using binary dilation, which as our data are the specific boundaries of the 'tumour' to begin with whereas the algorithm is designed for data where the observer instead creates a mask of the entire 'tumour' area, the first step was not required for our data. After the outlines for the GT and the observer have been obtained, a distance transform is calculated from the coordinates of the GT boundary using the 'distance_transform_edt' function from the Multidimensional image processing (scipy.ndimage) package in python, which creates a Euclidean distance transform matrix which is then applied to the observer's outline, illustrated in Figure 5.9. The result of this is a distance metric calculated by the relative 2D Euclidean distance, in pixels, between the two traces. A paired-samples t-test was run on the distance-to-GT measures found with this algorithm, finding a statistically significant ($t(9) = 4.559, p = .001$) difference in distance, where the WH condition had a significantly lower distance-to-GT than the NH condition, individual t-test results shown in Table 5.3. A side-by-side comparison

of the individual results from both algorithms, as well as their respective means, is shown in Figure 5.10.



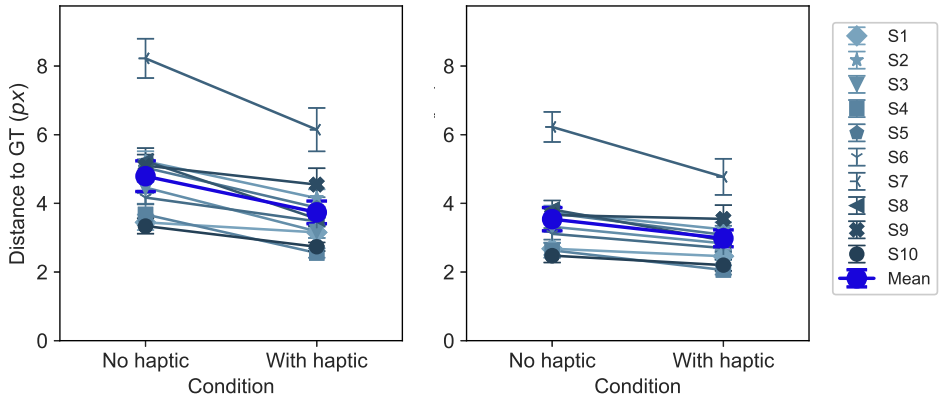
(a) NH outline (b) Ground Truth (c) Overlap

(d) NH Distance gradient (e) GT Distance gradient (f) Overlapped Distance gradients

Figure 5.9: The output outline of (a) drawn by participant on the non-haptic (NH) surface, (b) the outline of the GT tumour, (c) the overlap outlines as used in the medical method. (d) shows the distance gradient from the NH outline, (e) shows the distance gradient of the GT, and (f) shows the distance gradient of the overlapping outlines, where the region around (125, 50) is softer and less defined than the same region in (d) and (e), indicating higher variance and lower accuracy in that region.

5.4 Discussion

In this experiment we aimed to investigate the effect of added haptic height mapping would have on delineating a generated 'tumour' embedded in a synthetic



(a) Delaunay Triangulation (b) Medical Method

Figure 5.10: Comparative performance of distance-to-GT per condition, with error bars showing the standard error of the mean, comparing between the two methodologies of (a), Delaunay triangulation in Matlab, and (b), the medical contour analysis method.

	Distance		
	t	df	p
S1	1.2782	19	0.217
S2	2.087	19	0.051
S3	2.705	19	0.014*
S4	2.638	19	0.016*
S5	2.247	19	0.037*
S6	2.741	19	0.013*
S7	2.469	19	0.023*
S8	4.533	19	<.001*
S9	0.258	19	0.799
S10	2.837	19	0.011*
Mean	4.559	9	0.001*
Median	3.478	9	0.007*

Table 5.3: The values of the t-test per participant, and for the mean and median per participant per condition, medical contour analysis method. The results were confirmed with Wilcoxon signed-rank tests when normality was in doubt, tables shown in Appendix B, Table B.4 and assumptions in Table B.5.

medical image. Based on the previous research in the field of vision and touch improving precision in human observers (Adams et al., 2016; Burge et al., 2016; Ernst & Banks, 2002; Helbig & Ernst, 2007), we predicted that the addition of haptic height-mapping would improve the accuracy of manual 2D delineations

of simulated tumours – both on a per-individual level, as well as overall across all participants. Our results show that the overall distance-to-GT measurement across all participants did, as expected, significantly reduce (Figure 5.7), with all participants seeing some level of improvement. 8 participants had a statistically significant reduction in distance-to-GT, while two participants had non-significant reduction in distance-to-GT. A secondary line of enquiry looked at the volume of delineation – both for underdelineation where 'tumour' is left behind and overdelineation where 'healthy tissue' is included in the contour – on which a paired samples t-test run on the mean delineated volumes for each participant showed that the volume of tumour left out significantly decreased (Figure 5.8b), while the volume of 'healthy tissue' included was a non-significant increase (Figure 5.8a). Without the haptic signal, untrained undergraduate students significantly underdelineate volume, but with the additional haptic the underdelineated volume is no longer significant. While an amount of the haptic delineation is more 'healthy tissue', this is a non-significant increase, whereas the increase in 'tumour' volume is statistically significant. In essence, without the additional haptic signal more of the 'tumour' is missed and left behind, while the added haptic signal had significantly less of the 'tumour' missed, it had a non-significant increase in amount of 'healthy tissue' included which, when considered alongside the significant reduction in distance-to-GT, would translate to a more accurate overall contour.

As a post-hoc test, the contours drawn by the observers were also run through a python-based analysis method provided to us by our collaborator at Christie Hospital in Manchester, Dr Alan McWilliam, which uses multidimensional image processing functions to create Euclidean distance transforms which are applied to the contours. The result of this method showed a statistically significant reduction

in distance-to-GT for 7 of our participants.

These results clearly indicate that the addition of haptic height mapping increases performance in tumour-delineation for untrained non-clinicians, both for the standard clinical metric of distance-to-GT , as well as volume of tumour included.While there is an increase in the volume of included 'healthy tissue' , the distance to ground truth is greatly reduced overall, and the difference between the effect haptic topography has on the erroneous delineations shown in Figure 5.8 strongly suggests that it is not due to people being more cautious, but in fact that they are being more precise. However, it is worth noting that, as over- and underdelineation are not independent, these measures are difficult to compare using common statistical methods. The novel aspect of this experiment is the less-investigated effect of visuohaptic cue combination on observer accuracy, compared to the standard metric of precision. Whilst most psychophysical experiments mind less about accuracy and predominantly investigate overall precision, this approach does not lend itself well to the real-world task of tumour delineation which relies heavily on the accuracy of the position of the tumour itself.

5.4.1 Limitations

While the statistical results are solid, they are based on the delineation of single slice images of non-real tumours embedded in non-real medical images synthesised to the same statistical frequency of a subsection of a brain scan. The delineations are done with an in-lab custom experimental program and performed by undergraduates, PhD students and academics, none of whom had any medical background. Proper delineations are performed on a stack of consecutive images, by trained professionals using specialised software. Before any strong conclusions

can be made regarding the exact benefit and feasibility of using haptics in a real-world diagnostic setting requires further research on at least each of the following sections: medical students as observers; synthetic 'tumour' embedded in a real medical scan; 3D synthetic 'tumour' embedded in a full stack of medical images; real scans of human tissue with cancerous tumours; and interfacing the haptic device framework with the specialised delineation software.

5.5 Summary

In this chapter we examined the effect of haptic feedback on tumour delineation in simulated medical images. After testing 10 participants on outlining a simulated embedded tumour on either a flat surface or a haptic topographically rendered one, we found a positive effect of the additional haptic information where the median distance-to-ground-truth outline was significantly lower in the haptic condition, as well as the volume of missed tumour in the additional topography condition also being significantly lower than that of the flat condition. These findings are a positive indication for a pathway towards potential use in cancer diagnostics and treatment, if it holds for 3D images and real-life tumours.

Chapter 6

General discussion

6.1 Introduction

In this thesis we set out to investigate whether having haptic information available as well as visual information would improve delineations of tumours in medical images. To explore this we ran a series of visuohaptic tasks of increasing complexity. In the first of these experiments we successfully showed that combined visuohaptic cues improves performance for signal detection, where previous literature has predominantly focussed on combined visuohaptic cues for signal discrimination (Ernst & Banks, 2002; Helbig & Ernst, 2007). In the second experiment we investigated the effect of having linearly increasing dissimilarity between the visual and haptic surface textures on precision in a cross-modal 2AFC slant discrimination task, with the aim of investigating the potential viability of using medical images sourced from different imaging modalities such as from MRI and CT scans for each of the respective perceptual cues without incurring a large loss in precision levels in a medical setting. The results found that the increased dissimilarity

between textures did not translate into any uniform degradation or alteration of performance. In the third and final experiment we explored the effect of added haptic topography, by means of luminance-based normal mapping in a tumour delineation task, compared with the current industry standard of drawing on a flat surface with a stylus. The results for this experiment showed that being able to haptically explore the textures of the generated 'medical image' significantly improved the accuracy of the drawn outline compared to when drawing on a flat surface, with observers having a significantly lower distance-to-ground-truth metric with the additional haptics, as well as underdelineating significantly less of the 'tumour' volume.

Overall, we found that having haptic information available improves detection of Gaussian bumps in noisy backgrounds and improves the accuracy of delineations of 'tumours' in 'medical images', while the increased visuohaptic dissimilarity between still-similar textures does not systematically degrade precision in a slant discrimination task. In this chapter we will highlight the findings of each experiment and discuss the potential avenues for further research.

6.2 Experiment 1

In our first group of experiments we approach the main question from a simplified point of view: whether given a non-aligned rig (similar to what would be used in a hospital setting), the addition of a haptic cue would improve the detection of a hidden Gaussian bump compared to using the vision-only cue, as tested across four different levels of added visual noise. The results from Experiment 1.0 showed a clear, statistically significant improvement in bump detection in the

combined cue condition compared to the single cues in isolation. As predicted, the threshold of the Weibull psychometric function fits linearly decreased in the vision-only conditions in proportion to the linearly increasing noise, while in the combined visuohaptic condition the threshold of the function fits asymptoted towards the highest single-cue reliability as the reliabilities of the visual noise levels approached, and were surpassed by, the reliability of the haptic-only cue.

However, timing was not a factor that had been taken into account during the experimental planning stage, and time spent viewing stimuli is a contributor for sensitivity in the Signal Detection Theory paradigm (Prins & Kingdom, 2016), where an extended viewing time can improve the performance of an observer compared to restricted viewing time. As observers took longer in the haptic-only and combined-cue conditions than vision-only, the effect of trial durations on the results could not be confidently excluded due to the effect that longer time spent exploring during a trial could potentially lead to improved performance. Additionally, some observers from the first experiment reported being able to occasionally hear a faint scraping sound emitted from the motors of the haptic device during the active exploration phases. To control for these potential confounds, Experiment 1.1 was created as a control experiment.

Experiment 1.1 was run on the high visual noise condition (20%), replacing the haptic-only condition with a second vision-only condition with a set time limit, and containing additional auditory masking sounds to control for the sound of the haptic device motors. In this experiment, the observer first viewed the bump freely in their own time in the 'free exploration, vision-only' modality. They then explored using the 'combined visuohaptic' mode, again freely in their own time. Lastly, they viewed the stimulus for the median amount of time spent in the

'combined visuohaptic' condition, and once the median duration had passed the trial image went grey and the observer could now select the perceived image, this being the 'matched time, vision-only' modality. This series was repeated four times per person, with the noise playing for the extent of each individual trial for all three conditions.

The hypothesis of this experiment was that the results of the combined condition would again show a lower required signal level for detecting the bumps compared to the free-exploration visual-only condition, and that the signal required for the matched-time vision-only would fall somewhere between these two, erring towards the free visual-only condition. The matched-time vision-only condition did not show any improvement over the free exploration time vision-only condition, successfully ruling out time as the driving factor of the original effect from Experiment 1.0. Meanwhile however, the combined cue condition unexpectedly had a performance equal to the other two modalities. As Experiment 1.0 had a very robust effect, this result was very unexpected, and a second, final control experiment was run to disambiguate the reasons behind the lack of effect in Experiment 1.1. We hypothesised that this was either due to the lack of training or the inclusion of the auditory masking sound.

The second control experiment, the third experiment overall, Experiment 1.2 aimed to replicate the original effect from Experiment 1.0, while using the additional auditory masking. It featured another 10 naïve participants, with 20% visual noise, but unlike Experiment 1.1 it included the haptic-only condition as well as the haptic reliability matching performed prior to data collection. The results of the third experiment again followed our initial predictions; overall the observers had significantly improved detection at lower signal levels for the com-

bined visuohaptic condition compared to the visual-only condition. This finding rules out the errant noise from the haptic device as a deciding, errant auditory cue, as well as cementing the importance of observers having haptic-only experience such as achieved by having a haptic-only training task. One question remains for the purpose of future research, which is whether the haptic-only 'training' block acts as a training task for a signal already known to the observer, or whether it allows observers to recruit the cue as a novel, arbitrary link to the visual stimuli. The exact process as to how and why the haptic-only task is so fundamental is however beyond the scope of this thesis.

6.3 Experiment 2

Having established the positive effect of haptic addition to a simple Gaussian detection task in Experiment 1, we then move on to a more complex stimuli, a spatially coaligned set-up, and a slant discrimination task with realistic textures.

As it is well known that having different signal sources and incongruent signals can be detrimental to precision (Rohe & Noppeney, 2015; Welch & Warren, 1980), we are interested in quantifying to what extent perceptual similarity affects this. As different medical imaging modalities contain different levels of information depending on the area and tissue scanned, there exist already some techniques for viewing these different scans visually overlapped and side-by-side for comparison purposes (van Elmpt et al., 2014; Yadav & Yadav, 2020) – though we would be interested in the potential viability of using the different imaging modalities as the source-images for the separate visual and haptic cues. To this end, we opted to use a perceptual similarity database of naturalistic textures, the PerTex database

(Clarke et al., 2011), for which we created a dissimilarity rating from the known similarity ratings between the textures. Our hypothesis was that as the dissimilarity between the textures increased, the precision in the slant-discrimination task would in turn decrease linearly in proportion to the growing incongruence of the signals. Building on the findings of Experiment 1 we created a training task that both allowed the observers to attune to their haptic sense with a match-to-sample haptic task, and allowed us to gauge the haptic texture discrimination ability of our observers – requiring all participants to score at least 70% correct when matching between visual-only and haptic-only presented textures at the largest similarity difference Δ_3.

In the main task of the experiment, the observers were asked to perform a basic 2AFC slant discrimination task on a left-right slanted surface, where neutral is frontoparallel to the observer. Unbeknownst to the observers, the visual and haptic textures differed by a linearly increasing Euclidean distance in perceptual space, and while the visual texture was altered between modes the haptic texture remained the same in all conditions. The results of the experiment did not however match our hypothesis. Where we predicted a deterioration in precision of slant discrimination as the difference between the two textures increased, instead we found no significant decrease, increase, or overall any consistent change in the shape of performance across the ten participants. As the individual results had different and completely individual responses to the various Δ-levels, the mean precision across all participants presented as a flat line across Δ-levels. This null result was robust both on a group level as well as the individual level, where model comparisons were run to compare the trend of the changes in individual Δ-levels with polynomials at the 0^{th}, 1^{st} and 2^{nd} degree which showed no uniform trend,

after which linear regressions were run on the individual Δ-levels and again no uniform trend was found. However, as our hypothesis and motivation is regarding the feasibility of using different-but-similar signals, this null-result is an overall positive indicator that using different medical imaging modalities (e.g. MRI and CT) for each different sensory modality (e.g. haptic and visual) is less likely to incur a high 'cost' in a medical setting, as previously assumed.

6.4 Experiment 3

For the final experiment we again increased the complexity of the stimuli, looking at the delineation of 'tumours' in 'medical' images. Due to several reasons such as the inherent uncertainty of where healthy tissue ends and cancerous tissue begins, the clear benefit of using known, statistically matched images in perceptual tasks, as well as the concern that as real medical images have clear organ boundaries, it was deemed infeasible to randomly position a generated 'tumour' with regards to the constraints of arbitrary organ boundaries. We elected to generate a series of 'medical' images that are statistically matched as a selected subset of a medical scan, and to generate and embed simulated 'tumours' into these generated 'medical' images. In using the spatially coaligned visuohaptic rig the participants were positioned very similarly to how many clinicians delineate tumours on tablets in actual clinical settings. Our hypothesis for this experiment was that the addition of haptic topography would lead to a significant improvement in the accuracy of delineations and that the amount of volume over- or underdelineated would be greatly decreased.

In the experiment, observers were asked to outline the 'tumours' to the best of

their ability. Half of the trials were done with a simulated haptically flat surface, similar to drawing on a tablet, while the other half was done with the addition of haptic height mapping based on the luminance of the pixels. As in Experiment 2, a training phase was included. The training phase was the same task as the main experimental task of outlining a 'tumour', but with an additional answer key of what the specific 'tumour' embedded in this particular image was shaped like. The result of the experiment matched our overall hypothesis that the addition of haptic topography significantly improves accuracy when delineating 'tumours' in 'medical' images. The primary metric of overall distance-from-delineation-to-ground-truth, which is currently the key metric used in the medical industry, shows for our data that the with-haptic topography condition was significantly closer to the ground truth compared to the flat non-haptic height mapped condition. While several observers reported that they felt their performance was worse with the added haptic bumps, the statistical analyses show that observers included significantly more 'tumour' when the height mapping was available compared with only visual representation and the 'flat' haptic base. While the observers also did include more 'healthy' tissue, overall this was not a statistically significant amount.

6.5 Overall discussion

The main conclusions from this thesis overall are that the addition of a haptic signal helps in detection, discrimination and delineation tasks, and that in the case of simple detection tasks this holds true even for spatially misaligned visuohaptic signals. The additional benefit is not simply due to the extended exploration time,

nor is it due to any errant sounds from the motors of the haptic device during haptic exploration. However, the addition of haptic-only training tasks has been shown to be key in ensuring a uniform benefit for participants. In the slant task, the linear dissimilarity of the textures unexpectedly did not significantly impair performance or decrease precision, having no systematic effect on precision across the observer group. This null-result does however bode well for the possibility of using different medical image modalities for exploring with the different sensory cues. The third experiment found that, in a spatially coaligned setting, on simulated but statistically matched 'medical' images, observers were significantly more accurate at delineating an embedded 'tumour' when haptic height mapping was present, compared to drawing on a flat surface. They additionally included significantly more 'tumour' volume when using the haptic height-mapping than when using a 'flat' haptic base. While in the height-mapped condition it was found that the observers included more 'healthy' tissue as well compared to the 'flat' condition, this was not a significant increase, indicating that both accuracy and precision was increased when the haptic topography was present.

6.5.1 Haptic presentation of medical imaging data

As discussed in the literature review in Chapter 1, and the discussion in Chapter 4 and Chapter 5, there are many potential obstacles to implementing the addition of a haptic signal for the detection and delineation of tumours in medical imaging data. On a basic level, if the haptic device is not directly trained for, observers may fail to learn to integrate the haptic signal (Holmes et al., 2004, 2007; Holmes, Sanabria, et al., 2007; Maravita et al., 2002; Takahashi et al., 2009; Takahashi & Watt, 2014), which may be a contributing factor to how some of the newer tech-

nologies do not offer reliable improvement outside of the labs they were designed and tested in. Considering the raw data itself, the medical imaging modalities are density scans on a per-slice basis, whereas the proposed haptic rendering uses haptic normal-mapped topography, which is a type of height map display. It is possible that the clinician can either ignore or appropriately map the distinction between density and depth, but this is not a given. We have answers for some of these queries, though some are still left open to future studies. In all the experiments of this thesis, the visual signals used shape-from-shading, rather than full stereoscopic 3D depth, and the haptic signal was normal-mapped topography rather than using quantifiable units of haptic depth, such as in a 3D surface mesh.

From the first series of experiments, Experiment 1.0-1.2 (Chapter 3), we found that the addition of a haptic signal to a visually noisy signal significantly improves the detection of a 2D Gaussian signal in a 2AFC signal detection task. We also know that exploration time alone is not the primary reason for the improvement, nor is the improvement due to errant auditory cues from the haptic device. We also now know that it is crucial to 'train' the haptic-only signal for the benefit to occur in the combined visuohaptic condition. However, the experiments in Chapter 3 were all performed on a spatially misaligned setup, and we do not know whether the improvement will be the same for a spatially coaligned version of the experiment, or whether the haptic 'training' would be as crucial if the sensory signals were spatially coaligned as well as matched in temporal presentation and using a one-to-one movement mapping.

From Experiment 2 we know that, for textures with a small, predefined perceptual dissimilarity level, observers do not perform significantly differently depending on the dissimilarity level between the visual and haptic textures. Ad-

ditionally, it is known from the training task that all the observers were fully capable of differentiating between Δ_3-level texture-pairs when relying only on a haptic signal. However, it is not known whether the reliabilities of the four visual textures and the one haptic texture per observers were perceptually comparable for that individual, what effect the aperture had on the individual reliabilities of the respective visual textures, nor what the 8 different perceptual dimensions comprising the perceptual similarity matrix correspond to in terms of statistical or perceptual texture features. As such, it is unknown which texture features may have been more or less contributing to the results.

From the last experiment, Experiment 3, we know that after a period of training, and for artificially generated tumours embedded in statistically matched synthetic medical images, observers delineated significantly closer to the ground truth outline when the haptic topographic normal-mapping was present compared to when drawing on a haptically 'flat' surface, using our in-house custom delineation software. Analyses showed that observers included significantly more 'tumour' volume and not-significantly more 'healthy' volume with the added haptic signal compared to not having the additional haptic signal. Additionally, observers spent significantly longer time in the with-haptics condition compared to the flat condition, though this difference in exploration time is somewhat lower when comparing when the tumours in the order they were presented in compared to between the same tumour across conditions.

It is still unknown whether the benefit would transfer to 3D surface meshed pseudo-medical images, whether it transfers to genuine tumour shapes rather than the generated 'tumour' clusters or genuine medical images with real anatomical features and tumour constraints, and whether it occurs using more sophisticated

delineation software. There is also the question of whether the improvement in delineation distance and volume would occur for trained radiologists or untrained medical students who still have a higher level of anatomical knowledge than your average psychology undergraduate student.

6.5.2 Participant expertise

The observers who took part in these experiments were predominantly students in the School of Psychology and Clinical Language Sciences. Some postdoctoral researchers and other academics also took part, though none were medical students nor otherwise medical professionals. These observers as such did not have any training for expected anatomical structures, detection or delineation of tumours, nor expected tumour shape and size, other than what was provided as a training task in Experiment 3.

On one hand, this can be a benefit when testing novel technologies, as non-medical students are 'blank canvases' without prior experiences and lack expectations of anatomical structures or cancer pathologies, not having already been taught the different cost and risk functions associated with error margins and underdelineation. On the other hand, the simulated 'tumours' and 'medical' image backgrounds they were embedded in were intentionally designed to avoid organ boundaries and other defining anatomical features, meaning a trained clinician may not have been able to rely on their training for a given pathology or anatomical structure, likely forcing them to make a decision between applying the closest perceived pathology to which they had trained, or to delineate to risk/cost functions based on their instincts – the same way the non-medical students had to delineate.

Whether a trained radiologist would perform quicker or slower than the non-medical students would likely depend on the area of expertise, and whether they were able to put aside their training regarding anatomical features and tumour cost/risk functions. They may well be faster and more precise at delineating due to their overall training, or they may be less precise given the lack of clear cost/risk function provided. The opposite is also possible: the radiologists might be slower and much more precise, depending entirely on their training, personal caution and whether they perceive the 'tumour' to be requiring larger or smaller margins of error included. In the real-world medical cases, the clinicians are constrained with how much time they can allot per delineation, as there are a significant amount of scans to delineate, and only so many hours in a day. Lastly, they may well struggle with lack of functionality present in the in-house built delineation software used in Experiment 3, compared to the high-end delineation software available in hospitals.

