Stereoscopic Viewing, Roughness and Gloss Perception

Thomas Methven

Submitted for the degree of Doctor of Philosophy

Heriot-Watt University

School of Mathematical and Computer Sciences

September 2013

Abstract

This thesis presents a novel investigation into the effect stereoscopic vision has upon the strength of perceived gloss on rough surfaces. We demonstrate that in certain cases disparity is necessary for accurate judgements of gloss strength.

We first detail the process we used to create a two-level taxonomy of property terms, which helped to inform the early direction of this work, before presenting the eleven words which we found categorised the property space. This shaped careful examination of the relevant literature, leading us to conclude that most studies into roughness, gloss, and stereoscopic vision have been performed with unrealistic surfaces and physically inaccurate lighting models.

To improve on the stimuli used in these earlier studies, advanced offline rendering techniques were employed to create images of complex, naturalistic, and realistically glossy $1/f^\beta$ noise surfaces. These images were rendered using multi-bounce path tracing to account for interreflections and soft shadows, with a reflectance model which observed all common light phenomena. Using these images in a series of psychophysical experiments, we first show that random phase spectra can alter the strength of perceived gloss. These results are presented alongside pairs of the surfaces tested which have similar levels of perceptual gloss. These surface pairs are then used to conclude that naïve observers consistently underestimate how glossy a surface is without the correct surface and highlight disparity, but only on the rougher surfaces presented.

Dedication

This thesis is dedicated to all those who made it possible

Acknowledgements

I would like to acknowledge and thank those who have made this thesis possible.

First, I would like to thank my supervision, Professor Mike Chantler, for all his time, support and help during the last three and a half years. In particular, this thesis would never have been completed without the time he spent proof reading and explaining the frequency domain.

To Dr. Judy Robertson and Professor Patrick Green I would like to say that this thesis would never have been finished without your patience at my interruptions, questions and general confusion at several points throughout this process. Thank you. In addition, I would like to thank Professor Julie Harris and the Vision Labs at the University of St. Andrews for letting me visit and discuss stereoscope design.

My thanks also go to members of the Texture Lab for putting up with my silly questions throughout the last three and a half years, in particular to Dr. Stefano Padilla and Dr. Khemraj Emrith who were so welcoming to me when I first started and who were a constant source of help.

Finally, I would like to thank my family and friends who were so supportive throughout this process. In particular, thank you to Pete, Lou, Mum, and Dad for looking after me and feeding me for so many years, to Niko, Jenny, Neil C, Neil R, Claire, Nicola, Heather, Jack, James, Scotty and Michael for being so supportive throughout this process and to Dominik for being an unforgettable source of happiness in a dark year.

Research Thesis Submission Form

Table of Contents

List of Figures

List of Tables

List of Equations

Table of Terms

Term	Definition
$1/f^\beta$ Noise Surfaces	A particular type of textured surface.
Analysis of Variance	A collection of statistical models used to analyse differences between group means.
Bidirectional Reflectance Distribution Function	A function which defines how light reflects when it incidences with an opaque surface
Binomial Distribution	A discrete probability function used to determine whether a result was due to chance or experimental effect.
Bonferroni Correction	A method to adjust confidence intervals to ensure type I errors are not increased when considering multiple experiments.
Collimated Light	A light source with parallel rays which is, in general, considered to be at an infinite distance. In this thesis, we refer to it as a directional light source.
Crosstalk	Where one eye can see part of the image meant for the other.
Dendrogram	A type of tree diagram used to visualise the results of hierarchical clustering.
Disparity	In relation to stereoscopic viewing, the difference in point locations seen in the left and right eye resulting from a viewer's interoccular distance.
Dissimilarity Matrix	A matrix used to record the pairwise dissimilarity of a set of items.
Free Grouping	Where participants are asked to sort a set of items into groups of perceptual similarity with no limits on group size or number.
Fresnel Effect	The observation that reflective surfaces get apparently more reflective when viewed at grazing angles. In our case more specular reflection will be apparent.

Term	Definition
Front-Surface Mirror	A mirror where the reflective surface is the first surface light incidents, rather than a glass or acrylic layer. Used to avoid problems with ghosting.
Ghosting	A faint, offset second reflection in a mirror caused by light reflecting off a glass or acrylic layer placed before the reflective surface.
Ground Truth	The real result. For example, a rendering model which gives results close to ground truth would produce images indistinguishable from similarly staged photographs.
Interoccular Distance	The distance between the centre of the left and right eye.
Marginal Means	The mean values across a single experimental variable in experimental designs which include multiple variables.
Mauchly's Test of Sphericity	A statistical test to check whether the assumption of sphericity required for repeated measures analysis of variance has been violated.
Method of Adjustment	A classic psychophysical technique where a participant is asked to adjust the level of a stimulus to match a target stimulus presented to them.
Multidimentional Scaling	A set of statistical techniques which aim to find a low-dimensional space which represents a higher-dimensional space with as little error as possible.
Path Tracing	A rendering method which involves casting rays into three dimensional scenes to take account of phenomena such as indirect lighting and soft shadows.
Random Phase	The phase of a $1/f^{\beta}$ noise surface is what determines its appearance. By randomising this, one can generate surfaces of the same roughness, but different appearance.
Repeated Measures	An experimental design where each participant is used for every condition in the experiment.
Roll-Off Factor	β in the $1/f^{\beta}$ noise surfaces equation. This is what controls the apparent roughness of the surface. A higher roll-off factor makes the surface appear smoother.
Root Mean Square Roughness	A common measure of surface roughness.

Term	Definition
Stereogram	One of a pair of images which when viewed through a stereoscope allow one to see a scene with virtual depth.
Stereoscope	A device which directs the view of a viewer's eyes to two different stereograms. Often uses mirrors.
Stereoscopic Display	A display method which shows a viewer's eyes two different images, giving apparent virtual depth via stereopsis.
Stereopsis	The impression of depth apparent when each viewer's eye sees a scene from a slightly different view point and they have normal binocular vision.
Two-Alternate Forced Choice	An experimental method where a participant is asked to select which one of two stimuli better fulfils some requirement.

Table of Abbreviations

Term	Definition
2AFC	Two-Alternate Forced Choice
ANOVA	Analysis of Variance
AOI	Angle of Incidence
BRDF	Bidirectional Reflectance Distribution Function
CPI	Cycles per image width
FFT	Fast Fourier Transform
LUT	Look Up Table
MDS	Multidimensional Scaling
PNG	Portable Network Graphics
RMS	Root Mean Square
RGB	Red, Green, Blue triplet used to represent colours

Table of Symbols

Symbol	Description
$\partial_t u$	The value of the u component in a reaction-diffusion model at a certain generation t
$\partial_t v$	The value of the v component in a reaction-diffusion model at a certain generation t
$D_u \nabla^2$	The generalised diffusion process for the u component in a reaction-diffusion model.
$D_v \nabla^2$	The generalised diffusion process for the v component in a reaction-diffusion model.
$F(u)$	The generalised local reaction process for the u component in a reaction-diffusion model.
$G(v)$	The generalised local reaction process for the v component in a reaction-diffusion model.
$H(f)$	The magnitude spectrum of a $1/f^\beta$ noise surface
β	The roll-off factor of the magnitude spectrum of a $1/f^\beta$ noise surface
σ	The root mean square (RMS) height of a $1/f^\beta$ noise surface
$N(\beta)$	The normalising factor of a $1/f^\beta$ noise surface
θ	The seed for the random phase of a $1/f^\beta$ noise surface
C_λ	Reflection at wavelength λ in the Schlick BRDF
r	Microscale roughness factor in the Schlick BRDF
i	Isotropy factor in the Schlick BRDF
V	Outcoming direction of light
V'	Incoming direction of light

Symbol	Description
N	Surface normal vector
T	Surface tangent vector
H	Bisector vector of V and V'
$\bar{H}$	Projection of $H \perp N$
t	An input to the Schlick BRDF defined as $(H \cdot N)$
u	An input to the Schlick BRDF defined as $(H \cdot V)$
v	An input to the Schlick BRDF defined as $(V \cdot N)$
v'	An input to the Schlick BRDF defined as $(V' \cdot N)$
w	An input to the Schlick BRDF defined as $(T \cdot H)$
$R_\lambda(t, u, v, v', w)$	The end value of the Schlick BRDF
$R_\lambda(u)$	The spectral factor of the Schlick BRDF
$D_\lambda(t, v, v', w)$	The directional factor of the Schlick BRDF
$Z(t)$	The part of the directional factor dependant on the light zenith angle in the Schlick BRDF
$A(w)$	The part of the directional factor dependant on the light azimuth angle in the Schlick BRDF
P	The probability of a result occurring by chance as per a cumulative binomial test
p_{trial}	The probability of success in one trial (from 0 to 1)
k	The number of successes
n	The number of trials
$\propto$	The Bonferroni corrected confidence interval
$\propto_{uncorrected}$	The original, uncorrected confidence interval of 0.05

Chapter 1

Introduction

In this chapter, we will discuss the motivation of this thesis in *Section 1.1*, what is within and outwith the scope of this thesis in *Section 1.2* and the novel contribution of this work in *Section 1.3*. *Section 1.4* details the publications resulting from this thesis at time of print, before a brief overview of how this document is laid out is offered in *Section 1.5*.

1.1   Motivation and Goals

Perceptual judgements form a major part of our daily lives. Whether it is trying to judge the iciness of a road, the quality of a fabric, or the cost of a new smartphone, we rely on our perception of objects and materials every day to inform decisions and communicate them with others. For example, one might talk about the softness of a fabric or the glossiness of a sculpture to describe it to others who aren't there to experience it themselves.

With the proliferation of online shopping and digital representation of goods, however, this question of how to communicate these perceptual judgements has never been more topical. Most online shops only provide pictures of their goods, meaning that there is often a disconnect between what the buyer thinks they have ordered and what they receive. Ho et al., for example have shown that the perceived roughness of a surface (such as a jumper) is dependent on both the viewing and lighting angle, two things online shoppers cannot control [1, 2]. To be able to communicate these properties, however, you first need a language to discuss them.

Therefore, our initial motivation for this thesis was to investigate what property terms naïve participants[1] used to communicate their perceptual judgements. While there have been studies into what property terms people use to communicate and rate these judgements [3-5], there is little agreement on this topic. We wanted to create a simple

---

[1] As in, participants who are non-expert in the fields of Computer Science, Design or Psychophysics.

taxonomy which could be used to measure and rate different properties easily, allowing these properties to be transmitted through digital methods.

We found, however, that once we began to investigate the area that our motivation altered. After creating the taxonomy which is discussed later in this thesis, we found ourselves interested in investigating the interplay between two properties instead. As many early experiments into certain aspects of texture, such as roughness and gloss, were unable to use complex, realistic lighting models or surface topography due to limitations in computing power, those results might not be relevant to more realistic stimuli. We wanted, therefore, to investigate the perceptual interaction between two different surface characteristics in the most naturalistic way possible, while still being able to synthesise the surfaces to retain control.

The first goal of this thesis, thanks to our initial motivation, is to find the two surface characteristics which can be investigated in a fruitful way and which are readily understandable to naïve participants. This is intimately linked to the second main goal: investigating what presentation methods (such as haptic pens, head-tracking, or stereoscopic viewing) would allow us to present surfaces demonstrating these characteristics in the most realistic way. This naturally leads into the final goal of investigating how these two characteristics interact and whether the more realistically presented stimuli alter participants' perceptions of those characteristics. These three goals are presented in order as a linear argument in this thesis. The structure is described in more detail in *Section 1.5*, below.

1.2   Thesis Scope

When looking at a topic like texture, it quickly becomes apparent that it would be infeasible to investigate every feature which might have an effect upon a viewer's perception. Therefore, we decided to narrow down our search and look at the interaction of only two characteristics which were statistically controllable. This process is described in more detail in *Chapter 2* to *Chapter 4* but, in short, the outcome of this investigation was the decision to examine the interplay of gloss and roughness.

As there are a wide range of textured surface generation techniques available, we look only at a specific type, $1/f^{\beta}$ noise surfaces. We decided to use these after examining the literature as they have parsimonious parameters and have been used successfully in other experiments concerning roughness [6-8]. Similarly, there are many different

reflectance models available which can represent glossy surfaces, so we consider only one, the Schlick model. The detailed reasoning for both of these decisions is presented in *Chapter 5*.

Finally, we do not discuss the effect animation or movement of a surface might have in this thesis. Due to the high detail of the surface and the computational complexity of the reflectance model we have chosen, it was impractical to render the stimuli in real time. Instead, we created our stimuli via an offline rendering process using two dedicated rendering servers. As each stimulus still took (on average) 90 minutes to render, it was impractical for us to create animated stimuli for our experiments.

Any smaller limitations or assumptions required will be discussed in the chapters they are a part of, as they are required.

1.3   PhD Contribution

This thesis contains four main contributions to the fields of fabric categorisation, roughness, and gloss perception:

1. We present a two-level taxonomy which was determined by non-expert participants' decisions. From an original list presented by an expert in fabrics, we created a non-expert list of fabric properties, confirmed by naïve participants. Using different naïve participants, these properties were clustered into 11 groups, each with a representative word, providing a two-level taxonomy readily understandable by non-expert, English speakers. This taxonomy is the only one known to the author to be understandable by naïve participants and to cover such a wide range of different fabric types.

2. We show that the random phase of $1/f^{\beta}$ surfaces can alter the perceived strength of gloss. That is to say, two surfaces with the same global second order statistics, made of the same material, but with different phase spectra can be perceptually different in terms of gloss strength. To the author's knowledge this is the first time this result has been shown. This is discussed in detail in *Chapter 7*.

3. In response to point 2, pairs of surfaces were picked which exhibited similar gloss levels at five different roughness (β) levels which could be used in comparison experiments. These gloss-similar pairs of surfaces are provided for use by other researchers in *Appendix E*.

3

4. We show that the presence of highlight disparity on glossy surfaces increases the perceived strength of gloss, at least for rough surfaces. In other words, when a rough, glossy surface is presented without correct stereoscopic disparity, participants are unable to accurately estimate how glossy it is. We consider this finding to be the main contribution of this thesis. Although there has been previous work investigating gloss and disparity (such as Blake and Bülthoff's early work [9] or the more recent work of Wendt et al. [10, 11] which all used relatively smooth, unnatural surfaces) to the author's knowledge no one has used the same combination of naturalistic, rough surfaces and physically accurate lighting before. This finding is discussed in detail in *Chapter 8*.

1.4   Thesis Publications

While performing the work which forms the main backbone of this thesis, some of the experimental work performed was published independently, at different levels of completion. Specifically, two conference papers have been published which are related to the work presented in this document.

The first of these was based upon the investigation performed in *Chapter 3* and details creating a two-level taxonomy of fabric properties. It is entitled *A Comparison of Crowd-Sourcing vs. Traditional Techniques for Deriving Consumer Terms* [12]. The second publication was based on pilot experiments performed during this thesis, but which are not presented within. This publication is entitled *Problems of Perceiving Gloss on Complex Surfaces* [13].

1.5 Thesis Organisation

This thesis is divided into two major sections. In the first section (*Chapter 2* to *Chapter 4*) our aim was to choose two properties which we could investigate the interaction between. The second section of this thesis (*Chapter 5* to *Chapter 8*) describes the investigation performed on our selected terms, gloss and roughness.

As such, the structure of this thesis is perhaps different to what the reader is expecting. In particular, it contains two literature reviews. The one presented in *Chapter 2* is brief and wide ranging as it is used as the starting point of the thesis as a whole. *Chapter 5*, in contrast, contains a detailed and focused literature review into gloss and roughness, as

this chapter represents the point where we have chosen our two properties to investigate. A diagram of this structure is presented in *Figure 1.1*.

Beyond this, the argument of this thesis is mostly linear with the results of one chapter leading into the motivation of the next. As already mentioned, we begin in *Chapter 2* with a brief survey to see whether there was a common list or taxonomy of perceptual properties which was widely used. Discovering this not to be the case, in *Chapter 3* we detail the creation of a simple taxonomy of property terms. This uses techniques discovered in our survey to create a two-level taxonomy which was understandable by English speaking non-experts, representative of the whole property space, concise, and well distributed.

In *Chapter 4*, as we planned to present our stimuli via digital methods, we investigate which tools would be suitable for representing the different property groups. We identify three main categories which represent these property groups: tactile, physically simulated, and visual properties. In turn, we examine each category, before concluding that we will investigate gloss and roughness using a mirror.

Therefore, *Chapter 5* presents an in depth literature review of the gloss and roughness literature, using this knowledge to decide how best to create stimuli which would enable us to investigate these two properties effectively. Here we decide to use $1/f^{\beta}$ noise surfaces due to their isotropic, naturalistic appearance and parsimonious parameter set, lit with the Schlick reflectance model due to its good compromise between accuracy, accounting for the physical properties of light and computational complexity. Finally, due to the lack of research in this particular area, we decide to investigate the following question: Does highlight disparity strengthen perceived gloss on rough surfaces?

In *Chapter 6* we detail exactly how we constructed our stereoscope and our stimuli. We begin by deciding on the design of our stereoscope, choosing to build a single-plane mirror stereoscope. We then proceed to discuss in detail the construction of our stereoscope so the reader can reconstruct the system if necessary. We then discuss the stimuli created for our experiments, going into the detail of how we constructed and rendered these surfaces allowing the reader to recreate similar stimuli as required.

Chapter 1

Introduction

Section 2:
Investigating Gloss and Roughness

Chapter 5

Gloss and Roughness Survey

Chapter 2

Properties Survey

Chapter 3

A Taxonomy of Surface
Properties

Chapter 4

Digital Tools

Section 1:
Choosing the Properties

Chapter 6

Stereoscope and Stimuli
Creation

Chapter 7

Does Phase Affect Gloss
Perception?

Chapter 8

Does Highlight Disparity
Strengthen Gloss?

Chapter 9

Conclusions
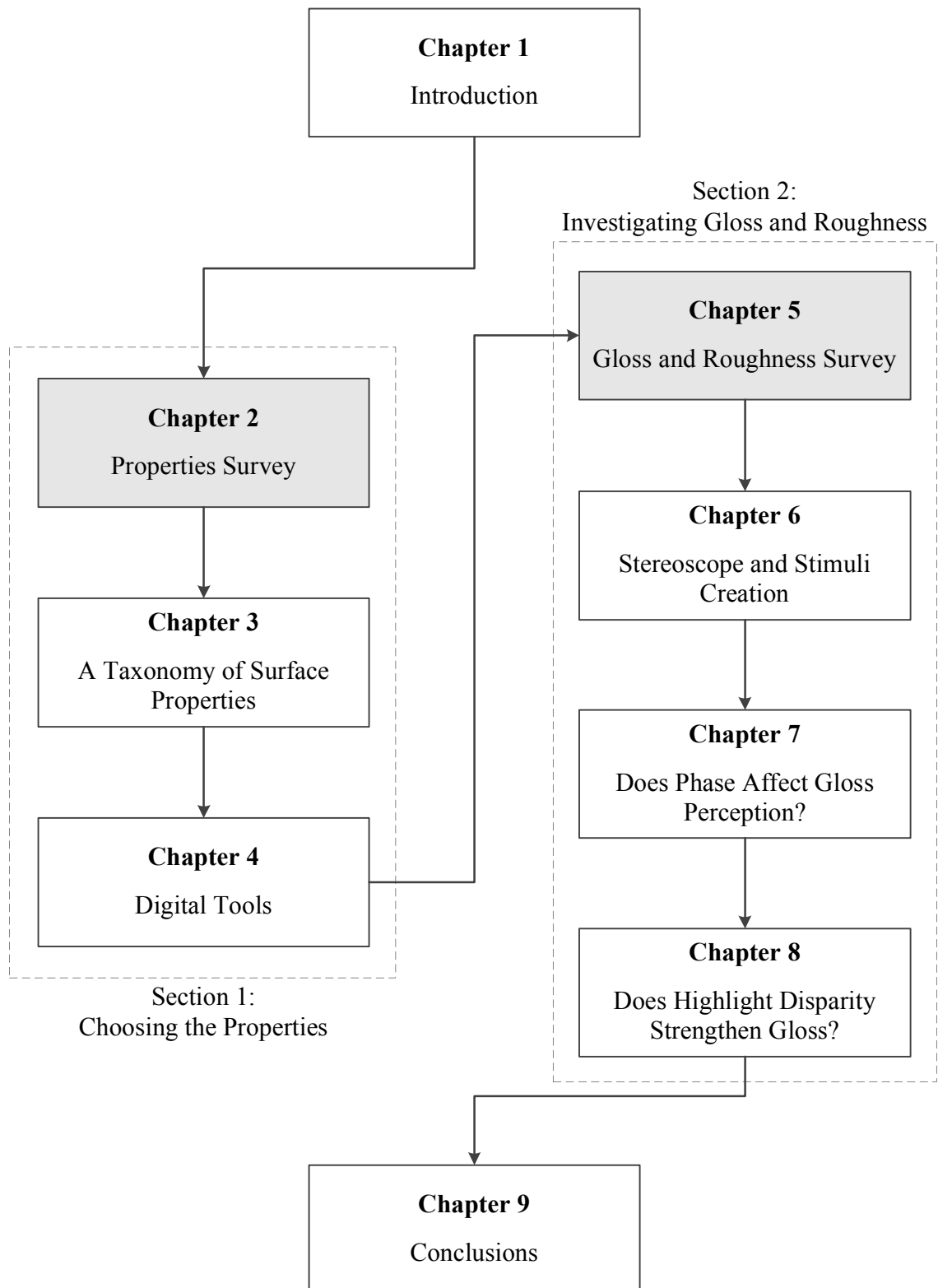
Figure 1.1 - Thesis organisation, showing the two thematic sections. The two chapters with the shaded background represent the location of the two contrasting literature reviews presented in this thesis.

Due to the type of stimuli used, in *Chapter 7* we present two psychophysical experiments designed to investigate whether random phase alters participants' perception of how glossy a surface is. As these results show that stimuli of the same root mean square roughness are perceived at different gloss strengths to others, we conclude this chapter by selecting pairs of surfaces which are of statistically similar gloss strength.

These pairs of surfaces are then used in *Chapter 8* where we answer the question posed earlier: Does highlight disparity strengthen perceived gloss on rough surfaces?

Finally, in *Chapter 9* we conclude the thesis with a detailed summary of the thesis as a whole, suggestions of future work and a reiteration of the main conclusions and contributions of our work.

Chapter 2

Properties Survey

2.1   Introduction

When going about our daily lives, materials such as glass, plastic, metal, or stone are all regularly interacted with and require us to be able to quickly and accurately delineate them. One of the most obvious cues we use to do this is gloss or 'shininess'. We can use this to tell plastics from metals, or use it to determine if a road or pavement is dangerous thanks to patches of water or ice, even if it might not be reliable [14].

We knew at this early stage that we wanted to investigate the interaction of multiple important properties which were easily understood by naïve participants, such as gloss. What we were unsure of, however, was just what properties were important to naïve, or non-expert, observers. To this end, we decided to investigate what work had already been done in this area to see if there was a common list or taxonomy of perceptual properties which were well understood and widely used. We are certainly not the first to look into how physical and perceived properties differ and how to go about bridging that gap. In fact, work has been being performed on this area for many years.

Therefore, in this chapter, we will discuss the brief review of the literature of surface properties which was the starting point for the rest of this thesis. This review of the literature is separated into four subsections. We start in *Section 2.2* by discussing the early work attempting to identify texture features which could be used for categorisation. Following this, we discuss influential experiments which built on the low-level feature work but instead gathered high-level features via free grouping techniques in *Section 2.3*. This led us to investigate work using fabrics and textiles in *Section 2.4*, before presenting the results of a wide-ranging survey which incorporated the results of 15 different fabric studies in *Section 2.5*. We end with a brief conclusion of our work in *Section 2.6*.

2.2 Texture Features and Categorisation

Much of the early work attempting to find properties looked at getting computers to identify texture features, often in the hope of categorising pictures of different stimuli automatically [15-18]. One of the earliest examples is a paper by Haralick et al. where they attempt to find common features in greyscale photomicrographs, aerial photographs and satellite imagery for categorisation purposes [19]. While they found 28 different texture features (such as *Angular Second Moment*, *Sum Average* and *Entropy*) which they suggest can be used for a wide range of different stimuli, the focus of the study was not on a participant's perception of the stimuli, although it is mentioned as possible future work. In addition, they state that a major unsolved problem with this type of categorisation comes from problems in gray-tone normalization. In other words, while humans will ignore differences in contrast and brightness when categorising images, most texture features will not.

While these types of study are of the utmost importance and have informed future experiments into texture properties, we are more interested in looking at the topic from a participant's point of view, rather than looking in detail at small texture features. In other words, we want the properties we use to be important to people, not automated categorisation algorithms.

As such, we looked at the work which followed on from Haralick et al., two important and influential psychophysical studies which attempted to bridge this gap between the texture features and the perceptual descriptions used by participants. The studies by Tamura et al. and Amadasun and King [20, 21] both use images of natural textures taken from the *Brodatz Texture* set, an album of 112 two-dimensional greyscale texture images taken of natural surfaces under controlled lighting conditions [22].

Tamura et al. decided to provide six high-level bipolar scales with which their participants could make their judgements. They sourced these from the literature at the time. These were *coarseness*, *contrast*, *directionality*, *line-likeness*, *regularity* and *roughness*. They then showed participants 16 of the *Brodatz Textures* in a pairwise comparison experiment and asked them to indicate which was 'more' in each of the six conditions, i.e. which was coarser, more directional etc. In comparison, Amadasun and King presented participants with ten textures from the same *Brodatz Texture* set and asked them to rank each on five different scales (again sourced from the literature): *coarseness*, *contrast*, *busyness*, *complexity*, and *texture strength*.

What was clear from these two similar experiments, performed eleven years apart, is that there is little agreement about what perceptual properties should be used to describe textures. Two independent experiments looking into it agreed on only two of the properties used. Also problematic for us was the fact that neither experiment source their properties from participants, so we had no way of telling if given a free choice naïve participants would use those properties to describe the textures presented to them. In addition, both experiments offer low-level texture features which we are uncertain if human observers would consider themselves when categorising images.

These considerations led us to investigate experiments which both involved participant's feedback at their core, and also aimed to offer more high-level texture properties. These are discussed in the next section.

2.3   High-Level Features and Free Grouping

One of the most influential studies into higher-level texture properties was performed by Rao and Lohse and aimed to counter some of the problems present in the earlier experiments [3]. In their experiment, they presented 30 textures which were drawn from the same *Brodatz Texture* set as earlier experiments and asked 20 subjects to group them into as many perceptually similar groups as they desired. This is often described as a 'free grouping' experiment. Following this they used a combination of hierarchical clustering analysis and multidimensional scaling (or MDS) to derive a three-dimensional property space, similar to the RGB colour model which is in common use in computer imagery. The three properties they forwarded as dimensions were *repetition*, *orientation* and *complexity*.

Again though, there is no agreement between studies about how many dimensions are required to accurately define the perceptual space. Long and Leow performed similar experiments to Rao and Lohse's to see whether they could find a texture space which was consistent with human perception [23]. Using both more people and more textures from the *Brodatz Textures* set (60 people and 50 textures) they concluded that a four-dimensional space is required.

Although these results are important and have influenced the course of research in this area, for us, it was the techniques used which were most interesting. The combination of 'free-grouping' experiments and hierarchical clustering provided a simple, relatively

low-cost[2] method of gathering complex, perceptual data. While it is true that MDS has been a common method of determining perceptual spaces [3, 4, 23-26], we were not interested in defining yet another *x*-dimensional space. Rather, we were interested in finding what terms participants use and creating a two-level taxonomy, instead.

More recently, Petrou et al. argue that there was an inherent bias present in the work performed by Rao and Lohse [5]. As the participants had already been asked to rank textures before grouping them, they state that the resultant groups were biased to rely on the properties previously presented to the participants. They go onto argue that it is impossible to give linguistic names to all the properties participants use to identify textures and rather than looking directly at high-level texture properties, one should use thousands of lower-level features instead and then select those which mirror human rankings. While this is an interesting counterpoint, it goes rather against our aim of elucidating what properties naïve participants use to talk about textures. As such, we decided to keep their points in mind when designing our experiments, but would follow a scheme more similar to the one used by Rao and Lohse. We felt confident in doing this as free-grouping experiments have continued to be widely used [4, 26-28].

At this point, we should note that we decided that it was unlikely that we would use images from the *Brodatz Texture* set at the basis for any of our work. Although clearly popular (as seen from the amount of experiments we have already reported which use it) it has some problems. For example, as explained by Emrith, nowhere in the literature are the illumination conditions for the set defined [29]. This is problematic, as Chantler proved that directional illumination can act as a directional filter and as such can vary a texture's characteristics significantly [30, 31].

It is because of the problems with the *Brodatz Texture* set that we turned to look at what work had been performed in this area using other stimuli. Perhaps unsurprisingly, there has been much work done in relation to fabrics and textiles, both in terms of visual and tactile perceptions. As this seemed a popular field of study we investigated this next. This investigation is briefly discussed in the next section.

---

[2] Low-cost in terms of people-hours required, rather than monetary cost.

2.4 Tactile Textures

A good example in this area is the work by Picard et al. who investigated the perceptual dimensions of 'everyday tactile textures' [4]. Picard et al. performed two experiments, the first of which asked participants to forward properties from simply imagining fabrics they interacted with every day. They then used linear regression analysis to examine the relationship between the frequency and order of the properties produced and found a simple relationship: The more frequently the properties occurred, the earlier they were produced. This experiment was interesting for us, as they confirmed in their conclusions that non-experts in fabrics still have access to a large number of properties which they produce in everyday use. In addition, it disagreed with what Petrou et al. state about linguistic properties not being enough to describe surfaces and textures we interact with everyday.

In the second experiment, a free-grouping task was used to sort 24 different car seat materials. These groups were then given a verbal label by each participant. This technique of labelling each group then allowed them to attempt to name the dimensions given when they found a three-dimensional space via MDS. As we've already mentioned, we will not be looking at using MDS, but this technique of labelling the groups was a useful technique for creating a taxonomy.

In a later experiment, Picard went on to see if there was any perceptual equivalence between visual and tactile properties when it comes to texture [32]. She found that there was some perceptual equivalence between the two. In the first experiment, for example, the results showed the same basic perceptual dimensions were used in touch and vision but were processed in different orders.

In the second experiment, participants again managed to produce a large number of properties (75) to describe texture. We believed that this showed that fabrics would be a good area to examine in order to create a wide-ranging taxonomy as participants seemed able to describe their properties with relative ease.

To this end, we found a survey which had been performed by Atkinson which collated results from several different property gathering studies into a single, coherent list. This survey is discussed in the next section.