It is possible that the results of Experiment 3 are not generalisable beyond brain cancer, as a brain scan was the source image from which the 'base' images were generated, though it is possible that it would not apply well to brain cancer due to the uncertainty of shape, size and placement of brain tumours. It is possible that the results of the untrained non-medical students would be comparable to untrained medical students, or more perceptually 'flexible' trained radiologists, who were able to and willing to apply the 'yes-and' philosophy to the generated non-real medical images and synthetic tumours. It is unlikely, but not impossible, that the results would extend directly to performance of real-world delineations of physical tumours in genuine medical images, delineation software aside. Currently there are a combinatorial explosion of potential variables that might influence the

end result, to which end further studies would need to help pare down the different sources of confounds.

6.6 Considerations for clinical adaptation

6.6.1 Challenges of adopting approach

In a real-world scenario, there is a multitude of sensory cues occurring all at once. A piece of cloth has shading, contours, visual texture, reflective properties, to name but a few – a matte surface is more likely to be pliable, rougher and temperate, while a shiny surface is more likely to be hard, smooth and cold (Adams et al., 2016). In order to be able to make accurate inferences, we have to be able to isolate what aspect of the sensory input is contributing to the different perceptual estimates of the object; a matte surface could be hard or pliable, but a glossy surface is rarely deformable, while a non-glossy fabric could be coarse and uncomfortable, or it could be soft and smooth, depending on the surface geometry and source material (Etzi et al., 2014).

To investigate the effects of these sensory cues in isolation, we use carefully constructed generated sensory signals. By generating our own signals, we have full control over what variables we are manipulating and as such can make direct connections between what is being changed in the stimulus compared to how this affects an observer's performance. For example, in Experiment 1 we used a simplistic 2D Gaussian embedded in Gaussian white noise, whereas the PerTex textures used in Experiment 2 are magnitudes more complex than the 2D Gaussian, having 8 differentiable perceptual dimensions. This large difference in perceptual complexity is present in spite of the fact that the textures in the

PerTex database were carefully collected, the physical features of the underlying textures normalised to ensure they had the same mean luminance and that the depths were of the same level and that all illumination followed the same Lambertian rendering. Had we instead used real-world textures as-is, it would have been impossible to infer what variables could be influencing similarity. Properties such as colour, texture depth, and illumination angle could all have affected the final percept of the texture itself.

In order to ensure we had 40 unique quasi-anatomical medical imaging backgrounds for each observer, which were matched in frequency, luminance, complexity and source anatomy, we had to generate our own based on the underlying statistics of the source image. With this method we are able to make inferences on the affect of haptic normal-mapped topography compared to drawing on a flat surface, irrespective of anatomical source and whether one observer had background images with unidentified abnormal anatomical features. Our generated backgrounds do not contain organ boundaries, they only contain synthesised soft-tissue grey matter, enabling us to randomly insert a tumour-like object without the concern of how tumours naturally grow in different types of tissue. In real medical images there are different organ boundaries, and there are different tissue-types, such as soft tissue, bones, cartilage and blood vessels. If one were to embed a tumour, generated or modelled from an actual scan, one would have to take into account the pathology of the simulated cancer type – bone cancer shows up as spikes growing on the underlying tissue, breast cancer shows as Gaussian-like lumps in surrounding 'lumpy' tissue containing fat, mammary glands and milk ducts, while lung cancer shows as growths in otherwise 'empty' spongy tissue. The shape, size and relative 'density' (rendered as luminance) of the 'tumour'

would have to be carefully matched to the anatomical site in question. In short, real medical images are magnitudes more complex than our generated stimuli, differing in local anatomy, how tumours present in the different cancer pathologies, and even differences in physical features of the individual patients – making it difficult to make direct inferences to how changes to sensory input might be affecting the final delineations. This is one of the challenges of comparing perceptual research as performed within the controlled confines of the perceptual laboratory to the active perception of objects in the real world.

Breast, prostate, and colorectal cancers can often be palpated manually, whereas lung, brain, and stomach cancers can only be detected through medical imaging. It is fully possible that the addition of a haptic signal is only sufficiently beneficial when used as a sensory substitution mechanism in cancer pathologies where the tumours themselves cannot be physically palpated by the physician, as clinicians are significantly better at locating tumours when using their physical finger compared to using a rigid tool (Greenwald et al., 2012), and the human finger has been found to have a textural resolution of up to 10 nm (Skedung et al., 2013).

The question also remains of whether the soft tissue contrasts in an MRI scan are haptically differentiable, or whether the texture-features themselves are haptically metameric (Kuroki et al., 2019). Whereas in a CT scan, the soft tissue contrast being much lower, it could be that the large difference in contrast between bone and soft tissue negatively adjusts the sensitivity of the haptic texture perception, leaving any anomalies being filtered out as simply 'noise'. If these issues are found to occur, they could of course be controlled for and adjustments made to the underlying medical images, for example amplifying texture features in an MRI or cropping out the source of 'distraction' such as task-irrelevant bone

in a CT image.

There is also the issue with sourcing participants. While psychophysical experiments typically do not require hundreds of participants, the participants recruited do perform the experiment for several sessions, usually spanning a number of hours. This makes it difficult to run experiments using trained medical experts, and even non-trained medical students, as both of these groups are generally very busy and would be difficult to get a sufficient number of participants for the required amount of data to be collected per observer.

To summarise, the proposed methodology is not currently directly usable as-is in a clinical delineation setting. A further review would need to be made for the macro groups of cancer pathologies, and further studies would have to be put in place to investigate haptic sensitivity and texture discrimination performance on realistic anatomical structures such as soft-tissue contrasts in both MRI and CT scan images, as well as other medical imaging modalities not mentioned in this thesis.

6.6.2 Ground truth

The crucial issue still remains that, even through delineation by several experts within specific cancer pathologies, the ground-truth tumour will be unknown for the majority of cancer pathologies (Njeh, 2008). There is the time factor between when the imaging scan is performed, when the delineation takes place, and when treatment can start (Nelms et al., 2012; Papiez & Langer, 2006). As the cancerous tissue grows out of otherwise healthy tissue, there is uncertainty of how far the mutation itself has already spread – which is usually mitigated by having relative 'safety margins' to include around the delineation, the size of which strongly differs

between the different cancer pathologies and the anatomical structures surrounding them; head and neck tumours have much lower 'safety margins' compared to skin cancer or colorectal cancer, due to the proximity to organs-at-risk (Leclerc et al., 2015).

Until further improvements can be made in the study of shape and developmental stages of different tumour pathologies, in the design and use of algorithms and methodology for extracting the most useful information from the medical images, in reducing the cost of running medical scans, and in reducing the time spent on delineations by clinicians, the ground truth of tumours cannot be fully known in living, human patients. This has negative implications for the current state of manual tumour delineation – where medical students are trained on slightly uncertain principles, for semi-automated and automated delineation algorithms which are trained on inherently imperfect data-sets – either through the size and shape of the tumour being unknowable or the non-natural use of synthetic tumours and a known ground truth, and lastly for our proposed methodology and other new methods of delineation – where the results can either be compared to best-guess delineations by talented but imperfect experts, or using a synthetic ground truth which cannot be fully comparable to real tumours found in the real world.

6.6.3 Benchmark tests

In order for a new method of delineation to be considered a significant improvement and evaluating whether or not to adopt it, it is important to have comparative data to the current methods of tumour delineation. While a novel method of delineating might be more accurate overall, if this accuracy comes at the expense

of significantly slower delineation time, or prohibitively costly equipmenet, it is unlikely to be offering enough of an improvement over the current method to overcome the costs associated with implementation, both monetary and training-wise.

The simplest method would be to compare the performance of the current standard of delineation, which is using delineation software to traverse through the image stacks provided by the imaging modalities, occasionally viewing different imaging modality scans in parallel or superimposed onto one another. It would first be required to test the difference in performance on a simulated tumour in a simulated medical image, though ideally this would be simulated as a 3D 'stack' of images, similar to the output of an MRI or CT scan. By using the simulated images and synthetic tumours, the ground truth would be known and used to compare the relative performance between the methodologies.

It would be interesting to compare relative performance differences between medical students without prior training and trained radiologists with a large amount of experience, as well as between different cancer pathologies. The performance would be quantified by several metrics, such as delineation time, accuracy of delineation, variability between clinicians, and correctly delineated volumes. If the visuohaptic signal method improves the accuracy of the delineation and the volume ratios, as well as the inter-clinician variance, with a small-to-medium increase in delineation time, the next step would be to compare delineation time with task exposure – whether the delineation time reduces at a sufficient rate to be considered feasible for the methodology to be implemented in the field. Additionally, if the improvement is significant in the medical students without prior training, but not in the trained radiologists with specific work experience, it would suggest that an early introduction to the methodology would offer greater benefit

than late adoption, as the training already undertaken by experienced clinicians may overshadow any potential benefit.

6.6.4 Further studies towards a real-world implementation

While the results of these three experiments come together to show the potential for haptic signals to beneficially be added for tumour delineation tasks, there is still much research to be done before this would be implementable in a real-world clinical setting. In order to establish more definitively whether the proposed approach would be potentially viable, there are several experiments which would follow on from Experiment 3. By gradually increasing the realism and complexity of the stimuli, such as running the experiment with a combination of synthetic tumours, real tumour-shapes, generated medical images, and genuine medical images, we would be able to establish whether the synthesized stimuli are comparable to the real-world medical images, having the control be a tumour-containing medical imaging scan – where the ground truth would be the delineation as performed by expert radiologists. This could be run on all of the three potential participant pools of non-medical students, medical students, and trained clinicians, or simply the last two, and would be comparing all permutations with drawing on a haptically 'flat' surface as well as using the normal-mapped topography.

While it would be scientifically rigorous to gradually increase the realism of the individual aspects such as tumours, medical images and observer expertise, it would be prohibitively slow to do so for all possible permutations, making it unfeasible to explore every aspect in isolation. However, by running an experiment as described above, we would be able to determine several interesting points, such as realism of the generated tumours embedded in generated medical im-

ages, realistic tumour-shapes embedded in real medical images, and the genuine tumour-containing medical image. If there are large differences between these, it would indicate that more studies are required to either improve the realism of the non-tumour-containing stimuli sets, or run multiple permutations on the tumour-containing medical images specifically.

Another potential direction would be to run Experiment 1, 2 and 3 using 3D surface meshes of the respective stimuli, where the 3D surface meshes are generated using the relative luminance of the source image as a height map. By doing so, we would have quantifiable haptic depth as well as stereoscopic visual depth, which is a more salient sensory cue than shape-from-shading (Harris, 2004; Lovell et al., 2012). If the performances are worse than the performances of the original experiments, this indicates that the haptic 'roughness' is more important or available as a cue than the haptic 'depth' per se. In Example 3 specifically, if the 3D surface meshes are performing worse than the normal-mapped haptics but better than delineating on a 'flat' surface mesh, then the height of the haptic stimuli may be too large, or it may be more suitable to modify digital pens in the same method as used by Evreinova et al. (2012), rather than husing a full 6DOF haptic feedback device. If the performance is the same or better, it would be worth exploring further the differences between the two rendering methods.

A third potential direction would be to use either synthetic backgrounds which were matched to, or generated from, different imaging modalities. This would effectively merge the stimulus presentation methodology of Experiment 2 and Experiment 3, and would allow us to investigate whether the discrepancy between the visual and haptic signals is disruptive, negligible, or beneficial. Additionally, by altering which imaging modality would be presented in the different sensory

modalities, it would be possible to establish which of the imaging modalities would be more suitable for the respective sensory ones, be it per anatomical region or overall as a general rule of thumb. These suggested directions are, of course, non-exhaustive – in addition to the experimental directions outligned above, there are several other proposed methodologies for improving delienation, some of which may well be synchretically combined with the suggested implementation of this thesis.

There are other avenues currently being researched to improve the delineation of tumours through the use of haptics, though most of these rely on semi-automation, which offers both its own benefits and drawbacks (Banerjee et al., 2017; Latifi-Navid et al., 2016; Nyström et al., 2009; Schulz-Wendtland et al., 2017; Vidholm et al., 2008). One of the studies that looks at the use of haptics for seeded semi-automation of delineation is one by Vidholm et al. (2008). The authors have developed a technique for using haptic exploration and stereoscopic vision as presented alongside with the ability for clinicians to 'seed' the respective areas of the liver to guide semi-automatic liver segmentation. Their proposed technique also utilises a deformable model in the shape of a 2-simplex mesh sphere that is initialised on a specific region of a flat slice image rendered in 3D. The user interactively reposition and scale this sphere before starting the deformation process, effectively modelling the sphere into the shape of the organ in question. The benefit of their proposed technique is the encompassed simplex mesh mitigates leakage problems that occur during traditional fast marching techniques of automatic organ delineation, which historically does not yield as smooth results.

6.7 Summary

In this chapter we have summarised and discussed the findings of Experiments 1.0-1.2, Experiment 2, and Experiment 3. We have found that a haptic signal can improve performance in detection and delineation of tumour-like objects embedded in perceptually 'noisy' backgrounds, and that an increasing but small level of perceptual incongruence between visual and haptic textures does not reliably impair performance in a slant discrimination task. Additionally, we have discussed some of the remaining concerns regarding the fully fledged clinical implementation of the proposed method and suggested some potential avenues for further research, and identified a possible branch for further interdisciplinary work.

Chapter 7

Contributions and implications

7.1 Context

This thesis aimed to investigate whether having the addition of a haptic signal as well as visual information would help improve the detection and delineation of tumours embedded in medical images. In order to explore this, three main experiments were run; gradually increasing the complexity of the stimuli.

There are several aspects of tumour delineation as it is in the field today which can be improved upon. One of the major bottlenecks of improving treatment outcomes is the accurate delineation of tumours. As tumours grow out of otherwise healthy tissue, it is often-times difficult to identify where the tumour ends and healthy tissue begins, and how much of the surrounding healthy tissue has already been 'infected', and how much will be cancerous by the time treatment can commence. Not only is the cancerous tissue commonly difficult to identify due to the background, there are also large differences between clinicians delineating the same tumour. This inter-clinician variance has many potential causes, ranging

from whether the radiologist was trained to delineate on the actual border of the tumour itself or on the outside of the border, to the cancer pathology and personal interpretation of the local anatomy of the scan.

While modern delineation software allows the clinicians to easily traverse the stack of images comprising a 3D medical scan such as an MRI or CT scan, it is still a 2D representation of a 3D object. By rendering the scans slice-by-slice in 2D, there is still a loss of information regarding the body as a whole.

The current medical imaging modalities themselves are also imperfect, where some scanning technologies are better suited for some cancers but not others. For a soft-tissue sarcoma, an MRI is usually the better choice. For bone cancer or ovarian cancer, CT scans show more information. It is not unusual to use a combination of these imaging modalities, which are commonly viewed in parallel side-by-side, using a toggle to switch between the two with one image overlaid onto the other, or gradually change the opacity between them to glean more specific information as needed. This is, however, not an ideal way of investigating the images.

In perceptual psychology as a whole, it has been shown many times that the use of multiple sensory signals can greatly improve a percept; such as using both vision and touch for estimating the size, height, or orientation of objects. When the signals are considered of comparable reliability, the sensory system can combine the information and reduce perceptual errors. These specific issues were condensed into three specific questions – for which, in order to answer them, three different experiments were run.

7.2 Experimental questions and contributions

Question 1. Can haptics help with detecting a signal, even when vision and touch spatially misaligned?

Even if the ideal addition of a haptic signal can improve performance, it is not certain that it would be feasible to have fully spatially coaligned vision and touch in a medical delineation setting. As such, it was important to establish that 1) the performance improved for the detection of a signal, as the majority of visuohaptic combination studies predominantly look at signal discrimination, rather than signal detection, and 2) that this was true for spatially misaligned vision and touch, as the majority of visuohaptic combination studies exclusively look at spatially aligned signals.

Using a signal which matches the features of mammographic tumour images, we showed that the addition of a haptic signal significantly improved the detection of a signal compared to when using either cue in isolation. Even when controlling for exploration time and potential spurious auditory cues, it was shown that haptic-only training was a key element for the cue integration to occur, and found that the overall improvement fits well with the maximum-likelihood-estimation model – even for spatially misaligned visuohaptic signals.

Question 2. If haptics and vision are mildly discrepant, how quickly would integration break down?

While there is a lot of complementary information available when comparing different medical imaging modalities, there is currently no elegant way of doing so. There are naturally occurring differences between the images, such as density

(where bone appears 'black' on an MRI scan but 'white' on a CT scan), spatial distortion of an MRI due to the magnetic field, and random movements of the body between scans such as breathing or 'chaotic' organ movement (Nelms et al., 2012; Papiez & Langer, 2006). It is important to establish that, if using different-but-similar source images for vision and touch, how different the source images can be before the integration breaks down and the additional information is no longer beneficial.

By using a database of textures with perceptually ranked similarities, we were able to explore the effect of a gradually increasing the perceptually ranked dissimilarity between the visual and haptic textures, seeing how the introduced incongruence affected the performance of observers in a slant discrimination task. After first controlling for whether the observers could reliably identify the different Δ_3 dissimilarity level texture pairs, a psychometric function was collected for each of the four texture pairs per observer. The results show that there was no reliable decrease in performance, nor any emerging trend in the effect of introduced dissimilarity for low levels of perceptual dissimilarity. This was tested for linear regression of all four Δ-pairs per observer and mean across all observers, and using only the extremes Δ_0 and Δ_3 per observer, as well as testing the best-fitting polynomial per observer, identifying whether the results were better fit to a 'flat' line (0^{th} degree), a 'rising' or 'falling' line (1^{st} degree) and a 'horse shoe' or 'u-shape' (2^{nd} degree), where the hypothesis suggested the precision would follow a decreasing 1^{st} degree polynomial shape. Again, there were no consistent results across the individual observers nor for the mean across observers.

Question 3. Would haptic height mapping of an image help delineating the tumour?

While the addition of haptic information can offer significant and statistically optimal improvement in many perceptual tasks, there is no guarantee that the addition of information will always be beneficial. For example, using augmented reality for surgeons can cause a higher rate of 'missed' objects due to distractions in the visual field. It is important to establish whether, for the same generated medical image with a synthetic tumour in both the visual and haptic modality, observers would be better or worse at delineating the synthetic tumours.

Experiment 3 used increased realism and complexity of stimuli, in order to look at the delineation of synthetic tumours embedded in artificially generated medical images. A series of 'medical' images were generated to be statistically matched to a source image being a selected subset of a medical scan, and to generate and embed simulated 'tumours' into these generated 'medical' images. By generating statistically matched 'medical images' and synthesising tumours, we could test the delineation of artificial tumours while comparing the results to a known ground truth. By using the same haptic mapping algorithm as in Experiment 1 and Experiment 2, we could compare the delineated tumours as done once on a flat surface, and once on a haptically 'bumpy' surface, where the order of presentation was counter-balanced and randomly chosen on a per-participant basis. Additionally, all 10 observers had 40 unique medical images with tumours generated on a per-observer basis. As such, there was no duplication of the medical backgrounds or tumours between observers.

The spatially coaligned visuohaptic rig had the participants positioned very similarly to how modern delineation stations are set up, where clinicians delineate

tumours on graphics tablets. The hypothesis for this experiment was that the addition of a haptic normal-mapped signal would lead to an improvement in the accuracy of the tumour delineations, and that the amount of volume over- or under-delineated decrease. The results showed a significant decrease in distance to ground-truth in the haptic normal-mapped condition compared to the haptically 'flat' condition, with a significant increase in volume of 'tumour' delineated, and a non-significant increase in volume of 'healthy' tissue.

7.3 Implications of findings

7.3.1 Experiment 1

Experiment 1 was required to lay out the ground-work of whether the visuohaptic cue integration would occur and show benefit in a signal detection task, whereas the literature predominantly focusses on signal discrimination, and whether this was the case when the signals were not presented as spatially aligned. Experiment 1.0 showed a clear, statistically significant improvement in signal detection of a 2D Gaussian bump when using combined visuohaptic signals compared to either single cue in isolation. The threshold of the Weibull psychometric functions decreased as expected in proportion to the linearly increasing noise in the visual-only condition, while in the combined visuohaptic condition the threshold of the psychometric function asymptoted towards the lowest single-cue variance as the reliabilities of the vision-only condition for the respective noise levels approached, and were surpassed by, the reliability of the noise-less haptic-only cue.

The first of the two follow-up experiments showed that, contrary to the predictions and findings of Experiment 1.0, no significant improvement was found for

either the combined visuohaptic condition or the matched time condition, despite having comparable exploration time and detection thresholds between Experiment 1.0 and 1.1. While ruling out exploration time as the primary contributor to the improved performance, a further experiment was required to establish whether the improvement was due to an errant auditory cue or whether the haptic-only condition unintentionally doubled as a form of training or familiarisation of the haptic device as a tool for haptic detection. The results from Experiment 1.2 matched the original results from Experiment 1.0, which indicates that the improvement found required experience with the haptic cue in isolation prior to collecting for the combined visuohaptic condition, in order for the observers to show improvement.

7.3.2 Experiment 2

Experiment 2 aimed to quantify the extent to which high levels of perceptual similarity, or rather low levels of perceptual *dissimilarity*, affected the integration of vision and touch for a slant discrimination task, aiming to break down the sensory correspondence between the two textures at higher levels of perceptual dissimilarity. This was done to emulate the small levels of naturally occurring incongruences which would be present if one were to use different imaging modalities as sources for the different sensory modalities.

Contrary to the hypothesis, the results did not show any consistent effect of increasing dissimilarity on slant discrimination performance; there was no consistent decrease, nor increase, in precision, across the observers. Overall, the results did not show any consistent effect, whether positive or negative. As such, a small increase in Euclidean distance between visual and haptic texture did not reliably

reduce slant discrimination performance for realistic textures. These results indicate that, for perceptually similar images such as perhaps the same areas in medical scans, there is the potential that the introduction of a haptic signal with a slightly different source image would improve the available information, but not detract from the improvement.

These findings are an important first step to avoid ruling out the effect of discrepant but similar signals. It is also an interesting consideration regarding visual and haptic texture perception. It would be interesting to run follow-up experiments with different haptic signals, keeping instead the visual signal constant, matching the stimulus reliabilities through the use of more tailored aperture widths, and to explore the haptic similarity ratings between the textures in the set. It would also be interesting to see if synthetic medical images, generated similarly to those used in Experiment 3, would have a similar lack of effect. It would also be an interesting experimental series to try and perceptually match textural subsets of the different medical imaging modalities.

7.3.3 Experiment 3

We examined the effect of haptic feedback on tumour delineation in simulated medical images. A positive effect was found on the haptic normal-mapping; the median distance-to-ground-truth outline showed a significant decrease of distance in the normal-mapped condition compared to the 'flat' condition, as well as the volume of missed tumour in the normal-mapped condition being statistically significantly less than that of the flat condition. As such, observers delineated significantly more and closer to the ground truth tumour when they can feel the topography, while some of the additional volume is non-tumour, the majority of

it is 'tumour'. This offers the potential use of haptic additional signal for use in cancer diagnostics and treatment, where further research will need to look at the use of genuine medical images, genuine tumours, and trained clinicians, to name but a few avenues.

7.4 Methodologies

Additionally, in designing and running the three experiments, we also applied a number of methodologies that are novel either in themselves or in their implementation. The first of these was the rigorous calibration of the spatially coaligned rig, as detailed in Chapter 2, which used a combination of Vicon motion tracking of objects, both directly and transformed using a vector-stick, and the object coordinates as rendered by the haptic framework. There is currently a paper underway regarding the rigorous calibration procedure.

Secondly, in Experiment 2, the stimuli were chosen to be naturalistic texture surfaces taken from the Edinburgh PerTex database of rendered real-world surface textures (Clarke et al., 2011), chosen in part for their numerically quantified similarity distance to one another in perceptual space. One texture was first selected to be the 'base texture' of a texture pair, where the haptic and visual textures were both set as initial texture, Δ_0. Each subsequent texture-pair $(\Delta_1, \Delta_2, \Delta_3)$ was selected to be visually different at predefined dissimilarities measured in distance within the 8-dimensional Euclidean space defining the perceptual similarity features. The PerTex database of perceptually similar textures has, to our knowledge, not been used with a haptic signal, nor to compare haptic and visual similarity or effect of dissimilarity.