2.5　Fabric Property Survey

In reference to properties used for fabrics in particular, Atkinson has compiled a comprehensive list of 145 different properties which had been either suggested by experts, or drawn from them via experiments [33]. This list of properties was taken from 15 different studies and is by far the most comprehensive that we came across. In addition, this core set of properties had already been used successfully to create a set of scales with which to rate how accurately participants rate textiles in different modalities [34].

The final list provided has 69 terms in it, with properties removed if they were not judged to be a property which could usefully communicate the fabric between people. For example, *expensive* was removed as two observers might judge it differently. Indeed, any properties judged to be hedonic, value-based, describing the fabric type rather than a property or describing function was removed. Usefully for us, all of the properties suggested in this list are what we would describe as high-level as they describe properties of the fabrics in more human terms. The full list of accepted and rejected terms is presented in *Table 2.1*.

There are some issues of this list presented by Atkinson, however. The list itself was compiled by an expert in fabrics and contains words that the author of this thesis (not being a fabric expert himself) had no knowledge of. As we wanted any taxonomy we were going to use to be fully understandable by naïve observers, this was concerning.

Still, as this was such a thorough study, it seemed like an excellent place to start. To this end, we decided to investigate this survey further. This investigation is the topic of the next chapter.

Accepted Properties		Rejected Properties	
Boardy	Loose	Absorbent	Non-Directional
Bumpy	Lumpy	Acute	Obstructed
Clingy	Malleable	Animal	Open
Coarse	Matte	Artificial	Ordinary
Cold	Mossy	Baby	Peach Like
Crisp	Non-Stretchy	Bonny	Pilling
Crumpling	Pliable	Bright	Plain
Crushable	Raised	Budge	Pleasant
Delicate	Raspy	Clammy	Plious
Dense	Relief	Classic	Plush
Downy	Resilience	Cloth	Practical
Dry	Ribbed	Compact	Prickly
Elastic	Rigid	Complicated	Protective
Even	Rough	Conservative	Quiet
Falling	Sheer	Cotton	Raw
Fine	Shiny	Cozy	Rib
Firm	Sleek	Crepe	Rugged
Flexible	Slippery	Decency	Rugous
Floating	Smooth	Directional	Satin
Flowing	Snagging	Dull	Satiny
Fluffy	Soft	Durable	Scratchy
Full	Solid	Elegant	Silk
Fuzzy	Sparse	Expensive	Silky
Grainy	Spongy	Exquisite	Simple
Granulous	Springy	Fancy	Sleazy
Greasy	Starchy	Fashionable	Somber
Hairy	Sticky	Flamboyant	Strong
Hard	Stiff	Flannelette	Technical
Harsh	Stretchy	Fleece	Tender
Heavy	Supple	Granite Like	Unbearable
Hot	Synthetic	Handwork	Unique
Irregular	Thick	Irritating	Unpleasant
Light	Thin	Jeans	Velvet
Limp	Tight	Luxurious	Velvety
	Tough	Mellow	Vulgar
		Modern	Weighting
		Natural	Wool
		Nervous	Woolly

Table 2.1 - All accepted and rejected terms proposed by Atkinson. The shaded columns show the terms from his literature review which were rejected for the reasons explained in *Section 2.5*.

2.6   Conclusions

In this chapter, we have briefly investigated what work has already been done to determine the properties which are important for people to describe and discuss textures and images.

We began by discussing early work attempting to identify texture features which could be used for categorisation in *Section 2.2*. Following this, we discuss influential experiments which gathered high-level features via free grouping techniques in *Section 2.3*. This led us to investigate work using fabrics and textiles in *Section 2.4*, before presenting the results of a wide-ranging survey performed by Atkinson which incorporated the results of 15 different fabric studies, in *Section 2.5*.

During this chapter we discussed two important studies which offered interesting techniques for gathering participant feedback. From Rao and Lohse we discussed the free grouping task and the hierarchical clustering analysis which could be used to find how participants relate different stimuli such as words or fabrics. From Picard et al. we discovered an extension to this work which allowed an experimenter to attempt to name the groups by asking participants to label their groups as they make them.

These techniques were useful, as our main conclusion after investigating the literature was there was no property list perfect for our use. While Atkinson's list was close, we were unsure if it would be suitable for use with English-speaking, naïve observers. In addition, 69 different terms would be impractical to investigate in any detail. To this end, in the next chapter we will investigate this list in more detail and try to overcome these problems.

Chapter 3

A Taxonomy of Surface Properties

3.1 Introduction

As we showed in the previous chapter, there are many different high-level properties, scales and taxonomies present in the literature, with much work done in the fabric domain. These are accompanied by just as many different techniques, methodologies and experimental designs which have been used to look into these properties. As we don't know at this stage which properties are important to non-expert observers we want to investigate where best to focus our attention and which of these techniques to investigate further.

We also discussed that Atkinson (an expert in fabrics) has performed a survey of this literature and provided a list of 69 property terms [33]. As this list was compiled by a fabric expert, we want to ensure it is satisfactory and complete from the perspective of a non-expert who might not use the same technical language, before creating a more focused set of words to narrow discussion in the next chapter.

As such, the goal of this chapter is to create a simple taxonomy of property terms which can be used to narrow the focus of our search. This will, in turn, inform the selection of which properties the rest of this thesis will investigate.

To formally define the criteria which we will use to direct and then evaluate the work presented in this chapter, we want our taxonomy to:

(C3.1)   be understandable by English speaking non-experts (as we will be using such participants in our experiments),

(C3.2)   be representative of the whole property space (so we don't miss any properties),

(C3.3)   provide a concise set of properties (to narrow our investigation), and

(C3.4)   be well distributed across the space (to avoid repetition from investigating two similar areas).

This chapter will therefore look at extending and grouping Atkinson's properties, for the reasons discussed in *Section 3.2*, to create a taxonomy which fulfils these criteria. In

Section 3.3 we look at criteria *(C3.1)* to ensure that the properties are understood and used by English-speaking, non-expert participants. Following that, in *Section 3.4* we ensure that the property list is representative of the whole space to fulfil criteria *(C3.2)* before presenting a non-expert property list in *Section 3.5*. Finally, in *Section 3.6* we present the process by which we clustered this non-expert property list to create a taxonomy which was both concise *(C3.3)* and well distributed *(C3.4)* using the grouping and verbal labelling techniques discussed in *Chapter 2*.

Before we look into creating our taxonomy, however, we must first discuss why Atkinson's list of properties is not sufficient for our use, which is outlined in the next section.

3.2   Atkinson's Property List

As mentioned in the previous chapter, Atkinson compiled a list of property terms [33]. These were derived from a survey of the literature concerning what property words different groups use to describe fabrics. Atkinson's properties are listed in alphabetical order in *Table 3.1*.

Atkinson Property List				
Boardy	Falling	Harsh	Raspy	Sparse
Bumpy	Fine	Heavy	Relief	Spongy
Clingy	Firm	Hot	Resilience	Springy
Coarse	Flexible	Irregular	Ribbed	Starchy
Cold	Floating	Light	Rigid	Sticky
Crisp	Flowing	Limp	Rough	Stiff
Crumpling	Fluffy	Loose	Sheer	Stretchy
Crushable	Full	Lumpy	Shiny	Supple
Delicate	Fuzzy	Malleable	Sleek	Synthetic
Dense	Grainy	Matte	Slippery	Thick
Downy	Granulous	Mossy	Smooth	Thin
Dry	Greasy	Non-Stretchy	Snagging	Tight
Elastic	Hairy	Pliable	Soft	Tough
Even	Hard	Raised	Solid	

Table 3.1 - The 69 fabric properties proposed by Atkinson

This list suffered from five important problems for use in the research presented in this thesis:

(i)   Some of the properties were elicited from expert groups such as designers or fabric creators. This might lead to terms not understood by non-experts, which is counter to criteria (C3.1).

(ii)   Four of the surveys used to form this list were in languages other than English (Japanese and French) which were then translated for publication. As all our future participants were to be English speakers, we wanted to ensure that no meaning was lost in translation. Again, this was counter to criteria *(C3.1)*.

(iii)   The studies which made up the list came from a wide variety of different fabric fields, such as shirt material, bag fabrics, or car seat coverings. We were concerned, therefore, that some of the words might be domain specific and not applicable to a wide array of fabrics and surfaces, counter to criteria *(C3.2)*.

(iv)   This list had too many properties to realistically investigate in a sensible timeframe, counter to criteria *(C3.3)*.

(v)   Several of the words in the list appeared, at least in an initial inspection, to be similar (i.e. *Furry* and *Fuzzy*). This was counter to criteria *(C3.4)*.

It was because of the above issues that we decided to create our own taxonomy which would be more suitable for our use. The first step of creating this taxonomy was to address the first two problems mentioned above. This is discussed in the next section.

3.3   Is Atkinson's List Understood by Non-Experts?

The purpose of this study was to make sure our taxonomy fulfilled the first criteria *(C3.1)*. That is, we wanted to ensure that any taxonomy we were to use was readily understandable by non-expert English speakers. We therefore had to ensure that the 69 words gathered from the literature survey (shown in *Table 3.1*) were suitable to use in our taxonomy.

3.3.1   Participants

In order to obtain a more balanced selection of participants (to make our results as representative as possible as per *(C3.2)*), this study was conducted at two different locations. Ten participants were sourced from University College London and twenty at Heriot-Watt University Edinburgh for a total of thirty participants.

The participants were all undergraduate university students, sourced from advertisements sent around both universities. We excluded anyone who had a background in design or manufacture of fabrics. In addition, all the participants' first language was English. All of the participants were aged between 18 and 25. Of the 20 Edinburgh participants, 8 were female and 12 male. Of the London participants, 4 were female and 6 were male, making a total of 12 female and 18 male participants.

Every observer was asked to read, complete, and sign a consent form similar to the one presented in *Appendix A* before they took part in the experiment.

3.3.2   Procedure

To make the study easy to present to a participant, the 69 words selected were printed in black onto both sides of identical pieces of heavy card. These were the same size as a standard business card at 85mm wide by 54mm tall. Each word was printed exactly the same size, in the same font and was centred on the piece of card to ensure there was no obvious difference between each card beyond the words themselves, as shown in *Figure 3.1*.
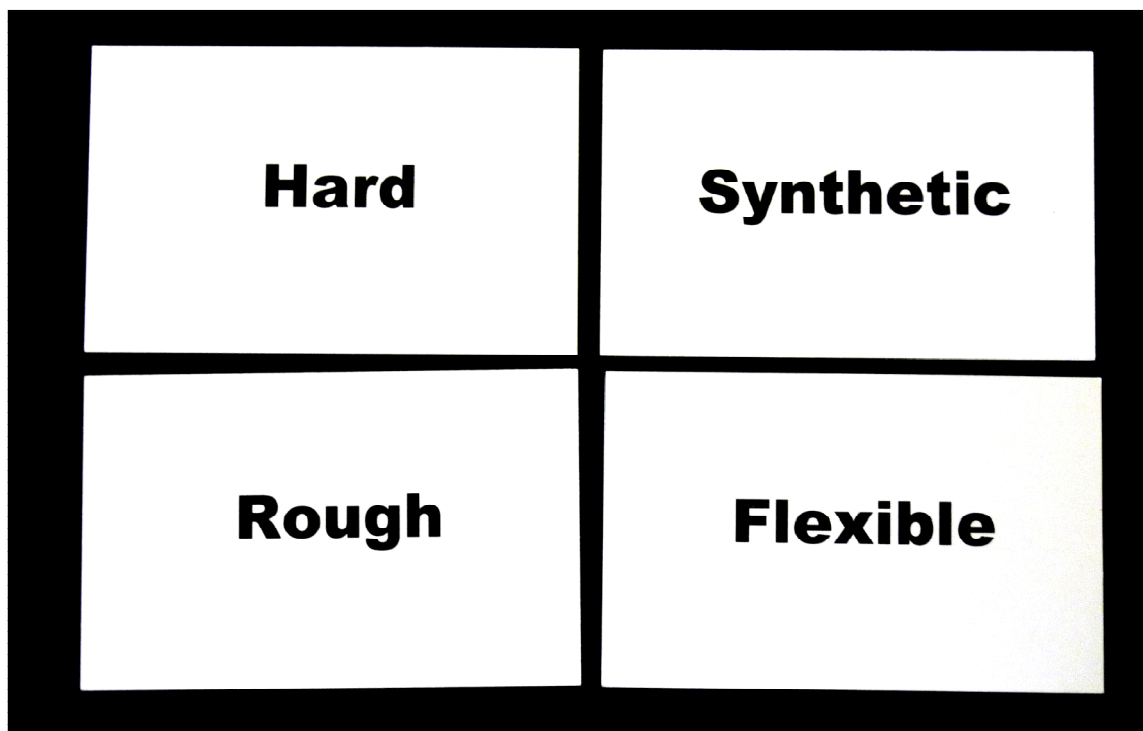
Figure 3.1 - Four example stimuli used in the property studies

This stack of 69 cards was then shuffled into a random order and presented to participants in a single stack. They were asked to sit down at a table which had three areas delineated by masking tape. These areas were labelled, from left to right, 'Words

You Know and Use Often', 'Words You Know but Don't Actively Use' and 'Words You Don't Know or Understand'. Their task was to then sort the stack of words into the three different areas of the table however they saw fit without any time limit. They had to make a decision for each word.

After participants made their selections, we recorded which category that they put each word in and gave them a score. Any words in the 'Words You Know and Use Often' group got a score of 1, those in 'Words You Know but Don't Actively Use' were scored as 0 and the 'Words You Don't Know or Understand' group words got a score of -1.

Once we had all of the participants' responses, we then totalled the score for each word to get a rating for each of Atkinson's 69 properties. We then ordered the properties by how well understood they were. Thanks to this simple scoring system, we knew that any word with a non-positive score was understood by fewer people than those who actively used it and could therefore be discarded.

3.3.3   Results

The results of this study are detailed in full in *Figure B.1* with each participant's scores for each word shown. For a more compact version of the results, *Figure B.2* shows the total scores of every word in decending order.

Using these results, we were able to decide which of these properties fulfilled criteria *(C3.1)* and were readily understandable by English-speaking non-expert participants and which were not.

The first thing to note is that there were only three words which every participant said they use often to describe fabrics. These were *Smooth*, *Soft* and *Thick*. In total, there were 52 words which got positive scores. In contrast, there were 17 words from Atkinson's list with non-positive scores, showing that few people understood their meaning or would use them to describe fabrics. These are shown in *Table 3.2*.

As stated in *Section 3.3.2*, we decided than any words which garnered a non-positive score in this study would be discarded from any future work. Although this is a pragmatic decision, we believe it was a sensible one as the weight of evidence shows that on average they are not well understood and this might cause participants problems in future. We therefore decided that the 17 words in *Table 3.2* would no longer be used.

Dry	Harsh	Floating	Malleable	Pliable	Resilience
0	0	-1	-1	-4	-4
Sheer	Granulous	Starchy	Full	Downy	Sparse
-4	-6	-8	-10	-11	-11
Mossy	Falling	Boardy	Raspy	Relief	
-13	-16	-18	-20	-22	

Table 3.2 - Properties from Atkinson's list which were removed (i.e. those with non-positive scores). The properties with shaded backgrounds were conflicting results with *Section 3.4*

Now we'd discovered which properties from Atkinson's initial set we needed to remove to make sure the list was understandable by English-speaking non-experts, we next needed to investigate whether any properties needed to be added, which we will discuss in the next section.

3.4   Is Atkinson's List Complete and Representative?

At the same time as the previous study was being performed, Orzechowski was performing a parallel study to discover if the initial word set mentioned in *Section 3.2* was missing any words which were used by naïve participants and not by experts [12]. This was used to fulfil criteria *(C3.2)*.

In his study, Orzechowski discovered 429 properties used by non-expert participants. Many of these words were considered to be hedonic, emotive or opinion based (i.e. *Pretty*, *Nice* or *Boring*), however, and were removed. Orzechowski therefore decided to accept only 65 of them. Of these, many overlapped with Atkinson's list (*Table 3.2*) so there were 26 new words to add to our list. These are shown in *Table 3.3*.

Bendable	Breezy	Brittle	Creasable	Crinkly	Crunchy	Dry
Floaty	Furry	Glossy	Gritty	Grooved	Harsh	Malleable
Natural	Noisy	Plain	Pliable	Ridged	Sandy	Scratchy
Stringy	Sturdy	Tangly	Textured	Warm		

Table 3.3 - 'Discovered' properties from Orzechowski which will be added to Atkinson's property list. The properties with grey backgrounds are conflicting results with *Section 3.3.3*. It is worth noting at this point that the properties with shaded backgrounds are those which conflicted with the results from *Section 3.3.3*.

One study was stating these four words should be removed while the other was stating they should be added.

We decided that Orzechowski's study should take precedence as a participant offering a word themselves was a stronger indicator of use than rating a list of pre-defined words.

3.5 Final Non-Expert Property List

Combining the two studies from *Section 3.3* and *3.4*, we now had a list of 78 words which we were able to demonstrate that most English-speaking, non-expert participants understood and used on a regular basis to describe fabrics and surfaces. This list therefore fulfilled criteria *(C3.1)* and *(C3.2)*.

Non-Expert Property List				
Bendable	Elastic	Hard	Raised	Sticky
Breezy	Even	Harsh	Ribbed	Stiff
Brittle	Fine	Heavy	Ridged	Stretchy
Bumpy	Firm	Hot	Rigid	Stringy
Clingy	Flexible	Irregular	Rough	Sturdy
Coarse	Floaty	Light	Sandy	Supple
Cold	Flowing	Limp	Scratchy	Synthetic
Creasable	Fluffy	Loose	Shiny	Tangly
Crinkly	Furry	Lumpy	Sleek	Textured
Crisp	Fuzzy	Malleable	Slippery	Thick
Crumpling	Glossy	Matte	Smooth	Thin
Crunchy	Grainy	Natural	Snagging	Tight
Crushable	Greasy	Noisy	Soft	Tough
Delicate	Gritty	Non-Stretchy	Solid	Warm
Dense	Grooved	Plain	Spongy	
Dry	Hairy	Pliable	Springy	

Table 3.4 - The list of 78 non-expert properties we proposed as they were demonstrably used by English-speaking, non-expert participants to describe fabrics, fulfilling criteria *(C3.1)* and *(C3.2)*

These 78 words are listed in *Table 3.4* and are the words which we used in to create the taxonomy as discussed in the next section.

3.6 Creating a Taxonomy via Grouping

The purpose of the study reported in this section was to take our non-expert property list (shown in *Table 3.4*) and create a simple taxonomy to fulfil criteria *(C3.3)* and *(C3.4)*.

To create this taxonomy, we decided to use clustering methods (discussed in *Section 3.6.5*) on our non-expert property list so we could determine groups which were well separated from each other with meanings that had little, if any, overlap. For each group, we then picked a single word to represent it, forming a simple two-level taxonomy.

We discuss this procedure in more detail in this section, beginning with the study design.

3.6.1   Participants

For this experiment, we endeavoured to obtain participants from a wide a background as possible, to make sure the groupings were representative, as per criteria *(C3.2)*. Seven of the participants were undergraduate students from the local Edinburgh universities (Heriot-Watt University, Edinburgh Napier University and The University of Edinburgh) and two were post-graduate students. The other eleven participants were sourced from the local area and were in full-time employment.

The participants were all sourced from advertisements around Edinburgh and none had a background in design or manufacture of fabrics. In addition, all the participants' first language was English, although five of them were not born in the UK. The participants' ages ranged from 18 to 35. Also, to avoid any possible prejudice in the grouping (as pointed out by Petrou et al. [5]) none of them had taken part in the study described in *Section 3.3* and as such had no properties suggested to them before taking part.

We also attempted to get an even balance of male and female participants. Of the final twenty participants, 12 were female and 8 were male.

Every observer was asked to read, complete, and sign a consent form similar to the one presented in *Appendix A* before they took part in the experiment.

3.6.2   Procedure

To gather the required clustering data we performed a free-grouping experiment because, as discussed in *Chapter 2*, it is a popular and successful way of creating groupings.

To this end, we updated the set of cards we created in *Section 3.3.2*. We removed the 17 properties cards which we had decided were not well understood (*Table 3.2*) and printed the new 26 properties onto identical cards (*Table 3.3*).

This new set of 78 property cards (*Table 3.4*) was then shuffled into a random order and presented to participants in a single stack. They were asked to sit down at a large, empty table and to group the properties however they saw fit, with as many or as few groups as they wanted. The groups could be of any size. They were told they would be asked a few questions about the task afterwards, but were not informed what form those questions would take

After they finished grouping all of the words, which took them between 30 and 45 minutes, they were asked to indicate their favourite and least favourite groups and a

single property in each group which was most 'representative' of that group. They did this using a set of pre-made symbols. An example grouping is provided in *Figure 3.2* to show how participants indicated their choices. This was inspired by the verbal label technique use in work by Picard et al. [4] which was discussed in *Chapter 2*.
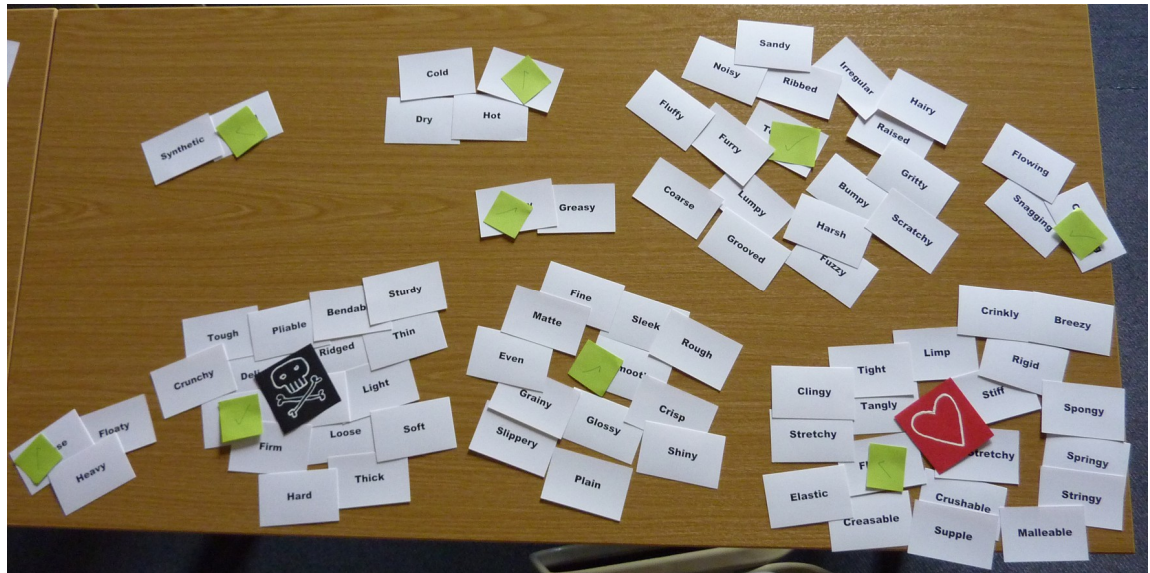
Figure 3.2 - Example property grouping by a participant. Note the small square cards with ticks show which words are representative and the cards with a heart or skull and crossbones show favourite and least favourite group, respectively

Once they were done, participants' answers were photographed for future reference and their groups recorded. Finally, the results of all participants were collated and a matrix created which allowed us to see the dissimilarity between each property and every other property in the list, so we could cluster them into groups. In addition each property had a score to show how representative it was of the group it was in.

3.6.3   Creating a Dissimilarity Matrix

To record our participants' grouping data, we used a dissimilarity matrix. A dissimilarity matrix (sometimes called a distance matrix) is used to describe pairwise dissimilarity between N items. It is a square *N*-by-*N* matrix with each element equal to the dissimilarity between its row and its column. In our case, as we are using properties, each element therefore defined how dissimilar the property labelled on its row ($D_{(i)}$) is to the one labelled on its column ($D_{(j)}$).

As we had 20 participants, we needed to aggregate the data so that each element in the matrix defines the collective the dissimilarity for all twenty participants. The procedure to do this was as follows:

(i)   For each participant a 78-by-78 matrix was created with each row and column representing one property in the full set. Each element was set to 0.

(ii)   This matrix was populated to create an Occurrence Matrix, a binary matrix where any property $P_{(i)}$ grouped together with any property $P_{(j)}$ had the value '1' inputted at both intersections. See *Figure 3.3* for an example.

(iii)   These twenty occurrence matrices were summed, element-by-element, to give a similarity matrix which showed how many people grouped each property with each other property.

(iv)   Finally, to convert this similarity matrix into a dissimilarity matrix, the value of every element was deducted from the total number of participants to give a dissimilarity matrix. In our case, this is $D_{(i,j)} = 20 - P_{(i,j)}$.

	P_1	P_2	P_3	P_i	P_{N-1}	P_N
P_1						
P_2						
P_3						
P_j				1		
P_N						

Figure 3.3 - Example Occurrence Matrix

A full dissimilarity matrix from this experiment is available in *Figure B.3* with a smaller zoomed in sub-section of only 15 words provided in *Figure B.4*.

As can be seen from this sub-section 'Bendable' and 'Breezy' was grouped into the same group by no participants (a dissimilarity of 20), where as 'Creaseable' and 'Crumpling' was put into the same group by 15 participants (a dissimilarity of 5).

3.6.4   Is Our Data Non-Metric?

Usually dissimilarity matrices are thought of as dealing with distances in Euclidian space and therefore being metric data. This, however, is not necessarily true when using psychophysical data. For data to be metric it must comply with four basic conditions:

(3.1) $D(i, j) \geq 0$

A distance cannot be less than zero (Non-negativity)

(3.2) $D(i, j) = 0$ if and only if $i = j$

Only elements on the diagonal can have a zero distance (Identity of indiscernibles)

(3.3) $D(i, j) = D(j, i)$

The matrix must be symmetric along the diagonal (symmetry)

(3.4) $D(i, k) \leq D(i, j) + D(j, k)$

Distances between any three elements must form a valid triangle, i.e. the sum of the lengths of any two sides must be greater or equal than the length of the remaining side (triangle inequality)

If there is a single instance of any condition being broken, the data is not metric. Our data, available in full in *Figure B.3*, obeys the first three conditions, in that we have no negative distances, only the diagonals have zero distances and the matrix is symmetric. It does, however, break the condition *(3.4)* in certain circumstances. One example is provided here:

(3.5) $D(i, k) > D(i, j) + D(j, k)$ where $i = 42, j = 1, k = 48$ as*:*
        $D(i, k) = 20, D(i, j) = 4$ and $D(j, k) = 1$

Therefore, our data is non-metric which has implications for the choice of clustering algorithm, as we discuss in the next section.

3.6.5   Clustering the Properties

There are many different approaches to clustering dissimilarity matrices, as mentioned in *Chapter 2*, and often the decision of which to use is pragmatic as they all give different clusters.

In our case, we decided to use an agglomerative clustering method as they have been used successfully in previous experiments into clustering complex perceptual spaces [3, 28, 29]. Agglomerative clustering uses a bottom-up approach to creating clusters,

starting with N singleton clusters (where N is the number of items being clustered). These clusters are then merged recursively, pair by pair, using a linkage function until all the items are in a single cluster [35].

It should be noted at this point that this produces what Ding and He define as 'hard clusters', meaning that each property can only be assigned to one group [35]. They state that in real life many points might be near the boundary between clusters and as such might be better assigned partially to different clusters. In our case, however, we wanted each property in a single cluster for our final taxonomy, so this clustering method was sufficient.

There are four commonly used linkage functions which are suitable for use with non-metric data. These are the *single link*, the *complete link*, the *group average link* (sometimes called the *Unweighted average distance* or *UPGMA*) and the *weighted average link*. As defined by Gordon [36], these four functions are generally defined as follows:

(3.6)
$$d(C_i \cup C_j, C_k) = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \gamma |d(C_i, C_k) - d(C_j, C_k)|$$

Where $d(C_i, C_k)$ is the dissimilarity between two different classes, and where α and γ define the different linkage mechanisms. Table 3.5 defines the different values of α and γ for the different link methods, with w_i being the weight associated with class C_i, which is usually the number of objects it contains.

Linkage Function	α_i	γ
Single Link	$\dfrac{1}{2}$	$-\dfrac{1}{2}$
Complete Link	$\dfrac{1}{2}$	$\dfrac{1}{2}$
Group Average Link	$\dfrac{w_i}{(w_i + w_j)}$	0
Weighted Average Link	$\dfrac{1}{2}$	0

Table 3.5 - Parameter values of the four linkage functions for non-metric data

The choice of linkage function is a pragmatic one, so for our data we chose to use the *group average link* function as it is the simplest method and has been used successfully by Halley with their data [28].

3.6.6   Displaying the Clusters with Dendrograms

Usually results from hierarchical clustering algorithms are displayed in a type of tree diagram called a dendrogram. These tree diagrams allow us to easily visualise the complex information which can result from clustering algorithms.

Each node at the right-most level of the diagram represents a single element, or node, which has been clustered with lines joining different elements representing how similar they are. Using this information, we can perform simple, qualitative checks about the structure of the clustered data. For example, the sample dendrogram in *Figure 3.4* shows what appear to be three clear groups.
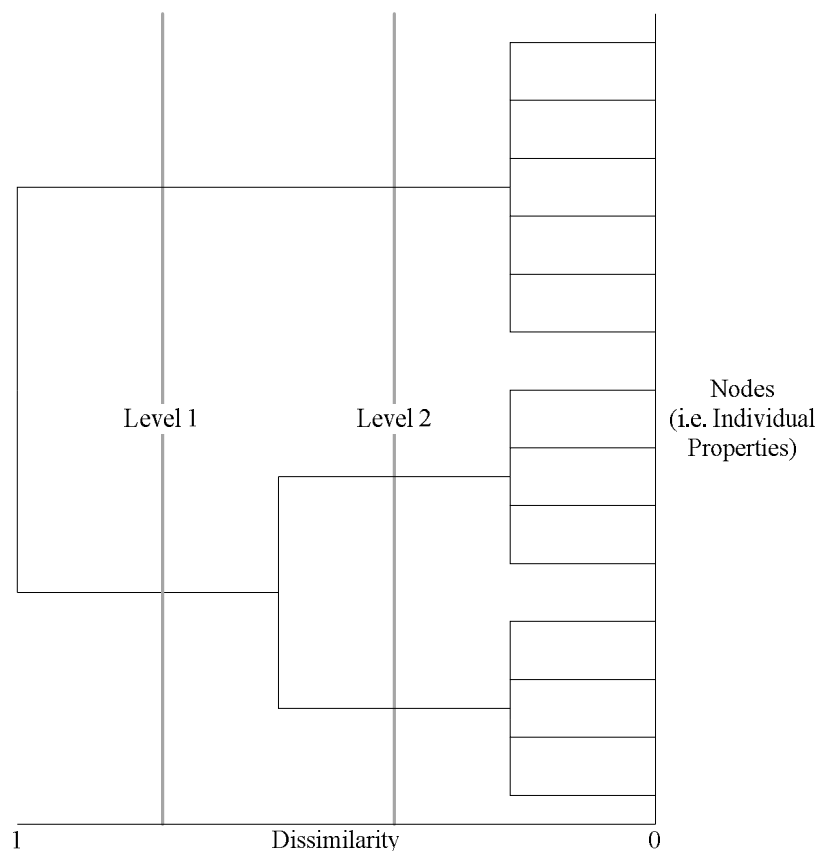
Figure 3.4 - Example Dendrogram, showing three apparent groups and two different 'cutting' lines where it could be split into different clusters

In the example above, all three apparent groups join at the same dissimilarity level, and then the lower two groups are more similar to each other than they are with the topmost group. So in our example dendrogram, if we decided we wanted two groups, we could cut the dendrogram at *Level 1*. If, however, we decided to split the dendrogram at *Level 2*, our clustering would return three groups instead. It is with this technique we split our grouping data to form the taxonomy as discussed in *Section 3.6.8*.