Lastly, the base of the stimuli for Experiment 3 was a synthetically generated medical image background. While the requirement included being comparably realistic to match real medical images, with unique-but-consistent backgrounds per individual tumour and observer, which were generated using the textureSynth Matlab toolbox by Portilla and Simoncelli (2000). The toolbox works by inputting a source image, which is then analysed and reconstructed statistically based on perceptual scales, frequencies and other features. The resulting images were both statistically and perceptually comparable to medical images, as verified in discussions with radiologists at The Christie NHS trust in Manchester. The method of generating artificial medical images using the textureSynth toolbox for arbitrary medical images has not to our knowledge previously been done, and was a subject of some interest when presenting the work in thesis to clinicians at The Christie NHS trust in Manchester.

7.5 Conclusion

The main conclusions of this thesis are that the addition of a haptic normal-mapped signal aids in the detection and delineation of tumour-like objects, does not reliably deteriorate slant-discrimination performance with small-but-increasing levels of visuohaptic texture incongruence, and that the improvement in a signal detection tasks also occurs when the sensory signals are presented spatially misaligned. Experiment 3 aimed to investigate haptic height mapping effect on tumour delineation, predicting based on existing literature that the accuracy of the manual delineations would improve on both a per-individual level as well as across the mean of all participants, on a group level. The novelty of Experiment 3 is in

the demonstrated effect of haptic signal for cue integration on observer accuracy in a delineation task, where most psychometric studies focus on precision. Experiment 3 did show statistically significant improvement in tumour delineation metrics when using haptic normal-mapping compared to drawing on a haptically 'flat' surface, however, it is worth noting that this was performed on single slices of images using synthetic tumours and generated medical backgrounds, performed by non-medical students and other academics, using custom, in-house built delineation software. Since the images did not exist as a 3D stack as a 'whole', the observers were untrained, the tumours and medical backgrounds synthesised, there are many possible permutations required in order to test whether the proposed method is feasible and beneficial for industry use.

The primary contributions are the confirmation that visuohaptic cue combination occurs for signal detection as well as the established signal discrimination, that it also holds true for spatially misaligned sensory signals. Using increased perceptual dissimilarity does not reliably deteriorate performance in a slant discrimination task, and using haptic normal-mapped surfaces improves the delineation of synthetic tumours compared to outlining on a haptically 'flat' surface, when performed by participants with a non-medical background. Additionally, there are novel implementations of several methodologies; performing detailed calibration of partially obscured objects with Vicon motion tracking and vector-objects, performing perceptual texture comparisons using haptics as well as visual similarity, and using a texture generation toolbox to generate quasi-realistic medical image backgrounds for controlled ground-truth stimuli.

Appendix A

Photos of rigging

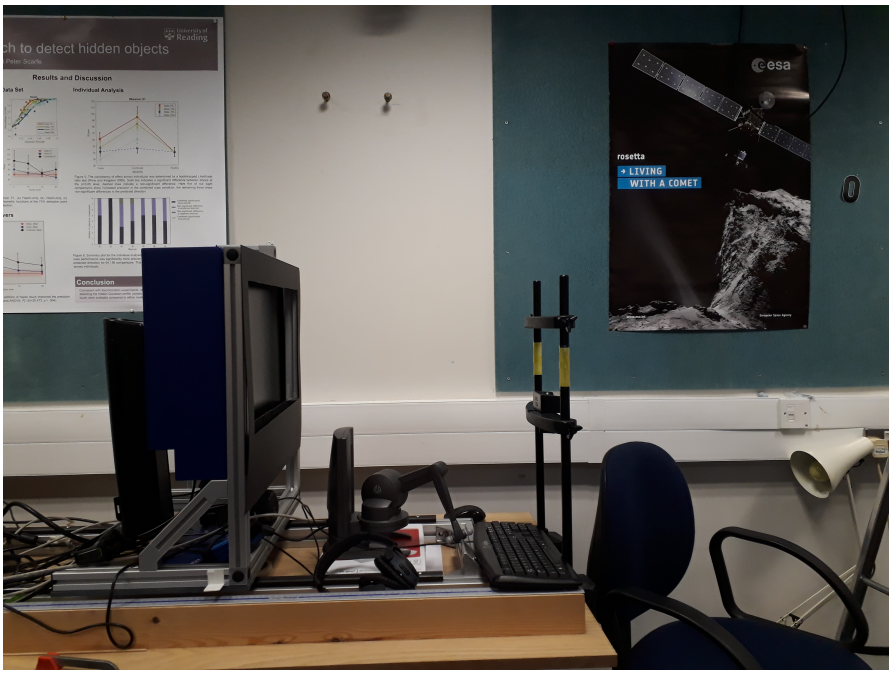


Figure A.1: Photo of the spatially misaligned experimental rig. The participant is placed in the chinrest, regulated to keep all participants' field of vision height consistent. The haptic device is placed at a comfortable distance from the participant, at 31.5cm from the base of the chinrest.

Figure A.2: Photo of the spatially coaligned experimental rig. The participant is placed in the chinrest, regulated to keep all participants' field of vision height consistent and angling the head to match the viewing angle. The haptic device is placed at a comfortable distance from the participant, underneath a front-silvered mirror.

Figure A.3: Photo of the first iteration of the 'vector stick' used to calibrate the spatially coaligned rig. On one end is a Vicon 'object' named 'Card', on the other is a single unlabelled marker on a precision point. The single marker is used to calibrate the length and relative location of the vector stick base and point, and is removed for collecting the points on the rig.

Figure A.4: Photo of the second and final iteration of the 'vector stick'. On one end is a Vicon 'object' named 'Caterpillar', on the other is a single unlabelled marker on a precision point. The 'Caterpillar' object has more points than the 'Card' object and is therefore easier to accurately locate by the Vicon cameras. They are made of 3D printed plastic components with curved bases that fit more smoothly around the tripod, both of these features contribute to much greater rigidity compared with the cardboard 'Card' object, with a flat base. The precision point has been upgraded from fairly unrigid cocktail sticks to a solid metal 3.5 mm knitting needle, which again increases rigidity. The tip of the needle has been covered in a thin layer of medical tape to avoid scratching the surface of the monitor and the mirror.

Figure A.5: Close-up photo of the 'Caterpillar' Vicon object. It consists of three separate 3D printed plastic components originally fabricated for an unrelated experiment, repurposed for the use in our calibration procedure.

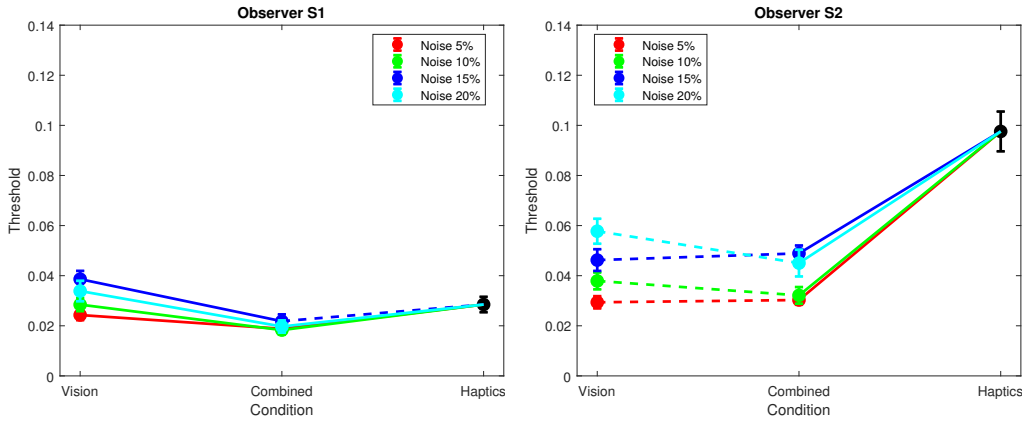
Figure A.6: Photo of the laser level setup used to visually confirm the calculated angle of the viewpoint. The laser level displays the angle on a digital screen and is able to emit a red laser cross out of one side. By placing the laser level at the calculated height and angle of the viewing vector, the position of the red laser cross could be used to illustrate where the hypothetical observer's gaze would hit the monitor. This was used to verify that the angle did point directly at the centre of the reflected monitor.

Appendix B

Individual observer graphs

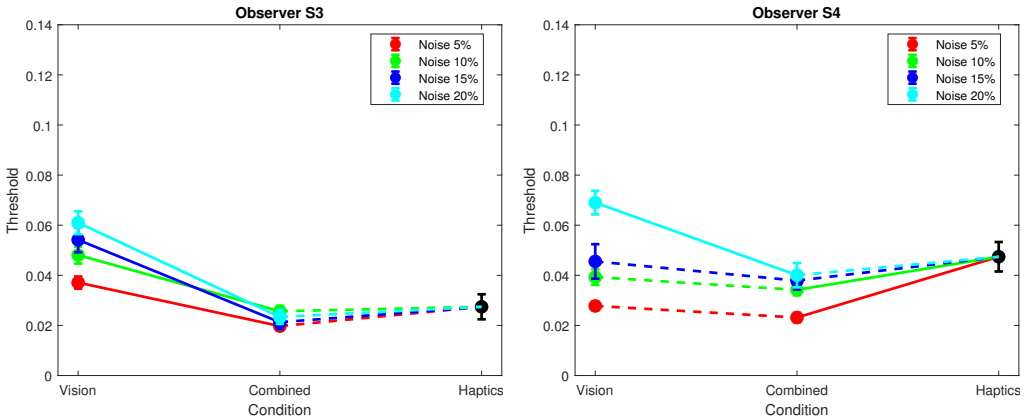
B.1 Experiment 1

B.1.1 Individual results, 1.0



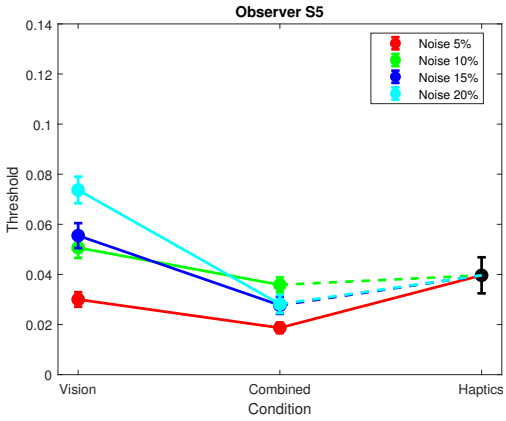
(a) S1 (b) S2

Figure B.1: Experiment 1.0 individual results pt 1, comparative slopes per condition, per participant. Error bars show the 95% confidence interval.

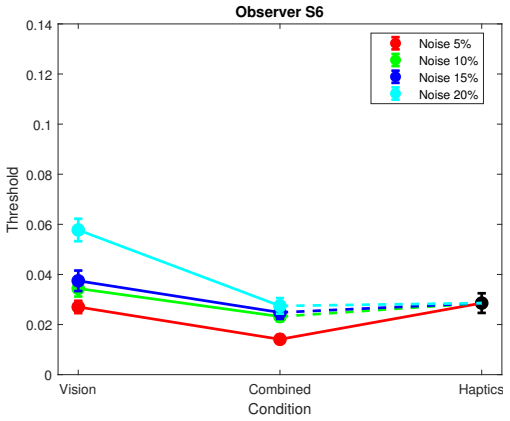


(c) S3

(d) S4

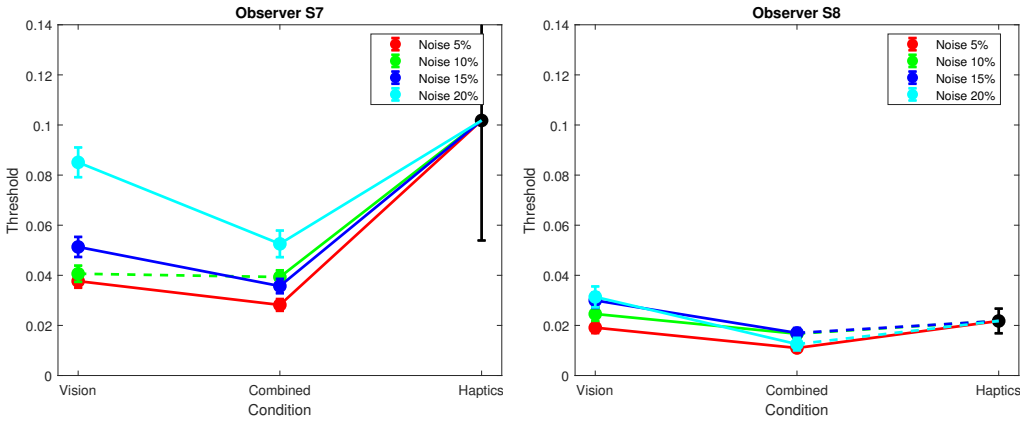


(e) S5



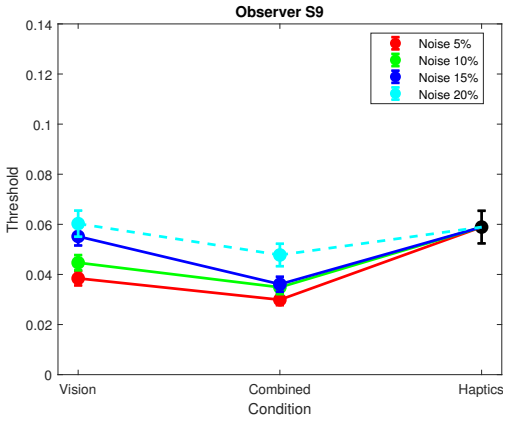
(f) S6

Figure B.1: Experiment 1.0 individual results pt 2, comparative slopes per condition, per participant. Error bars show the 95% confidence interval.

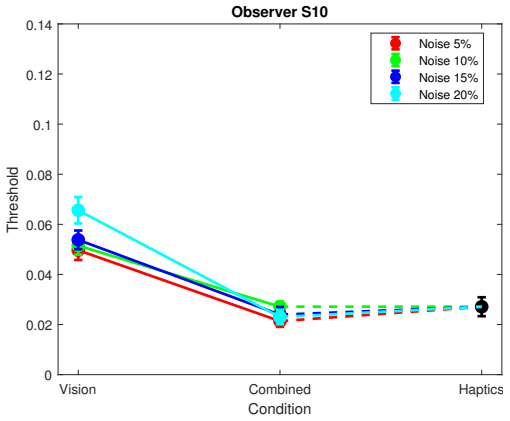


(g) S7

(h) S8



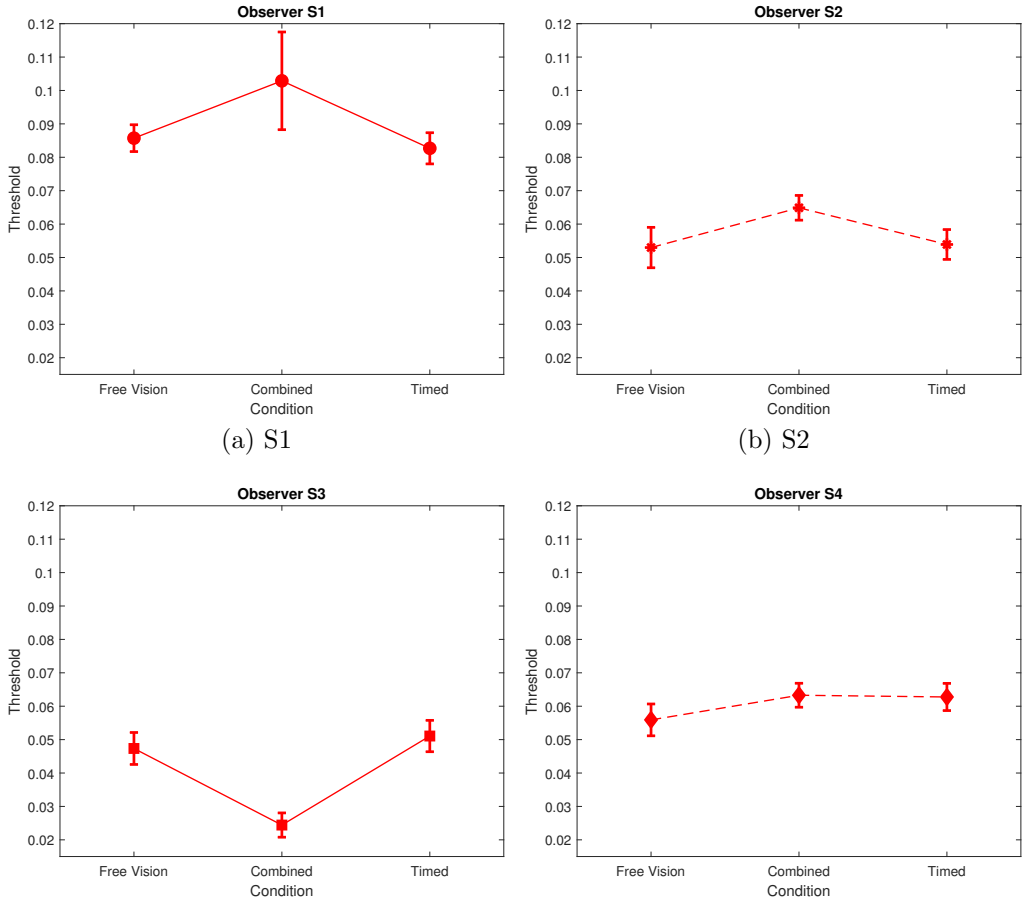
(i) S9



(j) S10

Figure B.1: Experiment 1.0 individual results pt 3, comparative slopes per condition, per participant. Error bars show the 95% confidence interval.

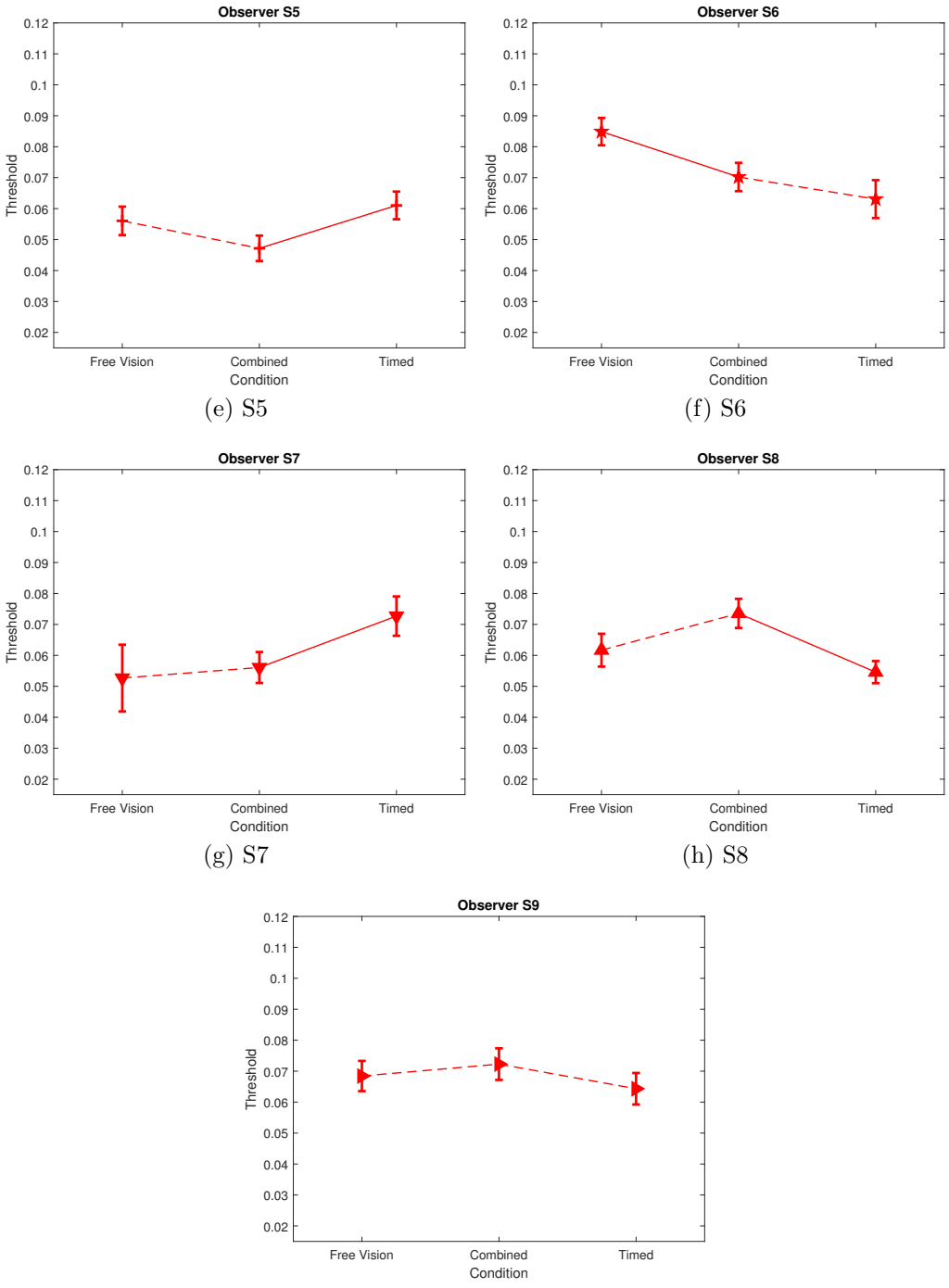
B.1.2 Individual results, Experiment 1.1



(a) S1 (b) S2

(c) S3 (d) S4

Figure B.2: Experiment 1.1 individual results pt 1, comparative slopes per condition, per participant, with error bars showing the 95% confidence interval.



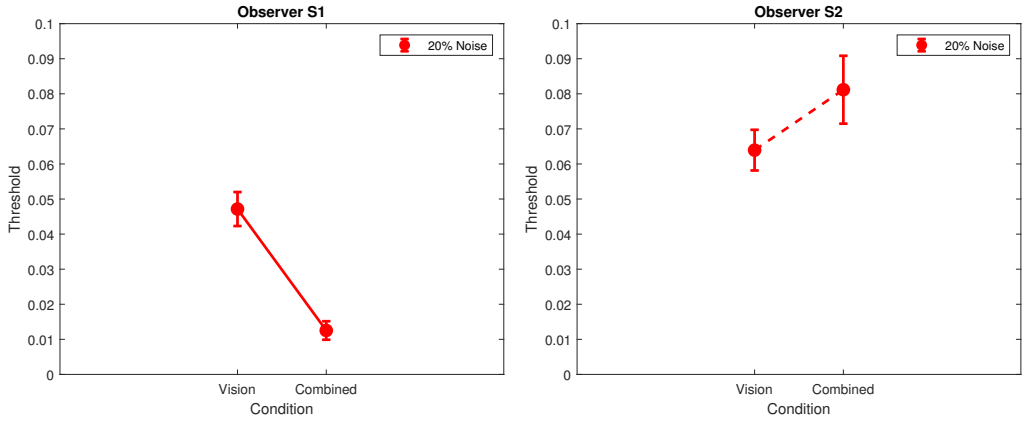
(e) S5 (f) S6

(g) S7 (h) S8

(i) S9

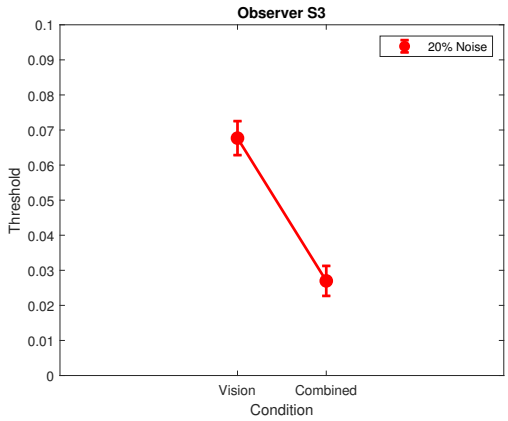
Figure B.2: Experiment 1.1 individual results pt 2, comparative slopes per condition, per participant, with error bars showing the 95% confidence interval.