3.6.7   Group Sizes

Before clustering the properties, we checked to see if there were any obvious patterns discernible from the number groups participants made. *Figure 3.5* shows a histogram of the different numbers of groups.

This showed that there was a wide range of different group sizes, as for 20 participants there were 13 different numbers of groups used. While the most popular number of groups was 18 only three participants used that many groups. It was clear, therefore, that there was no optimal number of groups or optimal group size for our participants which could inform how many groups we should return from our clustering analysis.
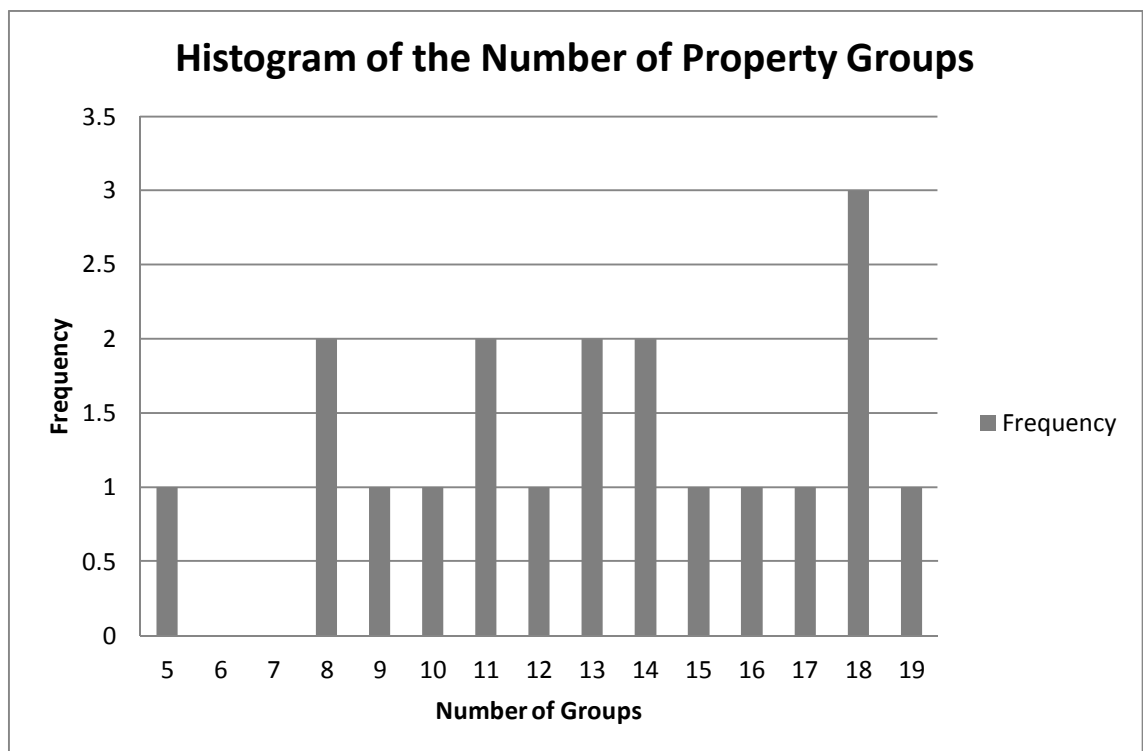
Figure 3.5 - A histogram of the number of property groups used by participants to categorise all 78 property words

As we could not find an optimal number of groups to use in our clustering analysis, we decided to make a pragmatic decision as discussed in the next section.

3.6.8   Results - Finalising the Taxonomy

The dendrogram in *Figure 3.6* shows how our non-expert properties have been clustered into different groups. Each node at the right side of the dendrogram represents a single property and each vertical joining line going across the diagram represents where a group is split in two.
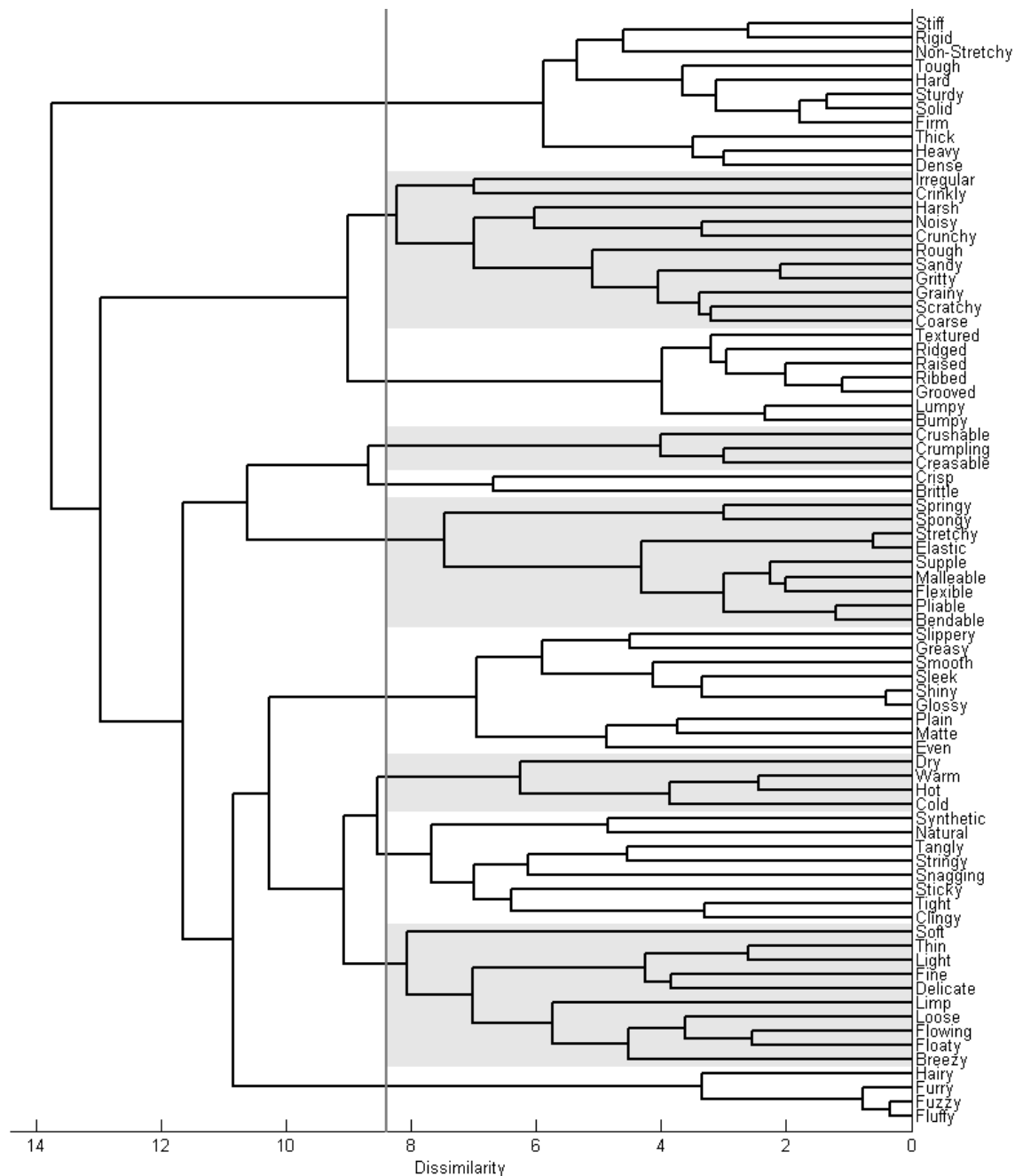
Figure 3.6 - A dendrogram showing participants property grouping data.The horizontal axis shows every property in the experiment and the vertical the dissimilarity at which the clustering occurs. For clarity, the clusters have been deliniated by alternating background shades below the 'cutting' line which determines the groups.

At this stage, we noted one important property of the dendrogram: there were no clean, dominant groups which could be easily identified, as per our ideal dendrogram example. This implied that there was no perfect grouping which our participants followed, which was to be expected from the lack of agreement on group sizes from our participants as discussed in the previous section.

We decided to split the dendrogram into the 11 groups which are present in *Figure 3.6* for two reasons:

a) If we used fewer groups, properties which appeared to be well separated would be put into the same group. For example, if we split the properties into ten groups, *Cold*, *Hot*, *Warm* and *Dry* would merge with the group containing *Tangly*, *Natural* and *Synthetic*.

b) If we split the properties into more than eleven groups, it would split a group we believed to be well clustered already. The next split would remove *Irregular* and *Crinkly* from the group containing *Noisy*, *Crunchy* and *Rough*.

As shown in *Figure 3.2*, in addition to the dissimilarity data we also gathered scores of how representative each word was in the eyes of our participants. This allowed us to pick a single word from each group to represent it as a higher-level description by simply counting how many times a word was labelled as representative. We then picked the word with the highest score from each group. The scores for the top two scoring words per group are show in *Table 3.6*.

	Score	Group
Solid	10	9
Heavy	4	9
Coarse	7	6
Irregular	7	6
Rough	6	6
Textured	12	4
Grooved	2	4
Crumpling	5	8
Creasable	2	8
Brittle	3	3
Crisp	2	3
Flexible	9	1
Malleable	6	1

	Score	Group
Smooth	8	10
Glossy	6	10
Warm	9	7
Hot	5	7
Natural	8	5
Sticky	8	5
Tangly	6	5
Delicate	8	2
Fine	7	2
Fuzzy	5	11
Fluffy	4	11

Table 3.6 - **Top two representative property scores for each property group from top to bottom on the dendrogram in *Figure 3.6*. The shaded groups match the shaded areas in *Figure 3.6*.**

Note that there are two groups with a tie for the most representative word. At this stage we decided to pick one of them, but kept a record of all these words as future work might enlighten which one is more suitable in the situation.

Figure B.5 shows a simplified version of the dendrogram in *Figure 3.6* where all the nodes below where the cut took place have been removed. Each bottom node therefore represents a single group of properties which is denoted by the eleven representative

words chosen: *Fuzzy*, *Delicate*, *Natural*, *Warm*, *Smooth*, *Flexible*, *Brittle*, *Crumpling*, *Textured*, *Coarse* and *Solid*.

Finally, our finished two-level taxonomy is presented in *Table 3.7*, below. It shows each of the eleven groups determined by our clustering techniques with the highest scoring representative word at the top level, in bold. The words are ordered by their scores, so that the lower in each group they are, the less representative of that group they are.

Solid	Coarse	Textured	Crumpling	Brittle	Flexible
Heavy	Irregular	Grooved	Creasable	Crisp	Malleable
Sturdy	Rough	Raised	Crushable		Pliable
Rigid	Noisy	Bumpy			Springy
Stiff	Crinkly	Lumpy			Stretchy
Dense	Scratchy	Ribbed			Elastic
Thick	Sandy	Ridged			Bendable
Hard	Crunchy				Spongy
Tough	Grainy				Supple
Firm	Harsh				
Non-Stretchy	Gritty				

Smooth	Warm	Natural	Delicate	Fuzzy
Glossy	Hot	Sticky	Fine	Fluffy
Plain	Cold	Tangly	Flowing	Hairy
Shiny	Dry	Clingy	Soft	Furry
Even		Synthetic	Loose	
Matte		Stringy	Floaty	
Sleek		Snagging	Thin	
Slippery		Tight	Light	
Greasy			Limp	
			Breezy	

Table 3.7 - Final non-expert taxonomy of properties. The bold words are the top-level, representative words for the properties that are shown under them. The shaded groups match the shaded areas in *Figure 3.6*.

This taxonomy, thanks to the free-grouping and clustering techniques we have described in this section, now fulfils the final two criteria presented in the introduction. Thanks to the representative words for each group we have a more concise set of properties to work with *(C3.3)* and, as the dendrogram in *Figure 3.6* shows, the groups are well distributed across the property space we have, fulfilling criteria *(C3.4)*.

3.7 Conclusions

In this chapter we have described in detail the process we used to create a two-level taxonomy of property terms (show in *Table 3.7*) which fulfils our four basic criteria. This comprises the main output of this chapter.

As discussed in the introduction, we wanted our taxonomy to:

(C.1)   be understandable by English speaking non-experts,

(C.2)   be representative of the whole property space,

(C.3)   provide a concise set of properties, and

(C.4)   be well distributed across the space.

We therefore went through a three-step process inspired by work performed by Rao and Lohse [3] and Picard et al. [4, 32] in an attempt to fulfil all of these criteria.

We updated the properties provided by Atkinson to be certain they were understood by English speaking, non-expert participants (*Section 3.3* and *(C3.1)*) and then used data provided by Orzechowski to make sure we were not missing any properties (*Section 3.4* and *(C3.2)*).

Finally, in *Section 3.6* we performed a free-grouping experiment and clustered the properties into 11 groups and found representative properties for each group. This allowed us to create the two-level taxonomy shown in *Table 3.7*. These 11 properties provide a much more concise set *(C3.3)* which are well distributed across the space we started with *(C3.4)*, as can be seen from the dendrogram presented in *Figure 3.6*.

As we intend to present stimuli via digital methods, it is with these 11 properties we can now go forward and look into what digital tools we can use to communicate them effectively. As such, the next chapter will look into the advantages and disadvantages of currently available technology and use this information to make a decision into which properties will be studied in more depth in the rest of this thesis.

Chapter 4

Digital Tools

4.1 Introduction

In the previous chapter we presented our two-level taxonomy of property terms aimed at narrowing the focus of our search for which property to investigate in this thesis. This provided us with a set of 11 different property groups which we could investigate. Our next aim was to find two of these groups which we could investigate the interaction between.

Therefore, as we had decided from the beginning to present stimuli via digital methods, it was important to investigate which of these eleven groups we could effectively communicate using digital tools.

The goal of this chapter is to present the advantages and disadvantages of different display and interaction methods, before coming to a decision about which of the properties we will investigate in more detail later in this thesis. There are four main criteria with which we will judge any presentation method. We want our presentation method (now or in the near future) to be:

(C4.1)   available,

(C4.2)   affordable,

(C4.3)   simple to use, and

(C4.4)   suitable for presenting a property.

Therefore in this section we will be looking at the different groups of properties and examining what tools we could use to represent these. In general we have broken these up into three categories which are discussed in *Section 4.2*. These three categories are then discussed in more detail in *Section 4.3* to *4.5*, first looking at tactile properties, then physically-simulated properties, before finally discussing visual properties.

To begin, the first section will look at the groups in *Table 3.7* and see into which of these three groups they fall.

4.2   Property Categories

As mentioned in the introduction, the purpose of this section is to determine how best to represent each of the eleven property groups outlined in *Table 3.7*. To this end, we decided to categorise each of these groups into one of three different presentation methods and discuss each in turn in the subsequent sections.

The three presentation categories we chose to investigate in detail were 'tactile' properties, or those that require some sort of physical, touch-based feedback, 'physically-simulated' properties, or those which require an active physical simulation, and 'visual' properties which can be represented adequately with an image or a solid three-dimensional surface.

At this stage, we were aware that not all of the property groups easily fit into one of these three categories. For example, the *Smooth* and *Coarse* groups could be represented either visually or with tactile feedback. In most cases, therefore, we picked a single category which best represented the property group, but in these two cases we put these two property groups into both categories as they both could represent the groups adequately. The categories are presented in *Table 4.1*.

Tactile	Physically-Simulated	Visual
Solid	Crumpling	Coarse
Coarse	Brittle	Textured
Flexible	Delicate	Smooth
Smooth	Fuzzy	
Warm		

Table 4.1 - A table showing how we have categorised the different property groups. These property groups are from *Table 3.7* and have been put into one of three groups, showing which would be best for representing them. The groups with the shaded backgrounds are those which are repeated over multiple groups. N.B. The *Natural* group has been excluded for reasons discussed below.

One group was not represented in *Table 4.1*, however. The *Natural* property group was not added to the table as a decision was made early into this process that we would not investigate this group further. This decision was due to the nature of the group itself. We believed *Natural* to be a more judgement-based property which would be hard to encapsulate in a meaningful way. In addition, the group was, in our opinion, the least coherent of the 11 groups and therefore the hardest to represent.

With this decision made, we then investigated the three categories generally, in order to represent the property groups within. We present discussions on these investigations in the next three sections, starting with the tactile category in the next section.

4.3 Tactile Properties

In this section we will discuss the representation methods we investigated to represent the possible tactile property groups from our list, *Warm*, *Solid*, *Flexible*, *Smooth*, and *Coarse*. This discussion is split into three main sections, thermal feedback, vibration feedback, and haptic pens.

4.3.1 Thermal Feedback

Briefly considered to represent the *Warm* group was the use of thermal feedback. In essence this would have been a small metal plate which we could have controlled the temperature of. This could then be touched by a participant and altered for different fabrics to give the participant an idea of the temperature of the item.

This idea, however, was quickly rejected. Our main objection to this representation method was its simplicity. We didn't believe there was enough scope for investigating this property in a novel manner. Beyond this, to represent the fabric this way would require more than just temperature, but also visual or other forms of haptic feedback as well.

Due to these concerns, we did not investigate this option in depth and considered the *Warm* property group to be outwith the scope of this thesis.

4.3.2 Vibration Feedback

There is a wide range of different vibration technologies available, with a huge increase in the ubiquity of this form of feedback with rise of video game controllers and mobile phones having vibration as standard. The Android operating system, for example, offers vibration feedback to improve the usability of its virtual keyboard. The research into this area predates the ubiquity of vibration in our daily lives, however, with topics such as whether vibration can represent roughness argued as early as 1925, according to Lederman et al. [37].

While we originally considered this to be an interesting topic for investigation, we decided that we would not investigate this form of feedback in depth. The use of vibration is a complex topic which does not lend itself to easy use or control without expertise in the area. As we had no experience in this area of research we decided to focus our time instead on another form of haptic feedback (haptic or force feedback pens) with which we had more experience both personally and within our research group. These haptic pens are discussed in more depth in the next section.

4.3.3 Haptic/Force Feedback Pens

The final investigation we performed in this area was into what are regularly called haptic or force feedback pens. These are most commonly devices with a pen-like pointer on the end of an arm. This arm contains a series of motors which allow the device to resist or provide motion in different axes. An example of the type we used is show in *Figure 4.1*.

Figure 4.1 - The haptic pen used for our investigation, a Phantom Omni

These devices have three main advantages over the usual mouse and keyboard used with computers. First, as the end of the pen is considered to be the pointing device and it is on a free-moving arm, it allows a user three-dimensional interaction with a scene as it can move through all three dimensions. In addition, as the pen can also be rotated, it allows for free rotation in the scene as well. Finally, due to the motors in the device, it allows users to 'feel' the scene. That is, the motors can resist the motion of a user to make it appear as if the pointer has stopped against a surface within the virtual scene.

This force feedback ability was what interested us the most and led us to investigate the practicalities of the device for psychophysical experiments. To perform this investigation, we created two different applications which allowed us to explore the device's capabilities.

The first was an application which would be useful for investigating the *Flexible* group. This modelled how flexible or springy a fabric was by using the motors in the haptic device to provide resistive force when a participant attempted to flex it. To provide visual feedback, a series of high-quality images of a real fabric captured by *totallytextures Ltd.* was animated, the animation synchronised with the cursor to show the fabric being flexed [38]. An example of this application in use is shown in *Figure 4.2*.

Figure 4.2 - The haptic pen test application for the *Flexible* group in use. N.B. The help information present around and on top of the fabric could be disabled once a participant understood how to use the application to enhance the connection between haptic and visual feedback.

This application was initially successful, as when the guides shown in *Figure 4.2* were disabled, the connection between haptic and visual feedback was anecdotally convincing.

This led us to create an application which allowed a user to 'feel' a rendered three-dimensional surface with the haptic pen in the hopes of representing the *Smooth* or *Coarse* group. The geometry of the surface could then be probed by the pen, the motors giving a participant feedback when the pointer collided with the surface, allowing them to easily feel the surface contours. We quickly found, however, that there was a problem with this design: the feedback was too weak to be accurate or reliable.

The device we were testing with (and the only one we had access to) was a Phantom Omni haptic device [39]. As quoted by the manufactures, it has a maximum exertable

force of 3.3 N. In other words, it could push with approximately a third of the force which gravity offers when one is holding a 1 kg bag of sugar. While this was noticeable when using the device and allowed users to explore surface contours, it caused two problems. First, it caused a simple, spherical, test surface we tried which should have felt hard to have a certain amount of give, due to the motors not being strong enough. This meant it was impractical to accurately represent the *Solid* group. Second, whenever a participant pushed too hard, the pointer penetrated the test surface, destroying the illusion of shape already created. While there are versions of this haptic pen with stronger motors, they are significantly more expensive and therefore not as readily available.

This motor strength problem was what convinced us that using haptic devices would not be the correct path for our research. As we wished to perform psychophysical experiments, we had to ensure that we had complete control over every part of our setup. With these haptic devices there would always be a factor beyond our control and subject to possible bias. As such, we decided not to investigate the *Flexible*, *Smooth*, *Coarse*, *Solid* groups, or any terms such as *Soft* using these devices.

4.3.4   Summary

In this section we have discussed the three main tactile feedback methods we considered using to represent our property groups. These were thermal feedback, vibration feedback, and haptic pens. We concluded, however, due to a lack of interesting research opportunities, expertise, or control, not to focus on any of the three feedback methods.

While thermal and vibration feedback were quickly rejected, the most promising feedback method, haptic pens, were investigated in more depth. Two different test applications were developed in order to see whether these haptic pens could be used to convey either the flexibility or relief of a surface. It was found, however, that the motors present in the haptic pens we had available were not strong enough to be used in psychophysical experiments without introducing bias.

We therefore decided not to rely on tactile feedback to represent any of our property groups from *Table 3.7*. Specifically, therefore, we decided not to investigate the *Warm*, *Solid*, or *Flexible* group further, and to not use tactile feedback to investigate the *Smooth* or *Coarse* group.

4.4 Physically-Simulated Properties

In this section we will briefly discuss our decision not to investigate the physically-simulated property groups from our list, *Delicate*, *Fuzzy*, *Brittle* and *Crumpling*. All the groups on this list rely on some sort of simulation being run to truly represent them accurately. For *Delicate* and *Brittle*, for example, we would need to model material breaking or tearing. For *Fuzzy* and *Crumpling* we would need to model the movement of individual fibres or how the fabric reacts to gravity and motion.

As modelling physical properties like this was more complex and computationally intensive than the applications discussed in the previous section, we decided it would be too time consuming to create our own test applications and would rather focus on creating test applications for the other categories of property.

Furthermore, rather than investigate examples from others, we made the decision not to investigate this category of properties in depth. We made this decision for three reasons. First, we had no experience in the physical modelling which would be required to make the fabrics or surfaces react realistically. Second, to accurately represent this class of fabrics, we would want to be able to measure a range of real fabrics and we did not have access to tools to measure properties of fabrics such as their stretch or drape. Finally, to model these properties properly, we would have to employ real-time rendering techniques which would be counter to our desire to present the most realistically rendered stimuli that were practical. Either we'd have to employ less complex lighting and reflectance calculations, or we'd have to render a complete video of a predetermined physical interaction offline, beforehand, which would be unfeasibly time-consuming[3].

We decided that we would rather render stimuli using more complex and realistic surface geometry with physically-based reflectance models than render in real-time. We believed this would give answers closer to the ground truth than modelling physical interactions.

These considerations meant we decided not to investigate any of the properties from the physically-simulated property category. In short, we would not investigate the *Delicate*, *Fuzzy*, *Brittle* or *Crumpling* property groups elucidated in the previous chapter. In the next section, we will therefore discuss the final category of properties, visual.

---

[3] The stimuli we created took 25 days to render without animation.

4.5 Visual Properties

In this section we will discuss what representation methods we investigated to represent the possible visual property groups from our list, *Smooth*, *Coarse*, and *Textured*. As the items in these property groups are mainly visual, they could be represented by a simple image on a monitor, but we wanted a display method which would provide a representation closer to the ground truth. With a brief review of the literature, we found that there were three common techniques of improving static images: accounting for head motion, stereoscopic viewing and motion [10, 40].

One important note at this point is that we decided not to consider animated or moving stimuli. As discussed in the previous section, due to the fact we were planning to use complex offline rendering techniques, animated stimuli would be unfeasibly time-consuming to create. Therefore, this discussion is split into two main sections, based on the two technologies we investigated, head tracking and 3D displays.

4.5.1   Head Tracking

First, we investigated head tracking techniques. In general, these are systems which detect through some method where a participant's head is in relation to the screen and then update the scene in real time to reflect this. So, for example, if a participant were looking at a cube's front face, if they moved their head to the left, they would be able to see the left face, also. Most of these systems work in a similar way and give the illusion that you are looking through a window, rather than viewing images on a monitor. Therefore, as we already had experience using the Wii Remote, we created a test system which would track a participant's head position.

In essence, the original Wii Remote has two ways of motion tracking. To detect fast movement and basic orientation, it uses a series of accelerometers to measure different axes. For finer uses, like using it as a pointer, it contains an infrared camera in the front which is used to detect infrared LEDs. This, in combination with the (inaccurately named) Sensor Bar which contains infrared LEDs a known distance apart, allows the Wii Remote to calculate the position in relation to the monitor. The Sensor Bar with its infrared LEDs is show in *Figure 4.3*.

Figure 4.3 - The Wii Sensor Bar with its infrared LEDs. N.B. The LEDs are not visible to the naked eye, but most digital cameras are sensitive to infrared sources, showing them as purple light. Image from [41].

It is the Wii Remote's infrared camera which we used to track a participant's head movements. First proposed by Johnny Lee, the system works by reversing the Wii Remote and Sensor Bar's position [42]. A Wii Remote is positioned under a monitor, and the Sensor Bar is attached to a participant's head. That allows the relative position between the two to be calculated to determine a participant's head position.

Attaching a Sensor Bar to a participant's head wasn't practical, however, so instead we used a pair of safety goggles with built in LEDs. The original white LEDs were replaced with infrared LEDs of a similar power, allowing the Wii Remote to detect the two infrared points. The average position of these two LEDs was then calculated to determine a participant's position relative to the screen, with the distance between them used to calculate distance from the screen. These details were then used to update an OpenGL scene every frame via the glFrustum function using the head position to alter the view of the scene.

With our simple test scene, we found that the effect was compelling. The system accounted for the slight, unconscious head movements made by participants and appeared to trick the visual system into considering the images on screen as a three-dimensional scene, rather than a two-dimensional projection. It was not without problems, however. The most obvious problem was that the system was unstable. If, as often happened when a participant rotated their head, the Wii Remote lost sight of one of the infrared LEDs, the view would move suddenly and the distance from the screen could no longer be calculated.

Still, we believed this system worth investigating further, especially if it could be combined with a system which would display the scene in a manner suitable for stereoscopic viewing. To this end, we decided to investigate what technologies and techniques existed for stereoscopic viewing, as discussed in the next section.

4.5.2   3D/Stereo Displays

The second method we investigated to enhance the visual display of our property groups was '3D Displays'. This broad heading covers many different techniques and devices designed to display images three-dimensionally, so in our case we narrowed it to those often classed as 'Stereo Displays'. That is, our investigation was limited to systems based on stereopsis, the method of displaying a different image to the left and right eye to give virtual depth to scenes. While there are real 3D displays on the market, we did not consider then in this investigation. Systems such as volumetric displays which rely on a fast spinning plane and synchronised, rapidly updating projection to display 'voxels' in 3D space are available but are too expensive to be considered.

Instead, we started by investigating a widely available commercial stereo display used in desktop gaming and visualisation, NVIDIA 3D Vision [43]. This system relies on active shutter glasses, glasses which have liquid crystal layers in each lens which become dark when a voltage is applied. Then, which lens is darkened is alternated while the monitor rapidly alternates between two projections of the same scene in synchronization with the glasses so each eye only sees the correct projection. In the case of the 3D Vision system we tried, this happens at 120 Hz, allowing a refresh rate of 60 frames per second for each eye.

To test the viability of this system, we created a simple, brightly coloured room with a spinning cube in the centre as a test scene. This was designed to allow us to quickly test whether this system suffered from any problems such as desynchronisation between the monitor and glasses. Our first impressions were positive, as we found that the refresh speed and synchronisation were both acceptable. The virtual depth expected was rapidly apparent and easily adjustable. In addition, the nature of the system allowed for limited head movement, meaning it could be easily combined with the head tracking discussed in the previous section. At the same time, however, there were problems with using this system for psychophysical experiments.

The first was an obvious flickering effect present in any part of a participant's view not covered by the monitor. This flickering effect rapidly caused headaches in both the author and a few naïve observers who were asked to use our test scene. The second problem was that when using the 3D Vision glasses, the monitor's contrast, brightness and colour settings were fixed, meaning we were unable to calibrate it for use in our experiments. This is possibly to counter the third problem, which is that the glasses caused a large drop in light transmission. Even when the liquid crystal layer is clear, the

level of light received by a participant is much lower than without it being present. In addition, only one eye is receiving light at a time, reducing the amount of light reaching the vision system as a whole. Finally, we noticed obvious crosstalk between the two eyes. This was most apparent when using bright colours, or white surfaces. This was problematic; if we wanted to represent the *Smooth* group (which contains terms such as *Glossy* and *Shiny*) bright points caused by specular highlights were essential. An example of this crosstalk problem is presented in *Figure 4.4*.
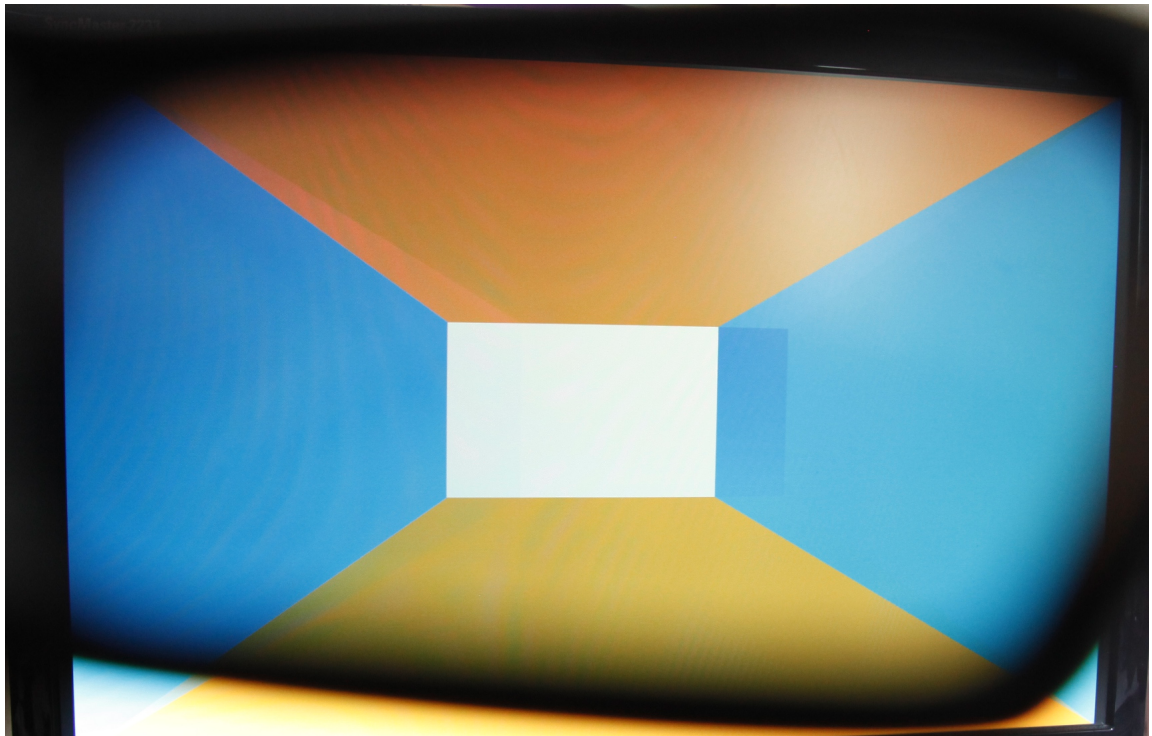
Figure 4.4 - An example of crosstalk in the NVIDIA 3D Vision system. The cross talk is most apparent to the right of the white rectangle in this example. The 'ghost' image is caused by part of the other projection of this scene leaking through the lens when it is supposed to be dark.

While attempting to solve these problems, we found that the flickering was simple to remove. All our testing was performed in a room lit by the fluorescent tubes common to office environments. These work at between 100 and 120 Hz depending on the type, close to the rate of the glasses. This similarity caused the flickering effects observed. When the glasses were used without the fluorescents, the flickering was negligible.

Sadly, however, we failed to solve the other problems. As we wanted to perform psychophysical experiments monitor calibration was essential to us, as was ensuring there was no crosstalk. To this end, we decided to investigate other stereoscopic display methods.

The next technique considered was a half-silvered mirror stereo display. This involves two monitors, polarised in different directions, and merging the two images with a half-silvered mirror. The two images are then separated between a participant's eyes using polarised glasses. Ideally, the participant will only see a single image projected on the mirror with correct virtual depth. This setup is show in Figure 4.5.
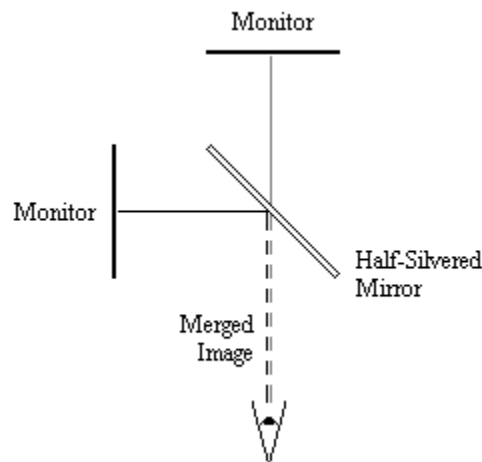
Figure 4.5 - A diagram of a half-silvered mirror stereo display. This system uses a single, half-silvered mirror to merge two images from different, linearly polarised monitors. The two images are then separated between a participant's eyes using polarised glasses.

This system was appealing as it again accommodated head movement (at least to a certain degree) and so could be used with the head tracking already discussed. Also, as we would be able to use a pair of matched monitors, we would be able to calibrate them as necessary. Sadly, however, it still suffered from two of the major problems present with the 3D Vision setup. Namely, by the very nature of this system there is decreased light transmission. Both the half-silvered mirror and polarising steps reduce the amount of light received by a participant. Worse still, however, we found when testing with an available system that cross talk was still readily apparent.

Due to the recurring crosstalk issue, we ended our investigation into stereo display technologies by considering a method guaranteed to have no crosstalk. A mirror stereoscope in general consists of a series of mirrors which direct a participant's right and left eye to two different images. As the system completely separates the view of both eyes, there is no possibility of crosstalk as long as there is no head movement. In addition, due to there being no polarising step or liquid crystal layer, the light transmission is much higher than the other systems discussed. This, of course, solves our main concerns with the other systems, but would simultaneously rule out the

possibility of combining this system with head tracking. This consideration is discussed more in the next section.

4.5.3   *Discussion and Summary*

To conclude this section into our visual property groups, we will discuss briefly the decision we had to make at this stage.

In *Section 4.5.1* we discussed in detail the method of using a Wii Remote to track a participant's head and update the scene accordingly. We found that, despite a stability problem, the system worked well and enhanced the scene. In *Section 4.5.2* we investigated three different 3D display techniques and found that only one of them, the mirror stereoscope, was suitable for us.

Ideally, we would have incorporated the head tracking and 3D display together, as previous experiments have shown that both head tracking and stereoscopic viewing can enhance a participant's perception of visual properties like gloss and make their judgements more consistent [10, 40], but the mirror stereoscope relies on a participant's head not moving in relation to the monitor. This forced us to make a decision about which representation method was best for our circumstances.

We were aware at this stage that if we wanted to create the most realistic stimuli we could, it was likely we'd have to render our stimuli offline. This was at odds with the real-time rendering required for head tracking. While we could have reduced the quality of our stimuli for use with head tracking, we decided that higher quality stimuli presented through a stereoscope would offer a situation closer to the ground truth.

Therefore, we decided that the use of a mirror stereoscope, rather than head tracking, would be best to present our visual property groups *Smooth*, *Coarse*, and *Textured*.

There was an interesting side effect of our investigation into mirror stereoscopes. As we were conducting a brief survey of the literature in this area, we found that there was a large body of work investigating both gloss and roughness using stereoscopes of varying types. Both of these terms appeared in our property groups as *glossy* and *rough* and in both cases they were the second most representative term of their respective groups, *Smooth* and *Coarse*.

As we had already come to the decision that we were going to investigate the interaction between two of these three visual property groups, we further decided that investigating

46

the interaction between *Glossiness* and *Roughness* would prove fruitful, especially when combined with stereoscopic display.

4.6   Final Discussion and Conclusions

As we've discussed in this chapter, there is a wide range different digital tools which are currently available which we could use to represent different properties. We decided on three categories which represented groups: tactile properties, properties requiring physical-simulation and visual properties. We then, in turn, discussed these three categories and suggested which techniques might be well suited to presenting the properties contained within, stating advantages and disadvantages of each as well as discussing any tests we performed.

First, we discussed different tactile properties and concluded we would not investigate this category of property. This was due to the nature of the haptic device, the *Phantom Omni*, which we tested during this chapter. While this system was certainly affordable it suffered from a lack of motor strength, meaning it altered the hardness of anything it was supposed to be representing. Surfaces felt soft, even if they were representing a completely solid object. While other systems would have provided better motor strengths they are much more expensive and therefore unlikely to be affordable in the near future.

We briefly discussed physically-simulated properties and concluded that we would not be investigating them further. The reason for this decision was threefold: we had no experience in the area, we lacked the ability to measure real fabric properties, and we wished to avoid real-time rendering.

Finally, for visual stimuli, there were two areas we investigated: head tracking and stereoscopic display methods. Both of these allow us to present much more realistic and information rich stimuli to participants. Head tracking allows us to present an object which realistically reacts to participants' head movement, giving the illusion of looking through a window onto a virtual scene. For stereoscopic displays, we investigated stereopsis, which instead shows a different projection of the scene to each eye giving the illusion of depth.

While ideally we would have used both head tracking and stereoscopic display, it was infeasible in our case. Two of the stereoscopic display methods we tested suffered from crosstalk between the two eyes which dampened the perception of depth. The only

system without crosstalk was the mirror stereoscope. As this required the participants' heads to be kept still, however, it would be infeasible to incorporate head tracking. In addition, if we were to use head tracking we would need to be able to render our stimuli in real time (meaning we could not use as physically accurate stimuli).

Therefore, we decided to use a stereoscope to display our stimuli. As discussed in *Section 4.5.3* this presentation method lends itself particularly well to two different groups found in *Chapter 3*, those being the *Smooth* and *Coarse* group. This combined with what we discovered during our investigation showing there is a large body of research into the interaction between *Gloss* and *Roughness* (*Glossy* and *Rough* are both terms second most representative of their respective groups), has led us to decide that these two properties would be fruitful to investigate further. Specifically, the use of a stereoscope allows us to investigate this interaction in a method closer to the ground truth.

Now that we've narrowed our focus, in the next chapter we discuss in more detail the literature available for both gloss and roughness, focussing on how best to create stimuli to investigate these properties effectively.

Chapter 5

Gloss and Roughness Survey

5.1   Introduction

In the *Chapter 4*, we decided to choose the properties *Glossy* and *Rough* from our taxonomy originally proposed in *Chapter 3*. This was due to our discovery of a large body of literature which has been examining both gloss and roughness for the past 20 years. We will examine this literature in more detail in this chapter. Furthermore, there has been recent research into the interplay of gloss, roughness, and stereoscopic disparity which we examine in more detail in *Section 5.5*. As, in the previous chapter, we presented a way in which we could examine the interaction of these properties in a more realistic way (using a mirror stereoscope) we believe this to be a fruitful avenue of research.

Now we have decided on a more focussed area in which to investigate in this chapter we discuss the available literature for these properties and decide on how best to create stimuli which will enable us to investigate them effectively.

Before beginning our survey of the literature, however, we first define five important criteria (*Section 5.2*) which shape both the literature survey and the stimuli which were created as a result. To begin our literature review, we discuss gloss on real samples in *Section 5.3*. As we decide to use synthetic stimuli, in *Section 5.4* we investigate the different techniques which can generate rough surfaces before discussing what work has already been performed with roughness, glossiness and stereoscopic vision in *Section 5.5*. This is followed by an investigation into different BRDF models available which allow for the creation of virtual glossy surfaces, as presented in *Section 5.6* before finishing the chapter by discussing different light sources in *Section 5.7*.

Before we begin, however, we start with a clarification on what is meant by the term 'roughness' in this thesis.

5.1.1 A Clarification of Roughness

Before we begin the literature review presented in this chapter, there is an important definition which needs to be clarified. Qi et al. defined two types of roughness in surfaces: mesoscale and microscale. The mesoscale roughness he defined as being distinct 3D surface texture and microscale roughness as a parameter in the microfacet reflection model [44]. In other words, mesoscale roughness is determined by the physical structure of a surface (grooves, ridges, bumps etc.) whereas microscale roughness is determined by how the material of the surface reflects light.

Simply put, surfaces can have the same mesoscale roughness, but a different microscale one. One example of this would be two similarly carved pieces of wood having different varnishes which might alter how the items reflect light.

Therefore, in the rest of this thesis, when we talk about 'roughness' we are discussing mesoscale roughness only, rather than microscale roughness. When microscale roughness is a consideration (such as when we discuss BRDF models in this chapter) we have ensured that all stimuli have identical microscale roughness parameters.

5.2 Stimuli Criteria

Before we investigated the literature for our two properties, we decided it was important to define some criteria to ensure our review remains focussed. The fields of computer graphics and vision science are both well subscribed and are used in many different areas. That meant that if we were not careful we would find ourselves overwhelmed by the literature.

To this end, we decided to define five criteria which would direct our literature review and directly determine what types of stimuli we would create and use. These are listed below, with each main point being emboldened before a short explanation of why this criterion is important following:

(C5.1) **The creation parameters of the stimuli should be parsimonious and predictable.** If this is true, it makes it easier to define, create and control our stimuli.

(C5.2) **The stimuli should appear as contiguous, acceptable surfaces.** We want the participants to believe they are real surfaces being shown to them in both monocular and binocular conditions.

(C5.3) **The surfaces should appear naturalistic.** We want to make sure that any stimuli we use are as realistic as possible so our results are applicable to the ground truth.

(C5.4) **The surfaces should be isotropic.** For simplicities sake, we are not considering surfaces which are predominately defined by their directionality as this will complicate the analysis of the results due to the increased number of variables.

(C5.5) **The stimuli should be created in a reasonable time frame.** One of the major problems with creating virtual stimuli is that more processing power or time will provide better stimuli. To reach a realistic compromise, we wanted to ensure that time was an important consideration in our choice of techniques.

With these criteria, we can now present our focussed investigation into the literature for our selected properties. We will start with discussing the work performed with real stimuli.

5.3 Gloss on Real Surfaces

At this stage we were aware than any technique of creating and rendering virtual stimuli would be unlikely to capture every intricacy of a real sample. Due to the complexities of lighting, representing the scene, capturing the subtleties of head movement and other phenomena, it is difficult to create the 'perfect' virtual stimuli. While we have stated up to this point that we always intended from the beginning to present digital samples, we thought it prudent to reconsider this point with this new information in mind.

One way of avoiding this problem is to use real surfaces and thus avoid the problems of rendering all together. This was of course the only way of performing gloss measurement and scaling experiments before computers were able to render realistic surfaces and has been in use for decades, with examples in paint [45, 46] and paper finishes [47]. This ensures that any surface will contain all the complexity of geometry and lighting which can be difficult to capture in computer simulations.

One major problem with using measured scales as described above is that the devices used in industry to measure glossiness, normally called gloss meters, require a flat sample. As stated by Qi, textured surfaces with mesoscale roughness, such as those which we want to use, cannot be measured with conventional gloss meters [48]. Beyond this fact, both Leloup et al. and Ji et al. have found that the assumption that gloss meter

measurements have a linear correlation with perceptual scales, doesn't hold true [49, 50].

With a similar aim, Obein et al. performed an experiment examining the relationship of stereo viewing and gloss using real samples of black coated paper [51]. They also concluded that the relationship between perceived gloss and measured gloss was non-linear. An interesting extension of their work, however, was that they examined the scales in both monocular and binocular conditions and found that *'the sensitivity of the observers is improved in binocular vision mainly for the judgment of very glossy samples.'* This implies that stereo viewing is important (at least in certain cases) in gloss perception when using real samples and that observers were able to perceive the small specular disparities involved.

Our main objection to using real samples, however, is that they do not comply with criteria *(C5.1)*. That is to say, the creation parameters of real stimuli are not parsimonious as there are many factors which would need to be taken into account. If we were just considering specular gloss (which is but the first of six common types of gloss as defined by Donnell and Billmeyer [52]) then the measurement of non-metallic gloss in paint has been standardised. Even for this specific definition of gloss, however, it has been standardised in several places by different standards institutes making it difficult to decide on a single solution [53-55].

Compounding this problem was the fact that the facilities to create and measure the gloss and roughness of real stimuli along controllable, parsimonious scales were not available to us. The laboratory in which this research was conducted has experience in computer graphics and programming, but lacks the facilities for stimuli manufacture.

This lead us to conclude that, for the reasons briefly presented in this section, we would instead create virtual stimuli as this aligned with the specialties available and allowed us more control over the entire creation process.

5.3.1 A Clarification of Gloss

As mentioned in the previous section, Donnell and Billmeyer formalised six different common types of gloss [52] from Hunter's original definitions [56]. This was in order to help them create methods for obtaining visual interval scales of gloss in real samples. The six types of gloss defined were from visual observations of real glossly samples presented in the literature at the time. At this stage then, it is important for us to be clear

and specific as to which types of gloss this thesis would be concerning. The six types of gloss as stated by Donnell and Billmeyer are defined in *Table 5.1*.

From these six types of gloss, the one we decided to investigate in this thesis was specular gloss. As such, we have not looked at surfaces with grazing angles of light to avoid participant confusion with sheen. As mentioned in *Section 5.1.1* we only considered mesoscale roughness (not microscale roughness) which discounted distinctness-of-image gloss as being a factor in our experiments. Finally, we controlled all other factors of the surface and reflection model so as to discount the effect of the other three types of gloss in our BRDF lighting model.

Type of Gloss	Brief Description
Specular Gloss	Perceived surface brightness associated with the luminous specular (regular) reflection from a surface.
Sheen	Perceived shininess at a large angle of incidence seen in otherwise matte specimens.
Distinctness-of-Image Gloss	The sharpness with which images are perceived after reflection from a surface.
Contrast Gloss (Luster)	Perceived relative brightness of brighter and less-bright adjacent areas of the surface of an object, resulting from selective reflection in directions relatively far from those of specular reflection.
Reflection Haze	Perceived scattering of light reflected from a surface in directions near those of specular reflection.
Macroscopic Surface Properties	Grouped under this heading are all the properties of a surface that can easily be seen without the aid of a high-power microscope.

Table 5.1 - The six types of gloss as defined by Donnell and Billmeyer

Therefore, whenever we mention 'gloss' in this thesis we are talking specifically about 'specular gloss' in Donnell and Billmeyer's terms, which manifests itself as white specular highlights which are present on reflective surfaces under white lighting such as fluorescents.

5.4 Roughness Literature

In this section we will present a survey of the literature which investigates the use and creation of stimuli for use in roughness experiments. Using our criteria listed in *Section 5.2*, we examined different techniques of generating rough surfaces, before selecting the one which we used in our experiments.

We examined common methods of generating virtual surfaces in both the spatial domain (*Section 5.4.1* to *Section 5.4.3*) and the frequency domain (*Section 5.4.4* and *Section 5.4.5*). While we are aware there are other possible domains (such as time and wavelet) with which we could have generated stimuli, these were outwith the scope of this thesis and as such are not discussed.

Finally, in *Section 5.4.6* we formalise our findings and present our conclusion on what type of creation process we used in this thesis and why.

5.4.1   *Perlin and Simplex Noise*

'Perlin Noise' is a common technique for procedurally generating textures and surfaces in the computer graphics literature. Originally introduced in a course at SIGGRAPH in 1984 [57], then formalised the next year [58], Perlin noise is a type of gradient noise which can be used to simulate a wide array of natural effects. Indeed, in Perlin's initial paper he discusses how he had created '*very convincing representations of clouds, fire, water, stars, marble, wood, rock, soap films and crystal*'.
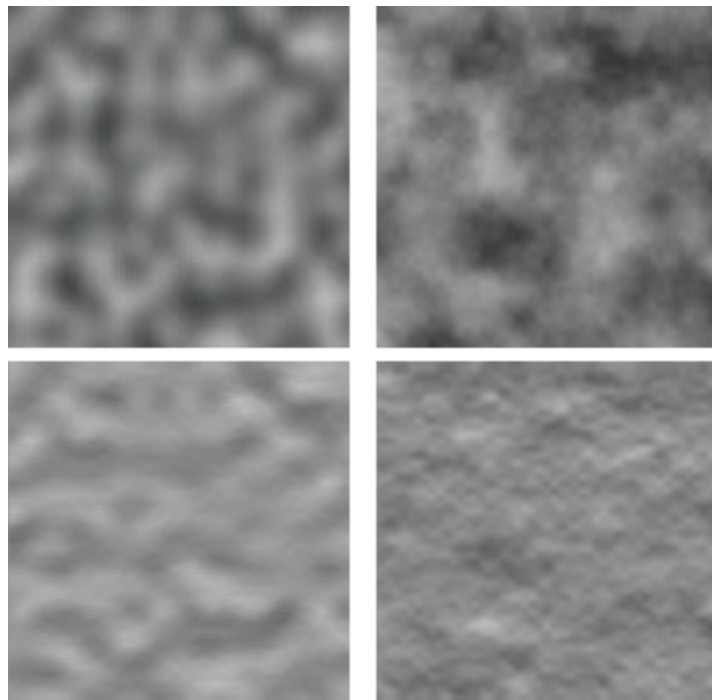
Figure 5.1 - Perlin noise height maps (upper row) and rendered surfaces (lower row). Left: A two-dimensional slice through three-dimensional noise. Right: Perlin noise simulating a 1/f$^\beta$ noise texture through scaling and addition. Both greyscale images are from [59]. The bottom images have been rendered with a simplified Lambertian reflectance model which assumes the light is coming from directly above the top of the image.

At its most simple, Perlin noise is a spatial creation technique which takes a three-dimensional vector as the argument, meaning it is a parsimonious creation method, as per *(C5.1)*. It is rotation and translation invariant and has a narrow size range of features which can be used to create contiguous, acceptable surfaces *(C5.2)* which, according to the author, appear naturalistic *(C5.3)*. Additionally, it was designed from the first to be computationally efficient, which fits in with our criteria *(C5.5)*. Two examples of different Perlin noise are shown in *Figure 5.1*.

There are, however, some problems with this technique. First, the original Perlin noise technique suffered from some deficiencies, although these were corrected in a more recent paper [60]. Second, while this technique appears to be naturalistic, its true nature is ill-defined. As stated by Perlin and Hoffert in a paper which extends Perlin noise into volumetric rendering [61], '*the model as described is highly empirical, leaving unanswered the disturbing question of why such simple techniques produce such visually convincing results.*' This might account for why this technique is so common in the computer science (as opposed to psychophysics) literature where speed is more important than accuracy. For example, Perlin noise was used by Ramanarayanan et al. to disturb the surfaces of spheres in the aim of defining a concept of 'visual equivalence' which aimed to be a less conservative measure of image difference [62].

A third, unexpected problem with Perlin noise was that Perlin himself stated that although the technique was supposed to be isotropic, because it consisted of adjoining patches there were significant directional artefacts [63], which is counter to our criteria *(C5.4)*. To combat this, he introduced a new type of noise, called 'Simplex Noise', a refinement of Perlin noise which is more computationally efficient, especially in higher dimensions, and is not susceptible to the same artefacts that prevent Perlin noise being isotropic.

One major problem still remains, however. As Simplex noise was designed to give the same general appearance as Perlin noise, it still suffers from the same 'disturbing question' mentioned above. That is, although it appears to be naturalistic, as per criteria *(C.5.3)*, how well it complies with natural processes is still under discussion.

One interesting point that Perlin makes, however, is that with his noise techniques '*a 1/f signal over space can be simulated by looping over octaves*' [58]. These $1/f^\beta$ processes are normally created in the frequency domain, rather that the spatial and as such will be covered later in *Section 5.4.4*. An example of this simulation technique is available in *Figure 5.1*.

To conclude, we decided not to use Perlin or Simplex noise to generate our rough surfaces mainly due to the lack of clarity over criteria *(C5.3)*. In addition, if as Perlin states $1/f^{\beta}$ textures can be simulated, why not instead use the process directly and avoid any problems with accuracy which are associated with simulation? To this end, we abandoned Perlin Noise as our preferred technique of generating rough surfaces and continued to investigate other generation techniques.

5.4.2 Reaction-Diffusion Models

'Reaction-Diffusion Models' are a set of mathematical models which describe a two-step, iterative process involving one or more co-located substances. In general, there is a diffusion step which cause the substances to spread spatially and a second local chemical reaction step where the substances change concentrations or type within a 'cell'. This same process is then repeated over many generations to allow the interaction of these two processes to define a pattern. By definition then, these processes occur in the spatial domain and while there is no requirement for such, many people relate the local 'cells' of chemical reaction to pixels of an image for simplicity.

In our case, we will be briefly discussing a subset of reaction-diffusion models defined by Turing in his paper ''*The chemical basis of morphogenesis'* in which he suggested a two-component system (with substances u and v) [64]. Turing himself states that his model '*gives rise to patterns reminiscent of dappling*' and others have talked about variations on the model giving rise to the patterns seen on leaves, zebra, leopards, giraffe and other mammals or even mollusc pigmentation [65-68].

In general this process is defined as follows for the two components:

(5.1)
$$\partial_t u = D_u \nabla^2 + F(u)$$

(5.2)
$$\partial_t v = D_v \nabla^2 + G(v)$$

Where $D_{[u,v]} \nabla^2$ represents the diffusion component and $[F, G]([u, v])$ represents a local reaction per cell in u or v which is often changed in different models to get different growth patterns. Examples of surfaces created using Turing's version of Reaction Diffusion Model are presented in *Figure 5.2*.

For us, however, we believe that Reaction Diffusion surfaces like those shown in *Figure 5.2* fail criteria *(C5.1)*. Although they have relatively few input variables, the definition between input and output is ill-defined in most models. For example, increasing one of the variables cannot be said to make a surface which is 'more stripy' or 'less spotty'.

There have been some efforts to categorise the space (such as one by Munafo for the Gray-Scott Model [69]) but these simply show how chaotic the reaction-diffusion model output space truly is. This simple fact makes it difficult to create a series of images along a single, meaningful axis.
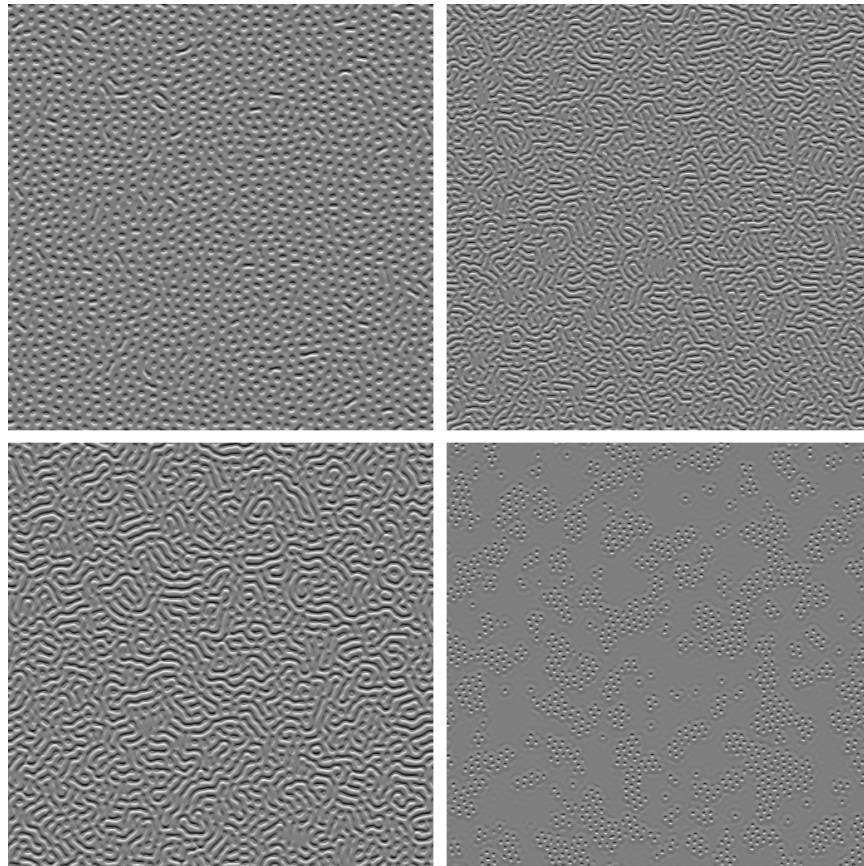
Figure 5.2 - Reaction Diffusion model examples, based on Turing's original work. Each of the four pictures are generated by the same model propsed by Turing using different input values and seeded with pseudo-random white noise. They have been rendered with a simplified Lambertian reflectance model which assumes the light is coming from directly above the top of the image in order to make them appear as a surfaces rather than height maps.

It was for this reason we decided not to use reaction-diffusion models to create our rough stimuli as we wanted more explicit control over the output. Instead we continued our investigation into spatial generation techniques before looking into generation methods in the frequency domain.

5.4.3   *Near-Regular Textures and Placement Techniques*

In this section we will briefly discuss the group of techniques for generating textures and surfaces which are created by placing or modifying many (usually repeating) texture elements, known as 'textons' on a surface. Clarke describes these textures thus:

'a near-regular texture is a regular texture with a degree of randomness added' [70]. This randomness can come from variation in size, colour, shape, lighting or position of individual textons [71]. This technique then falls within the spatial domain.

In general there are two main ways of producing this type of texture. The first, as discussed and by Liu et al. [71-73] and Nicoll et al. [74] involves taking an already irregular texture sample and deforming it into a 'near-regular' version which can be synthesized into larger textures. While this technique is interesting, it is not suitable for our purposes. For one, while the synthesised textures created appear naturalistic (as per criteria *(C5.3)*) due to their random abnormalities, most of the examples provided by Liu et al. are highly anisotropic as they are structured and periodic which violates criteria *(C5.4)*. More importantly, the generation technique relies on already available irregular textures as an input. This creates an infinite regression problem in our decision, as we are then forced to question how to generate or source this input if we use this technique, meaning we find ourselves no closer to a solution[4]. For this reason, we rejected this near-regular texture technique before investigating it in depth.
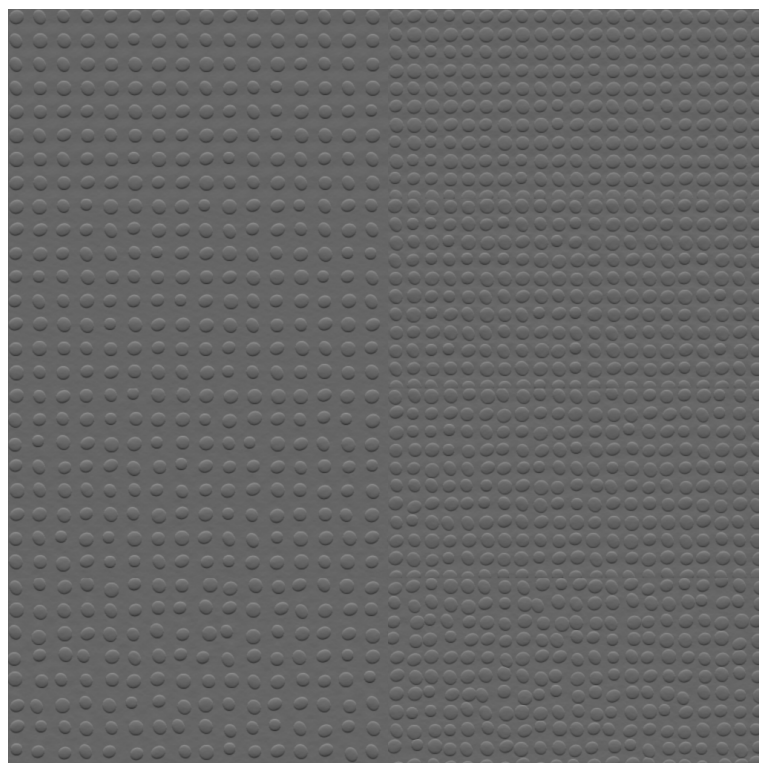
Figure 5.3 - Examples of near-regular surfaces derived from placement techniques. This image was presented by Clarke on p.26 in [70].