B.1.3 Individual results, Experiment 1.2

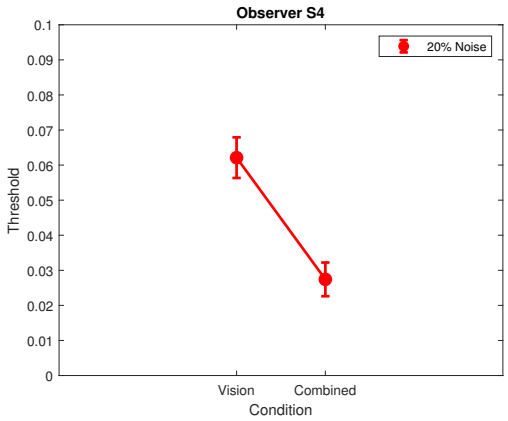


(a) S1

(b) S2



(c) S3



(d) S4

Figure B.3: Experiment 1.2 individual results pt 1, comparative slopes per modality, per participant, error bars showing the 95% condifence interval.

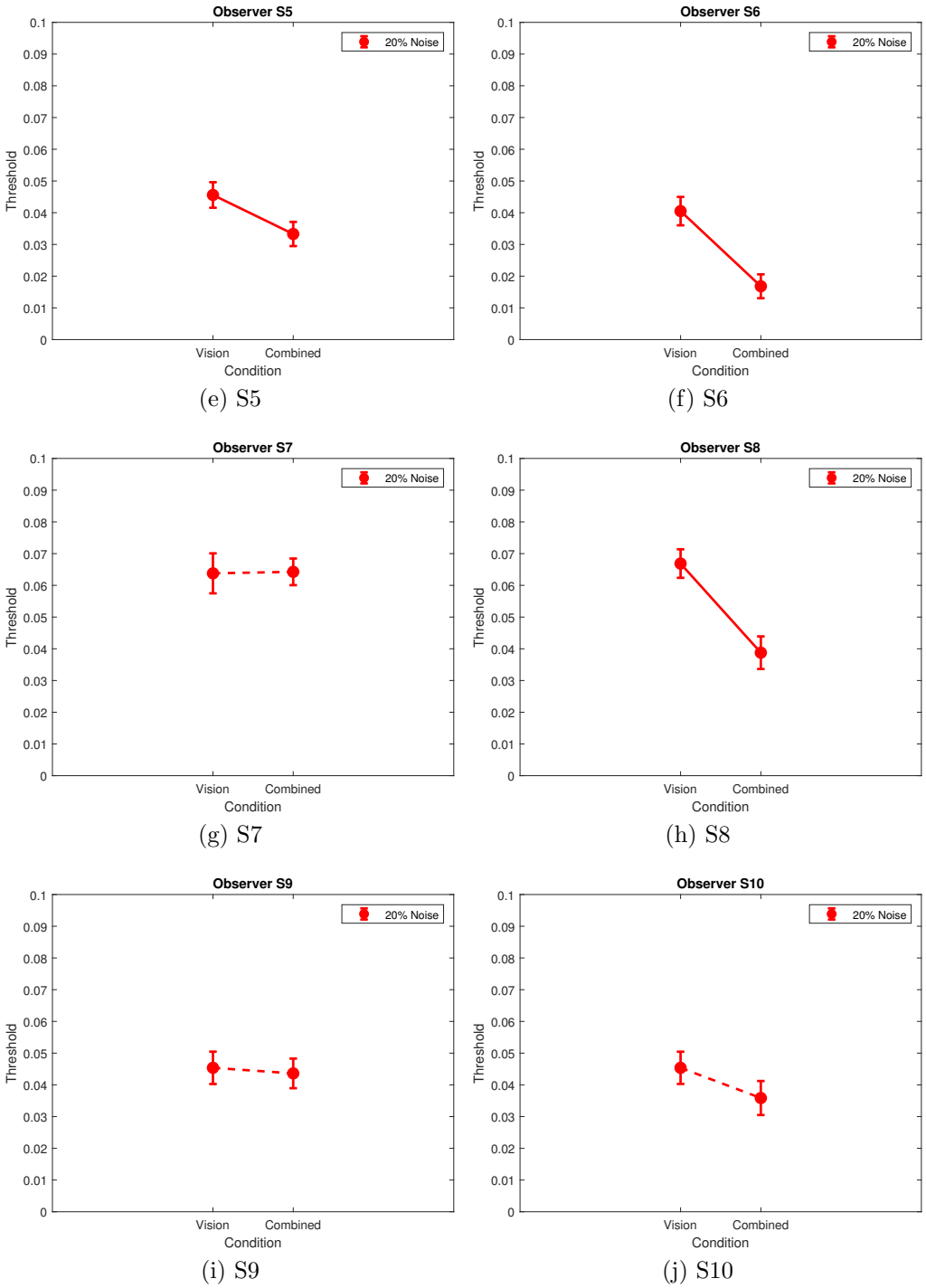
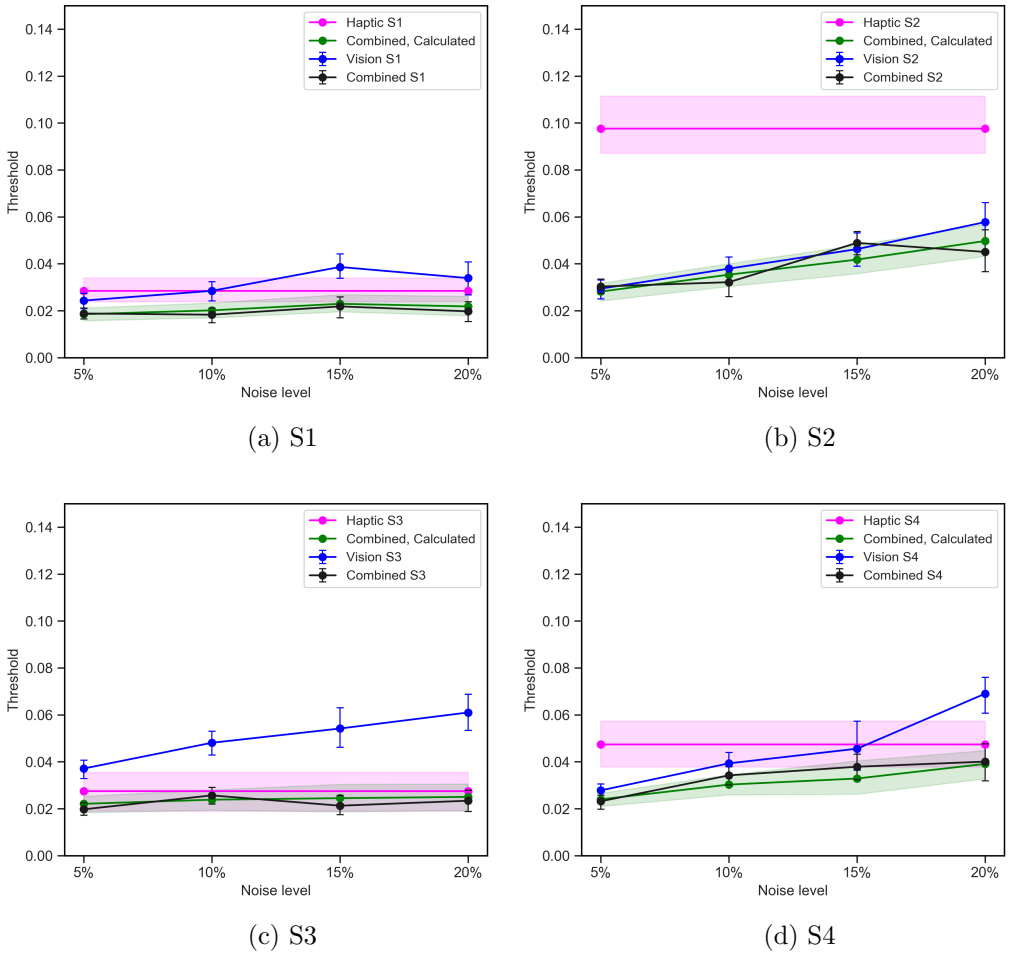


Figure B.3: Experiment 1.2 individual results pt 2, comparative slopes per modality, per participant, error bars showing the 95% condifence interval.

B.1.4 MLE calculations, Experiment 1.0



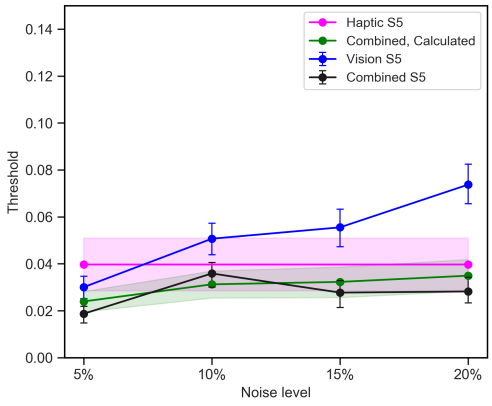
(a) S1

(b) S2

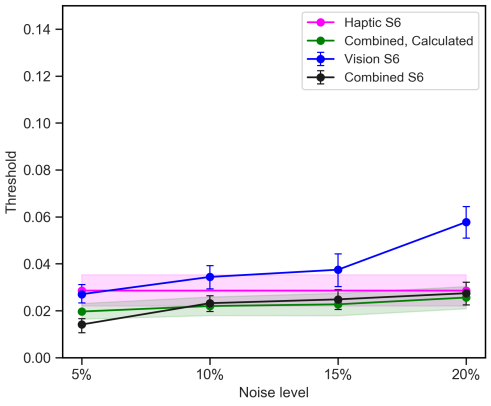
(c) S3

(d) S4

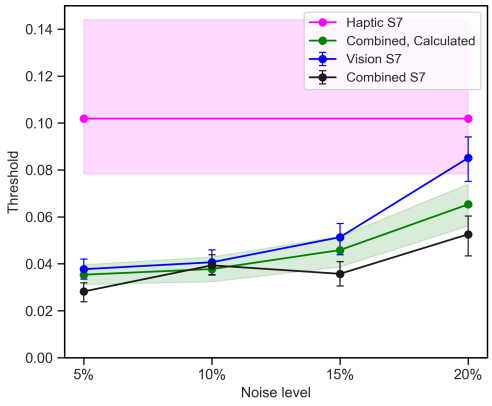
Figure B.4: Experiment 1.0 individual MLE predictions pt 1, with constrained y-axis for between-observer comparison.



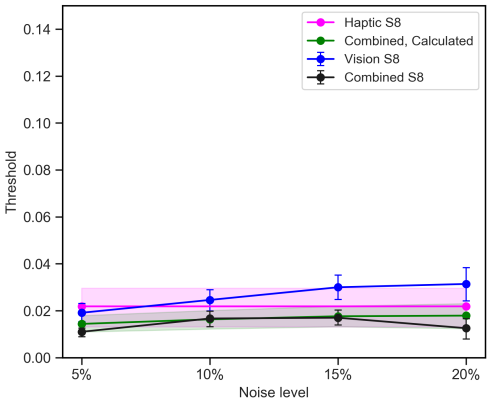
(e) S5



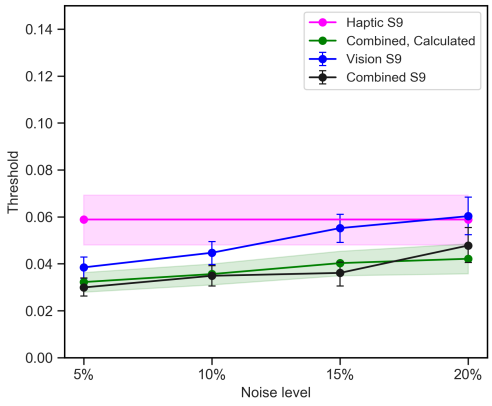
(f) S6



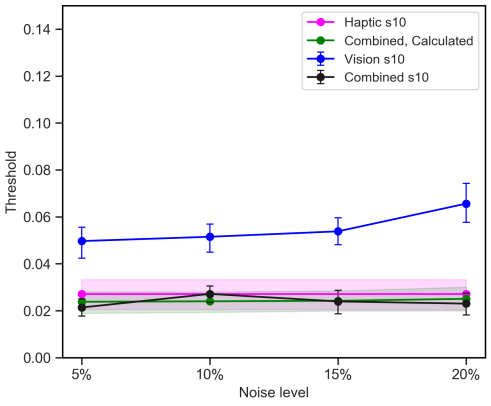
(g) S7



(h) S8



(i) S9

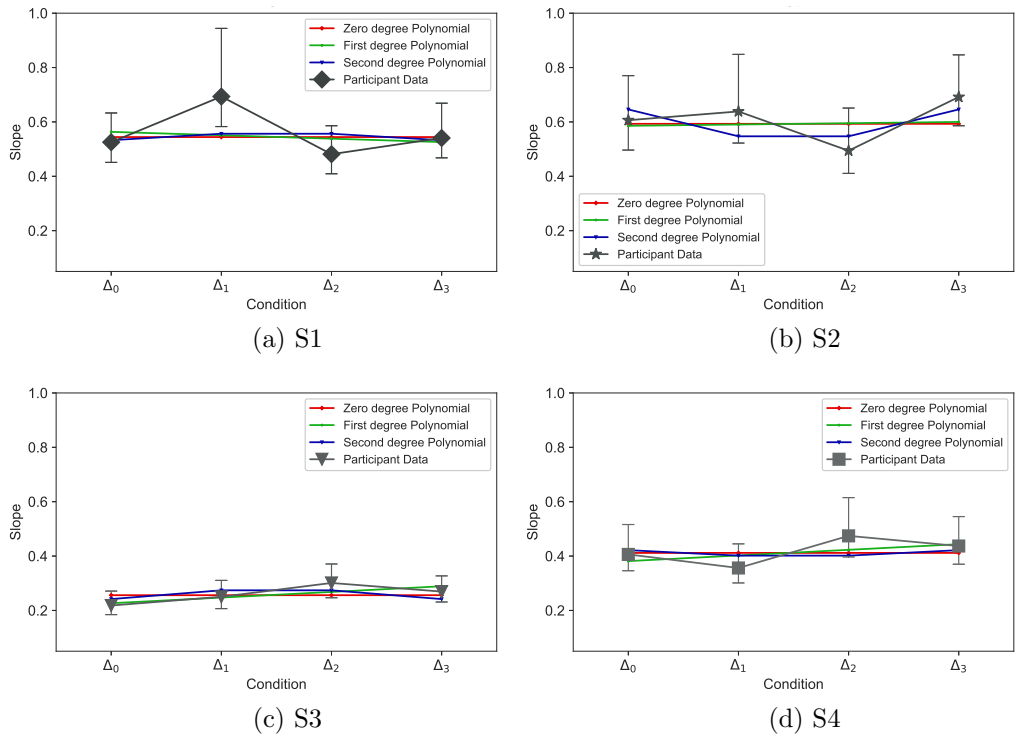


(j) S10

Figure B.4: Experiment 1.0 individual MLE predictions pt 2, with constrained y-axis for between-observer comparison.

B.2 Experiment 2

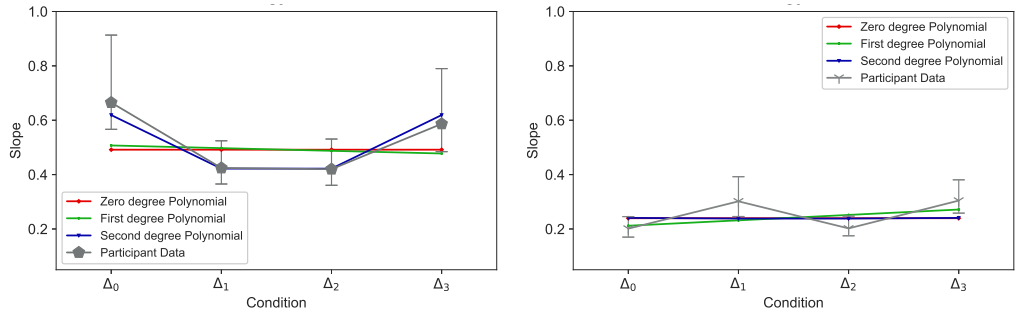
B.2.1 Individual results



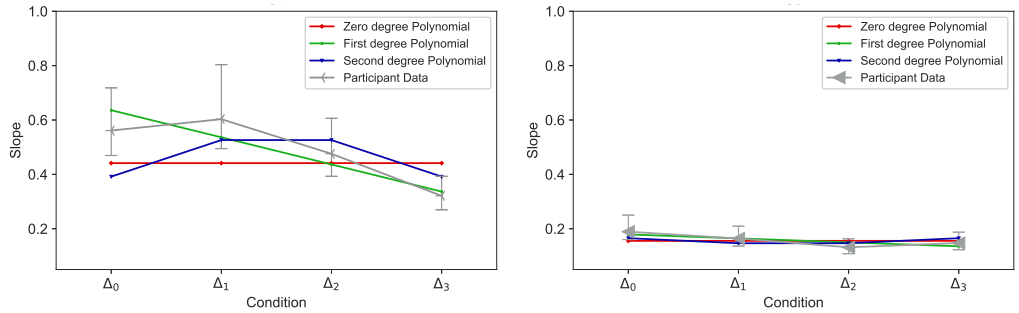
(a) S1 (b) S2

(c) S3 (d) S4

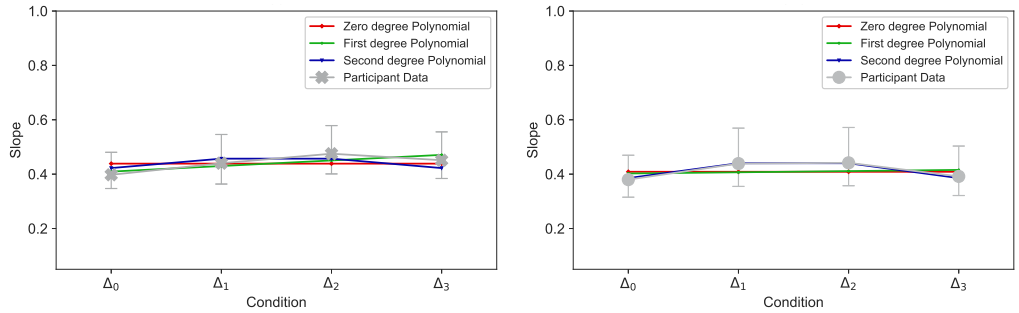
Figure B.5: Experiment 2 individual results pt 1, comparative slopes at all Δ-levels, for all participants, with error bars showing the 95% confidence interval.



(e) S5 (f) S6



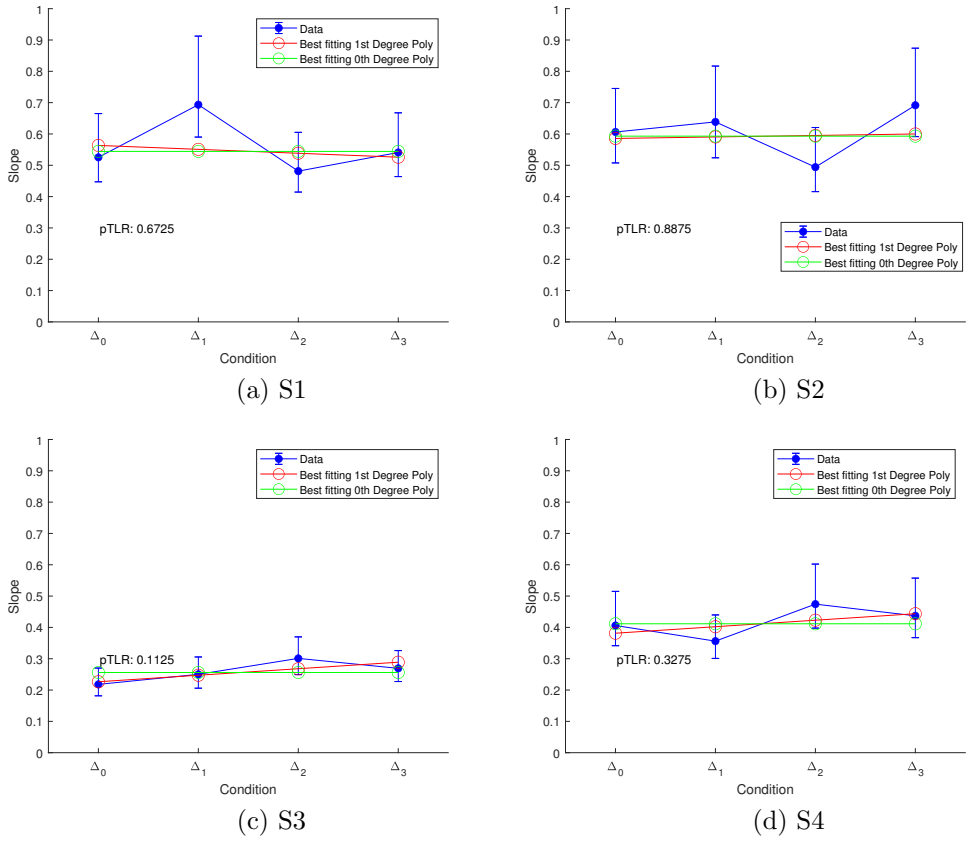
(g) S7 (h) S8



(i) S9 (j) S10

Figure B.5: Experiment 2 individual results pt 2, comparative slopes at all Δ-levels, for all participants, with error bars showing the 95% confidence interval.

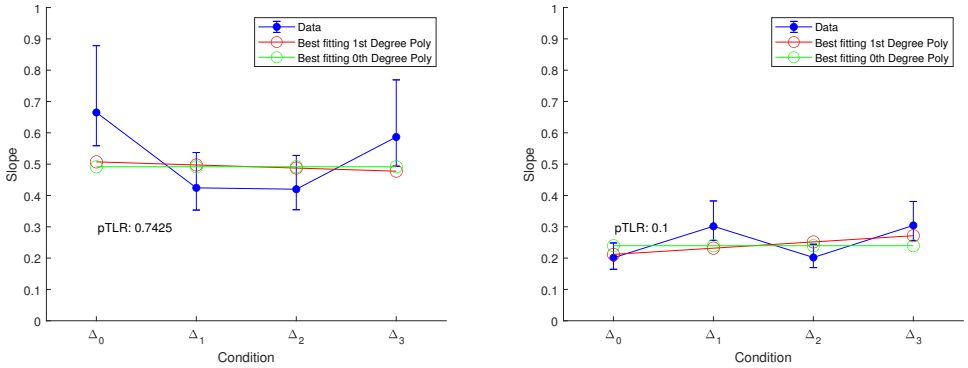
B.2.2 Best-fitting polynomial, all Δ-levels



(a) S1 (b) S2

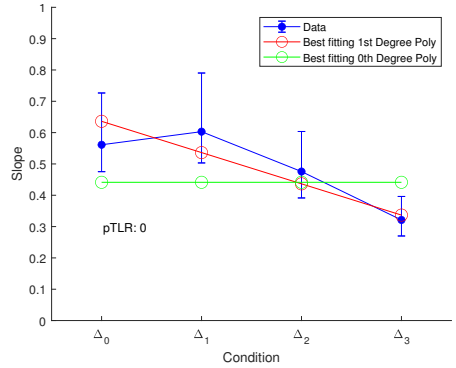
(c) S3 (d) S4

Figure B.6: Experiment 2 individual results pt 1, model comparison of polynomial fits to slope over dissimilarity Δ-level, for all participants, with error bars showing the 95% confidence interval. The pTLR value is the model comparison's equivalent of a p-value, where the first order polynomial is only considered to be a significently better fit to the data for a $pTLR < 0.05$, as is the case only in (g).

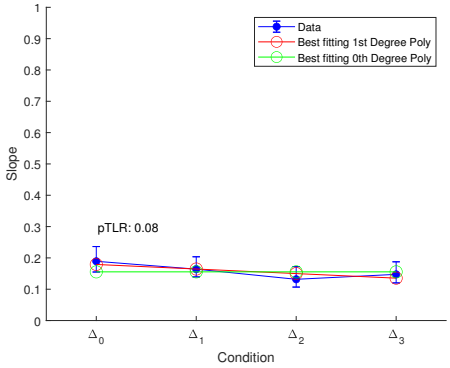


(e) S5

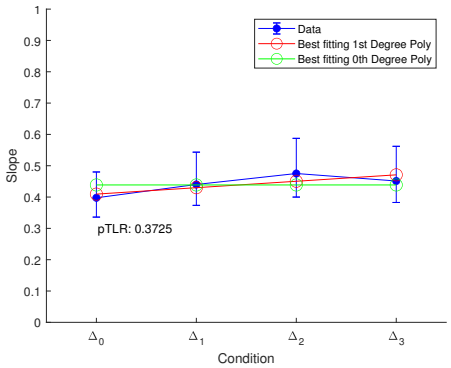
(f) S6



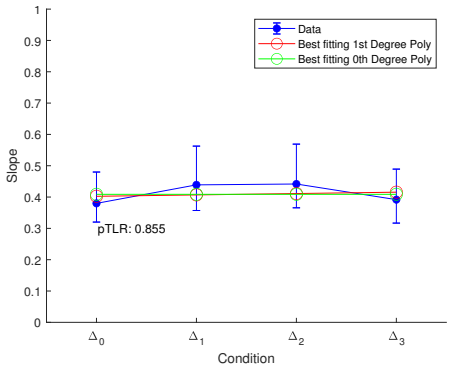
(g) S7



(h) S8



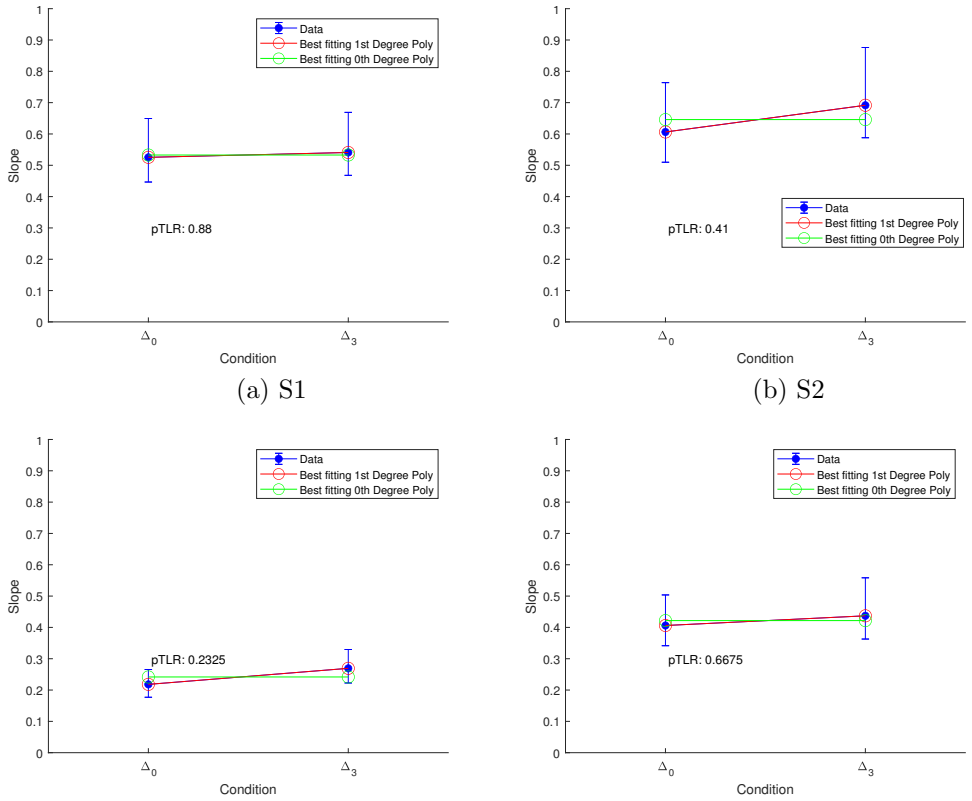
(i) S9



(j) S10

Figure B.6: Experiment 2 individual results pt 2, model comparison of polynomial fits to slope over dissimilarity Δ-level, for all participants, with error bars showing the 95% confidence interval. The pTLR value is the model comparison's equivalent of a p-value, where the first order polynomial is only considered to be a significently better fit to the data for a $pTLR < 0.05$, as is the case only in (g).

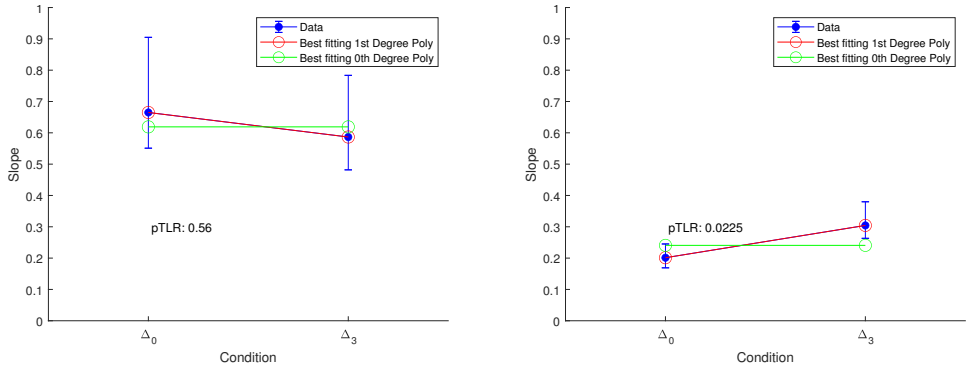
B.2.3 Best-fitting polynomial, Δ_0-Δ_3 only



(a) S1 (b) S2

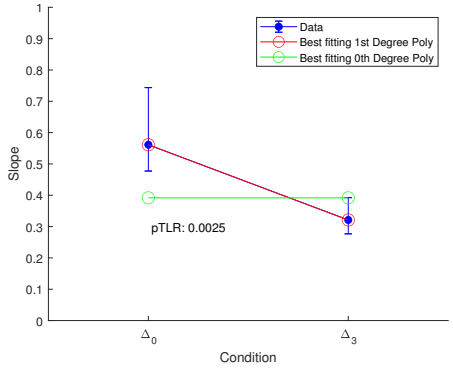
(c) S3 (d) S4

Figure B.7: Experiment 2 individual results pt 1, comparative slopes at Δ_0 and Δ_3, for all participants, error bars showing the 95% bootstrapped confidence interval. The $pTLR$ value is the model comparison's equivalent of a p-value, where the first order polynomial is only considered to be a significently better fit to the data for a $pTLR < 0.05$, as is the case in (f) and (g).

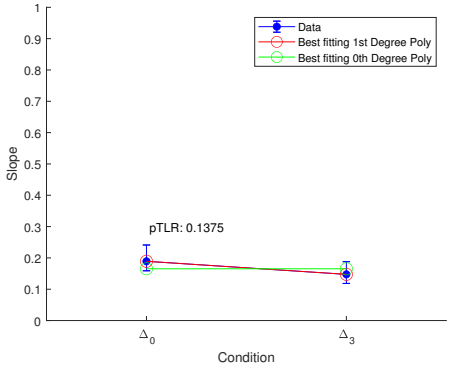


(e) S5

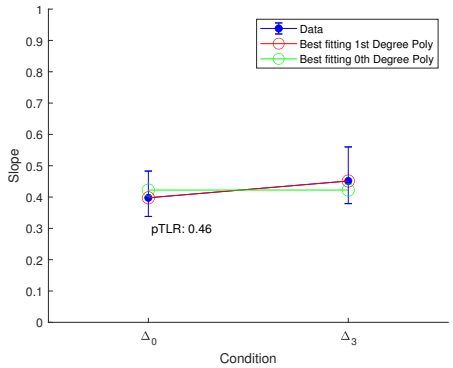
(f) S6



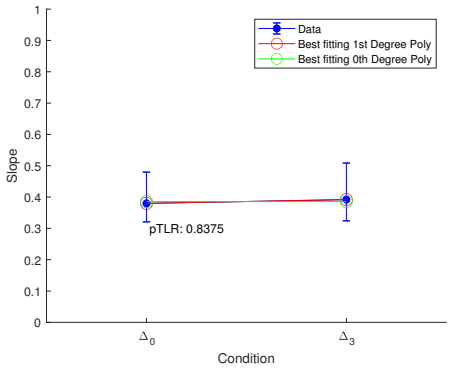
(g) S7



(h) S8



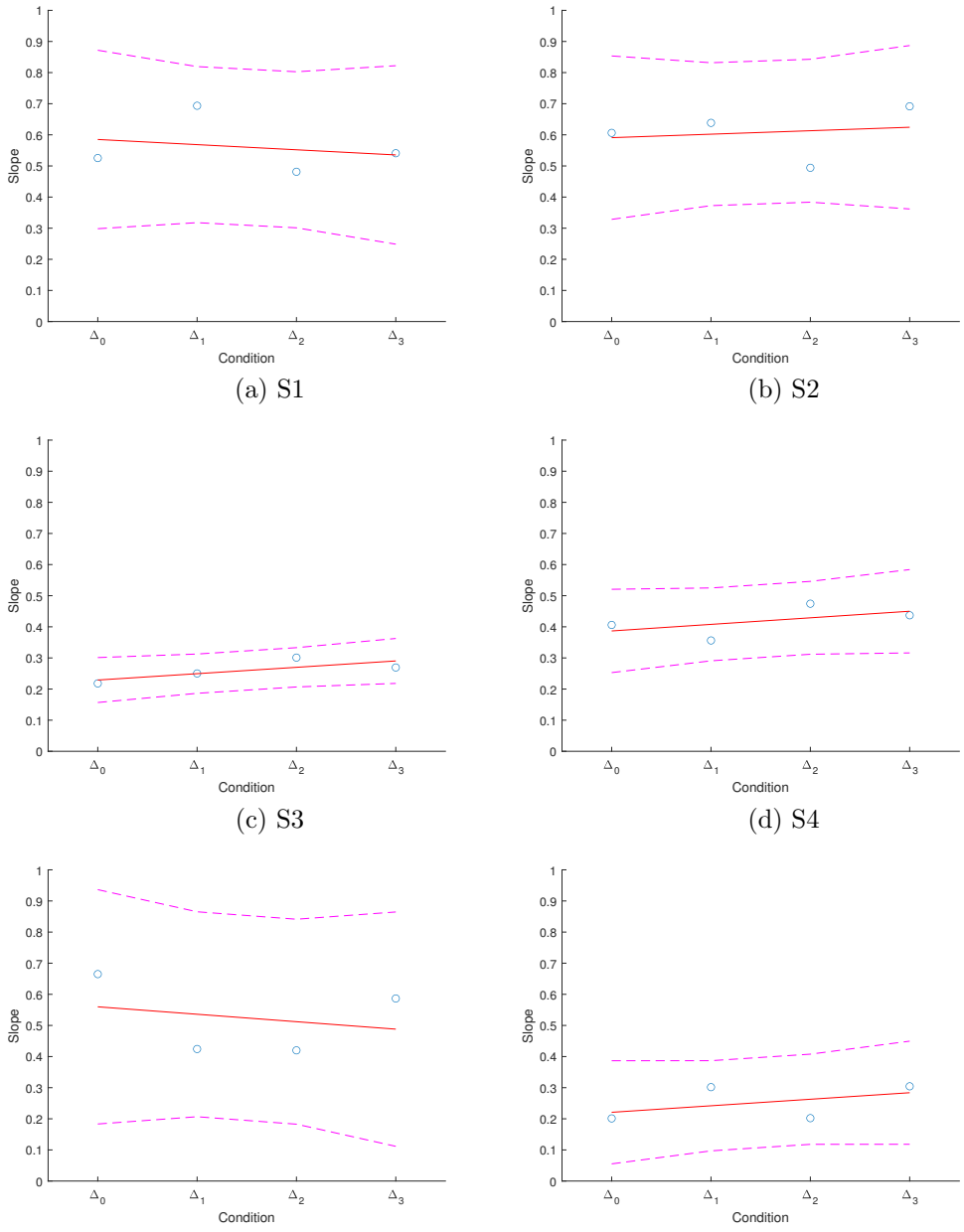
(i) S9



(j) S10

Figure B.7: Experiment 2 individual results pt 2, comparative slopes at Δ_0 and Δ_3, for all participants, error bars showing the 95% bootstrapped confidence interval. The pTLR value is the model comparison's equivalent of a p-value, where the first order polynomial is only considered to be a significently better fit to the data for a $pTLR < 0.05$, as is the case in (f) and (g).

B.2.4 Linear regression

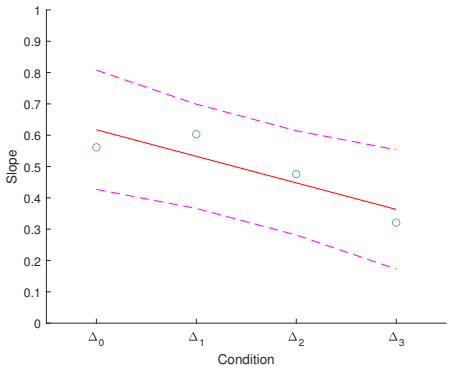


(a) S1 (b) S2

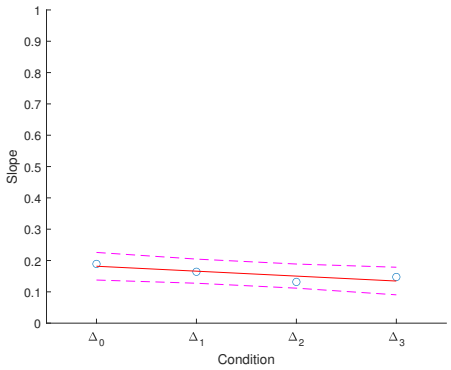
(c) S3 (d) S4

(e) S5 (f) S6

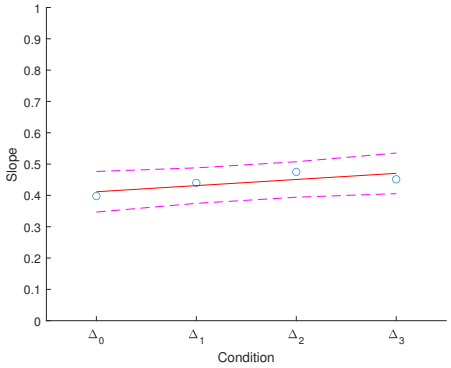
Figure B.8: Experiment 2 individual results and mean, linear regression of slope and dissimilarity for all participants, pt 1. Error regions show the 95% prediction interval.



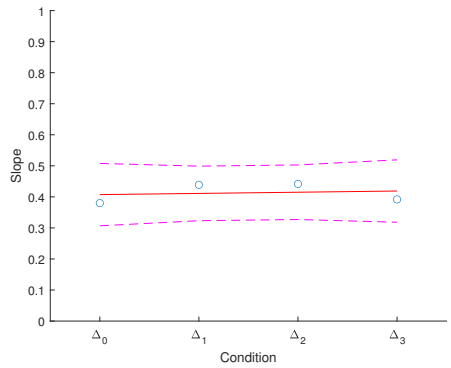
(g) S7



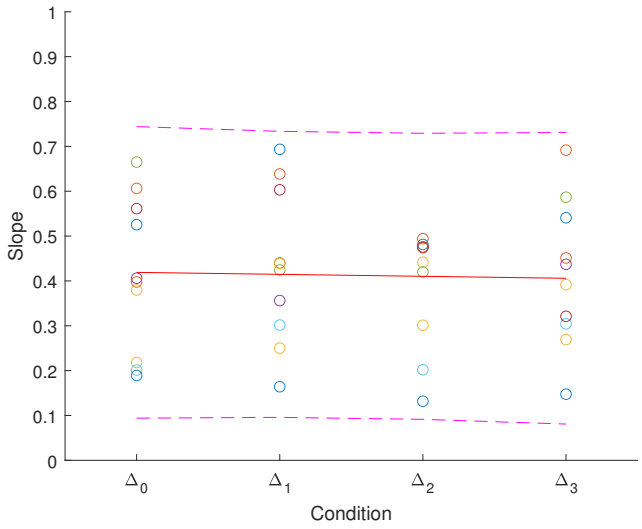
(h) S8



(i) S9



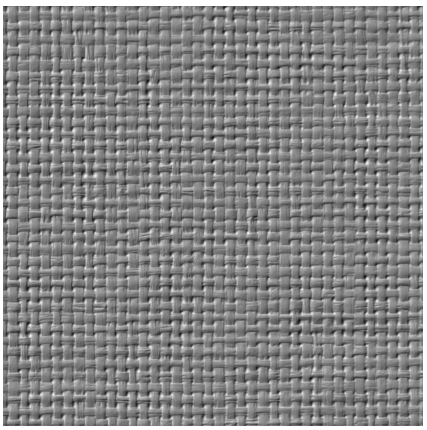
(j) S10



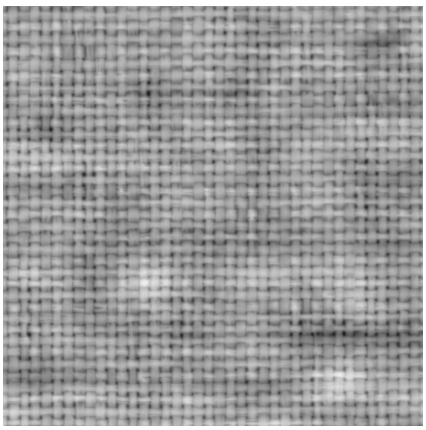
(k) Mean

Figure B.8: Experiment 2 individual results and mean, linear regression of slope and dissimilarity for all participants, pt 2. Error regions show the 95% prediction interval.

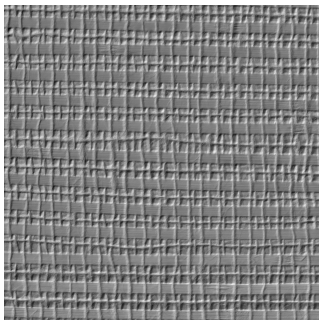
B.2.5 Performance and texture pairs, per observer



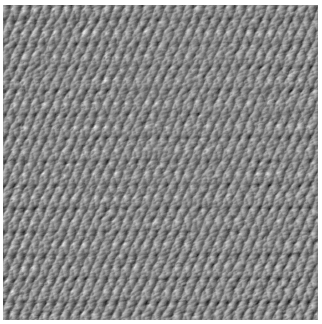
(a) Δ_0, Rendered texture 168



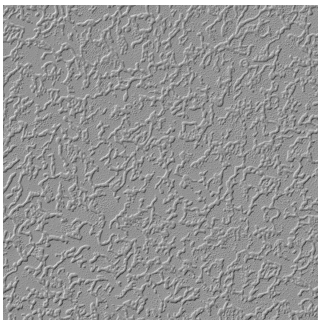
(b) Δ_0, Height-mapped texture 168



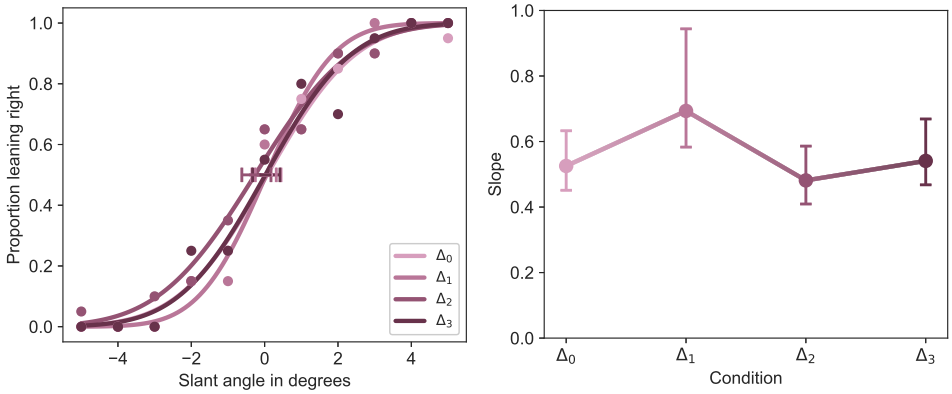
(c) Δ_1, Rendered texture 177



(d) Δ_2, Rendered texture 88

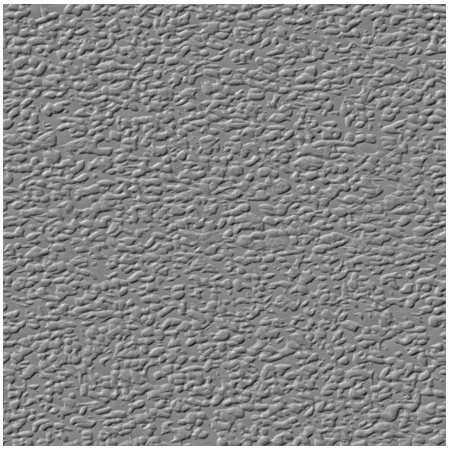


(e) Δ_3, Rendered texture 313

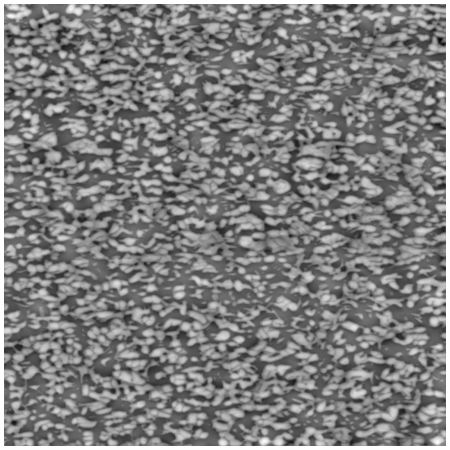


(f) Performance S1

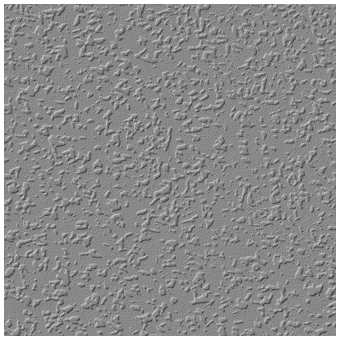
Figure B.9: Observer S1 was good regardless of texture pair, with a slight improvement for texture pair Δ_1 (Visual 177, haptic 168). For observer S1, there was no notable difference between Δ_0 and Δ_3, even though the former has more clear horizontal geometric structure while the latter is much more abstract. They had remarkably low bias and stayed centred around 0.

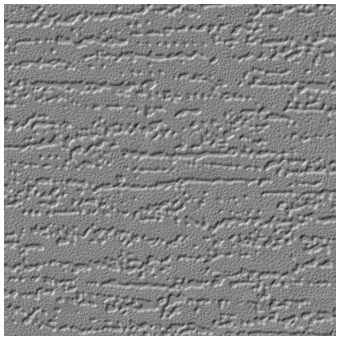


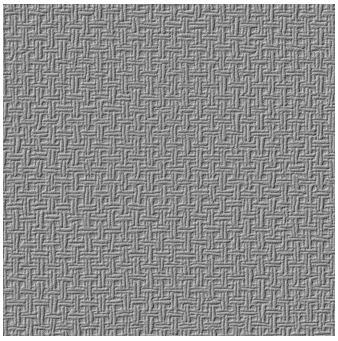
(a) Δ_0, Rendered texture 41



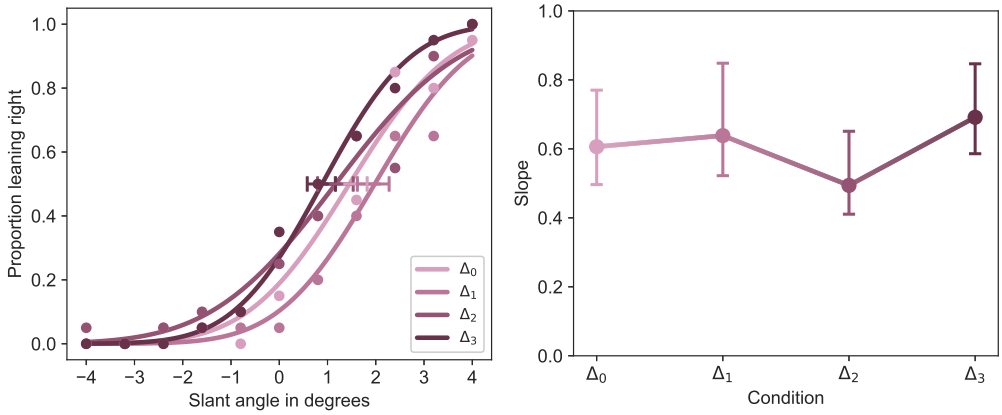
(b) Δ_0, Height-mapped texture 41





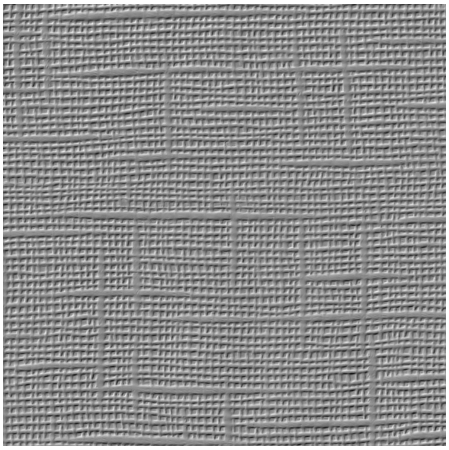


(c) Δ_1, Rendered texture 239 (d) Δ_2, Rendered texture 123 (e) Δ_3, Rendered texture 220

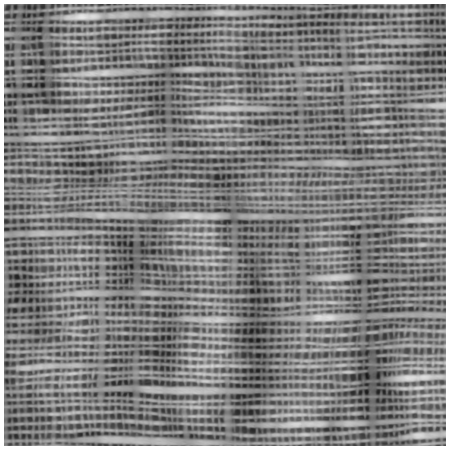


(f) Performance S2

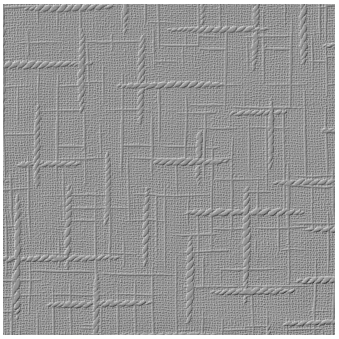
Figure B.10: For observer S2, one would have expected the slight increase in geometric linearity in Δ_2 to have a positive effect, given perspective cues and horizontal stripes, but on the contrary it was the only pair which performed worse than the other three Δ-pairs. (Visual 123, haptic 41). They had a notable bias to the right.