---

[4] This problem is sometimes referred to as 'turtles all the way down', as popularised by Stephen Hawking

The second way of generating this type of texture is to start with a flat surface and 'place' textons on it to create mesoscale surface texture. This was demonstrated by Clarke who used slightly randomised ellipsoidal-shaped textons placed at regular intervals to generate his surfaces [70]. Examples of his surfaces are presented in *Figure 5.3*. Ho et al. also used a similar process which involved 'jittering' a regular grid of 400 points and then placing an ellipsoid (scaled in the z-axis between a specific range determined by how bumpy the surface was designed to look) at each of these randomised points in such a way that they intersected and combined to create what they called a 'bumpy' surface [75]. Neither of these examples, however, produced a surface which could be easily called naturalistic as per criteria *(C5.3)*. In both cases the repeating nature of the textons used is apparent to observers and gives the generated image a readily apparent synthetic appearance. Due to this synthetic appearance, we rejected placement methods as suitable for our experimental stimuli.

It was at this stage that we decided that spatial generation techniques were unlikely to meet all the criteria we defined in *Section 5.2*. Rather than investigate more techniques in this domain, we instead elected to investigate the frequency domain to see if we would find a process more suitable for our needs. Therefore, in the next two sections we present our investigation into common texture and surface generation techniques which rely on the frequency domain, beginning with $1/f^\beta$ noise processes.

5.4.4 $1/f^\beta$ Noise

$1/f^\beta$ is a general term for any noise process with the power spectral density of that form and is often considered a random fractal function generated in the frequency domain [76, 77]. It is a more generalised description of processes such as white noise and Brownian noise where the β component determines the slope of the magnitude spectrum and therefore the type of noise created. For example, white noise can be created by using β of 0 and Brownian noise with β of 2. It is interesting for us as $1/f^\beta$ noise processes appear regularly in nature, such as the example of Brownian noise which is the noise produced by Brownian motion (first observed in 1827) [78] as an approximation of the power spectra present in images of natural scenes [79] or in the structure of rocks [80].

This technique of generating naturalistic noise has been widely used [7, 8, 48, 70, 81, 82] and in particular both Padilla et al. and Qi have used surfaces created from this noise process for conducting experiments into different effects of perceived roughness

[48, 82]. Padilla et al. defined this model as having two main components which they called 'Root Mean Square (RMS) Height' and the 'Roll-Off' factor.

Both Padilla et al. and Qi created $1/f^{\beta}$ noise surfaces ('$1/f^{\beta}$' is a simplification of *Equation (5.3)*) by first generating a height map with a magnitude spectrum $H(f)$ scaled with spatial frequency as shown in *Equation (5.3)* below where β is the roll-off factor of the surface height magnitude spectrum, σ is the RMS height of the surface and $N(\beta)$ is a normalising factor:

(5.3)
$$H(f) = \frac{\sigma}{N(\beta)} f^{-\beta}$$

Following this, the height maps were used to construct and render solid surfaces (albeit through different means with Padilla et al. and Qi) which could be presented to participants in their experiments.

Finally, there is a third parameter θ which is the seed value used to generate the random phase of the surface. As stated by Padilla [7]: '*Phase controls the relative positions of the sinusoidal basis functions that when added together produce an instance of a surface. It does not affect the variance of the surface or the shape of the power spectrum*'. In other words, changing the phase alters the appearance of the surface, but not the other statistics as defined by β or σ.

This is a useful feature of $1/f^{\beta}$ noise surfaces as this means they can be randomised without altering other important statistics of the surface. Indeed, Padilla showed that changing the random phase used to generate these surfaces does not change the perceived roughness [7]. Therefore, it is possible to generate several surfaces of the same perceived roughness which are visually distinct, a useful feature when performing experiments as it stops participants performing pixel-per-pixel comparisons.

It is important to note that the technique we are discussing here assumes that the roll-off is the same in all directions, centred on the mean component ($f = 0$). It is this behaviour in the frequency domain which causes the height maps created to be isotropic. Linnet discusses this in more detail and shows different types of surfaces it is possible to create by not assuming this [83]. For example, Linnet shows that if you shift the peak away from $f = 0$ (while maintaining the required symmetry) then you can instead create naturalistic looking anisotropic textures. These textures look like the sand waves found on the sea floor. Any of these techniques, however, introduced additional

parameters which we were not interested in. As such, we did not investigate any of these added complications in depth.

The idea of creating $1/f^{\beta}$ noise surfaces was appealing as they complied with all of our criteria. The technique for generating the height maps is parsimonious (as there are only two parameters), the noise process can be rendered as surfaces which by their nature will be isotropic and the height maps are quick to produce. In addition, the surfaces which the process produces appear naturalistic. For example, *Figure 5.4* shows a comparison between a noise surface (where $\beta = 2$) rendered using the same simplified Lambertian reflectance model we've been using to present our surfaces in this section, compared with a 'fracture' surface created by McGunnigle by cracking a block of plaster [84].
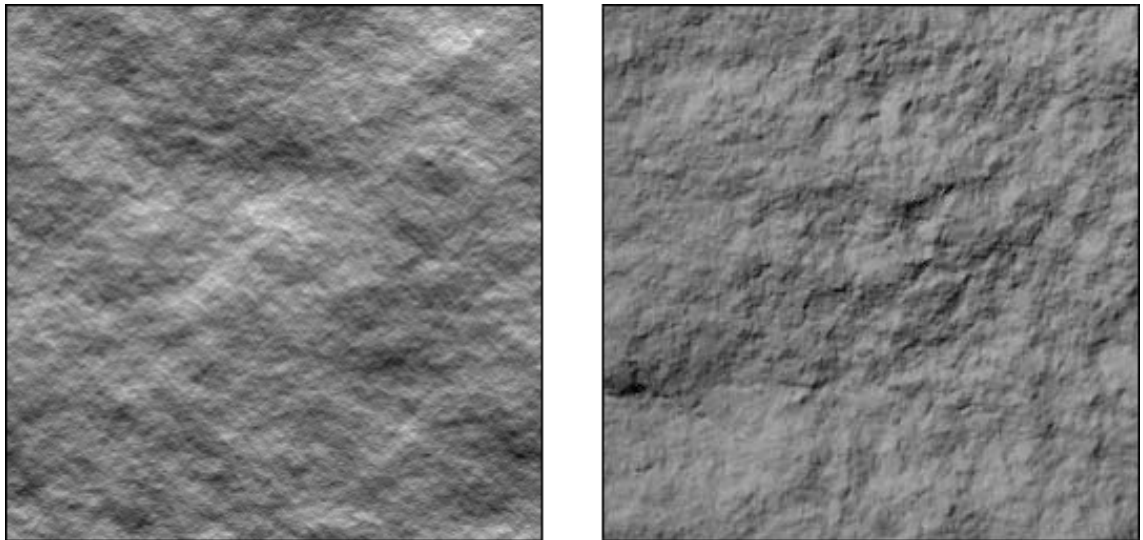
Figure 5.4 - A comparison of a 'fractal' surface and a 'fracture' surface. Left: A $1/f^{\beta}$ fractal noise surface (where $\beta = 2$) rendered with a simplified Lambertian reflectance model which assumes the light is coming from directly above the top of the image. Right: A photograph of a fracture surface created by cracking a solid block of plaster, as presented by McGunnigle [84].

At this point, it was clear to us that $1/f^{\beta}$ noise surfaces were suitable for the experiments already planned. They fulfilled all of our criteria, had been used for similar experiments in the past, were simple to create and allowed us great flexibility in how to render them. As such, we decided that we would use these surfaces in our work. In the next section we briefly discuss connected techniques before summing up our thoughts.

5.4.5   *A Brief Note on Sayles, Mulvaney and Ogilvy Surfaces*

In the previous section, we discussed the use of $1/f^{\beta}$ noise as a way of creating height maps which could be used for generating solid surfaces. They are attractive because of the few parameters (only two) and the fact they've been used by many contemporaries to conduct similar work into roughness. $1/f^{\beta}$ noise, however, is not the only frequency domain generation technique available.

McGunnigle and Chantler, for instance, discussed three different frequency domain generation techniques from Sayles [85], Mulvaney [86] and Ogilvy [87] and provided examples on the effect of altering one of dominant or unique parameters in the different models [84].

While these techniques of generating rough surfaces were briefly considered, they were quickly rejected. The Ogilvy surfaces have a directional component which not only makes the generation method more complex, but is directly contrary to criteria *(C5.4)*. Mulvaney surfaces have an extra cut-off parameter which is not necessary for our investigation and thus was discarded also. Finally, McGunnigle states that Sayles surfaces are similar to the $1/f^{\beta}$ surfaces we have already discussed [88]. As we had both more experience with $1/f^{\beta}$ surfaces and more evidence that they can be used successfully for roughness experiments (as described in the previous section), we decided against investigating Sayles surfaces further.

5.4.6   *Conclusions*

In this section we have discussed four different techniques based in the spatial domain and four in the frequency domain, discounting most of them due to criteria set out in *Section 5.2*. Simplified versions of the criteria are as follows:

(C5.1)   **The creation parameters of the stimuli should be parsimonious and predictable.**

(C5.2)   **The stimuli should appear as contiguous, acceptable surfaces.**

(C5.3)   **The surfaces should appear naturalistic.**

(C5.4)   **The surfaces should be isotropic.**

(C5.5)   **The stimuli should be created in a reasonable time frame.**

During this survey we decided to use $1/f^\beta$ noise surfaces as our stimuli generation method of choice. Therefore, to conclude this section, we present the succinct reasoning for our decision in *Table 5.2*.

In the next section we discuss the work which has already been performed investigating gloss, rough surfaces and stereoscopic vision and justify why the work presented in the rest of this thesis is novel and important.

Generation Technique	Succinct Reasoning	Criteria Failed
Perlin/Simplex Noise	Empirical techniques with little proof that they accurately model natural processes. They can be made to replicate $1/f^\beta$ noise surfaces, but this is more complex than using them directly.	*(C5.3)*
Reaction Diffusion Models	The definition between input and output is ill-defined in most models. The possible output space is therefore chaotic and difficult to control. Also as the process is generational there is no upper time limit on generation.	*(C5.1) (C5.5)*
Near-Regular Textures	Most techniques require an input texture, thus not solving our initial problem of generating synthetic surfaces	*(C5.1)*
Placement Techniques	The surfaces created by these techniques appear artificial even when randomness is added to the system.	*(C5.3)*
$1/f^\beta$ Noise Surfaces	This technique of generating surfaces is already commonly in use and complies with all of our criteria.	*None*
Sayles Surfaces	Sayles surfaces are similar to the previously discussed $1/f^\beta$ noise surfaces, but are less commonly used.	*None*
Mulvaney Surface	Mulvaney surfaces have an extra parameter over the $1/f^\beta$ surfaces and as such have an extra, unnecessary level of complication.	*(C5.1)*
Ogilvy Surfaces	Ogilvy surfaces have a directional component in their definition, making them anisotropic.	*(C5.4)*

Table 5.2 - A brief summation of the creation techniques surveyed. The succinct reasoning of why a technique was rejected is presented in each case expect for the shaded row which we chose to use. In this case the reasons for accepting it are presented. Where a technique failed to meet one of our criteria, the criteria failed is offered also.

5.5 Gloss, Roughness and Disparity Literature

It was now important to investigate the work which had already been done investigating gloss and roughness so we could ensure our work was interesting within this field.

Perhaps the most apparent trend in the gloss literature is that most of the work already performed in studying gloss has used simple surfaces or lighting models. For example, most of the work currently available using computer-generated surfaces use surfaces

such as spheres or hemispheres [9, 89, 90], those created from sinus gratings [10, 11], or other techniques which generate relatively smooth (and we would say) unrealistic surfaces [75, 91].

One of the earliest papers which investigated the effect which stereoscopic viewing had upon gloss was presented by Blake and Bülthoff and used curved surfaces, either spheres or hemispheres recessed into a flat plane [9]. They found that *the three-dimensional appearance of a highlight on a computer-simulated stereoscopic curved surface affects the observers' judgement of surface gloss'*. They concluded that correct highlight disparity increased the realism of the stimuli. This result has been confirmed by more recent research by Wendt et al. who found that the presence of highlight disparity increased both the authenticity and the strength of perceived gloss [11] and by Sakano and Ando who found that both head motion and stereo disparity enhanced the perceived strength of gloss [40]. It is also well reported in the literature that binocular cues and "naturalness" (as defined by Hartung and Kersten, which we call being closer to the ground truth) are important for accurate gloss perception [92-94]. As Kerrigan and Adams put it, *'binocular vision provides information that could distinguish specular highlights from other luminance discontinuities…'* [90].
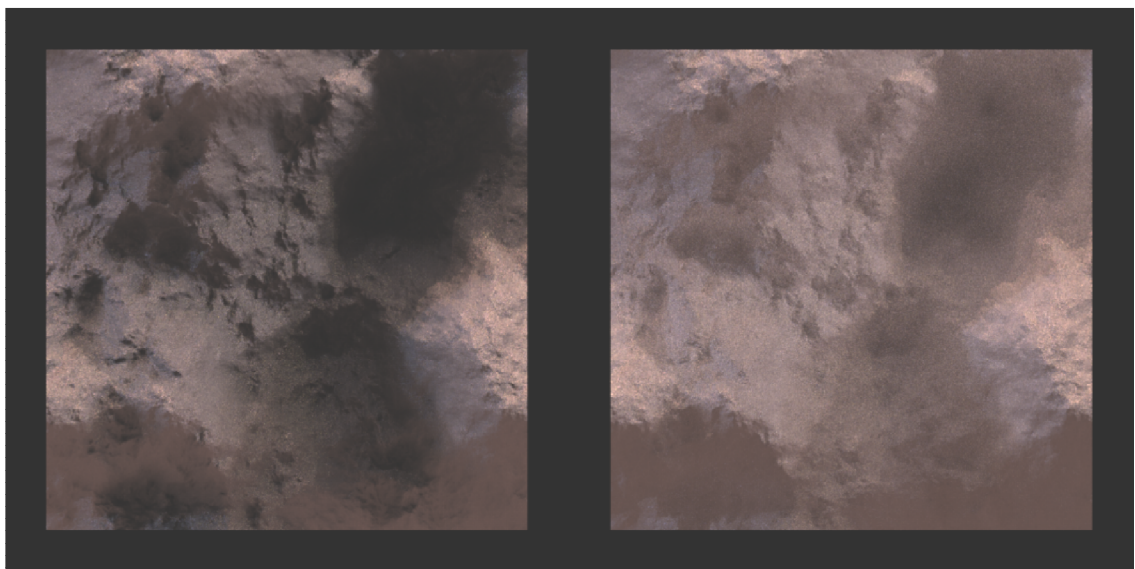
Figure 5.5 - An example from Qi showing the importance of interreflections. Both images were rendered identically from the same height map using a path tracing algorithm, apart from the number of bounces. The left image was rendered with a maximum bounce of 1 and the right with a maximum bounce of 10. The surface with a higher maximum bounce shows brighter areas because of inter-reflections. Images from [48].

While these studies are interesting, they used both relatively simple surfaces and simplistic lighting to create the stimuli. For example, they each used the Phong lighting model [95] which is an empirical model which has since been improved on, but is quick to render and can be calculated in real time. In addition, none of these papers take into account the interreflections of light (the cause of indirect illumination) which Qi has shown alters the final image significantly [48]. An example of this phenomenon is presented in *Figure 5.5*.

There has been some work performed with surfaces which appear more rough, for example Ho et al. have used surfaces composed of randomly oriented Lambertian-shaded facets in order to investigate the effect of illumination direction and viewpoint on perceived roughness [1, 2]. While these surfaces are visually rough, they are not particularly complicated as the 'jittered grids' which were used to give randomised surface relief were 20 points square, meaning a resolution of 400 changes in relief across the surface. The work presented by Qi (which used the types of surfaces proposed in the last section) by comparison had height maps of 512 pixels square, meaning a total number of 262,144 changes in relief instead, allowing for much more complex stimuli [48]. In turn, Qi's work while looking into gloss and roughness did not consider how stereoscopic viewing altered participants' perceptions.

Additionally, there has been work using more realistic lighting models. For example, Vangorp et al. investigated how shape influenced the perception of reflectance on various types of surfaces, some glossy [96]. They rendered realistic looking scenes using the Ward BRDF model [97] and path tracing techniques to take account of soft shadows and interreflections.

Indeed, the Ward model is popular in psychophysical experiments concerning realistic rendering as it obeys many desirable common criteria such as being parsimonious, obeying Helmholtz reciprocity and conserving energy. Helmholtz reciprocity states that at a surface, incoming and outgoing rays of light can be considered reversals of each other. Therefore, you could swap a sensor and light source's position without changing the sensor's measurement. This is important as it is a fundamental property of light and as such, any rendering method which does not obey this is not physically accurate.

In addition, the Ward model is the native BRDF model in Radiance [98], a popular suite of rendering tools. As such, an isotropic version of Ward's model was used by Pellacini et al. [99] and Ferwerda et al. [100] as the basis to create a psychophysically-based gloss model with parameters which would be more meaningful for users. Interestingly

though, the stimuli used by Pellacini and Ferwerda were simple spheres and those used by Vangorp et al., though much more complex, would not be described as particularly rough.

We therefore believed that there was still the interesting question of whether the results first presented by Blake and Bülthoff and then Wendt et al. held true for rough, more complex surfaces when path tracing and more complex lighting models are considered. That is: Does highlight disparity strengthen perceived gloss on rough surfaces? This was the area we decided to investigate and as such what the rest of this thesis will therefore cover.

5.6   Gloss BRDF Investigation

In the previous section, we discussed the literature on roughness, glossiness and stereoscopic vision. We found that there was a lack of research in each of these three topics that used more realistic lighting models and more complex, rough, naturalistic surfaces. To that end, in this section we will discuss seven of the more common bidirectional reflectance distribution functions (BRDFs) and justify why we selected the one we used in our experiments.

We note at this stage that there are many more BRDF models than those we will discuss here. It would be too time consuming to consider every possible lighting model available, and as such there have been pragmatic decisions made into which models will be talked about here, using both the criteria set out in *Section 5.2* and what is common in the literature.

5.6.1   A Note on Path Tracing, Bounces, and Sampling

As discussed in *Section 5.5* (and shown in *Figure 5.5*) interreflections and indirect illumination are important when creating complex, rough stimuli such as the ones we used in this thesis. One way of taking account of this is to use path tracing with multiple bounces instead of a single bounce. This technique has been used by Vangorp et al. and Qi to account for self-shadowing, soft shadows and interreflections [48, 96].

As such, we intend to use multiple bounces in our rendering process (10 multiple bounces, as used by Qi). This consideration was important as the BRDF model we used would have to be calculated a maximum of ten times for each light probe, thus increasing our rendering time by up to a factor of ten. We limited the bounces to a

maximum of ten to fulfil criteria *(C5.5)* so that our stimuli were created in a reasonable time frame.

Finally, it is important to note at this stage that we will be using a derivative of the Monte-Carlo method to perform the required sampling in our rendering process. In our case, we will be using the 'Metropolis Light Transport' method, first proposed by Veach and Guibas [101]. Details of the method used are available in *Section 6.3.5.*

5.6.2   *Lambertian Model*

The Lambertian model of reflectance defines a perfect diffuse reflection surface using a cosine law [102]. It is one of the simplest lighting models as it defines that any surface irradiated by light will reflect it in all directions equally. It is completely devoid of specular highlights or glossy behaviour, and as such causes any surface rendered with this model to appear perfectly matte.

Obviously, as we were considering glossy surfaces this model would not be suitable for our work, but it was necessary to mention briefly as many models of gloss attempt to create a spectrum between matte and glossy surfaces where 'Lambertian' surfaces are often considered to be the perfect matte surface.

5.6.3   *Phong/Blinn-Phong Models*

The Phong model is lighting model originally suggested for more realistic offline rendering which is now commonly used in real-time graphics rendering [95]. It is an empirical model of local illumination which simulates specular reflections by considering the viewing direction in relation to the light source. It was published in conjunction with a new interpolation method which allows the reflection calculations to be performed per pixel, rather than per polygon. This enables much more realistic curved surfaces to be generated, as seen in *Figure 5.6* and is now common (along with the less accurate Gourand shading) in most real time rendering systems as it is easily available at the shader level in modern graphics hardware.

The Phong model is additive with an ambient, diffuse and specular component being calculated separately before being combined for the final image. In general, however, the Phong model is not considered to be physically correct. For example, the ambient component is used to account for the effects of interreflection and light scattering in the scene but is obviously a basic approximation of the real phenomenon.
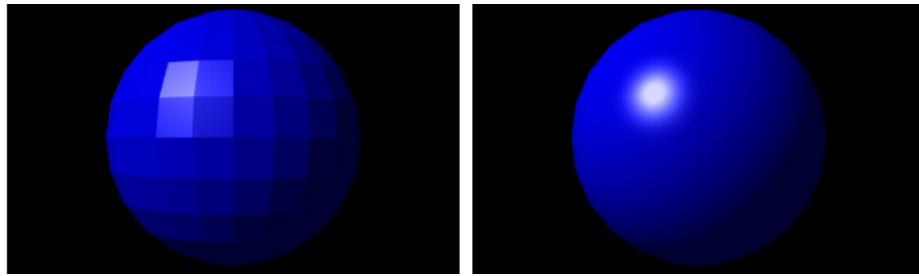
Figure 5.6 - Flat shading vs. Phong shading . Both spheres have the same number of faces, but the one with Phong shading (right) appears much more spherical than the one without (left). The image is public domain.

The Blinn-Phong (or modified Phong) model [103] is also popular. It is the default shading model in both the OpenGL and Direct3D fixed-function pipeline. Instead of calculating intensity components using a scalar product calculation as in the Phong model, the Blinn-Phong calculates the halfway vector. This leads to slightly different, larger specular highlights at the same exponent, but more importantly performs better when flat, glossy surfaces are viewed from a steep angle. Indeed, Ngan et al. found that in general the Blinn-Phong behaved more accurately than the Phong model when compared to measured data [104].

In our case, however, we wanted to use a more complex lighting model than either of these models can provide. One of the reasons the Phong and Blinn-Phong models can be calculated in real time on modern hardware is they are mathematically simple and don't take account of many of the physical laws of light such as Helmholtz reciprocity or conservation of energy. Therefore, if we refer back to our criteria at the beginning of this chapter, we find both models fail *(C5.3)* in that they don't appear naturalistic.

Instead, we will present our investigation of models which perform better in the upcoming sections, starting with the commonly used Ward model.

5.6.4   *Ward Model*

As discussed in *Section 5.5*, the Ward model [97] is a popular choice in much of the more recent psychophysical literature where the advances in computing power have made it possible to render more realistic stimuli in less time. In addition, thanks to it being the rendering model of the Radiance software [98] and the work by Pellacini et al. and Ferwerda et al. [99, 100] it is freely available and has physically meaningful parameters.

Indeed, the Ward model does have apparent significant advantages over the Phong and Blinn-Phong models. It takes account of reciprocity and energy conservation and is relatively fast to render stimuli with, as it is less complex than some of the models we will discuss later. It is, however, lacking in other areas. For one, it does not account for the Fresnel effect in the formula, unlike some other reflection models. In addition, Ngan et al. found that when measuring the model against real materials (100 BRDFs which were captured with approximately 4 million measurements each by Matusik [105]) it performed similarly to the Blinn-Phong model [104].

This was obviously a worrying result, as it shows that the Ward model is incapable of creating the level of realism we were looking for in our stimuli. If Blinn-Phong rendered stimuli failed criteria *(C5.3)* so must any model which performs similarly compared to real data.

We therefore decided to investigate a more computationally expensive but accurate BRDF model. As Ngan et al. found that the Cook-Torrance model consistently outperformed the other models we have already discussed, we investigated that next.

5.6.5   Cook-Torrence Model

The Cook-Torrence model was the first theoretical reflectance model, proposed by Cook and Torrence in 1981 [106, 107]. It was based on work investigating how electromagnetic waves reflect off rough surfaces and the concept of 'microfacets' [108]. These are small, perfectly flat, planar elements which make up a seemingly smooth surface and help to explain why specular reflection on many surfaces is blurred and not a perfect mirror.

Qi goes into this concept in more detail [48], but in general microfacets are considered as having their own normals, several of which are distributed around the local mesoscale surface normal. As only the microfacets with the 'correct' normals are considered in the specular reflection calculation, the degree to which these normals differ from the surface normal is defined by the microscale roughness of the surface and thus determines the distribution of the specular highlights. Put in simple terms, the smoother the microscale roughness, the tighter the distribution of microfacets normals and therefore the more 'specular' the reflection and vice versa.

As a brief side note at this point, He et al. suggested a more complete (and therefore complex) BRDF model known as the He-Torrance-Sillion-Greenberg model which accounted for all known physical phenomena at the time [109]. In general, it is similar

to the Cook-Torrence model, but introduces an additional 'coherent reflection' term which He et al. showed was important. As mentioned in the previous section, however, Ngan et al. found in their experiments that the Cook-Torrence model matched real data 'quite well' and outperformed the other models discussed (Blinn-Phong and Ward) consistently over all their comparisons [104].

While these models both seemed appealing to us due to their accounting for so many physical phenomena, coupled with the results from Ngan et al., they both suffered from a similar problem: complexity. Both of these models are time consuming to render, even on modern rendering hardware, as they are computationally complex. This problem was exacerbated when considering the complex surfaces we decided to use in this thesis and the fact we were using a high number of light bounces to account for interreflection and soft shadows. It was this consideration which led us to decide that the Cook-Torrence (and by extension the He-Torrance-Sillion-Greenberg) model failed criteria *(C5.5)*. In other words, our stimuli would not be created in a reasonable time frame.

Therefore, in the next section we will discuss the next BRDF we investigated which is designed to be as similar as possible to the Cook-Torrence model while being much more computationally efficient through use of 'rational fraction approximation'.

5.6.6 *Schlick Model*

The Schlick model is presented as an intermediary model between the faster, less physically accurate models (such as Blinn-Phong) and the accurate, but computationally complex theoretical models (such as Cook-Torrence) [110]. It observes the main physical considerations of light, such as Helmholtz reciprocity, energy conservation and microfacets but is designed to be both efficient and to have understandable, parsimonious parameters.

Ideally, as discussed in the previous section, the Cook-Torrence model is to be closest to real, measured reflectance data but it was too computationally complex for our purposes. The Schlick model solves this problem by using 'rational fractional approximation'. In simple terms, they suggest approximations for three of the most time consuming elements of the Cook-Torrence model using only basic arithmetic operations and then measure the error and speed-up factor between the approximated and original function.

The Schlick model offers three of these approximated functions. One for approximating the Fresnel factor, one for approximating the geometrical attenuation coefficient and

one for the slope distribution function. Overall, Schlick calculates that with less than 3% total error, these three approximations offer a speed-up of 20 times. To give that speed-up context, if our rendering were to take a year using the Cook-Torrence model, it would take only 18 days with these approximations[5]. Indeed, it is common in computer graphics to use the Schlick approximation for the Fresnel term in the Cook-Torrence model by default as it was the most accurate approximation (0.6% error) and gave the best speed-up (3180% or ~32 times faster).

These points already made the Schlick model appealing, but the final model is more than a simple approximation of the Cook-Torrence model. The Schlick model provides three simple, understandable parameters which was important to us stated in criteria *(C5.1)*. The first is a reflection factor, the second a roughness factor (in this case microscale roughness, which as discussed in *Section 5.1.1* is outside the scope of this thesis) and a third isotropy factor. As we were only interested in isotropic surfaces and we were not considering microscale roughness, this left us with one easily understandable parameter related to gloss strength. Furthermore, this dimension is understandable as it ranges from 0, which is a perfectly matte (Lambertian) surface, to a physically impossibly glossy one[6]. One final useful feature of the Schlick model is that the authors also provide a method of using the BRDF with Monte-Carlo sampling techniques like those used in most path tracing methods. More details on the selected path tracer are provided in *Section 6.3.5*.

It was for these reasons that we decided to use the Schlick model to render our stimuli for this thesis. Not only are the results similar to the Cook-Torrence model, which has been shown to be the most accurate when compared to real data, but it performs much quicker and has a single parameter which we can manipulate to control the glossiness of the stimuli.

5.6.7   *Ashikhmin-Shirley Model*

In the previous section we argued the case for using the Schlick model as our BRDF due to its close approximation to the Cook-Torrence model while being much less computationally complex and having a single, understandable parameter which controls strength of gloss. Before continuing, however, we wish to briefly discuss the other

---

[5] In actuality, our 780 stimuli took approximately 25 days to render rather than an estimated 500 days or 1.36 years with the Cook-Torrence model.
[6] A value of 1 for this parameter would mean the index of refraction is infinite, which is impossible.

model which Ngan et al. found was as good as the Cook-Torrence model [104], the Ashikhmin-Shirley model [111].

Using our own reasoning from these BRDF sections, it was clear that if the Ashikhmin-Shirley was a good approximation of the ground truth then it should have been a contender in our decision of lighting model. Indeed, Qi used this model to render stimuli similar to our own in his work [48]. Qi's decision, however, was affected by the choice of path tracing software and therefore somewhat pragmatic. The same was true in our case as we were planning to use LuxRender [112] as our path tracing software and the BRDF model used in the version we chose was the Schlick model. As both the Schlick model and Ashikhmin-Shirley model were viable for use, we decided to use the Schlick model for simplicity of implementation.

5.6.8   Conclusions

In this section we discussed seven of the most common BRDF models in use today and decided that due to the good compromise of accuracy and computational complexity we would use the Schlick model for lighting our rough stimuli. As before, we made these decisions in reference to the criteria set out in *Section 5.2* simplified versions of which are as follows:

(C5.1)   **The creation parameters of the stimuli should be parsimonious and predictable.**

(C5.2)   **The stimuli should appear as contiguous, acceptable surfaces.**

(C5.3)   **The surfaces should appear naturalistic.**

(C5.4)   **The surfaces should be isotropic.**

(C5.5)   **The stimuli should be created in a reasonable time frame.**

To conclude this section, we present the succinct reasoning for our decision in *Table 5.3* to provide the reader with a concise overview of this section and our decision.