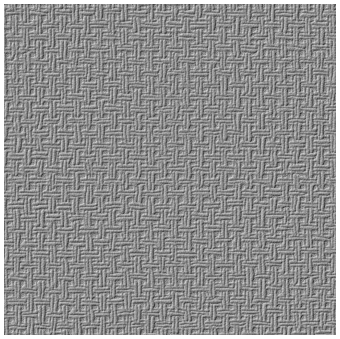
(a) Δ_0, Rendered texture 191



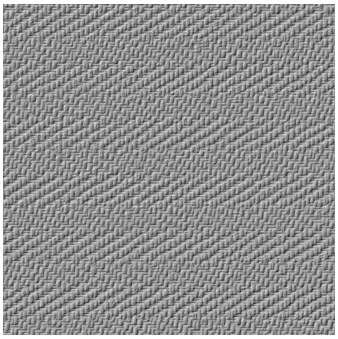
(b) Δ_0, Height-mapped texture 191



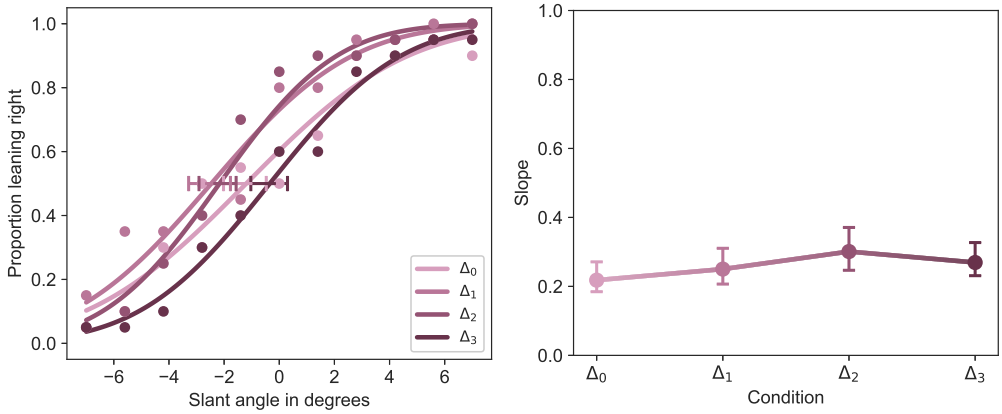
(c) Δ_1, Rendered texture 58



(d) Δ_2, Rendered texture 220

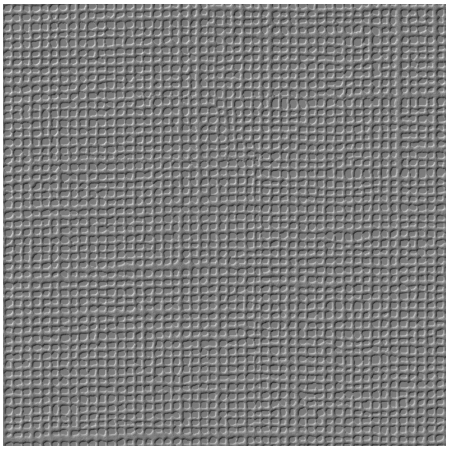


(e) Δ_3, Rendered texture 86

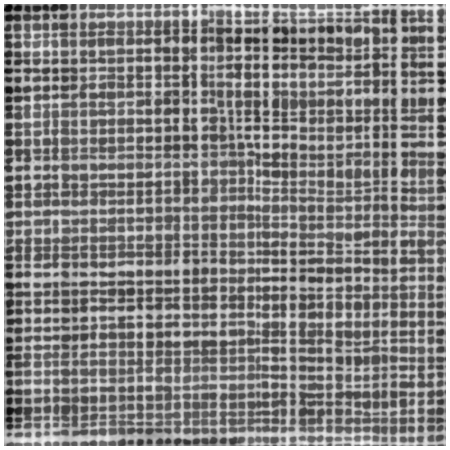


(f) Performance S3

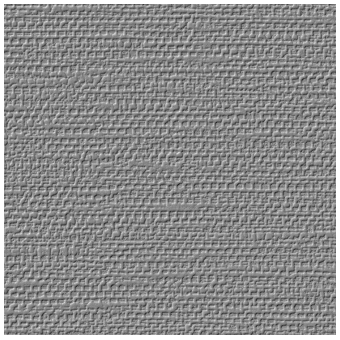
Figure B.11: For observer S3, all texture pairs have a notable geometric regularity, though the observer still performed relatively poorly, though at least this was very consistent across all four texture pairs. They had a notable bias to the left.



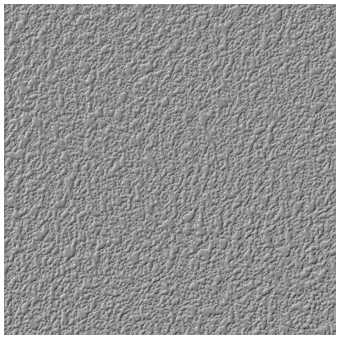
(a) Δ_0, Rendered texture 230



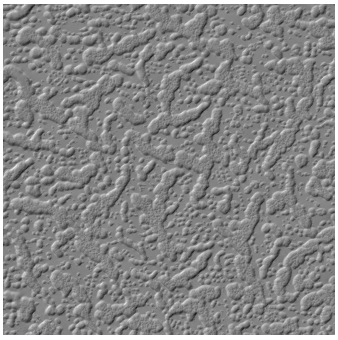
(b) Δ_0, Height-mapped texture 230



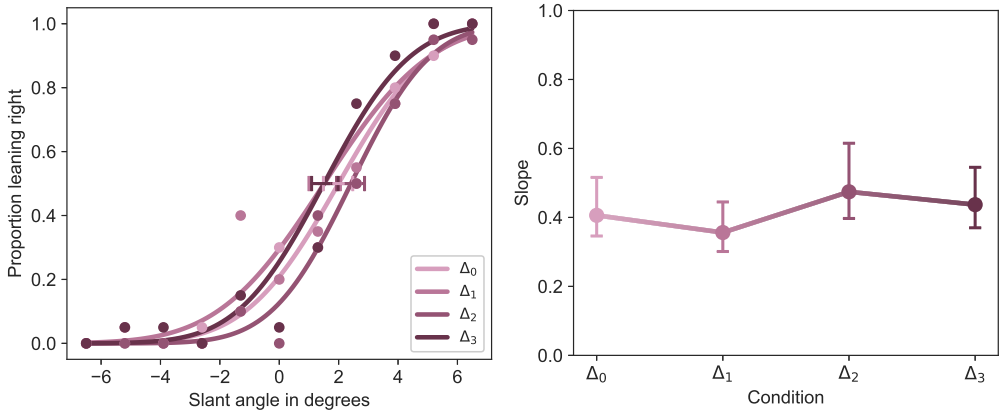
(c) Δ_1, Rendered texture 112



(d) Δ_2, Rendered texture 161

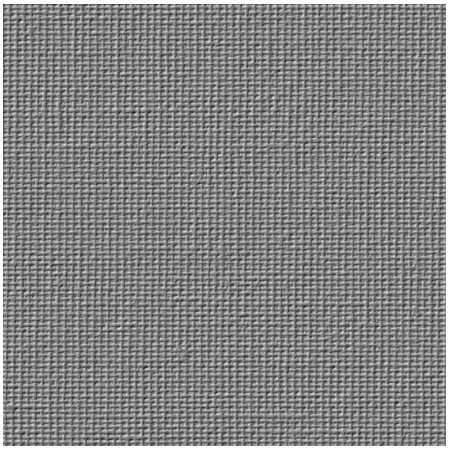


(e) Δ_3, Rendered texture 29

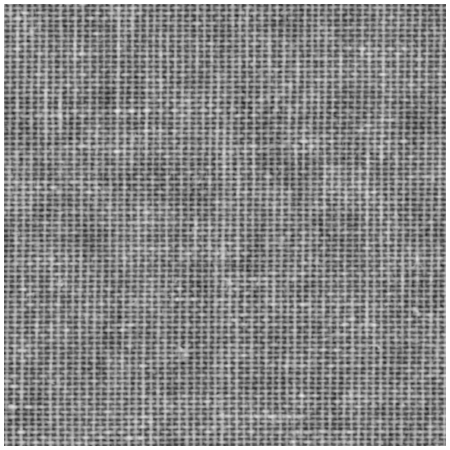


(f) Performance S4

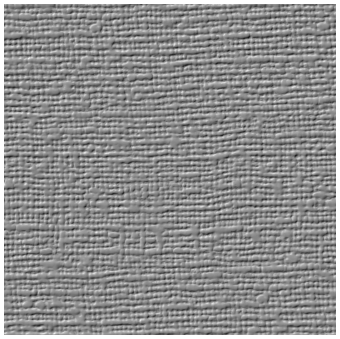
Figure B.12: For observer S4, there were notable geometric regularity in Δ_0 and Δ_1, while performance was fractionally better for Δ_2 and Δ_3. They had a notable bias to the right.



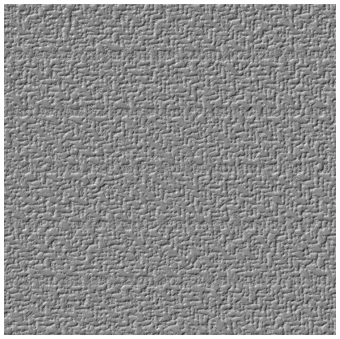
(a) Δ_0, Rendered texture 144



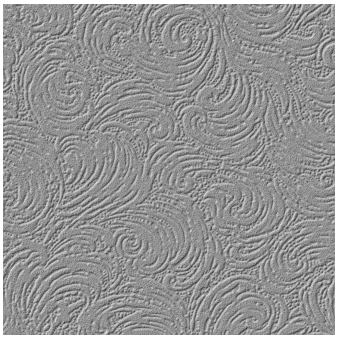
(b) Δ_0, Height-mapped texture 144



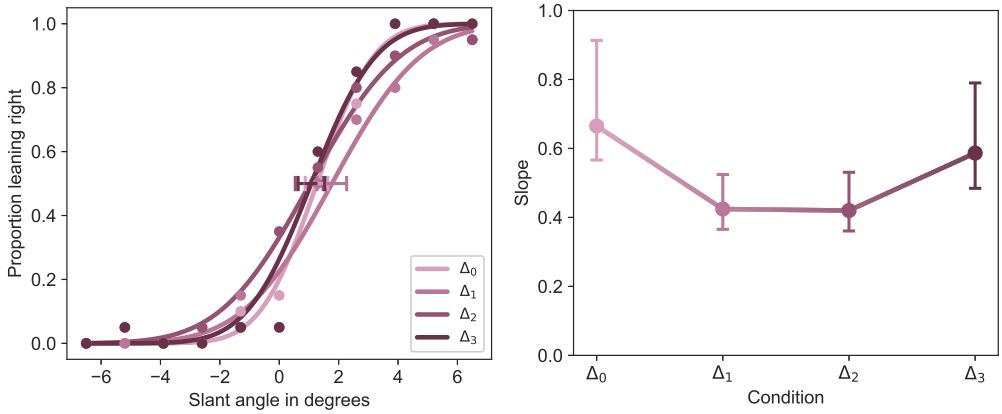
(c) Δ_1, Rendered texture 114



(d) Δ_2, Rendered texture 113

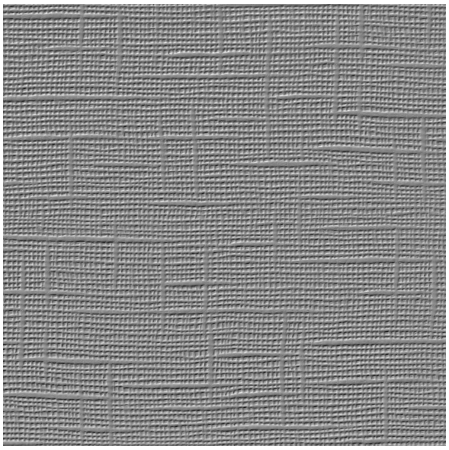


(e) Δ_3, Rendered texture 212

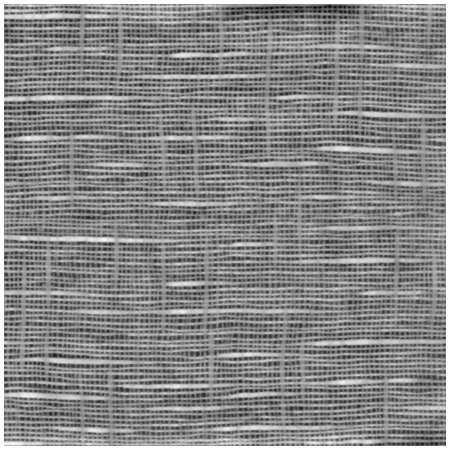


(f) Performance S5

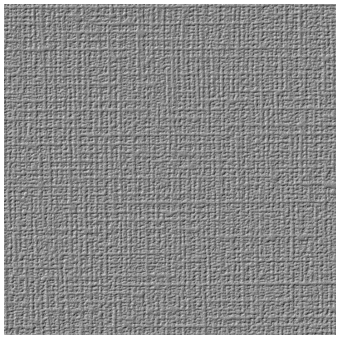
Figure B.13: For observer S5, texture pairs Δ_0 and Δ_1 showed high geometric regularity, while they performed better in Δ_0 and Δ_3, where Δ_3 had a notable circular geometry. They had a slight bias to the right.



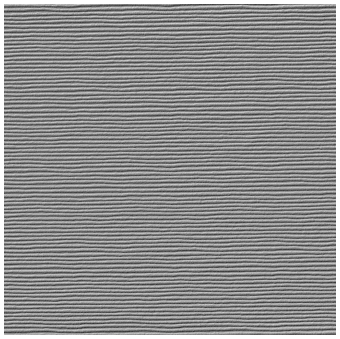
(a) Δ_0, Rendered texture 240



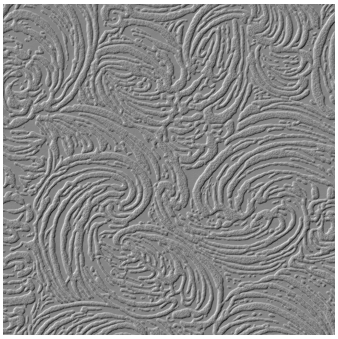
(b) Δ_0, Height-mapped texture 240



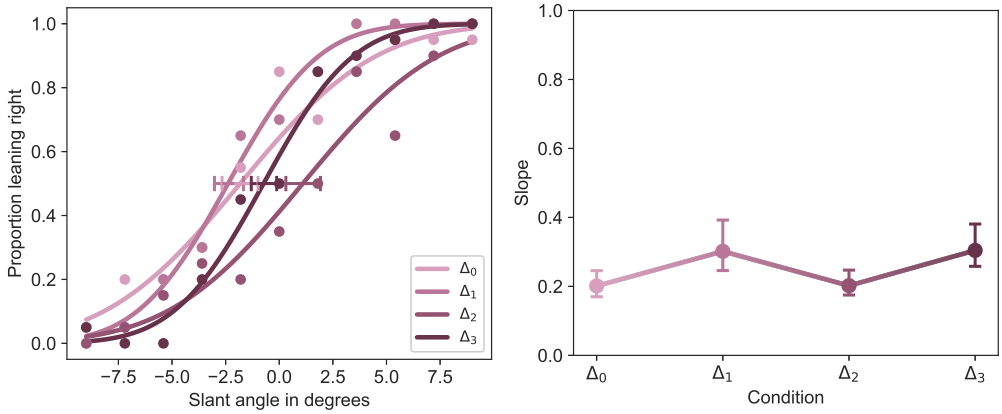
(c) Δ_1, Rendered texture 320



(d) Δ_2, Rendered texture 47

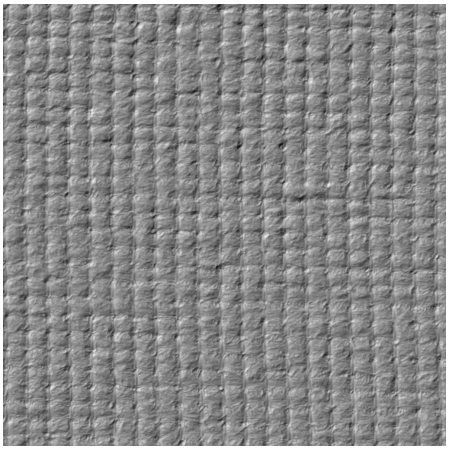


(e) Δ_3, Rendered texture 26

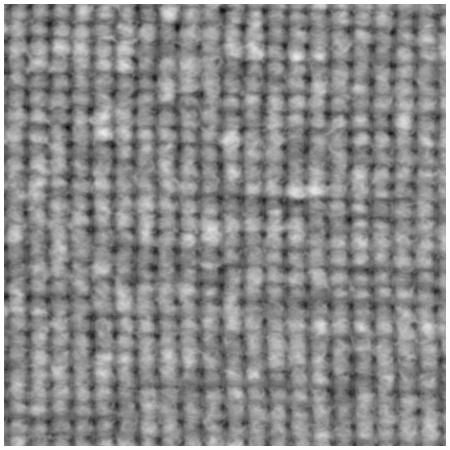


(f) Performance S6

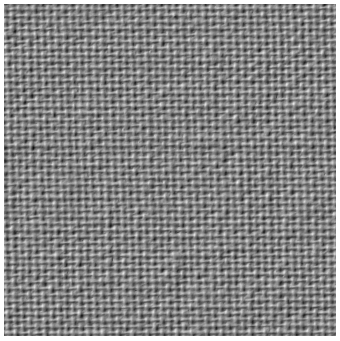
Figure B.14: For observer S6, texture pairs Δ_0 and Δ_1 had general geometrical regularity in both the horizontal and vertical regions, while Δ_2 had a strong horizontal features. Given the task, one would have expected an improved performance in texture pairs Δ_0 and Δ_2, while the results show a slight improvement in Δ_1 and Δ_3. They had a slight bias to the left.



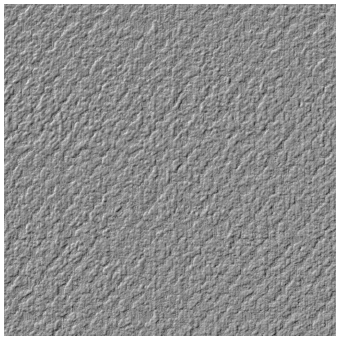
(a) Δ_0, Rendered texture 150



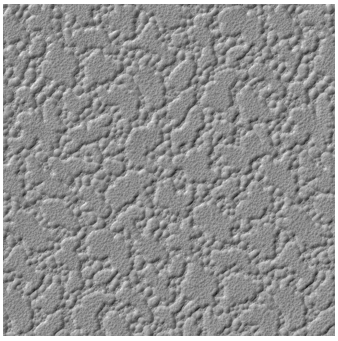
(b) Δ_0, Height-mapped texture 150



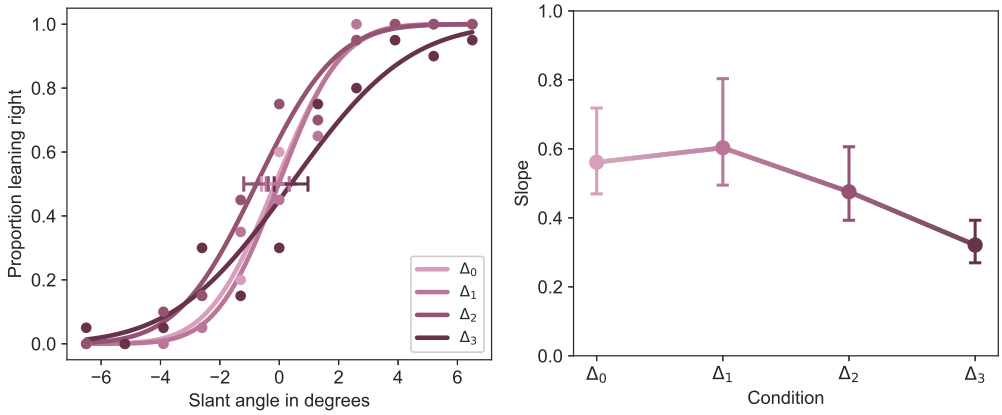
(c) Δ_1, Rendered texture 139



(d) Δ_2, Rendered texture 260

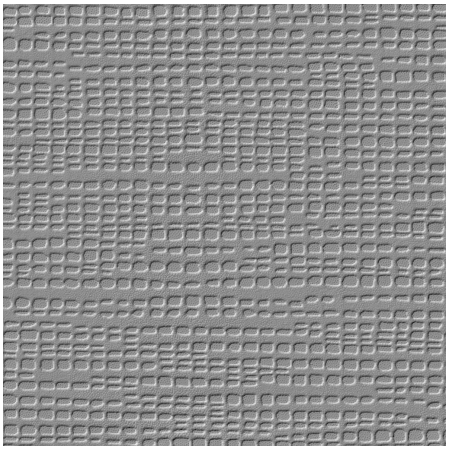


(e) Δ_3, Rendered texture 129

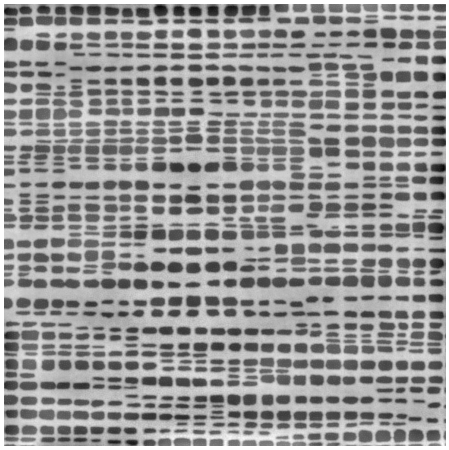


(f) Performance S7

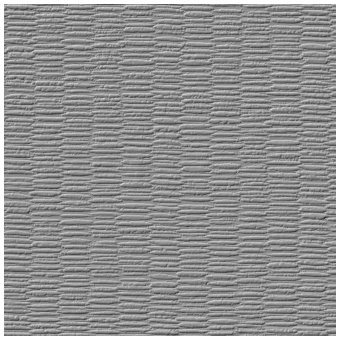
Figure B.15: For observer S7, texture pairs Δ_0 and Δ_1 had stronger geometric features in the horizontal and vertical direction, while Δ_2 and Δ_3 were more abstract and random. As expected, performance was best in Δ_0 and Δ_1, though Δ_0 is insignificantly better than Δ_2. Overall they did not show a strong bias in either direction.