BRDF	Succinct Reasoning	Criteria Failed
Lambertian	Not designed for glossy surfaces, mentioned as it is the model used for perfect matte reflection.	*None*
Phong/Blinn-Phong	Doesn't take account of many of the physical laws of light such as Helmholtz reciprocity or energy conservation and found by Ngan et al. to be a poor approximation of the ground truth.	*(C5.3)*
Ward	Was found my Ngan et al. to be no more accurate than the Blinn-Phong model when compared to the ground truth, despite the additional complexity.	*(C5.3)*
Cook-Torrence	While shown to be a close approximation of the ground truth and therefore desirable, it is computationally complex and highly time consuming to use in conjunction with multiple bounces in path traced rendering	*(C5.5)*
Schlick	A close approximation of the Cook-Torrence model, but simpler and quicker. In addition it has parsimonious and simple parameters including one for gloss strength.	*None*
Ashikhmin-Shirley	While also a valid choice of BRDF model, we chose Schlick over this for simplicity as Schlick is native to our path tracing software.	*None*

Table 5.3 - A brief summation of the BRDF models surveyed. The succinct reasoning of why a model was rejected is presented in each case, expect for the shaded row which we chose to use. In this case the reasons for accepting it are presented. Where a model failed to meet one of our criteria, the criteria failed is offered also.

5.7   Light Source

The final decision we needed to make concerning our stimuli was what type of light source we should use. The lighting of our surfaces was vital as it is the lighting in combination with the BRDF model which actually determines the colours in the final image. In this section we will briefly discuss the five different classes of light source we considered.

The five classes of light source discussed in this section can be easily sorted into order of complexity, as shown in *Figure 5.7* below. We will proceed to discuss these light source types in order.

Ambient	Directional	Point/Spot	Area	Environment Map
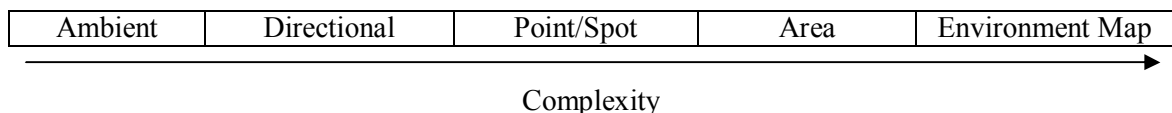
Complexity

Figure 5.7 - Light source complexity

Ambient light sources by their very nature are not physically accurate. They are used to counteract a lack of indirect lighting in many rendering processes. In our case we do not

need to consider using ambient light sources as the use of path tracing and a BRDF which considers microfacets and light scattering will light the scene more accurately.

Directional (or 'collimated') light sources are more physically accurate in some circumstances but are, in general, considered to be at an infinite distance. This type of light was rejected, however, as it will provide little or no light variation across the surface thanks to the source being at an infinite distance and collimated.

Point and spotlight were also rejected. Neither is realistic as there are no real light sources which originate from an infinitely small, single point. This assumption leads to unrealistic hard shadows being cast by objects or surface contours.

The area light source (sometimes called an 'extended' source) is a surface which emits light uniformly in a hemisphere centred on the surface normal. The result is an emitted light pattern similar to florescent lights often found in offices. This type of light source is much more complex to render and thus has not found much use in real time computer graphics. In contrast, it is well suited to realistic offline rendering as it is analogous to a real light source and allows for more realistic soft shadows.

Finally, we considered using environment maps. This technique is common in physically accurate rendering as there are freely available, high dynamic range light probes, such as those from Debevec et al. [113]. For our work, however, we decided not to use environment maps like those used by Qi in his work [48] as we did not wish to consider the colour variation present in environment maps in this thesis.

Therefore, we decided to use white area light sources to light our stimuli as they offered a good balance of control and physical realism.

5.8   Summary

In this chapter we have examined literature concerning roughness and glossiness and decided to investigate the following question: Does highlight disparity strengthen perceived gloss on rough surfaces?

After defining important criteria for our stimuli in *Section 5.2*, we began by investigating what work has already been performed on gloss perception using real stimuli in *Section 5.3*. We decided not to use real stimuli in our experiments as we did

not have the facilities to either create or measure the glossiness of real surfaces with any accuracy and we desired more control over the creation process.

As we were not using real stimuli, we investigated methods of generating synthetic surfaces instead. Using our criteria we investigated eight different surface generation techniques (four spatial, four frequency-based) and decided to use $1/f^\beta$ noise surfaces in *Section 5.4* due to their isotropic, naturalistic appearance and parsimonious parameters.

After selecting a method of creating synthetic surfaces, we investigated what work had already been performed using rough, glossy surfaces in *Section 5.5*. We found that while there has been work done looking at the interaction between surface shape, glossiness and stereoscopic vision, little has been done with rough surfaces and physically accurate lighting and reflectance models. This led us to decide on the question presented at the beginning of this summary.

Following on from this decision, we then investigated the different common bidirectional reflectance distribution functions (BRDFs) which have been used in the literature in *Section 5.6*. We decided to use the Schlick model due to its good compromise between accuracy, accounting for the physical properties of light and computational complexity. In addition, we discussed how we will use path tracing with multiple bounces to address the issues of interreflection and soft, self-shadowing. An important consequence of this is that due to the length of time this model takes to render, it was infeasible to update stimuli in real time, so all stimuli were rendered offline beforehand.

Finally, in *Section 5.7* we decided to use an area light source to light our stimuli as it allows for good control over the lighting, offers a realistic representation of a well known light source and allows for soft shadowing when used in conjunction with path tracing techniques.

Now we have finalised all the salient points required for generating our stimuli, in the next section we will discuss in detail how we created the stereoscope and how the stimuli were created using $1/f^\beta$ noise surfaces, rendered under an area light source with path tracing and the Schlick lighting model.

Chapter 6
Stereoscope and Stimuli Creation

6.1   Introduction

In the previous chapter, we investigated the literature on roughness, glossiness and stereoscopic vision and decided to investigate the following question: Does highlight disparity strengthen perceived gloss on rough surfaces?

During this literature review, we decided to use $1/f^\beta$ noise surfaces as our rough surfaces and to use the Schlick BRDF with path tracing. This, in addition to the decision to use a mirror stereoscope in *Chapter 4* allowed us to begin creating both the stimuli and presentation method required. In this chapter, therefore, we will go into detail concerning both how we constructed our stereoscope and our stimuli.

In *Section 6.2* we describe the decision process taken to determine what type of mirror stereoscope to build. Following this, we describe how we overcame some problems inherent in the design of stereoscopes, before providing enough details for the reader to build their own. In *Section 6.3* we detail the steps taken to create our glossy surfaces and provide all the information for similar stimuli to be recreated.

6.2   Building a Stereoscope

In *Chapter 4* we discussed how stereoscopic displays would be useful for presenting stimuli to our participants. We concluded that although there are many consumer versions of this technology in the marketplace at the moment (such as Nvidia's 3D Vision system), they were not suitable for use in psychophysical experiments such as those we planned to use due to their poor separation between the left and right eye images and low light transmission. We therefore decided to build a stereoscope for use in our experiments as they allow perfect separation and better light transmission. In this section we discuss in detail how it was designed and constructed and how this had a bearing on the creation on the stimuli.

6.2.1   Choosing the Design

Before we could begin we needed to decide which design of stereoscope we would use. The initial stereoscope design, as proposed by Sir Charles Wheatstone uses two mirrors and two panels upon which he affixed aligned images [114] as shown in *Figure 6.1*, below.
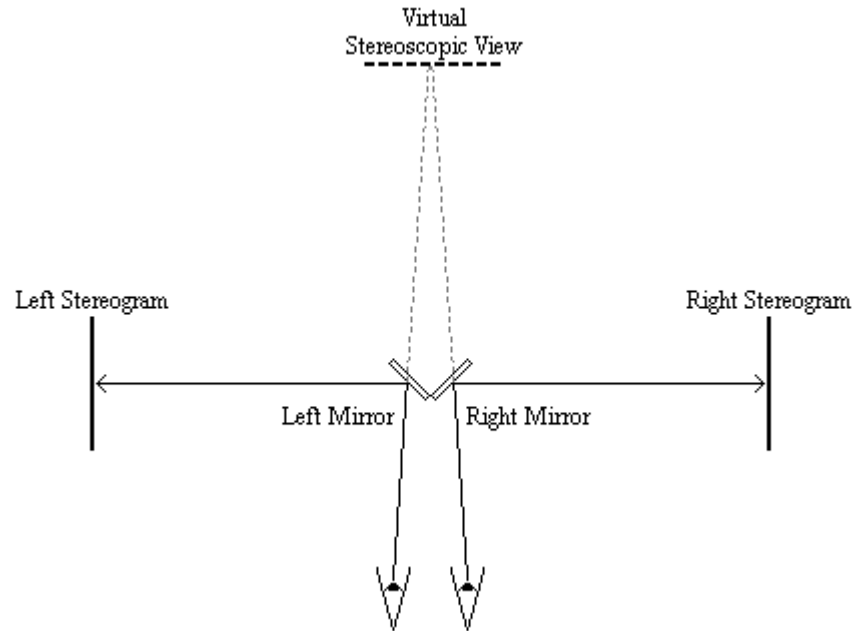
Figure 6.1 - A diagram of a Wheatstone stereoscope. This system uses two central mirrors positioned at 90° to each other to redirect the left and right eye to two different stereograms which in our case would show two projections of the same object or surface.

As we were planning to present participant-controlled stimuli we needed to be able to display them via a computer monitor. Therefore, for us, this design would mean using two synchronised monitors, positioned at the left and right stereogram positions.

While this is the simplest design of stereoscope, it did cause three main problems for our planned experimental design:

(i)   Our proposed experimental space was narrow, meaning it would be difficult to find the required width for this design.

(ii)   As it would require two monitors we would have to guarantee that both monitors were perfectly matched and calibrated to the same settings. While certainly possible, this would add unwanted complication to the design.

(iii)   We would need to synchronise the output to two monitors perfectly, guaranteeing no delay between both, again adding unwanted complication.

To combat these problems we decided to use a different stereoscope design which relies on a single display plane. This design is shown in *Figure 6.2* and consists of the two stereograms positioned in front of a participant with two pairs of mirrors redirecting the participant's view to them.
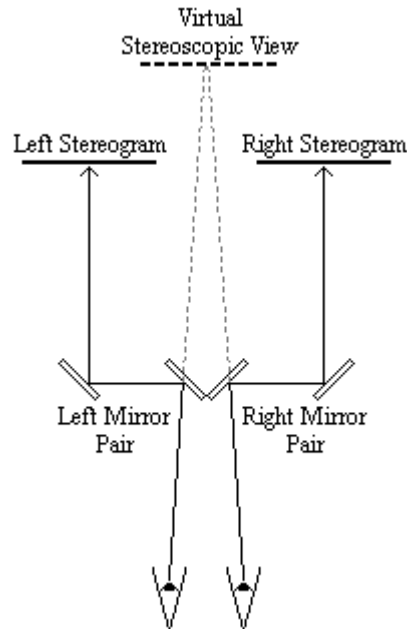
Figure 6.2 - A diagram of a single-plane stereoscope. This system uses two pairs of mirrors to redirect the left and right eye to two different stereograms placed in front of the viewer. In our case the stereograms were shown on a single monitor. We treated the monitor as if it were split in two vertically with the view redirected to the quarter and three-quarter positions.

The major advantage of this design is that it allowed us to use a single monitor instead of two. As the two stereograms are on the same display plane, with a sufficiently large monitor we could split it in two vertically and present the stimuli side-by-side on it. This meant we only had to calibrate and output to a single screen solving problem *(ii)* and *(iii)* easily. This configuration was also smaller, and shaped to suit our proposed experimental space, solving problem *(i)*.

It did, however, introduce its own problems to be considered, as detailed below:

(iv)   As each mirror reflects less than a 100% of the light it receives, adding more mirrors decreases the amount of received by an observer.

(v)   The maximum size of the stimuli is reduced as they are presented side-by-side on a single monitor.

(vi)   The calibration of this system is more difficult as the geometry of the four mirrors is more complex.

It is important to note at this stage that both designs of stereoscope rely on participants keeping their head still as their eyes must remain in the same relative position to the mirrors used. As such, we also added a fourth consideration as stated below:

(vii)   Participants must remain still in relation to the mirror pairs for the stereoscope to be effective.

We still believed, however, that this design was more suitable for our needs. Therefore, we solved these problems in the design of our own stereoscope, as described in the next section.

6.2.2   Single-Plane Stereoscope Considerations

In the previous section we decided to use a single-plane stereoscope instead of the Wheatstone design as it was more suitable for our needs. We also defined four problems with this design which needed to be considered. In this section we specify how we solved these problems.

For point *(iv)* we decided to use mirrors with high reflectance. In our case we chose front-surface aluminium mirrors [115] as they offered a high reflectivity in the visible spectrum (between 400 and 700 nm) as shown in *Figure 6.3*. The mirrors we used had a typical reflectivity in the visible spectrum of 89.01% at a 45° angle of incidence (AOI) as measured by the manufacturers [116]. The fact they were front-surface meant there was no possibility of 'ghosting' (a faint, offset second reflection in the mirror) which occurs in back-surface mirrors due to a small amount of light reflecting off the front glass layer before encountering the reflective coating.

The only remaining consideration with these front-surface mirrors was that the maximum size we could source for this experiment was 50.8mm (or 2") square. This limited the maximum size of stimuli we could present to a participant (depending on the distances used) and dictated that the mirror pairs (as seen in *Figure 6.2*) had to be placed close to the participant's eyes to give the maximum possible viewing cone.

Once we had decided upon the mirrors, point *(v)* became less of a consideration as the maximum size of the stimuli was defined by the mirrors rather than the monitor. Even so, we decided to use a large monitor with a high resolution in our experiments. The monitor selection was, however, most affected by the need for colour-correctness and ease of colour calibration (described in more detail in *Section 7.2.2*). We used a 24-inch

NEC PA241W monitor which had a resolution of 1920 by 1200 pixels [117]. This monitor was sufficiently large and had a resolution which allowed stimuli of a maximum of 600 pixels square at a distance of 85 cm from the viewer when the mirrors were taken into account.
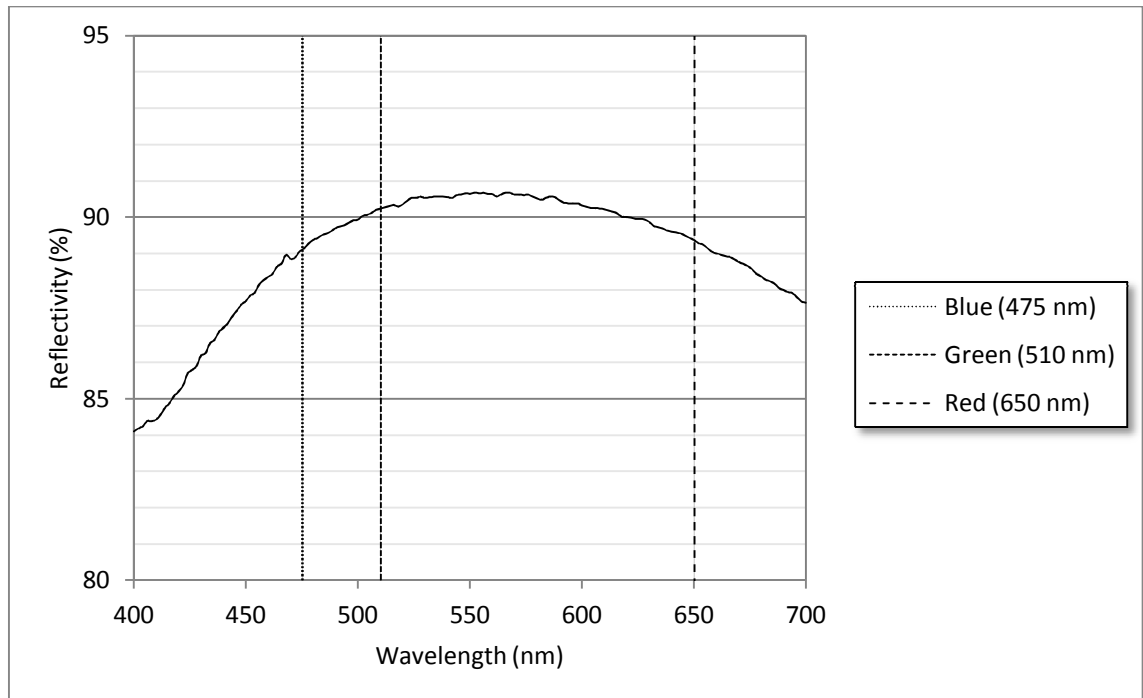
Figure 6.3 - Typical reflectivity at 45° AOI for the visible spectrum (400-700 nm) for the front-surface aluminium mirrors used in the stereoscope. Measurements from the manufacturers [116].

The solution we chose for point *(vi)* was to create a custom calibration program which we ran before experiments. This consisted of two different scenes which could be used to adjust the mirror setup for different participants. In general, the author would calibrate the system before a participant was asked to partake in the experiment as he was able to get the best calibration from experience. Then for each new participant, the mirror setup was adjusted for their own unique and slightly varying interocular distances.

The first scene used was the one shown in *Figure 6.4* and consisted of two different coloured, square targets split by a white line. This allowed simple questions to be asked to confirm that the participant had working binocular vision. The targets were different colours to allow easy differentiation between the left and right eye and had different rings for different parts of the calibration. For example, the smallest squares were used for fine adjustment of the mirrors, while the middle square was used for general

positioning. The participants could never see the outer squares or the white line bisecting the screen under normal conditions.
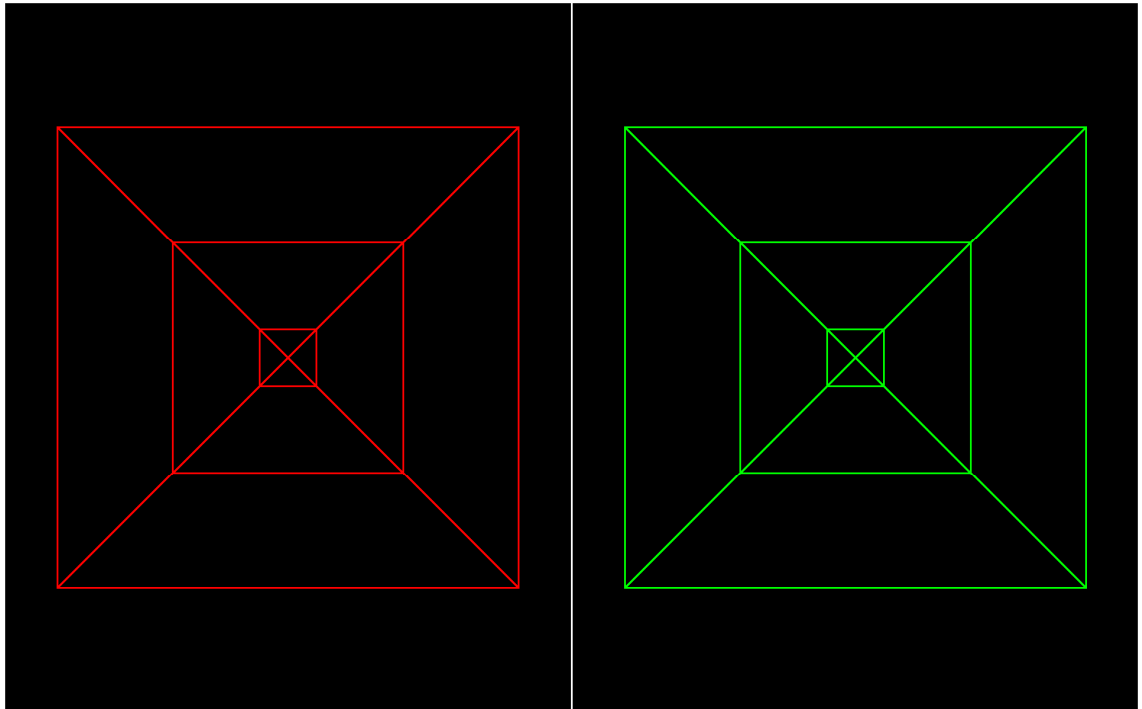
Figure 6.4 - The calibration screen used for our stereoscope. The left (red) target was where a participants left eye was adjusted to, the right (green) target likewise for the right eye. The participant was never able to see the outer square or bisecting white line under normal conditions.

One problem caused by calibrating for each participant was that the author was not able to easily confirm the calibration due to the differing interocular distances and alignments present between humans. Therefore we had a second scene consisting of two stereograms of a sample surface rendered in the same style as the surfaces participants would be seeing in the experiment. The participant was then asked questions about the surface to make sure they were seeing it with the required disparity to give virtual depth to the surface. Once the author was happy the participant was seeing the sample stimuli correctly, the configuration program was closed and the experiment begun.

Finally, for point *(vii)* we decided to use a standard adjustable chinrest in order to keep participants' heads in the same relative alignment with and distance from the mirror pairs used. This was mounted in front of the mirrors and would keep the participant's head still during use of the stereoscope.

The four solutions detailed in this section gave us the confidence that our stereoscope was the best method of presenting virtual three-dimensional surfaces to participants for

our experiments. In the next section we detail the actual construction of the stereoscope used, based on the design decisions and considerations already discussed.

6.2.3   Construction Details

In this section, we will briefly discuss the final build specifications of the stereoscope created based on the design considerations discussed in the previous section.

The four front-surface aluminium mirrors (which are described in more detail in *Section 6.2.2*) were mounted on compact kinematic mirror mounts by affixing the glass back of the mirror directly to the metal front of the mount with adhesive. This meant there was no surround on the mirror, and as such all of the 50.8mm square reflective area was useable.

The mounts used were two-adjuster mounts which occupied a volume of 25.4mm x 28.7mm x 33.02mm. The adjusters allowed fine angular adjustment during calibration in both the vertical and horizontal axis of ±4° for a total travel of 8°. These mounts had M4 threaded holes on the bottom which were used to attach them to stainless steel optical posts which had M4 mounting studs at the top.

The optical posts used had a diameter of 12.7mm and were 75mm long. They featured a metric scale engraved every 1mm down the length which allowed us to accurately align the mirrors on the same vertical axis. Coarse angular adjustments around a mirror's vertical axis were performed by adjusting the rotation of the optical posts in relation to the post holders.

The post holders used were 76.2mm in length with an external diameter of 25mm and had a square relief cut down their length for more stable post mounting. They also featured spring-loaded, locking thumbscrews. These put pressure on the optical post while still allowing us to rotate them for coarse adjustments without altering the height of the post. Once the coarse adjustment was finished, the thumbscrews would be tightened and the optical post would lock in position and rotation. These post mounts had a M6 threaded hole on the bottom which allowed us to affix them to rail carriers.

The rail carriers had a 25mm square platform on which the base of the post holder could sit perfectly flat and had a M6 counterbored hole through it so we could join the two without impeding the movement of any of our components. These rail carriers were

designed to attach onto a dovetail-shaped optical rail without requiring access to the rail ends which allowed us to swap the mirror's positions easily.

The optical rail used was a 600mm long, compact dovetail rail. This rail was CNC machined from aluminium to ensure good dimensional stability during use. This allowed us to keep accurate lateral and vertical alignment of our mirrors as the rail would not bow or flex, making calibration simpler and more accurate.
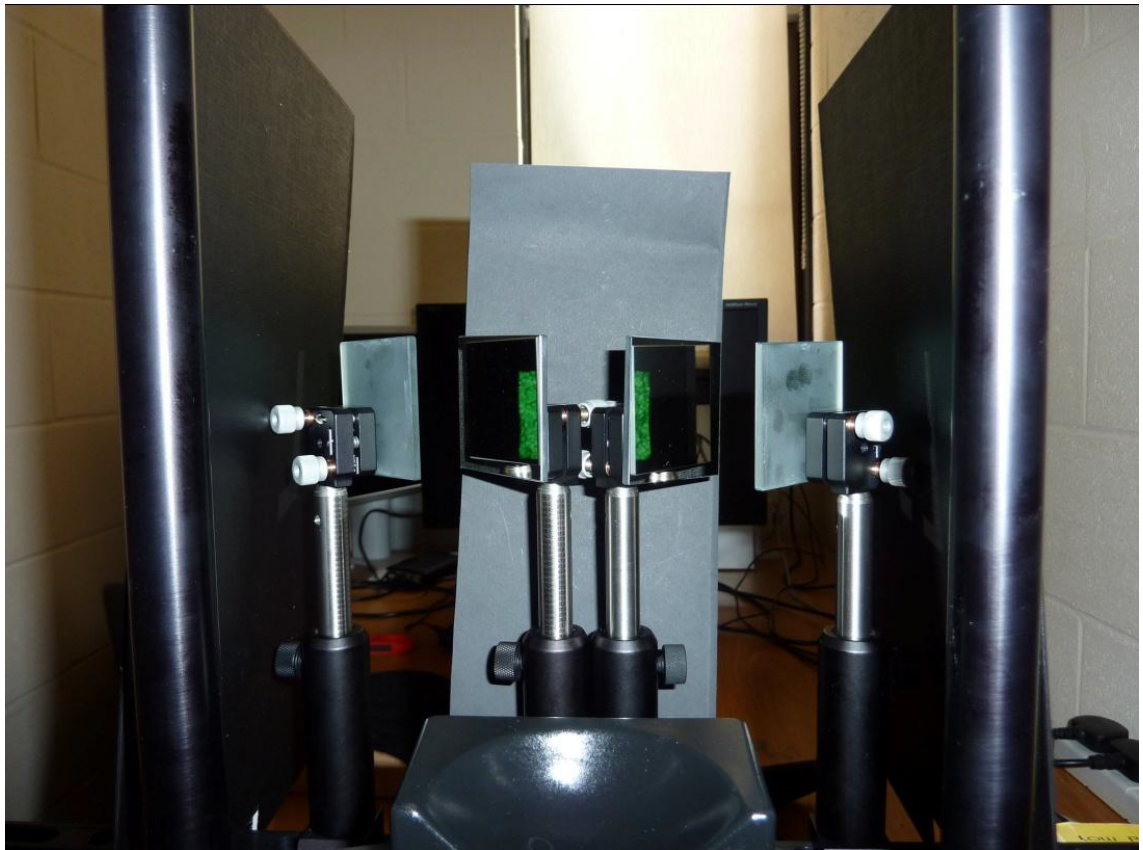
Figure 6.5 - The finished, constructed stereoscope. This picture was taken from behind the in-use viewpoint so the chin rest could be seen in relation to the rest of the stereocope. N.B. When in use the window behind the monitor was covered in a blackout curtain, it has been removed here to improve the lighting for the photograph.

Finally, this optical rail was attached to a thick, flat board so it could be positioned in the correct position above the table. The chinrest used was also clamped to the same board. As both the chinrest and optical rail were now attached together, it ensured the relative position between them was unchanged during the use of the stereoscope. The last component on this board was a 'Bullseye Level' (sometimes known as a 'Circular Bubble Level') which allowed us to ensure the entire stereoscope assembly was level in

both horizontal axes. This was matched by a similar spirit level on our monitor which ensured it was also level and therefore on the same horizontal plane as our stereoscope.

The final stereoscope is pictured in *Figure 6.5* and was used in the experiment detailed in *Chapter 8*. This picture was taken from slightly behind a participant's viewing position and shows the chinrest used to keep their head still during the experiment. When in use, the participant's eyes were less than 50mm from the mirrors to allow a participant to see as much as possible of the monitor.

6.2.4   Conclusions

In this section we have discussed the design and construction of a mirror stereoscope in detail. We believed it was important to finalise the design of this system before creating our stimuli as it would affect how we created the stimuli for our experiments.

We knew that due to the size of the screen and mirrors used in our stereoscope, our stimuli would have to be smaller than 600 pixels square to ensure the entire surface was visible. In addition, we knew what wavelength of light was attenuated least by our mirrors, allowing us to take this into consideration when designing the stimuli.

Now we have presented the design of our stereoscope system, we can discuss in detail the creation of the images used in our experiments.

6.3   Creation of the Stimuli

In *Chapter 5* we decided on both the type of rough surface we used to create our stimuli and how we were to light it. In this section we will discuss how we went about creating these stimuli in more detail.

We begin by explaining how we created the $1/f^\beta$ noise height maps, before describing how these were converted into solid surfaces through a process of subdivision and per-vertex displacement. We then discuss in more detail the Schlick reflectance model proposed in *Chapter 5* which we used to make our surfaces appear glossy. During this we discuss how, as we had to render these stimuli offline, we quantised our gloss measure to discrete levels. Finally, we discuss the details of the rendering process which don't fit into other sections (such as the path tracer used) before giving physical details on the finished stimuli.

6.3.1   Creating the Height Map

The first step in creating our stimuli was to generate a $1/f^\beta$ noise image which could be used as a height map. These height maps had a magnitude spectrum $H(f)$ scaled with spatial frequency as shown in *Equation (6.1)* below, where β is the roll-off factor of the magnitude spectrum, σ is the RMS height of the surface and $N(\beta)$ is a normalising factor:

(6.1)
$$H(f) = \frac{\sigma}{N(\beta)} f^{-\beta}$$

As with the height maps used by both Padilla and Qi, our height maps had random phase complex conjugate symmetry in the frequency domain which was then converted to the spatial domain through an inverse discrete Fast Fourier Transform (FFT) [7, 48]. The imaginary part of the FFT was zero for all stimuli created due to the complex conjugate symmetry and was ignored.

We decided at this stage to use five mesoscale roughness categories, as defined by the roll-off factor β. We chose to use β-values of 1.8, 2.0, 2.2, 2.4, and 2.6 (rough to smooth), informed by the work performed by Qi [48]. He examined surfaces of β-value 1.5 to 2.8 and found that any surface with a β-value of less than 1.8 had a dramatic drop in their apparent glossiness. We took 1.8, therefore, as our lowest β-value. As we wanted to examine a wide range of surface roughness levels, but were limited by the rendering time available, we decided to sample five categories with steps of 0.2 from our roughest stimuli up to what Qi considered was a smooth surface.

It should be noted at this stage, we only consider the effect of the roll-off factor in this thesis. Qi has shown that illusory gloss is not only apparent on Lambertian (matte) surfaces with high σ values, but also the perceived strength is altered by changing the σ value [48]. As Padilla showed that participants could consistently rate the change in roughness caused by increasing σ or β [7], we decided to standardise σ and concentrate on β to avoid possible interactions between illusory gloss and real gloss.

As we weren't interesting in discussing the effect of changing the RMS height of a surface, we standardised this in the same way as Qi. Before a surface was created from a height map (as discussed in the next section), the height map was standardised to unit RMS height and then scaled to the desired RMS height σ during the surface creation step. This was done by calculating the normalising $N(\beta)$ value from the standard

deviation of each individual height map, ensuring the final rendered surfaces all had identical, zero-mean RMS height.

Finally, the magnitude spectrum for the height maps was sampled at a frequency of 256 cycles per image width (cpi), giving a final height map (and hence image) resolution of 512-by-512 samples square. These height maps were saved in 16-bit per colour channel, Portable Network Graphics (PNG) format, allowing a possible 65,536 different height values for each sample. In tests with 8-bit per channel height maps (offering 256 height values per sample), a 'layering' effect was clearly pronounced due to the quantisation of the height values, as can be seen in *Figure 6.6*, below.
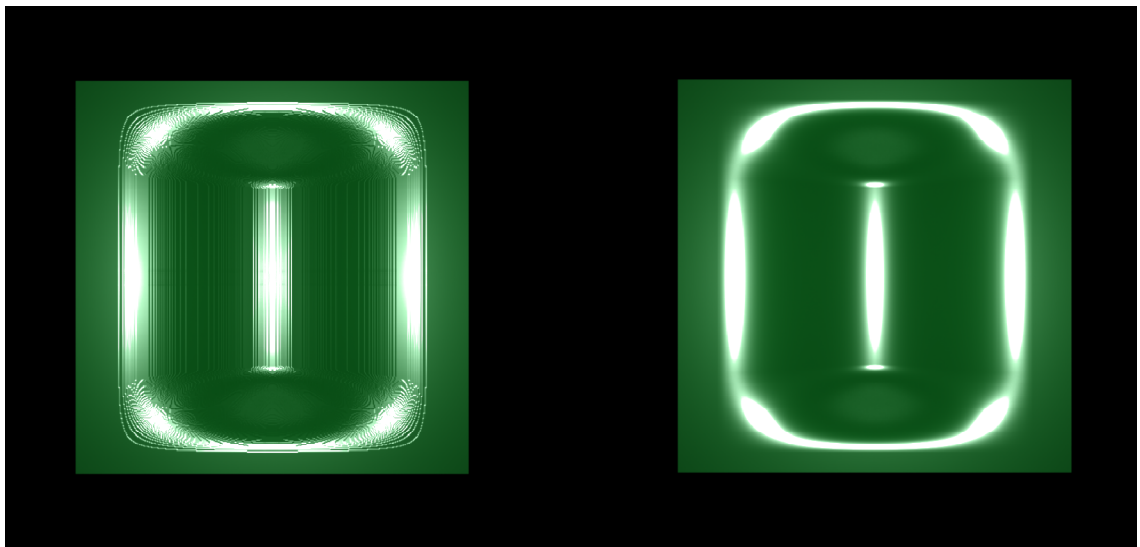
Figure 6.6 - A comparison of 8-bit vs. 16-bit rendered height maps. Left: 8-bit height map, Right: 16-bit Height Map. N.B. This surface did not appear in the experiments and is presented here as an example of the layering effect produced by the quantisation present in an 8-bit height map.

At each roughness level we created four different height maps each with different random phase as Padilla showed that random phase does not change perceived roughness [7]. These twenty height maps (along with our numbering scheme for them) are available in *Table C.1*.

6.3.2   Creating the Surface

The next step in creating the stimuli was to create the three-dimensional representation of the surface which would be rendered, using the height map discussed in the previous section to give it naturalistic relief.

We chose to use a process of per-vertex displacement as Padilla found this to be more accurate than per-pixel displacement techniques [7]. This process was performed in the following steps:

1. Create a simple plane, defined only by its four corner vertices.

2. Subdivide this plane nine times to give a high-resolution plane defined by 512-by-512 vertices. This allowed for 262,144 different points to displace.

3. For each vertex, sample the height map at the correct location and displace it based on the value found where a height map value of 0 gives maximum displacement and a value of 1 is no displacement.

An example plot of a plane which has undergone the per-vertex displacement step in this procedure (step 3) is presented in *Figure 6.7*, below. It shows a simplified example using a 32-by-32 pixel $1/f^2$ noise height map and a plane subdivided five times to give 32-by-32 vertices.
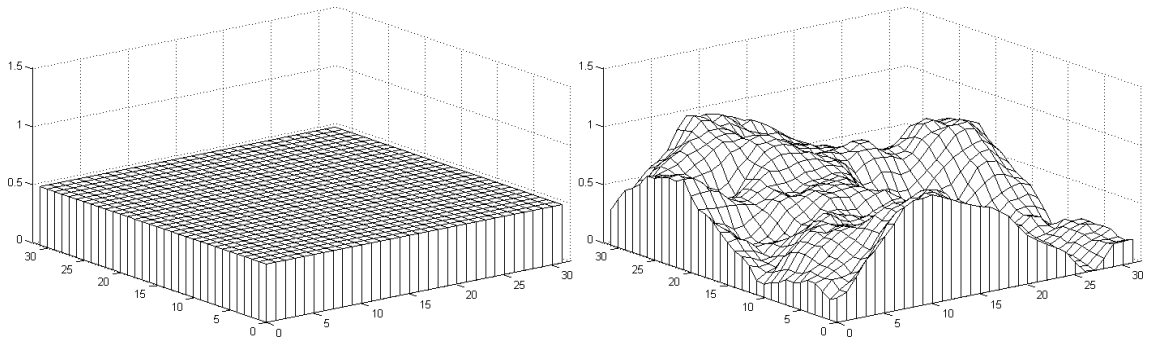
Figure 6.7 - An example plot of a plane subject to vertex displacement. Left: A surface from a mid-grey height map. Right: A surface generated from a 32 x 32 pixel $1/f^2$ height map. N.B. This is just a representation of the technique used. While this plot shows quads, we used triangles in our final rendering process.

Following this procedure, we created twenty surfaces of varying roughness defined by 262,144 XYZ triplets, based on the height maps created in the previous section. It should be noted at this point that the amount each surface was displaced in the z-axis varied slightly across the surfaces to ensure each had the same RMS roughness. The displacement amount was calculated by dividing our target σ-value of 0.2 by the normalising value, $N(\beta)$ which was the standard deviation of the height map used to displace the surface. This ensured each surface had unit, zero-mean RMS roughness.

These surfaces were then ready to be placed into a scene so they could be rendered, as discussed in the next section.

6.3.3   Creating the Scene

Now we had our surfaces defined, we placed them into a simple scene so they could be rendered. For this we used Blender, a popular piece of 3D modelling software [118].

The scene we created was very simple and consisted of three components: our displaced surface, a virtual camera, and a square, white area light source. Our surface was 6 units square and was positioned 10 units from the camera. Our area light source was 2 units square and positioned centred at the central camera position. The maximum displacement for a surface in this scene was 1 unit. Every surface used was positioned flat to the camera and lit by exactly the same light source, in the same relative position to make sure there was no perceived difference in roughness which can occur from differing illuminant or viewing angles [1, 2].

At this point, we note that for every stimulus we created, we required three versions of it, seen from three different perspectives. This was to allow us to show the stimuli in both monocular and binocular conditions. To do this, we had a left, right and central camera position. The left and right camera positions were offset ±0.5 units from the central one to mimic an observer's interoccular distance. The two images resulting from this were then used as left and right stereograms in our stereoscope (described in detail in *Section 6.2*) to give virtual depth. The central camera position acted as an average view of the surface and was presented as both left and right stereogram to give an image which appeared without any virtual depth, as if it had been viewed on a monitor without use of a stereoscope. This was our monocular condition as both eyes saw the same image even thought it was through a stereoscope.

Now we had created the scene, in the next section we discuss our use of the Schlick BRDF model.

6.3.4   Lighting the Surface

Now we have described both how we created the surface geometry and how we laid out the scene, we will discuss in more detail the Schlick reflectance model we used, as chosen in *Section 5.6*. In addition, we will state the exact settings we used with this model so others can recreate similar surfaces as desired.

In general, the Schlick BRDF has three parameters:

- $C_\lambda \in [0, 1]$: Reflection at wavelength λ

- $r \in [0, 1]$: Microscale Roughness factor (0 being perfectly specular)

- $i \in [0, 1]$: Isotropy factor (1 being perfectly isotropic)

As already stated in *Chapter 5*, we are only examining the effect of mesoscale roughness in this thesis, and only considering isotropic surfaces. Therefore, both r and i were kept constant for all the stimuli we created. This left us only a single parameter C_λ which was altered for our surfaces, which according to Schlick is analogous to '*the reflectivity at normal incidence*' or the strength of the specular reflection.

Before we can discuss the Schlick BRDF in detail, we must first define the required vectors and angles. Schlick's original definitions of these are reproduced in *Figure 6.8*, below [110].

$V =$	Outcoming direction of light
$V' =$	Incoming direction of light
$N =$	Surface normal vector
$T =$	Surface tangent vector
$H =$	Bisector vector of V and V'
$\bar{H} =$	Projection of $H \perp N$

$t = (H \cdot N)$	$\alpha = \cos^{-1} t$
$u = (H \cdot V)$	$\gamma = \cos^{-1} u$
$v = (V \cdot N)$	$\theta = \cos^{-1} v$
$v' = (V' \cdot N)$	$\theta' = \cos^{-1} v'$
$w = (T \cdot H)$	$\varphi = \cos^{-1} w$
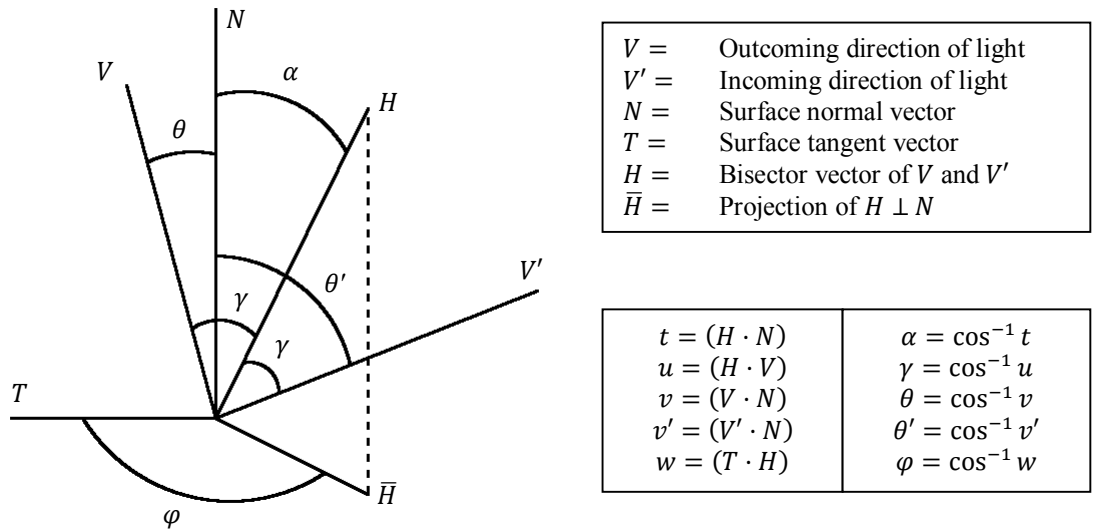
Figure 6.8 - Required angles and vectors for defining the Schlick BRDF

The Schlick model at its most simple is defined as a multiplication between a spectral and directional factor, as in *Equation (6.2)*:

(6.2)
$$R_\lambda(t, u, v, v', w) = S_\lambda(u) \, D(t, v, v', w)$$

The spectral factor is dependent on the incident angle and obeys Schlick's approximated Fresnel law, as in *Equation (6.3)*:

$$S_\lambda(u) = C_\lambda + (1 - C_\lambda)(1 - u)^5$$

The directional factor is dependent on the light's zenith angle α and the azimuth angle φ and is expressed by two factors $Z(t)$ and $A(w)$, as in *Equation (6.3)*:

(6.4)
$$D(t, v, v', w) = \frac{1}{4\pi v v'} Z(t) A(w)$$

Finally, $Z(t)$ and $A(w)$ are defined as polar equations of ellipses (with differing pole locations) as in *Equation (6.5)*:

(6.5)
$$Z(t) = \frac{r}{(1 + rt^2 - t^2)^2} \quad A(w) = \sqrt{\frac{i}{i^2 - i^2 w^2 + w^2}}$$

At this stage, it is worth remembering that due to the complexity of the Schlick BRDF and the high detail of the surface, it was infeasible to update the stimuli in real time[7]. Therefore all the stimuli had to be rendered offline, meaning to create the illusion of a participant changing the glossiness of a surface, we had to render all possible gloss levels beforehand. This, of course, required us to quantise a continuous parameter into a discrete one. This was done by selecting a linear range of C_λ values from perfectly diffuse ($C_\lambda = 0$) to the glossiest our monitors were capable of displaying. In the end, however, we did not use perfectly diffuse stimuli in our experiments, instead starting at the lowest value rendered with some apparent gloss. This gave us 13 different gloss levels for each surface.

Below, for clarity, we state the exact settings we used for rendering the glossy component of our stimuli with the Schlick formula already defined:

- $C_\lambda \in [0, 1]$: Ranged from 0.01 to 0.13 in steps of 0.01 in all colour channels

- $r \in [0, 1]$: Fixed at 0.14 for all stimuli

- $i \in [0, 1]$: Fixed at 1 for all stimuli ensuring isotropic reflectance

Now we have defined in detail the reflectance model used for our stimuli, in the next section we will cover the final rendering details pertaining to the path tracing methods used.

---

[7] In fact, depending on the roughness of the surface, each image took approximately 90 minutes to render.

6.3.5   *Rendering Details*

In the previous sections, we have defined all the main details of how we constructed our stimuli. In this section, we discuss the final details which are not easily covered in the previous sections.

We used LuxRender, an open-source, cross-platform rendering package as the path tracer for our stimuli [112]. As discussed in *Section 5.5* and *5.6.1*, to account for interreflections and soft shadows we calculated 10 bounces in our path tracing. In addition, each pixel was sampled 5000 times to ensure our stimuli had no 'fireflies' (rendering artefacts which appear as bright specks in the final image) which can occur with too few samples.

As mentioned in the previous chapter, we used the 'Metropolis Light Transport' method (first proposed by Veach and Guibas [101]) to perform our sampling. Specifically, we used an improvement of this technique which improves the speed of convergence proposed by Kelemen et al. [119]. This required two variables to be standardised, called 'Maximum Consecutive Rejects' and 'Large Mutation Probability' within LuxRender. The first defines how many samples are discarded before a change in the light path is enacted. The second defines how likely it is this change will be large (a new sample from elsewhere in the image) or a small, local change. We set these values to 512 samples and a probability of 0.4 respectively.

The surface itself is a regular grid, with each quad within defined as two triangles, joined at their hypotenuse and therefore sharing two vertices. To determine the normal at each vertex, we calculate the cross product of the surface tangents at that point. The tangent components are calculated from a simple weighted sum of the heights of the neighbouring vertices, not including the centre vertex under consideration. Finally, the normal at the point of incidence is calculated by simple normal interpolation across the face.

As the aim was to allow participants in our experiments to be able to alter the glossiness of the surface, we wanted to ensure that only the strength of the specular highlights was changing with the C_λ value, as discussed in the previous section. We did this by creating each stimulus in two layers, a diffuse layer and a specular highlight layer. For each of the surfaces, we rendered a perfectly diffuse version of the surface ($C_\lambda = 0$). To make the difference between diffuse and specular reflection apparent, this diffuse base surface was given a green colour (RGB: 0.1, 0.5, 0.1). Green was selected for two reasons.

First, it was the primary colour most reflective in the mirrors used in our stereoscope (see *Figure 6.3*). Second, it has been used as the colour of choice in other experiments looking at specular reflections [10, 89].

Then, for each surface, 13 different specular highlight layers containing no diffuse component were calculated with the values discussed in the previous section. This allowed for complete separation between the diffuse and specular components of the stimuli. An example of these two layers is available in *Figure 6.9*. Then each of these specular highlight layers were added to the diffuse layer, giving 13 different levels of gloss for each surface. An example of a finished series of images is available in *Table C.2*.
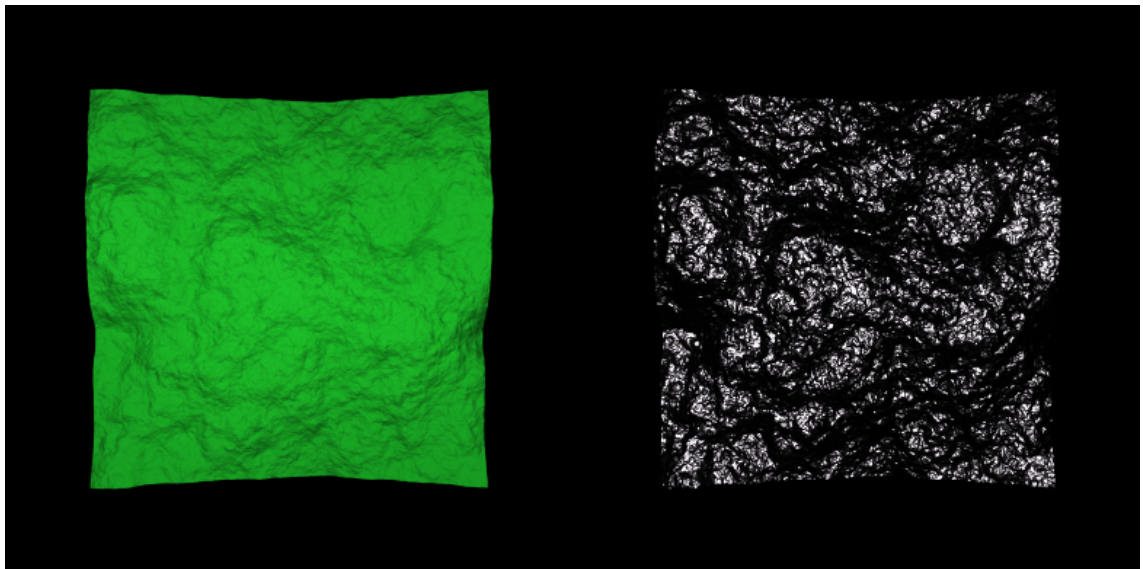
Figure 6.9 - An example of the two-layer rendering technique used. Left: The diffuse layer. Right: The specular layer. The finished stimuli is #5 in *Table C.2*.

Finally, as all of these rendered images were output in OpenEXR format (meaning the stimuli were stored as 32-bit per colour channel, high dynamic range images) we needed to tone map them so they could be displayed on a low dynamic range monitor. This was done with a simple linear tone mapping calculation which clamped any pixels of high intensity to the same, maximum intensity value which was manually tuned to lose as little detail as possible, as performed by Qi with his samples [48].

Now all the details of how these stimuli were created have been discussed, in the next section, we will outline the details of the final images created.

6.3.6   Final Stimuli Details

Using the details discussed in the preceding sections, we created 780 separate images (or stereograms) for use in our experiments. For each of our 5 roughness level we had 4 random phase surfaces, each of which had been rendered at 13 different glossiness levels, from 3 different viewpoints.

The finished stimuli had a maximum possible resolution of 568-by-568 pixels, and therefore a maximum possible size on the 24-inch NEC PA241W monitor used of 152mm square. The maximum difference in depth was 25.3mm in real terms. These stereograms were then presented in the middle of the screen, for the experiments in *Chapter 7*, or side-by-side on a black background with a minimum separation of 10cm (to make sure that there was no cross-talk between an observer's eyes) and viewed through our custom stereoscope for the experiment in *Chapter 8*).

6.4   Summary

In this chapter we have discussed how we created both our stimuli and the stereoscope with which to display them in depth.

To begin with, in *Section 6.2* we detail the design and construction of a single-plane mirror stereoscope. We first decided on the type of stereoscope to use, before discussing the design considerations which needed to be accounted for. We then presented in detail how we built the stereoscope so the reader can reconstruct the system we used as necessary. This was presented first as it had some bearing on the maximum size of stimuli we could use.

We then discussed the stimuli we created for our experiments in *Section 6.3*. Our stimuli were $1/f^{\beta}$ noise surfaces as selected in *Section 5.4*. These surfaces were rendered as static images, using a path tracer and the Schlick BRDF, from three different angles allowing them to be displayed as stereogram pairs with virtual depth, or as simple monocular images. Our stimuli were high detail and thanks to the use of path tracing, considering 10 bounces for each light probe and the use of an area light, take into account the effects of interreflections and soft shadowing. We go into detail about how we constructed and rendered these surfaces allowing the reader to recreate similar stimuli as required.

With these stimuli and our constructed stereoscope we could now investigate if the results found in studies discussed in *Section 5.5* hold true for these more complex, naturalistic stimuli. In particular, we are interested to see whether binocular vision strengthens the perceived gloss on rough surfaces, as the literature discussed in *Section 5.5* states it does with smoother surfaces. In other words, does highlight disparity strengthen perceived gloss on rough surfaces?

Before that, however, in the next chapter we discuss preliminary experiments designed to see whether the differing phases of our surfaces alter the perceived strength of their gloss.

Chapter 7

Does Phase Affect Gloss Perception?

7.1   Introduction

In the previous chapter we described how we created our high quality, naturalistic stimuli and a mirror stereoscope to present them to participants both with and without disparity. In this chapter, we answer an important question: Does random phase affect gloss perception?

As mentioned in *Section 6.3*, we are using surfaces each generated from a different random phase so as to avoid participants comparing two stimuli on a pixel-per-pixel basis. This is to force them to make judgements on the nature of the entire surface, rather than single pixels or features. While previous experiments have shown that random phase has no affect on perception of roughness [6], to our knowledge no-one has confirmed that there is no effect on gloss perception.

To this end, in this chapter we present two psychophysical experiments to see whether random phase alters a participant's perception of how glossy a surface is. First, we investigate a single roughness level to see if there was any effect on glossiness perception from changing the random phase of a surface. This is reported in *Section 7.3*. A second experiment was then carried out so perceptually similar surfaces from each of the five roughness levels at which we have stimuli could be selected. This is reported in *Section 7.4*. These results were used to select two pairs of surfaces at each roughness level to use in other experiments.

First, however, is a short description of the experimental setup we used for these experiments.

7.2   Experimental Setup

Both of the experiments in this chapter had the same setup, the only difference being what stimuli were presented to the participant in each case. Therefore we will only describe this setup once and describe the stimuli used in *Section 7.3* and *7.4*. In both

cases, the stimuli were $1/f^\beta$ noise surfaces constructed and rendered as described in *Section 6.3*.

7.2.1 Participants

For both experiments presented in this chapter we used ten participants. All participants had normal or corrected-to-normal vision. All twenty participants we used over the course of these two experiments were undergraduate or postgraduate students at Heriot-Watt University. All participants were naïve as to the purpose of the experiment and none worked with texture, gloss or human vision.

Every observer was asked to read, complete, and sign a consent form similar to the one presented in *Appendix A* before they took part in the experiment.

7.2.2 Apparatus

Both experiments were performed in a small room with no one but the participant present. The room had a single small window which was covered with a blackout curtain so there was no external light in the room and the lighting conditions could be controlled to be identical for each participant. The only light in the room came from a single florescent bulb directly above the participant and from the monitor used to display the stimuli.

The stimuli were presented on a single 24-inch NEC PA241W monitor. This monitor was selected for its colour correctness (provided by a 14-bit per colour 3D LUT), wide viewing angles with minimal colour deviation and luminance uniformity ($\leq 3.5\%$) across the panel [120]. This monitor also has the ability to internally alter the gamma settings which allows for much more accurate colour calibration.

In order to avoid the problems with un-calibrated displays as stated by Roehrig et al. [121] we have calibrated the monitor to linear gamma. Most Windows PC's (such as the one we are using) default to a gamma of 2.2 which means that colours are not displayed accurately on the monitor. As can be seen in *Figure 7.1*, a display with a gamma of 2.2 varies significantly from a monitor properly calibrated to output using a linear gamma profile. *Figure 7.1* also shows the calibration used in these experiments, with the three colour channels linear and overlapping.

The monitor was calibrated with an Eye-One Pro spectrometer to ensure that the gamma and colour was as close to linear gamma as possible. The colour temperature was set to 6500k and the max luminance to 120cd/cm^2. The calibrated gamma profile can be seen

in *Figure 7.1* and shows that all the colour channels exhibit a linear behaviour. *Table 7.1* further details the accuracy of the calibration profile used for this experiment, showing that the average error across the entire colour gamut was < 0.3% and the maximum error was < 0.8%.
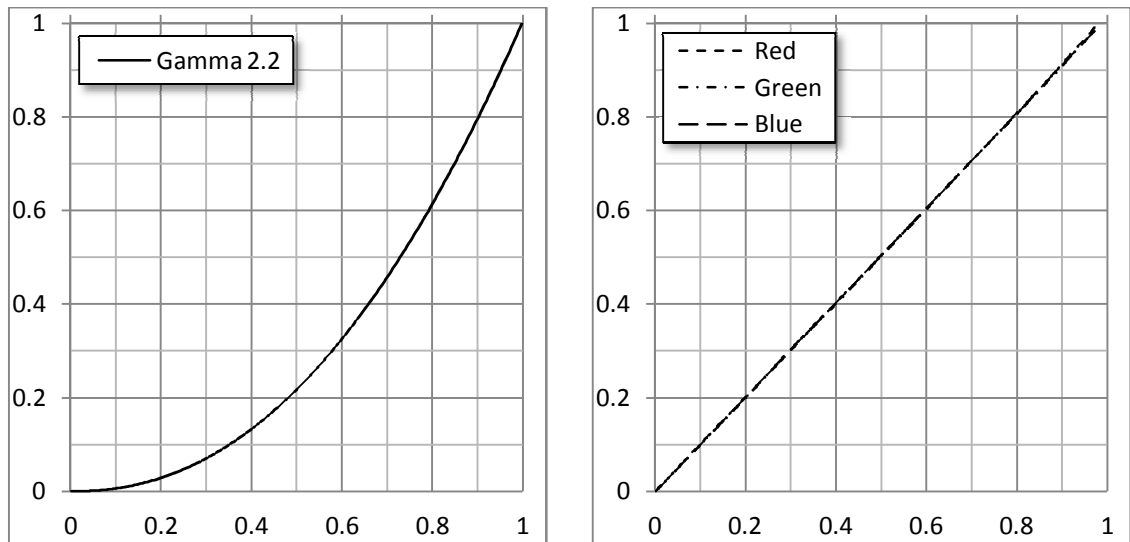
Figure 7.1 - Uncalibrated and calibrated gamma curves for comparison . Left: An ideal 2.2 gamma curve. Right: The gamma profile of our calibrated monitor used in our experiments showing each of the three colour channels

The monitor displaying the stimuli was positioned 85 cm in front of the participant, and centred in relation to the seat. No chin rest was used so the participants were free to move their head.

Colour Channel	Average Error	Maximum Error
Red	0.20%	0.78%
Green	0.29%	0.56%
Blue	0.12%	0.25%

Table 7.1 - A summation of the errors with the linear gamma calibrated monitor

Finally, a standard keyboard was positioned on the desk in front of the participants, allowing them easy access to the cursor keys. This was for participants to interact with the software used to present the stimuli to them, as is described in the next section.

7.2.3   Procedure

Both experiments reported in this chapter were presented as self-contained applications which participants were asked to use. After a short introduction from the author, the application was run and they were left alone to perform the experiment. After being asked to input their initials, a short instruction screen was shown to the participants in white text on an otherwise blank, black screen. This was to ensure each participant received the same instructions as outlined in *Figure 7.3*.
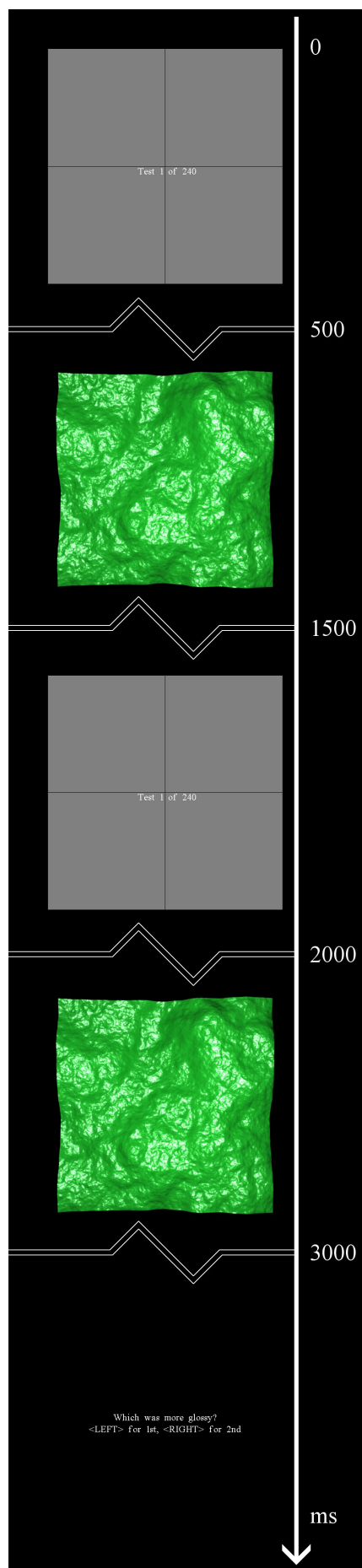
Figure 7.2 - 2AFC Experimental Design

Once a participant pressed the space bar, the experiment began.

For both experiments we used a two-alternative forced choice (or 2AFC) experimental design. As stated in the instructions, participants were shown two rendered surfaces sequentially and were asked to make a judgement of which was glossier.

For every participant, the order in which the trials were presented was randomised to avoid possible learning effects. The number of times each surface appeared first and second was perfectly balanced to avoid ordering effects.

Each trial followed the same pattern. It started with a grey masking image being shown for 500ms, before the first stimuli was shown for 1000ms. Once the presentation time was up, a second masking image was shown for 500ms before the second stimuli was shown for 1000ms. Finally, the screen returned to black and the following question was displayed to the participant:

Which was more glossy?
<LEFT> for 1st, <RIGHT> for 2nd

The application paused at this point until the participant answered the question. There was no way to skip a trial meaning a participant was required to answer even if unsure. Once a participant supplied an answer the next trial begun instantly. This process is summarised in *Figure 7.2*.

This design meant that participants were able to perform a trial every five to ten seconds, allowing

us to repeat the same stimuli multiple times per participant.

> *In this experiment, you will be shown pairs of glossy surfaces.*
> *You will see one for a set time, then the second.*
> *You will then be asked to indicate which was MORE glossy.*
> *To indicate the FIRST SURFACE was glossier, press <LEFT>.*
> *To indicate the SECOND SURFACE was glossier, press <RIGHT>.*
> *If you need to take a break, you can before answering.*
> *When you're ready, PRESS THE SPACE BAR to start.*

Figure 7.3 - Experiment instructions show to each participant

In both experiment designs we used two levels of glossiness, numbered level *06* and *07* in our rendering system. Examples are available in *Table C.2*.

The trials were split into two groups, ones where the participants were shown two surfaces of different statistical gloss levels (i.e. level *06* vs. *07*) and ones where the participant were shown two surfaces rendered with the same specular strength or glossiness. The first group of trials was not used for our final results but was present to both obfuscate the true nature of the experiment and to ensure the participant had some trials where they could answer correctly in order to help combat participant fatigue.

It was this second group of cases in which we were interested. If a participant was unable to see a difference between two different random phase surfaces of the same gloss level then we would expect responses from our participants to be close to chance as they simply guessed which was glossier. It is this 'closeness-to-chance' measure which we used to determine perceptual similarity for gloss.