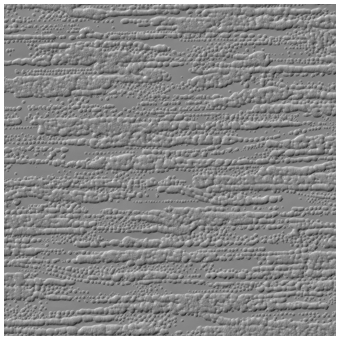
(a) Δ_0, Rendered texture 222



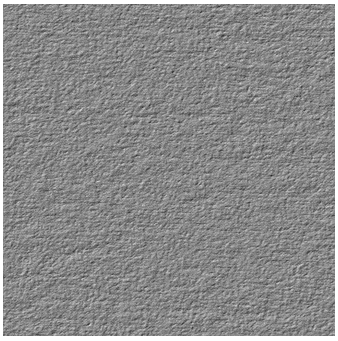
(b) Δ_0, Height-mapped texture 222



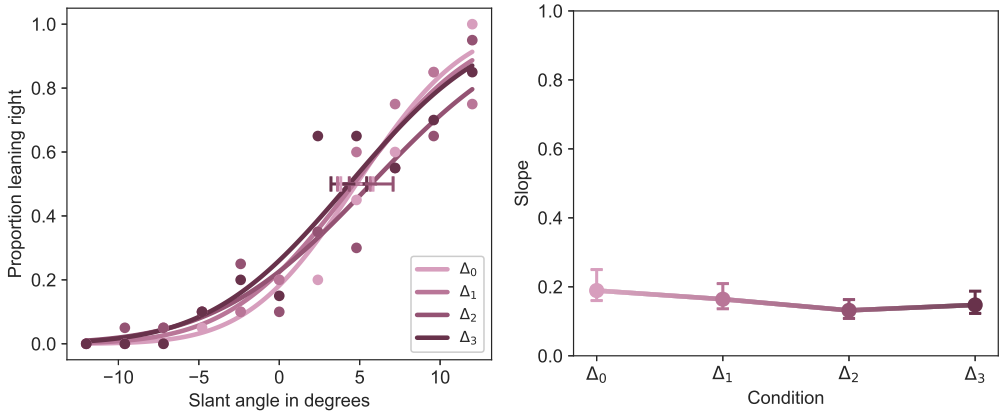
(c) Δ_1, Rendered texture 124



(d) Δ_2, Rendered texture 133

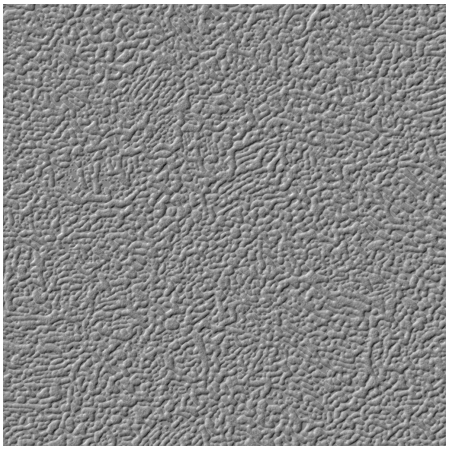


(e) Δ_3, Rendered texture 154

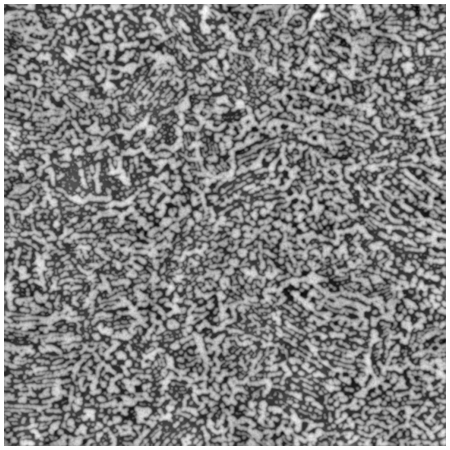


(f) Performance S8

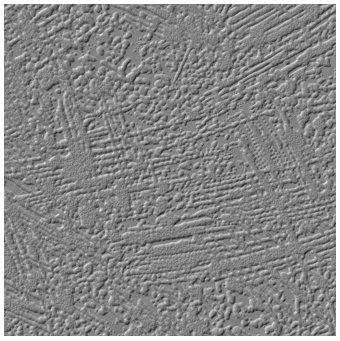
Figure B.16: For observer S8, performance was overall quite low. However, texture pairs Δ_0, Δ_1 and Δ_2 all had strong horizontal features and would be expected to give more visual perspective cues. Contrary to expectations, the observer did not differ between any of the texture pairs. They had a strong bias to the right.



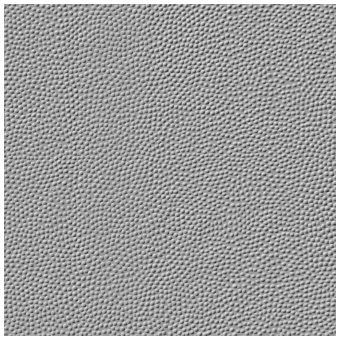
(a) Δ_0, Rendered texture 24



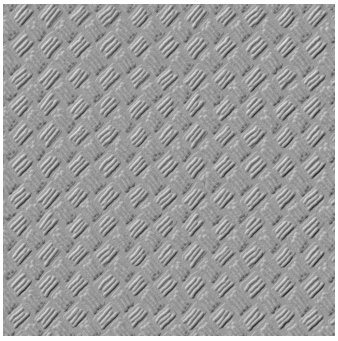
(b) Δ_0, Height-mapped texture 24



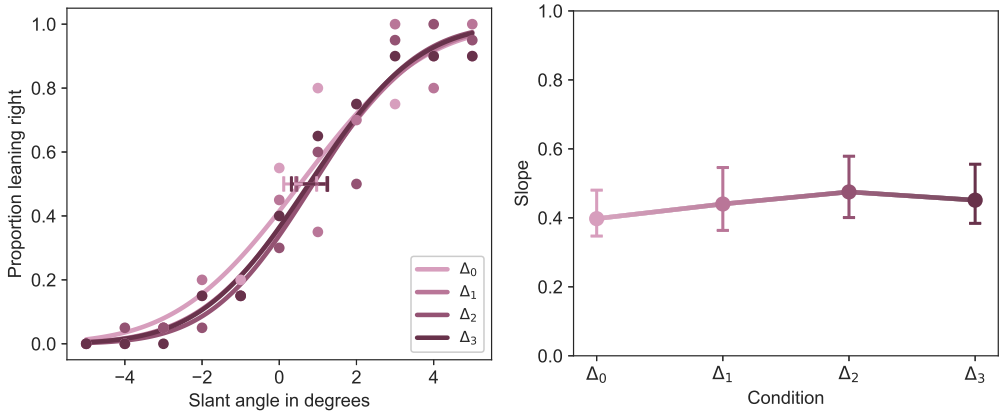
(c) Δ_1, Rendered texture 27



(d) Δ_2, Rendered texture 20

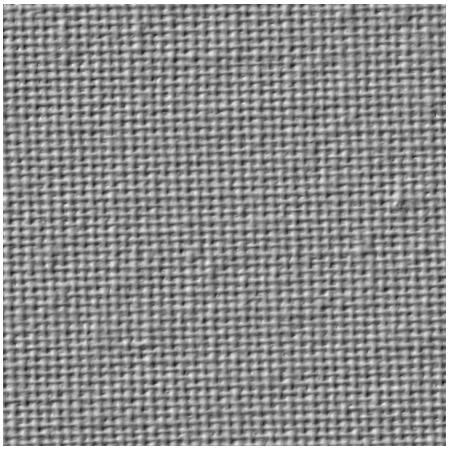


(e) Δ_3, Rendered texture 3

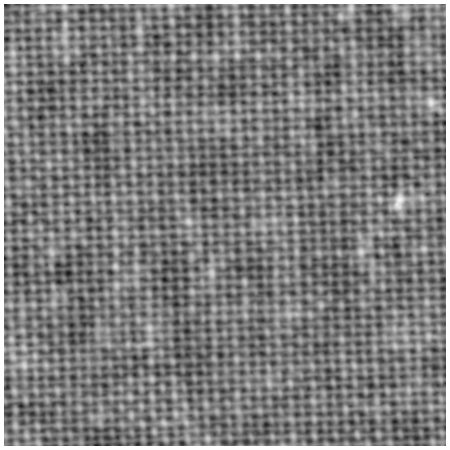


(f) Performance S9

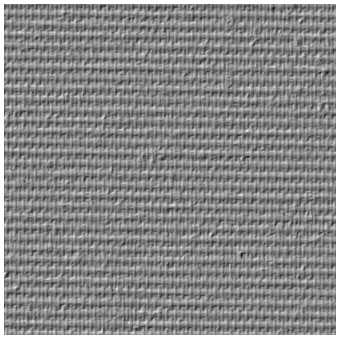
Figure B.17: For observer S9, performance was overall quite good. None of the texture pairs had strong vertical or horizontal statistical features, and there was no notable difference in performance across the texture pairs. The observer was slightly biased to the right.



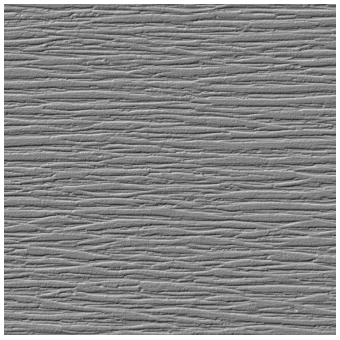
(a) Δ_0, Rendered texture 200



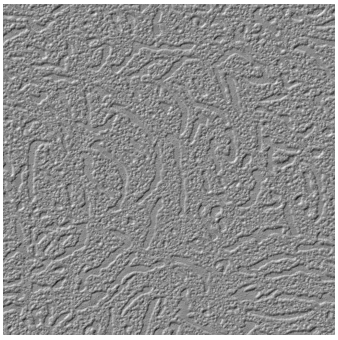
(b) Δ_0, Height-mapped texture 200



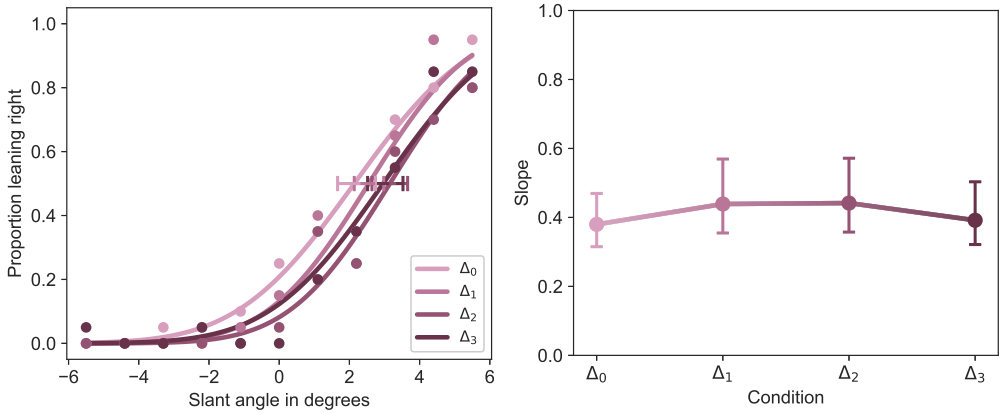
(c) Δ_1, Rendered texture 155



(d) Δ_2, Rendered texture 2



(e) Δ_3, Rendered texture 18

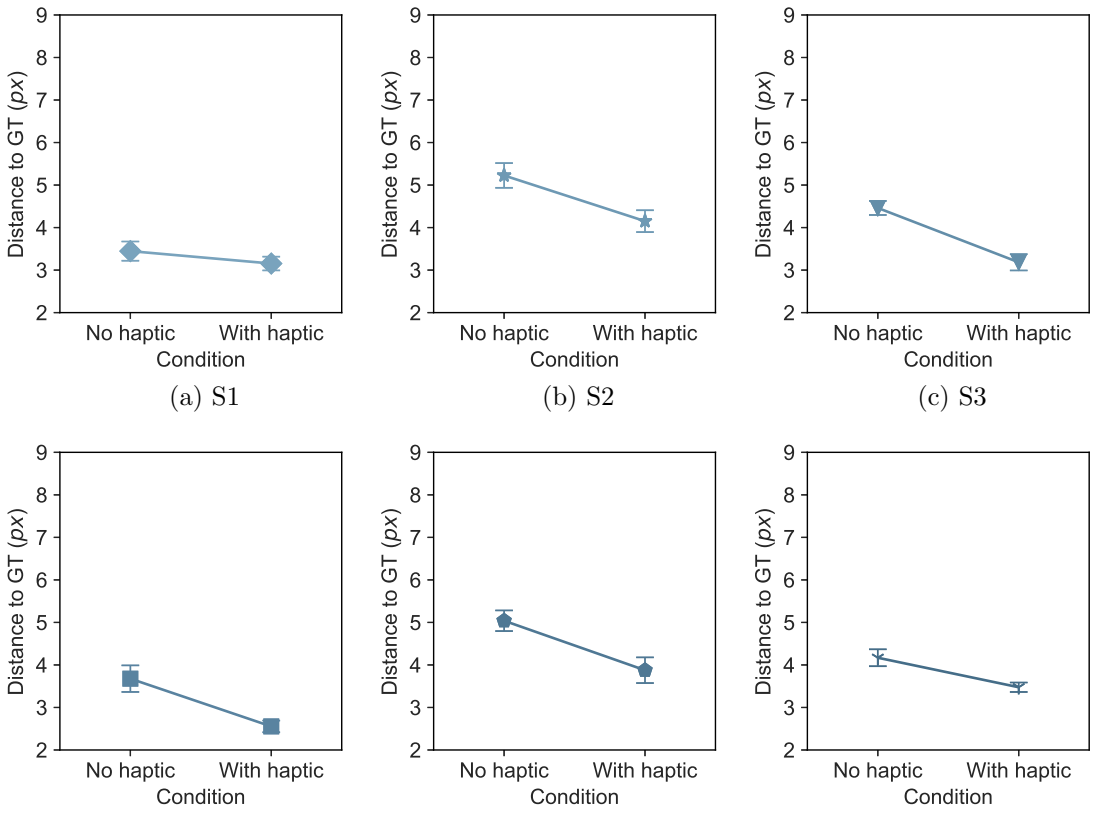


(f) Performance S10

Figure B.18: For observer S10, performance was overall average (where average means not good but also not bad). Texture pairs Δ_0, Δ_1 and Δ_2 all had strong horizontal statistical features while Δ_0 and Δ_1 also had strong vertical features. In spite of this marked difference, there was no notable difference in performance between the different texture pairs. The observer was markedly biased to the right.

B.3 Experiment 3

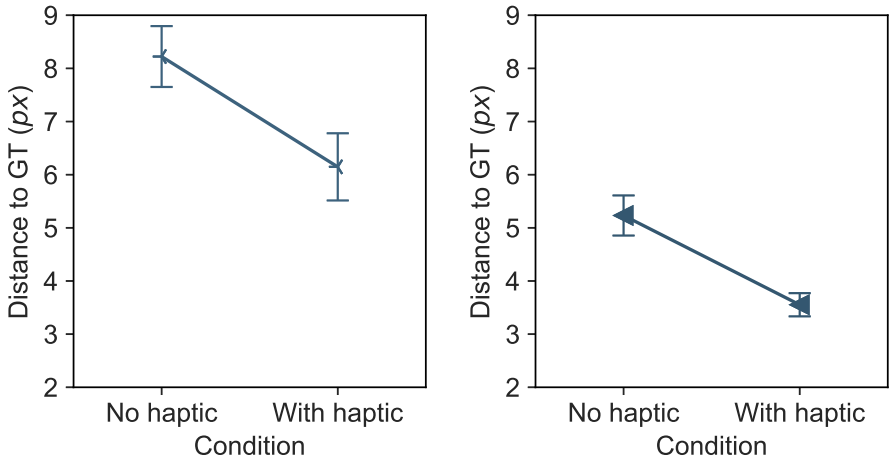
B.3.1 Individual results



(a) S1 (b) S2 (c) S3

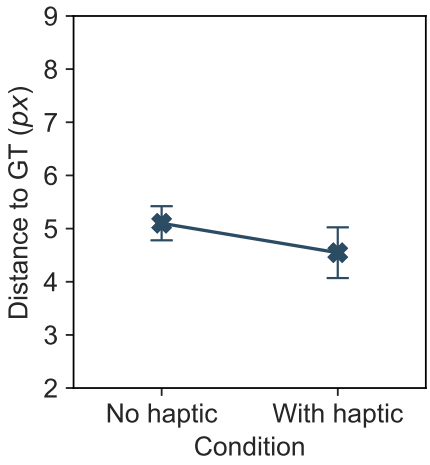
(d) S4 (e) S5 (f) S6

Figure B.19: Experiment 3 individual results, comparative performance of Distance-to-GT per condition, per participant, pt 1, with error bars showing the standard error of the mean.

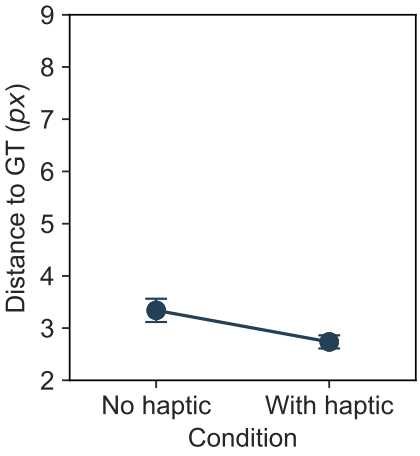


(g) S7

(h) S8



(i) S9



(j) S10

Figure B.19: Experiment 3 individual results, comparative performance of Distance-to-GT per condition, per participant, pt 2, with error bars showing the standard error of the mean.

B.3.2 Wilcoxon signed-rank test tables

Distance			
	W	df	p
S1	135.5	19	0.262
S2	155.0	19	0.017*
S3 206.0	19	<.001*	
S4	160.5	19	0.001*
S5	178.5	19	0.006*
S6	164.5	19	<.001*
S7	171.0	19	0.012*
S8	184.0	19	<.001*
S9	132.0	19	0.045*
S10	113.5	19	0.003*
Mean	55.0	9	0.002*
Median	55.0	9	0.002*

Table B.1: The values of the Wilcoxon signed-rank per participant for the Delaunay contours, and for the mean per participant, as well as the median per participant – per condition.

Assumption		
	W	p
S1	0.947	0.328
S2	0.947	0.318
S3	0.958	0.497
S4	0.866	0.010*
S5	0.966	0.668
S6	0.937	0.210
S7	0.937	0.207
S8	0.961	0.555
S9	0.798	<.001*
S10	0.919	0.096
Mean	0.958	0.759
Median	0.946	0.621

Table B.2: The values of the Shapiro-Wilk normality test per participant, and for the mean per participant per condition.

Volumes

	t	df	p_t	W	p_W
Overdelineation	-2.620	9	0.028	8.0	0.049
Underdelineation	8.004	9	<.001*	55.0	0.002*
Tumour included	-7.997	9	<.001*	0.0	0.006*
Overall contour volume	-5.621	9	<.001*	1.0	0.004*

Table B.3: Table of volumetric parametric and non-parametric comparisons, for the Student's t-test and Wilcoxons signed-rank for the mean delineated volumes per participant, per condition, significance occurs at $p < 0.025$. There was a non-significant increase in overdelineation (healthy tissue included), with a significant decrease in underdelineation (tumour missed out), with a significant increase in volume of delineated tumour and total delineated volume. Shapiro-Wilk normality checks were insignificant.

Distance

	t	df	p
S1	91.5	19	0.232
S2	142.5	19	0.059
S3	141.0	19	0.017*
S4	92.0	19	0.014*
S5	149.0	19	0.031*
S6	118.0	19	0.010*
S7	174.0	19	0.011*
S8	164.5	19	<.001*
S9	116.0	19	0.065
S10	61.5	19	0.012*
Mean	55.0	9	0.002*
Median	45.0	9	0.009*

Table B.4: The values of the Wilcoxon signed-rank per participant, and for the mean and median per participant per condition, medical contour analysis method.

Assumption		
	W	p
S1	0.919	0.096
S2	0.947	0.317
S3	0.951	0.376
S4	0.876	0.015*
S5	0.951	0.378
S6	0.923	0.115
S7	0.926	0.129
S8	0.951	0.381
S9	0.742	<.001*
S10	0.792	<.001*
Mean	0.882	0.137
Median	0.881	0.135

Table B.5: The values of the Shapiro-Wilk normality test per participant, and for the mean and median per participant per condition, medical contour analysis method.

References

Abbey, C. K. (2013). Classification images aid understanding of visual task performance and diagnosis. *SPIE Newsroom*.

Abbey, C. K., & Eckstein, M. P. (2002). Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments. *Journal of vision*, *2*(1), 5–5.

Abbey, C. K., & Eckstein, M. P. (2006). Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *Journal of Vision*, *6*(4), 4–4.

Abbey, C. K., & Eckstein, M. P. (2007). Classification images for simple detection and discrimination tasks in correlated noise. *JOSA A*, *24*(12), B110–B124.

Abbey, C. K., & Eckstein, M. P. (2009). Frequency tuning of perceptual templates changes with noise magnitude. *Journal of the Optical Society of America A*, *26*, B72.

Adams, M. A. (2019). *The integration of vision and touch for locating objects* (Doctoral dissertation). University of Reading.

Adams, W. J., Kerrigan, I. S., & Graf, E. W. (2016). Touch influences perceived gloss. *Scientific reports*, *6*, 21866.

Alfaro, A., Bernabeu, Á., Agulló, C., Parra, J., & Fernández, E. (2015). Hearing colors: An example of brain plasticity. *Frontiers in Systems Neuroscience*, *9*, 56.

Ames Jr, A. (1951). Visual perception and the rotating trapezoidal window. *Psychological Monographs: General and Applied*, *65*(7), i.

Backus, B. T. (2002). Perceptual metamers in stereoscopic vision. *Advances in neural information processing systems*, 1223–1230.

Backus, B. T., & Banks, M. S. (1999). Estimator reliability and distance scaling in stereoscopic slant perception. *Perception*, *28*(2), 217–242.

Banerjee, P., Hu, M., Kannan, R., & Krishnaswamy, S. (2017). A semi-automated approach to improve the efficiency of medical imaging segmentation for haptic rendering. *Journal of digital imaging*, *30*(4), 519–527.

Bertram, C., & Stafford, T. (2016). Improving training for sensory augmentation using the science of expertise. *Neuroscience & Biobehavioral Reviews*, *68*, 234–244.

Bologna, G., Deville, B., Vinckenbosch, M., & Pun, T. (2008). A perceptual interface for vision substitution in a color matching experiment. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1621–1628.

Bresciani, J.-P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of vision*, *6*(5), 2–2.

Broder, J. S. (2011). *Diagnostic imaging for the emergency physician e-book*. Elsevier Health Sciences.

Burge, J., Girshick, A. R., & Banks, M. S. (2010). Visual–haptic adaptation is determined by relative reliability. *Journal of Neuroscience*, *30*(22), 7714–7721.

Burge, J., McCann, B. C., & Geisler, W. S. (2016). Estimating 3d tilt from local image cues in natural scenes. *Journal of vision*, *16*(13), 2–2.

Carcedo, M. G., Chua, S. H., Perrault, S., Wozniak, P., Joshi, R., Obaid, M., Fjeld, M., & Zhao, S. (2016). Hapticolor: Interpolating color information as haptic feedback to assist the colorblind. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3572–3583.

Castella, C., Eckstein, M., Abbey, C., Kinkel, K., Verdun, F., Saunders, R., Samei, E., & Bochud, F. (2009). Mass detection on mammograms: Influence of signal shape uncertainty on human and model observers. *JOSA A*, *26*(2), 425–436.

Chen, Y.-T., & Chuang, M.-C. (2014). The study of tactile feeling and it's expressing vocabulary. *International Journal of Industrial Ergonomics*, *44*(5), 675–684.

Chit, S. M., & Yap, K. M. (2012). An investigation into virtual objects learning by using haptic interface for visually impaired children. *Sunway Academic Journal*, *9*, 29–42.

Clark, J. J., & Yuille, A. L. (2013). *Data fusion for sensory information processing systems* (Vol. 105). Springer Science & Business Media.

Clarke, A. D., Dong, X., & Chantler, M. J. (2012). Does free-sorting provide a good estimate of visual similarity.

Clarke, A. D., Halley, F., Newell, A. J., Griffin, L. D., & Chantler, M. J. (2011). Perceptual similarity: A texture challenge. *BMVC*, 1–10.

Conti, F., Barbagli, F., Balaniuk, R., Halg, M., Lu, C., Morris, D., Sentis, L., Warren, J., Khatib, O., & Salisbury, K. (2003). The chai libraries. *Proceedings of Eurohaptics 2003*, 496–500.

Cooper, J. S., Mukherji, S. K., Toledano, A. Y., Beldon, C., Schmalfuss, I. M., Amdur, R., Sailer, S., Loevner, L. A., Kousouboris, P., Ang, K. K., et al. (2007). An evaluation of the variability of tumor-shape definition derived by experienced observers from ct images of supraglottic carcinomas (acrin protocol 6658). *International Journal of Radiation Oncology* Biology* Physics*, *67*(4), 972–975.

Corke, P. (2017). *Robotics, vision and control: Fundamental algorithms in matlab® second, completely revised* (Vol. 118). Springer.

Debette, S., & Markus, H. (2010). The clinical importance of white matter hyperintensities on brain magnetic resonance imaging: Systematic review and meta-analysis. *Bmj*, *341*, c3666.

Delazio, A., Israr, A., & Klatzky, R. L. (2017). Cross-modal correspondence between vibrations and colors. *2017 IEEE World Haptics Conference (WHC)*, 219–224.

Dixon, B. J., Daly, M. J., Chan, H., Vescan, A. D., Witterick, I. J., & Irish, J. C. (2013). Surgeons blinded by enhanced navigation: The effect of augmented reality on attention. *Surgical endoscopy*, *27*(2), 454–461.

Drewing, K., & Ernst, M. O. (2006). Integration of force and position cues for shape perception through active touch. *Brain Research*, *1078*, 92–100.

Elmore, J. G., Jackson, S. L., Abraham, L., Miglioretti, D. L., Carney, P. A., Geller, B. M., Yankaskas, B. C., Kerlikowske, K., Onega, T., Rosenberg, R. D., et al. (2009). Variability in interpretive performance at screening

mammography and radiologists' characteristics associated with accuracy. *Radiology*, *253*(3), 641–651.

Ernst, M. (2006). A bayesian view on multimodal cue integration. *Human body perception from the inside out*, *131*, 105–131.

Ernst, M. O. (2007). Learning to integrate arbitrary signals from vision and touch. *Journal of Vision*, *7*(5), 7–7.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429.

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in cognitive sciences*, *8*(4), 162–169.

Ertl-Wagner, B. B., Blume, J. D., Peck, D., Udupa, J. K., Herman, B., Levering, A., Schmalfuss, I. M., members of the ACRIN 6662 study group, et al. (2009). Reliability of tumor volume estimation from mr images in patients with malignant glioma. results from the american college of radiology imaging network (acrin) 6662 trial. *European radiology*, *19*(3), 599–609.

Etzi, R., Spence, C., & Gallace, A. (2014). Textures that we like to touch: An experimental study of aesthetic preferences for tactile stimuli. *Consciousness and cognition*, *29*, 178–188.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of vision*, *10*(1), 6–6.

Evreinova, T., Evreinov, G. E., & Raisamo, R. (2012). Evaluation of effectiveness of the stickgrip device for detecting the topographic heights on digital maps. *Int. J. Comput. Sci. Appl.*, *9*(3), 61–76.

Gepshtein, S., Burge, J., Ernst, M. O., & Banks, M. S. (2005). The combination of vision and touch depends on spatial proximity. *Journal of Vision*, *5*(11), 7–7.

Greenwald, D., Cao, C. G., & Bushnell, E. W. (2012). Haptic detection of artificial tumors by hand and with a tool in a mis environment. *IEEE transactions on haptics*, *5*(2), 131–138.

Gueorguiev, D., Bochereau, S., Mouraux, A., Hayward, V., & Thonnard, J.-L. (2016). Touch uses frictional cues to discriminate flat materials. *Scientific reports*, *6*, 25553.

Guest, S., & Spence, C. (2003). What role does multisensory integration play in the visuotactile perception of texture? *International Journal of Psychophysiology*, *50*(1-2), 63–80.

Haijiang, Q., Saunders, J. A., Stone, R. W., & Backus, B. T. (2006). Demonstration of cue recruitment: Change in visual appearance by means of pavlovian conditioning. *Proceedings of the National Academy of Sciences*, *103*(2), 483–488.

Halley, F. (2012). *Perceptually relevant browsing environments for large texture databases* (Doctoral dissertation). Heriot Watt University.

Hamilton-Fletcher, G., & Ward, J. (2013). Representing colour through hearing and touch in sensory substitution devices. *Multisensory research*, *26*(6), 503–532.

Hamilton-Fletcher, G., Wright, T. D., & Ward, J. (2016). Cross-modal correspondences enhance performance on a colour-to-sound sensory substitution device. *Multisensory research*, *29*(4-5), 337–363.

Harms, C., & Lakens, D. (2018). Making'null effects' informative: Statistical techniques and inferential frameworks. *Journal of Clinical and Translational Research*, *3*(Suppl 2), 382.

Harris, J. M. (2004). Binocular vision: Moving closer to reality. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *362*(1825), 2721–2739.

Helbig, H. B., & Ernst, M. O. (2007). Optimal integration of shape information from vision and touch. *Experimental Brain Research*, *179*(4), 595–606.

Helbig, H. B., & Ernst, M. O. (2008). Visual-haptic cue weighting is independent of modality-specific attention. *Journal of vision*, *8*(1), 21–21.

Held, R., Ostrovsky, Y., de Gelder, B., Gandhi, T., Ganesh, S., Mathur, U., & Sinha, P. (2011). The newly sighted fail to match seen with felt. *Nature neuroscience*, *14*(5), 551–553.

Held, R. T., & Hui, T. T. (2011). A guide to stereoscopic 3d displays in medicine. *Academic radiology*, *18*(8), 1035–1048.

Heron, S., & Lages, M. (2012). Screening and sampling in studies of binocular vision. *Vision research*, *62*, 228–234.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science*, *298*(5598), 1627–1630.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of vision*, *4*(12), 1–1.

Ho, C., & Spence, C. (2014). Effectively responding to tactile stimulation: Do homologous cue and effector locations really matter? *Acta psychologica*, *151*, 32–39.

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

Holmes, N. P., Calvert, G. A., & Spence, C. (2004). Extending or projecting peripersonal space with tools? multisensory interactions highlight only the distal and proximal ends of tools. *Neuroscience letters*, *372*(1-2), 62–67.

Holmes, N. P., Calvert, G. A., & Spence, C. (2007). Tool use changes multisensory interactions in seconds: Evidence from the crossmodal congruency task. *Experimental Brain Research*, *183*(4), 465–476.

Holmes, N. P., Sanabria, D., Calvert, G. A., & Spence, C. (2007). Tool-use: Capturing multisensory spatial attention or extending multisensory peripersonal space? *Cortex*, *43*(3), 469–489.

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

Holmes, N. P., Spence, C., Hansen, P. C., Mackay, C. E., & Calvert, G. A. (2008). The multisensory attentional consequences of tool use: A functional magnetic resonance imaging study. *PLoS One*, *3*(10), e3502.

Hong, F., Badde, S., & Landy, M. (2021). Causal-inference regulates audiovisual spatial recalibration via its influence on audiovisual perception. *bioRxiv*.

Horan, G., Roques, T. W., Curtin, J., & Barrett, A. (2006). "two are better than one": A pilot study of how radiologist and oncologists can collaborate in target volume definition. *Cancer Imaging*, *6*(1), 16.

Howard, I. P., Rogers, B. J. et al. (1995). *Binocular vision and stereopsis*. Oxford University Press, USA.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jager, E. A., Ligtenberg, H., Caldas-Magalhaes, J., Schakel, T., Philippens, M. E., Pameijer, F. A., Kasperts, N., Willems, S. M., Terhaard, C. H., & Raaijmakers, C. P. (2016). Validated guidelines for tumor delineation on magnetic resonance imaging for laryngeal and hypopharyngeal cancer. *Acta Oncologica*, *55*(11), 1305–1312.

JASP Team. (2019). JASP (Version 0.11.1)[Computer software]. https://jasp-stats.org/

Jones, P. R. (2016). A tutorial on cue combination and signal detection theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology*, *73*, 117–139.

Kang, Z., & Kim, K. (2018). Multimodal perception study on virtual 3d curved textures with vision and touch for interactive multimedia systems. *Multimedia Tools and Applications*, *77*(2), 2209–2223.

Kingdom, F. A. A., Baldwin, A. S., & Schmidtmann, G. (2015). Modeling probability and additive summation for detection across multiple mechanisms under the assumptions of signal detection theory. *Journal of Vision*, *15*(5), 1.

Kinkel, K., Lu, Y., Both, M., Warren, R. S., & Thoeni, R. F. (2002). Detection of hepatic metastases from cancers of the gastrointestinal tract by using noninvasive imaging methods (us, ct, mr imaging, pet): A meta-analysis. *Radiology*, *224*(3), 748–756.

Klatzky, R. L., Lederman, S. J., & Reed, C. (1987). There's more to touch than meets the eye: The salience of object attributes for haptics with and without vision. *Journal of experimental psychology: general*, *116*(4), 356.

Knill, D. C., & Richards, W. (1996). *Perception as bayesian inference*. Cambridge University Press.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision research*, *43*(24), 2539–2558.

Kohler, A., Welsch, T., Sturm, A.-K., Baretton, G. B., Reissfelder, C., Weitz, J., & Riediger, C. (2018). Primary choriocarcinoma of the liver: A rare, but important differential diagnosis of liver lesions. *Journal of surgical case reports*, *2018*(4), rjy068.

Kompaniez-Dunigan, E., Abbey, C. K., Boone, J. M., & Webster, M. A. (2015). Adaptation and visual search in mammographic images. *Attention, Perception, & Psychophysics*, *77*(4), 1081–1087.

Körding, K. P., & Tenenbaum, J. B. (2007). Causal inference in sensorimotor integration. *Advances in neural information processing systems*, 737–744.

Kuroki, S., Sawayama, M., & Nishida, S. (2019). Haptic metameric textures. *BioRxiv*, 653550.

Lacey, S., & Sathian, K. (2014). Visuo-haptic multisensory object recognition, categorization, and representation. *Frontiers in psychology*, *5*, 730.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision research*, *35*(3), 389–412.

Lareau, D., & Lang, J. (2012). Haptic rendering of photographs. *2012 IEEE International Workshop on Haptic Audio Visual Environments and Games (HAVE 2012) Proceedings*, 107–112.

Latifi-Navid, M., Bilen, M., Konukseven, E. I., DOĞAN, M., & Altun, A. (2016). Fast and accurate semiautomatic haptic segmentation of brain tumor in 3d mri images. *Turkish Journal of Electrical Engineering and Computer Science*, *24*(3), 1397–1411.

Leclerc, M., Lartigau, E., Lacornerie, T., Daisne, J.-F., Kramar, A., & Grégoire, V. (2015). Primary tumor delineation based on 18fdg pet for locally advanced head and neck cancer treated by chemo-radiotherapy. *Radiotherapy and Oncology*, *116*(1), 87–93.

Lederman, S. J., & Klatzky, R. L. (2009). Haptic perception: A tutorial. *Attention, Perception, & Psychophysics*, *71*(7), 1439–1459.

Lim, J., Yoo, Y., & Choi, S. (2019). Guidance-based two-dimensional haptic contour rendering for accessible photography. *2019 IEEE World Haptics Conference (WHC)*, 401–406.

Louw, S., Kappers, A. M., & Koenderink, J. J. (2000). Haptic detection thresholds of gaussian profiles over the whole range of spatial scales. *Experimental brain research*, *132*(3), 369–374.

Lovell, P. G., Bloj, M., & Harris, J. M. (2012). Optimal integration of shading and binocular disparity for depth perception. *Journal of vision*, *12*(1), 1–1.

Machilsen, B., & Wagemans, J. (2011). Integration of contour and surface information in shape detection. *Vision Research*, *51*(1), 179–186.

Maidenbaum, S., Abboud, S., & Amedi, A. (2014). Sensory substitution: Closing the gap between basic research and widespread practical visual rehabilitation. *Neuroscience & Biobehavioral Reviews*, *41*, 3–15.

Maidenbaum, S., Arbel, R., Buchs, G., Shapira, S., & Amedi, A. (2014). Vision through other senses: Practical use of sensory substitution devices as assistive technology for visual rehabilitation. *22nd Mediterranean Conference on Control and Automation*, 182–187.

Maloney, L. T., & Landy, M. S. (1989). A statistical framework for robust fusion of depth information. *Visual communications and image processing IV*, *1199*, 1154–1163.

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. *Probabilistic models of the brain*, 13–36.

Maravita, A., Spence, C., Kennett, S., & Driver, J. (2002). Tool-use changes multimodal spatial interactions between vision and touch in normal humans. *Cognition*, *83*(2), B25–B34.

Marks, L. E. (1974). On associations of light and sound: The mediation of brightness, pitch, and loudness. *The American journal of psychology*, 173–188.

Marks, L. E. (2014). *The unity of the senses: Interrelations among the modalities*. Academic Press.

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Cross-modal interactions in auditory and visual discrimination. *International Journal of Psychophysiology*, *50*(1-2), 125–145.

Marks, L. E., Hammeal, R. J., Bornstein, M. H., & Smith, L. B. (1987). Perceiving similarity and comprehending metaphor. *Monographs of the Society for Research in Child Development*, i–100.

Martí-Bonmatí, L., Sopena, R., Bartumeus, P., & Sopena, P. (2010). Multimodality imaging techniques. *Contrast media & molecular imaging*, *5*(4), 180–189.

MATLAB version 9.4.0.813654 (R2018a). (2018). The Mathworks, Inc. Natick, Massachusetts.

McDonnell, P. M., & Duffett, J. (1972). Vision and touch: A reconsideration of conflict between the two senses. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *26*(2), 171.

Misselhorn, J., Daume, J., Engel, A. K., & Friese, U. (2016). A matter of attention: Crossmodal congruence enhances and impairs performance in a novel trimodal matching paradigm. *Neuropsychologia*, *88*, 113–122.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature*, *447*(7141), 206.

Nelms, B. E., Tomé, W. A., Robinson, G., & Wheeler, J. (2012). Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer. *International Journal of Radiation Oncology* Biology* Physics*, *82*(1), 368–378.

Njeh, C. (2008). Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of medical physics/Association of Medical Physicists of India*, *33*(4), 136.

Nowee, M. E., Voncken, F. E., Kotte, A., Goense, L., van Rossum, P., van Lier, A., Heijmink, S., Aleman, B., Nijkamp, J., Meijer, G. J., et al. (2019). Gross tumour delineation on computed tomography and positron emission tomography-computed tomography in oesophageal cancer: A nationwide study. *Clinical and translational radiation oncology*, *14*, 33–39.

Nyström, I., Malmberg, F., Vidholm, E., & Bengtsson, E. (2009). Segmentation and visualization of 3d medical images through haptic rendering.

Okamoto, S., Nagano, H., & Yamada, Y. (2012). Psychophysical dimensions of tactile perception of textures. *IEEE Transactions on Haptics*, *6*(1), 81–93.

Oruç, I., Maloney, L. T., & Landy, M. S. (2003). Weighted linear cue combination with possibly correlated error. *Vision research*, *43*(23), 2451–2468.

Osinski, D., & Hjelme, D. R. (2018). A sensory substitution device inspired by the human visual system. *2018 11th International Conference on Human System Interaction (HSI)*, 186–192.

Papiez, L., & Langer, M. (2006). On probabilistically defined margins in radiation therapy. *Physics in Medicine & Biology*, *51*(16), 3921.

Plaisier, M. A., Kappers, A. M., Tiest, W. M. B., & Ernst, M. O. (2010). Visually guided haptic search. *IEEE Transactions on Haptics*, *3*(1), 63–72.

Plaisier, M. A., van Dam, L. C., Glowania, C., & Ernst, M. O. (2014). Exploration mode affects visuohaptic integration of surface orientation. *Journal of vision*, *14*(13), 22–22.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*(1), 49–70.

Prins, N., & Kingdom, F. A. A. (2016, June 22). *Palamedes: Matlab routines for analyzing psychophysical data.* (Version 1.8.2). http://www.palamedestoolbox.org

Prins, N. et al. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the palamedes toolbox. *Frontiers in psychology*, *9*, 1250.

Purves, D., Cabeza, R., Huettel, S. A., LaBar, K. S., Platt, M. L., Woldorff, M. G., & Brannon, E. M. (2013). *Cognitive neuroscience* (2nd ed.). Sinauer.

Richardson, M., Thar, J., Alvarez, J., Borchers, J., Ward, J., & Hamilton-Fletcher, G. (2019). How much spatial information is lost in the sensory substitution process? comparing visual, tactile, and auditory approaches. *Perception*, *48*(11), 1079–1103.

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

Rohde, M., van Dam, L. C., & Ernst, M. O. (2016). Statistically optimal multisensory cue integration: A practical tutorial. *Multisensory research*, *29*(4-5), 279–317.

Rohe, T., & Noppeney, U. (2015). Sensory reliability shapes perceptual inference via two mechanisms. *Journal of vision*, *15*(5), 22–22.

Rosas, P., Wagemans, J., Ernst, M. O., & Wichmann, F. A. (2005). Texture and haptic cues in slant discrimination: Reliability-based cue weighting without statistically optimal cue combination. *JOSA A*, *22*(5), 801–809.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2004). Some observations on the effects of slant and texture type on slant-from-texture. *Vision research*, *44*(13), 1511–1535.

Rosas, P., Wichmann, F. A., & Wagemans, J. (2007). Texture and object motion in slant discrimination: Failure of reliability-based weighting of cues may be evidence for strong fusion. *Journal of Vision*, *7*(6), 3–3.

Schulz-Wendtland, R., Harz, M., Meier-Meitinger, M., Brehm, B., Wacker, T., Hahn, H. K., Wagner, F., Wittenberg, T., Beckmann, M. W., Uder, M., et al. (2017). Semi-automated delineation of breast cancer tumors and

subsequent materialization using three-dimensional printing (rapid prototyping). *Journal of surgical oncology*, *115*(3), 238–242.

Schwerdt, H. N., Tapson, J., & Etienne-Cummings, R. (2009). A color detection glove with haptic feedback for the visually disabled. *2009 43rd Annual Conference on Information Sciences and Systems*, 681–686.

Seibert, T. M., White, N. S., Kim, G.-Y., Moiseenko, V., McDonald, C. R., Farid, N., Bartsch, H., Kuperman, J., Karunamuni, R., Marshall, D., et al. (2016). Distortion inherent to magnetic resonance imaging can lead to geometric miss in radiosurgery planning. *Practical radiation oncology*, *6*(6), e319–e328.

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. *Trends in cognitive sciences*, *14*(9), 425–432.

Shrikhande, S. V., Barreto, S. G., Goel, M., & Arya, S. (2012). Multimodality imaging of pancreatic ductal adenocarcinoma: A review of the literature. *HPB*, *14*(10), 658–668.

Skedung, L., Arvidsson, M., Chung, J. Y., Stafford, C. M., Berglund, B., & Rutland, M. W. (2013). Feeling small: Exploring the tactile perception limits. *Scientific reports*, *3*(1), 1–6.

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, & Psychophysics*, *73*(4), 971–995.

Spence, C. (2013). Just how important is spatial coincidence to multisensory integration? evaluating the spatial rule. *Annals of the New York Academy of Sciences*, *1296*(1), 31–49.

Spick, C., Herrmann, K., & Czernin, J. (2016). 18f-fdg pet/ct and pet/mri perform equally well in cancer: Evidence from studies on more than 2,300 patients. *Journal of Nuclear Medicine*, *57*(3), 420–430.

Takahashi, C., Diedrichsen, J., & Watt, S. J. (2009). Integration of vision and haptics during tool use. *Journal of vision*, *9*(6), 3–3.

Takahashi, C., & Watt, S. J. (2014). Visual-haptic integration with pliers and tongs: Signal "weights" take account of changes in haptic sensitivity caused by different tools. *Frontiers in psychology*, *5*, 109.

Tantau, T. (2013, December 20). *The tikz and pgf packages: Manual for version 3.0.0*. http://sourceforge.net/projects/pgf/

Tapson, J., Diaz, J., Sander, D., Gurari, N., Chicca, E., Pouliquen, P., & Etienne-Cummings, R. (2008). The feeling of color: A haptic feedback device for the visually disabled. *2008 IEEE Biomedical Circuits and Systems Conference*, 381–384.

Tateishi, U., Hosono, A., Makimoto, A., Nakamoto, Y., Kaneta, T., Fukuda, H., Murakami, K., Terauchi, T., Suga, T., Inoue, T., et al. (2009). Comparative study of fdg pet/ct and conventional imaging in the staging of rhabdomyosarcoma. *Annals of nuclear medicine*, *23*(2), 155–161.

Unger, B., Hollis, R., & Klatzky, R. (2011). Roughness perception in virtual textures. *IEEE Transactions on Haptics*, *4*(2), 122–133.

Vadakkumpadan, F., & Sethi, S. (2018). Biomedical image analytics using sas® viya®. *Proceedings of the SAS Global Forum 2018 Conference*, 1961–2018.

Van de Steene, J., Linthout, N., de Mey, J., Vinh-Hung, V., Claassens, C., Noppen, M., Bel, A., & Storme, G. (2002). Definition of gross tumor volume in lung

cancer: Inter-observer variability. *Radiotherapy and oncology*, *62*(1), 37–49.

van Dam, L. C., & Ernst, M. O. (2010). Preexposure disrupts learning of location-contingent perceptual biases for ambiguous stimuli. *Journal of Vision*, *10*(8), 15–15.

van Elmpt, W., Zegers, C. M., Das, M., & De Ruysscher, D. (2014). Imaging techniques for tumour delineation and heterogeneity quantification of lung cancer: Overview of current possibilities. *Journal of thoracic disease*, *6*(4), 319.

Vidholm, E., Golubovic, M., Nilsson, S., & Nyström, I. (2008). Accurate and reproducible semi-automatic liver segmentation using haptic interaction. *Medical Imaging 2008: Visualization, Image-Guided Procedures, and Modeling*, *6918*, 69182Q.

Waskom, M., Botvinnik, O., O'Kane, D., Hobson, P., Ostblom, J., Lukauskas, S., Gemperline, D. C., Augspurger, T., Halchenko, Y., Cole, J. B., Warmenhoven, J., de Ruiter, J., Pye, C., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Bachant, P., ... Qalieh, A. (2018). *Mwaskom/seaborn: V0.9.0 (july 2018)* (Version v0.9.0). Zenodo. https://doi.org/10.5281/zenodo.1313201

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological bulletin*, *88*(3), 638.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. fitting, sampling, and goodness of fit. *Perception & psychophysics*, *63*(8), 1293–1313.

Yadav, S. P., & Yadav, S. (2020). Image fusion using hybrid methods in multimodality medical images. *Medical & Biological Engineering & Computing*, 1–19.

Yip, S. S., & Aerts, H. J. (2016). Applications and limitations of radiomics. *Physics in Medicine & Biology*, *61*(13), R150.

Yuille, A. L., & Bülthoff, H. H. (1993). Bayesian decision theory and psychophysics.

Zeki, S. (1993). *A vision of the brain*. Blackwell Scientific Publ.