The only difference between the designs of the two experiments was what stimuli they used and how many trials a participant was asked to perform. These details will therefore be detailed in *Section 7.3* and *7.4* as appropriate.

7.3 Single Roughness Experiment ($\beta = 2.2$)

For the first experiment, we wanted to determine whether changing the random phase of a surface altered the perception of how glossy that surface was. As a similar previous experiment by Padilla into roughness and random phase found no link [6], we expected a comparable result for gloss.

Before discussing the results, however, we first discuss the stimuli and number of trials specific to this experiment.

7.3.1   Stimuli and Number of Trials

In order to design the experiment to have the maximum number of repeated trials per participant, for this experiment we narrowed the focus to a single roughness level, $\beta = 2.2$. This roughness level was chosen as it was our central level of the five which we rendered and was therefore the most representative.

As discussed in *Section 6.3*, for each roughness level we had rendered four different random phase surfaces. These surfaces each had statistically identical *Root Mean Square* (RMS) roughness and had been rendered at 13 different glossiness levels along a linear scale. Examples are available in *Appendix C* (labelled 9 to 11). As previously discussed, we selected the two central gloss levels to use in this experiment.

This meant that for a single roughness, at a single level of glossiness (level '06' in *Appendix C*) we had four different surfaces and therefore 6 different comparisons we could make. To ensure that we showed the stimuli an even number of times first and second, we also displayed these comparisons the other way around, doubling this number to 12. Finally, we also showed these same comparisons to a participant at a second glossiness level (level '07' in *Appendix C*) meaning that a single set of trials for a participant was 24.

In addition to the trials we were interested in, as discussed in *Section 7.2.3*, we added in an equal number of comparisons between the two gloss levels. In the end, we therefore had 48 unique trials which were repeated five times per participant, giving a total number of 240 trials which took participants 12 to 15 minutes. Participants were offered a 100g chocolate bar to reimburse them for their time.

This trial design allowed us to gather 10 repeats of each pair comparison per participant (5 per display order) and as we had ten participants we garnered 100 different responses for each pair.

7.3.2   Results – Binomial Distribution Analysis

As mentioned in the previous section, we now had 100 responses for each of the 12 conditions in which we were interested in (the six different pair comparisons at the two different levels of gloss). If, as we believed, a participant was unable to see a difference between two different random phase surfaces of the same gloss level then we would

expect responses from our participants to be random as they simply guessed which was glossier.

In our case this would show results similar to those of a coin-toss, where participants randomly picked which of the pair appeared most glossy and as such their results would fall close to chance. As we only had two possible outcomes there results would therefore fall close to a perfect 50/50 split in which was picked first with no skew either way.

To examine this we used a cumulative binomial test which is designed to look for the probability that an observed result's deviation from the expected distribution is statically significant. In our case, was a certain pair of surfaces' 'closeness-to-chance' measure likely to be through random noise, or was it significantly different from chance. This is defined as follows:

(7.1)

$$P = \sum_{i=0}^{k} \binom{n}{i} p_{trial}{}^{i}(1 - p_{trial})^{n-i}$$

Where k is the number of successes (in our case, the first of a pair picked over the second), n is the number of trials, p_{trial} is the probability of success in one trial (from 0 to 1), P is the probability of this occurring by chance and:

(7.2)

$$\binom{n}{i} = \frac{n!}{(n-i)!\,k!}$$

Using this, we can compare the results to the standard 0.025 (5%, two-tailed) confidence interval to determine whether our results are likely to be by random chance, or if they are due to a perceptual difference noticed by our participants.

One important note is that as we will be looking at 12 different binomial tests, we need to adjust our 0.025 confidence interval to make sure that we don't get an increased likelihood of type I errors. To ensure our confidence level remains valid, we applied the Bonferroni correction. This states that we should divide our standard confidence interval by the number of trials, which in our case is 0.025 / 12 or 0.00208. This is the value we will be comparing our P values to in order to determine if we should reject our null hypothesis.

Our null and alternative hypotheses are therefore as follows:

$$H_0 : p = 0.5 \qquad \propto \; = \; 0.00208$$

$$H_1 : p \neq 0.5$$

The results of our experiment are documented in *Table 7.2*, below:

Gloss Level 06				
First Surface	Second Surface	First Picked	Binomial	Skewed?
9	10	46	0.242059	No
9	11	33	0.000437	Yes
9	12	47	0.30865	No
10	11	35	0.001759	Yes
10	12	48	0.382177	No
11	12	67	0.999796	Yes

Gloss Level 07				
First Surface	Second Surface	First Picked	Binomial	Skewed?
9	10	50	0.539795	No
9	11	30	< 0.001	Yes
9	12	46	0.242059	No
10	11	29	< 0.001	Yes
10	12	47	0.30865	No
11	12	66	0.999563	Yes

Table 7.2 - Results of the Bonferroni corrected binomial tests at β = 2.2. The rows with grey backgrounds are statistically significantly skewed from chance. The height maps and final rendered images 9 to 12 are presented in *Table C.1*

As can be seen from this table, it is clear that for three of the surface pairs (the same at both levels of gloss) we reject the null hypothesis, meaning that there is a statistically significant skew away from chance caused by random phase in these cases. Therefore, the random phase of a surface does have a statistically significant affect on a participant's perception of how glossy the surface is when compared to another.

Due to this finding, we were now required to test the other random phase surfaces at the other four roughness levels to make sure we didn't use stimuli which would systematically bias results in other experiments. This is discussed in *Section 7.4*. Before that, however, we wanted to confirm this with another statistical test. This is discussed in the next section.

7.3.3   Results – One-Way Repeated-Measures ANOVA

As shown in *Table 7.2*, we found that random phase did have a statistically significant effect on a participant's perception of the gloss. This analysis, however, only showed

that certain pairs were skewed, not that changing which pair was presented to a participant effected how close to chance their responses were. While this can be inferred, we wanted to confirm it with a second statistical test. Therefore, we analysed our results with a one-way repeated measures ANOVA to see if the different pair comparisons had a significant effect on gloss perception directly.

To do this, we took the average rating from each participant across both levels of gloss which they were asked to respond to. We believed this was justifiable as both of the gloss levels gave similar results (with an average error between them of 2.67%) and both gave identical responses for which pair comparisons were statistically skewed, as described in the previous section. This average was then normalised to between 0 and 1, with the chance level being at 0.5. The results of this process are shown in *Table 7.3*, below.

Normalised Values		
First Surface	Second Surface	First Picked
9	10	0.48
9	11	0.315
9	12	0.465
10	11	0.32
10	12	0.475
11	12	0.665

Table 7.3 – Mean normalised values for both gloss levels shown in *Table 7.2*

A one-way repeated-measures ANOVA was then conducted on these values.

Before reporting the results, we calculated the sphericity of our data, which is explained in detail by Field on page 428 [122], the results of which are shown in *Table 7.4*. Mauchly's test indicated that the assumption of sphericity had not been violated for the effect of the phase pairs: $\chi^2(14) = 18.07$, $p = 0.227$. Therefore we assumed sphericity and did not use a correction for the degrees of freedom.

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Pairs	.078	18.072	14	.227	.552	.820	.200

Table 7.4 - Mauchly's test of sphericity showing that we can assume sphericity for our experiment

The within-subjects effects showed that the 'closeness-to-chance' ratings of random phase surface pairs was affected by which phase pairs the participants were asked to compare $F(5, 45) = 9.316$, $p < 0.001$. In other words, not all the phase pairs can be at chance, as at least one pair is significantly different to the others. An abridged version

of the within-subjects effects is presented here in *Table 7.5*, with the full version available in *Table D.1*.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Pairs	Sphericity Assumed	.830	5	.166	9.316	.000	.509	46.580	1.000

a. Computed using alpha = .05

Table 7.5 - Abridged within-subjects effects showing that phase does have a significant effect upon the perception of gloss $p < 0.001$ (Actual: 0.000004)

As we have shown that changing which phase pairs are presented to a participant does have a statistically significant effect on their 'closeness-to-chance' measure, we finally examined the pairwise comparisons to see if we could determine which pairs were significantly different to the others. As in the previous section, we have adjusted for multiple comparisons using the Bonferroni correction, although this time the correction is applied on the significance value rather than the alpha value. These results show that there are two cases where the differences are statistically significant. These are presented in abridged form in *Table 7.6* and complete form in *Table D.2*.

(I) Pairs	(J) Pairs	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference Lower Bound	Upper Bound
9 vs. 11	9 vs. 10	-.165	.056	.251	-.388	.058
	9 vs. 12	-.150	.039	.063	-.306	.006
	10 vs. 11	-.005	.047	1.000	-.192	.182
	10 vs. 12	-.160	.069	.670	-.431	.111
	11 vs. 12	-.350[*]	.056	.002	-.571	-.129
10 vs 11	9 vs. 10	-.160	.078	1.000	-.469	.149
	9 vs. 11	.005	.047	1.000	-.182	.192
	9 vs. 12	-.145	.053	.343	-.354	.064
	10 vs. 12	-.155	.042	.073	-.320	.010
	11 vs. 12	-.345[*]	.048	.001	-.535	-.155

*. The mean difference is significant at the .05 level.
a. Computed using alpha = .05

Table 7.6 - Abridged pairwise comparisons showing two significantly different pairs. The two pairs are indicated with a shaded background and all p-values have been corrected using the Bonferroni correction method.

These pairwise comparisons back up the results from the previous section, as the three pairs of surfaces which are significantly different from each other (9 vs. 11, 10 vs. 11 and 11 vs. 12) are the three which were shown to be significantly skewed from chance in *Table 7.2*.

7.3.4   Discussion

In the previous two sections we have shown through both binomial distribution analysis (*Section 7.3.2*) and one-way repeated-measures ANOVA with pairwise comparisons (*Section 7.3.3*) that our four random phase surfaces were not perceived to be of the same glossiness by our participants.

We were still keen, however, to use different random phase surfaces for comparisons and thus ensure participants had to make judgements based on an entire surface rather than simply comparing two identical features. While at this stage we could select three pairs of surfaces which we could show were perceptually similar for a single roughness, as shown by the white rows in *Table 7.2*, we needed to extend this result to cover all our stimuli.   Therefore we decided to run an expanded experiment with the other four roughness levels (β = 1.8, 2.0, 2.4, and 2.6).

The details of this experiment are discussed in the next section.

7.4   Expanded Experiment (β = 1.8, 2.0, 2.4, 2.6)

For this expanded experiment, we wanted to determine which of our surfaces were the most perceptually similar in regards to gloss. At this stage we had this information for a single level of gloss (β = 2.2), but needed to repeat the experiment in order to get this information for the other four levels (β = 1.8, 2.0, 2.4, and 2.6).

Our aim was to find the two most similar pairs of surfaces at each roughness level for comparisons in other experiments. This would allow us to show multiple comparisons at each roughness level and avoid the problems mentioned in the introduction which can occur when comparing identical features.

One problem with repeating the experiment, however, is that if we do this naïvely, we either require four times the number of people or four times the number of trials per person. As discussion with participants from the previous experiment showed us that the first experiment was starting to become fatiguing, we wanted to make sure the new experiment wasn't too long.

We therefore decided to compromise and reduce the number of repeats for each unique pair in order to keep the experiment short enough to be completed in approximately 30 minutes with a short break in the middle. Details on this are presented in *Section 7.4.1*.

7.4.1 Stimuli and Number of Trials

As discussed in the introduction to this section, for this experiment we wanted to expand the results of *Section 7.3* to cover all of the stimuli we had created, rather than just a single roughness level.

Therefore, for this experiment our stimuli were the four surfaces from each of the roughness levels not previously examined (β = 1.8, 2.0, 2.4, and 2.6). As before, these surfaces each had statistically identical RMS roughness and had been rendered at 13 different glossiness levels along a linear scale. Examples of the surfaces used are available in *Appendix C* (labelled surface 1 to 8 and 13 to 18) and we selected the two central gloss levels, as before.

As already described in *Section 7.3.1*, for each roughness a single set of trials for a participant was 24. As we were using four roughness levels that meant a single set of counter-balanced, unique trials totalled 96. Combined with the trials containing comparisons between the two gloss levels, that meant we would need to present 192 trials to a participant for just two repeats of the stimuli we were interested in.

If we were to perform the same amount of repeats as the previous experiment, the experiment would run approximately an hour long. As participants from the previous experiment were already complaining of fatigue, we decided that making the experiment an hour long would cause the participants' results to deteriorate. To combat this, we decided to reduce the number of repeats of the unique trials to 3, rather than the 5 we used before.

This reduction meant that most participants finished this experiment in between 30 and 40 minutes with a short five minute break to rest their eyes half way through, in order to combat eyestrain and fatigue. Each participant was given a £5 Amazon gift voucher to reimburse them for their time.

This trial design allowed us to gather 6 repeats of each pair comparison per participant (3 per display order) and as we had ten participants we garnered 60 different responses for each pair at each of the four new roughness levels.

As the results between the two gloss levels in the previous experiment were so similar, we decided to use the average result between the two gloss levels for our analysis of these results both to increase the number of repeats per participant and reduce the number of categories. The results from this experiment are discussed in the next two sections.

7.4.2   Results – Binomial Distribution Analysis

First, as before, we wanted to see if any of our surface pairs were significantly skewed from chance. The procedure for doing this is described in detail in *Section 7.3.2* and so is not repeated here.

As we mentioned in *Section 7.4.1*, we took the average results for a participant between the two gloss levels, giving us a total of 24 different categories. Applying the Bonferroni correction as before on the standard 0.025 (5%, two-tailed) confidence interval meant we would be comparing our *P* values to 0.00104.

Our null and alternative hypotheses are therefore as follows:

(7.5) $$H_0 : p = 0.5 \qquad \propto = 0.00104$$

(7.6) $$H_1 : p \neq 0.5$$

The results of our experiment are documented in *Table 7.7*, below:

Averaged Values				Averaged Values			
β	Surfaces	First Picked	Binomial	β	Surfaces	First Picked	Binomial
1.8	1 vs. 2	32	0.740521	2.4	13 vs. 14	37.5	0.974053
	1 vs. 3	23	0.046230		13 vs. 15	29	0.448711
	1 vs. 4	18.5	0.001335		13 vs. 16	32	0.740521
	2 vs. 3	29	0.448711		14 vs. 15	23	0.046230
	2 vs. 4	22.5	0.025947		14 vs. 16	30	0.551289
	3 vs. 4	25.5	0.122530		15 vs. 16	35.5	0.922499
2.0	5 vs. 6	37.5	0.974053	2.6	19 vs. 20	24	0.077501
	5 vs. 7	34	0.877470		19 vs. 21	28.5	0.349442
	5 vs. 8	31	0.650558		19 vs. 22	30	0.551289
	6 vs. 7	28	0.349442		20 vs. 21	27	0.259479
	6 vs. 8	22.5	0.025947		20 vs. 22	30.5	0.551289
	7 vs. 8	25	0.122530		21 vs. 22	34	0.877470

Table 7.7 - Results of the binomial tests at β = 1.8, 2.0, 2.4 and 2.6. No pairs of surfaces are statistically significantly skewed from chance when the Bonferroni correction was applied. The height maps and final rendered images are presented in *Table C.1*

As can be seen from the table above, due to the Bonferroni correction reducing the size of our $\propto$-value we cannot reject our null-hypothesis, meaning we cannot say that any of the pairs from this experiment were statistically skewed from chance.

As we believed this result is due to the lack of power in this statistical technique when dealing with many categories, we performed an ANOVA on the results as before. In this

case, however, as we are changing both the roughness of the pairs and the pair type, we will require two-way repeated-measures ANOVA instead. This is described in the next section.

7.4.3   Results – Two-Way Repeated-Measures ANOVA

As shown in *Table 7.7*, we were unable to show that random phase at these four roughness levels had a statistically significant effect on a participant's perception of gloss using binomial distribution analysis. To corroborate this result we decided to use an analysis method which was more sensitive when using many different categories, the repeated measures ANOVA.

As we now had two different variables which we were changing in our experiment (roughness and pair type) we would not be able to perform exactly the same test as in *Section 7.3.3*. Instead, we used the similar two-way repeated-measures ANOVA which would check to see whether either variable had a significant effect on our 'closeness-to-chance' scores for each pair of surfaces.

We took the average rating from each participant across both levels of gloss which they were asked to respond to and then normalised it to between 0 and 1, with the chance level being at 0.5. The result of this process is shown in *Table 7.8*, below.

Normalised Values			Normalised Values		
β	Surfaces	First Picked	β	Surfaces	First Picked
1.8	1 vs. 2	0.533333	2.4	13 vs. 14	0.625
	1 vs. 3	0.383333		13 vs. 15	0.483333
	1 vs. 4	0.308333		13 vs. 16	0.533333
	2 vs. 3	0.483333		14 vs. 15	0.383333
	2 vs. 4	0.375		14 vs. 16	0.5
	3 vs. 4	0.425		15 vs. 16	0.591667
2.0	5 vs. 6	0.625	2.6	19 vs. 20	0.4
	5 vs. 7	0.566667		19 vs. 21	0.475
	5 vs. 8	0.516667		19 vs. 22	0.5
	6 vs. 7	0.466667		20 vs. 21	0.45
	6 vs. 8	0.375		20 vs. 22	0.508333
	7 vs. 8	0.416667		21 vs. 22	0.566667

Table 7.8 - Normalised values for the results shown in *Table 7.7*

A two-way repeated-measures ANOVA was then conducted on these values.

Before reporting the results, we calculated the sphericity of our data, the results of which are shown in *Table 7.9*. Mauchly's test indicated that the assumption of sphericity

had been violated for the effect of roughness $\chi^2(5) = 14.70$, $p = 0.012$ but had not been violated for the effect of the phase pairs: $\chi^2(14) = 14.134$, $p = 0.467$. Therefore the degrees of freedom for roughness were corrected using the Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.69$), but for phase pairs we assumed sphericity and did not use a correction for the degrees of freedom. We have used the Greenhouse-Geisser correction as $\varepsilon < 0.75$ as recommended by Field and Hole ([123], p187).

| Within Subjects Effect | Mauchly's W | Approx. Chi-Square | df | Sig. | Epsilon[a] | | |
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Roughness	.149	14.699	5	.012	.692	.904	.333
Pairs	.137	14.134	14	.467	.623	.992	.200
Roughness * Pairs	.000	.	119	.	.307	.670	.067

Table 7.9 - Mauchly's test of sphericity. For roughness we cannot assume sphericity and therefore the degrees of freedom were corrected using the Greenhouse-Geisser estimates of spericity. For phase pair we can assume sphericity.

The within-subjects effects showed that the 'closeness-to-chance' ratings of random phase surface pairs was affected by which phase pairs the participants were asked to compare $F(5, 45) = 2.96$, $p = 0.021$, in other words, at least one pair is significantly different to the others. Conversely, the 'closeness-to-chance' rating was not significantly affected by the roughness of the surface $F(2.08, 18.70) = 3.24$, $p = 0.06$. An abridged version of the within-subjects effects is presented here in *Table 7.10*, with the full version available in *Table D.3*.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Roughness	Greenhouse-Geisser	.336	2.077	.162	3.241	.060	.265	6.733	.555
Pairs	Sphericity Assumed	.311	5	.062	2.964	.021	.248	14.821	.810

a. Computed using alpha = .05

Table 7.10 - Abridged within-subjects effects showing that phase does have a significant effect upon the perception of gloss $p = 0.021$, but roughness does not 0.060.

These results confirm what we discovered in *Section 7.3* that different phase pairs are indeed significantly different from each other, showing they are not seen to be the same glossiness by participants.

Conscious of this knowledge, in the next section we will select two pairs of surfaces at each roughness which are the same perceptual glossiness for use in other experiments.

At this stage, as our two-way repeated-measures ANOVA had shown there was an effect, we consulted the pairwise comparisons as before to see whether we could determine any pairs of surfaces whose 'closeness-to-chance' measure differed significantly. The results for this are available in *Table D.4*. As with our binomial distribution analysis, however, there were no significant results.

7.5   Pair Selection

The final step of this chapter was to combine the results from the experiment in *Section 7.4* with those from the previous one, described in *Section 7.3*.

As we want there to be minimal effect to the perception of gloss from the difference in phase, we have decided to select the two pairs of surfaces at each roughness level which are closest to chance. We have already normalised the results from both experiments to between 0 and 1, so we can simply look at the values from both experiments and pick those which are closest to 0.5 (the normalised chance value).

The table below (*Table 7.11*) shows the two surfaces which are closest to chance for each roughness level, along with how offset from chance each is. The maximum error from chance is 3.33% with the average error across all surface pairs being 1.87%.

Roughness	First Image Number	Second Image Number	Offset from Chance	
$\beta = 1.8$	1	2	0.03333	3.33%
	2	3	-0.01666	-1.67%
$\beta = 2.0$	5	8	0.01666	1.67%
	6	7	-0.03333	-3.33%
$\beta = 2.2$	9	10	-0.02	-2.00%
	10	12	-0.025	-2.50%
$\beta = 2.4$	13	16	0.03333	3.33%
	14	16	0	0.00%
$\beta = 2.6$	18	20	0.00833	0.83%
	17	20	0	0.00%

Table 7.11 - A table of which pairs of random phase surfaces are the most similar. The height maps and final rendered images 1 to 20 are presented in *Table C.1* with the similar pairs show in Appendix E. N.B. A positive offset from chance shows the skewed towards the second image, a negative shows it was towards to the first.
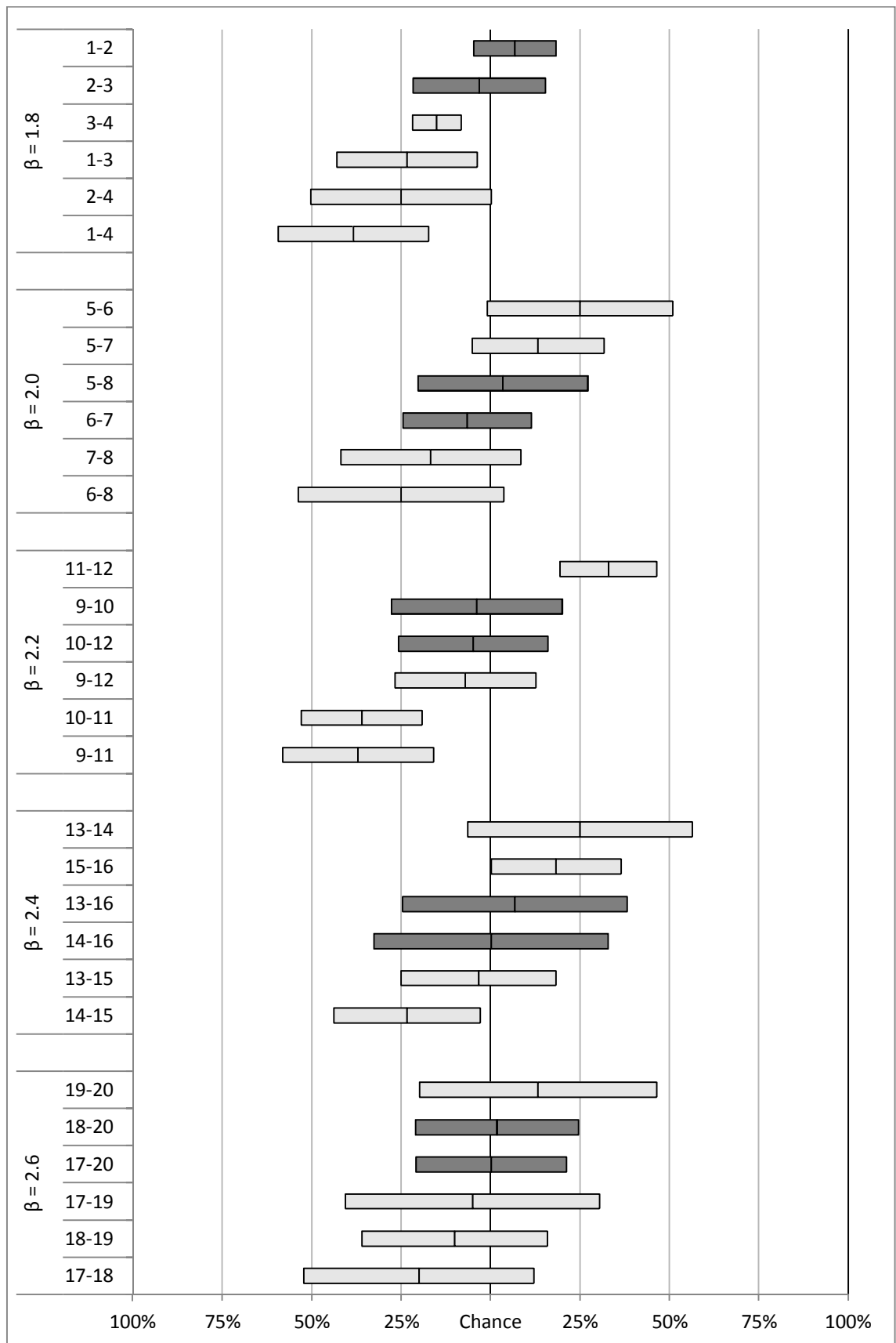
Figure 7.4 - A graph showing how similar each pair of random phase surfaces is. The closer to chance each pair is, the more similar the two surfaces were perceived as. The bars are uncorrected 95% confidence intervals to indicate which pairs were statistically skewed, with the centre line being the mean value. The height maps and final rendered images 1 to 20 are presented in *Table C.1* with the similar pairs show in *Appendix E*. The darker bars represent the pairs chosen.

Figure 7.4 shows the results for every pair of surfaces from both experiments, with their mean values represented by a vertical line and uncorrected 95% confidence intervals represented by a bar. The darker bars represent the pairs shown in *Table 7.11* which are those closest to 0.5 or chance.

7.6   Conclusions

In this chapter we have described in detail the process we used to confirm that random phase does, indeed, have an effect on perceived glossiness. By asking people to choose which of a pair of surfaces is glossier we were able to compile a 'closeness-to-chance' value for each. As these surfaces are statistically the same glossiness, we would expect these values to be close to chance, but we found that in some cases they were statistically significantly skewed from chance. This comprises the first main output from this chapter.

Our first experiment presented in this chapter, in *Section 7.3*, examined a single roughness level ($\beta = 2.2$) and found via binomial distribution analysis that three of the pairs were statistically skewed from chance. A one-way repeated-measures ANOVA then confirmed that changing the pairs did indeed have a significant effect on their 'closeness-to-chance' ratings.

We then extended this result to the other four roughness levels at which we have rendered surfaces ($\beta = 1.8, 2.0, 2.4$ and 2.6) in *Section 7.4*. We found that while we could not identify any specific pairs which were skewed from chance (possibly due to the high amount of confidence interval correction required from the large number of categories), a two-way repeated-measures ANOVA did show that changing phase again had a significant effect on 'closeness-to-chance' ratings.

We then used these normalised 'closeness-to-chance' ratings to select two pairs of rendered surfaces from the five different roughness levels which we can use for comparison experiments without statistical bias in our results. These pairs are presented in *Table 7.11* (and visually in *Appendix E)* and comprise the second output of this chapter.

These pairs of surfaces will now be taken forward to the next chapter where they will be used for participants to make comparative judgements between them in a method of adjustment experiment to see whether highlight disparity strengthens the perceived gloss of a surface.

Chapter 8

Does Disparity Strengthen Gloss on Rough Surfaces?

8.1   Introduction

In the previous chapter we found that two surfaces of the same root mean square (RMS) roughness and glossiness could be perceived by participants as being perceptually different in terms of gloss. In this chapter we will answer the question: Does highlight disparity strengthen perceived gloss and is this affected by roughness?

As described in *Chapter 6* we have a series of surfaces each generated from different random phase height maps. These are high fidelity, naturalistic and have been rendered with a physically accurate light source and reflectance function. In *Chapter 7* we went on to discover that even though these surfaces are statistically identical in both roughness and glossiness, they are perceived differently. To counter this we found pairs of these surfaces which were perceptually similar in terms of gloss.

These pairs of surfaces were used for a psychophysical experiment to see whether highlight disparity strengthens a participant's perception of how glossy a surface is and whether the roughness of the surface alters the result. In this chapter we discuss this experiment and the background behind it.

In *Section 8.2* to *8.6* we discuss the setup of the experiment which includes the utilisation of the stereoscope described in detail in *Section 6.2* which we used to present stimuli in both monocular and binocular situations. Specifically, we describe the new experimental design required by the complications of comparing stimuli which appear three-dimensional to those which appear as flat images in *Section 8.5*.

Following this we outline the results from the experiment which we analysed with both a two-way repeated-measures ANOVA (*Section 8.7.1*) and by comparing marginal means (*Section 8.7.2*). These results form the main output of this chapter and one of the main outputs of the thesis. Before that, however, we present a description of the experimental setup.

8.2 Participants

For the experiment presented in this chapter we invited 33 participants to take part. All of the participants we used were either undergraduate or postgraduate students at Heriot-Watt University or worked on campus. All participants were naïve as to the purpose of the experiment and none worked with texture, gloss or human vision. All participants had normal or corrected-to-normal vision and were tested to make sure that they possessed functioning binocular vision.

Of the initial 33 participants, we had to reject four participants due to the lack of binocular vision caused by vision abnormalities. In addition, a single participant was rejected after all participants had performed the experiment because they completed it in an abnormally short time when compared to the rest of the participants (approximately 15 minutes).

The remaining 28 participants finished the experiment in 45 minutes to 1 hour and all that finished were given a £10 Amazon voucher to reimburse them for their time.

Every observer was asked to read, complete, and sign a consent form similar to the one presented in *Appendix A* before they took part in the experiment.

8.3 Apparatus

The experiment was performed in a small room with no one but the participant present. The room had a single small window which was covered with a blackout curtain so there was no external light in the room and the lighting conditions could be controlled to be identical for each participant. The only light in the room came from a single florescent bulb directly above the participant and from the monitor used to display the stimuli.

The stimuli were presented side-by-side on a single 24-inch NEC PA241W monitor. More details on why this monitor was chosen are available in *Section 7.2.2*. As described in that section, the monitor was again calibrated to a linear gamma output as shown in *Figure 7.1*.

Participants viewing stimuli in this experiment did so through a custom made mirror stereoscope which consisted of four front-surface aluminium mirrors mounted on a kinematic mirror mount which were positioned on an optical rail. These mirrors were arranged in pairs to redirect the view of each of a participant's eyes to different points

on a single monitor. The left eye was directed to the left half of the screen, while the right eye was directed to the right half. This concept is described in more detail in *Section 6.2*. This mirror setup is pictured in *Figure 8.1*.

To display the stimuli with this system, the monitor was treated as if were in fact two monitors by splitting it in two vertically. Therefore, instead of showing stimuli centred on the monitor as we did in *Section 7.3* and *7.4*, two images were always shown, centred at the quarter and three-quarter mark on the screen. These correspond to the centre points where a participant's eyes were redirected.
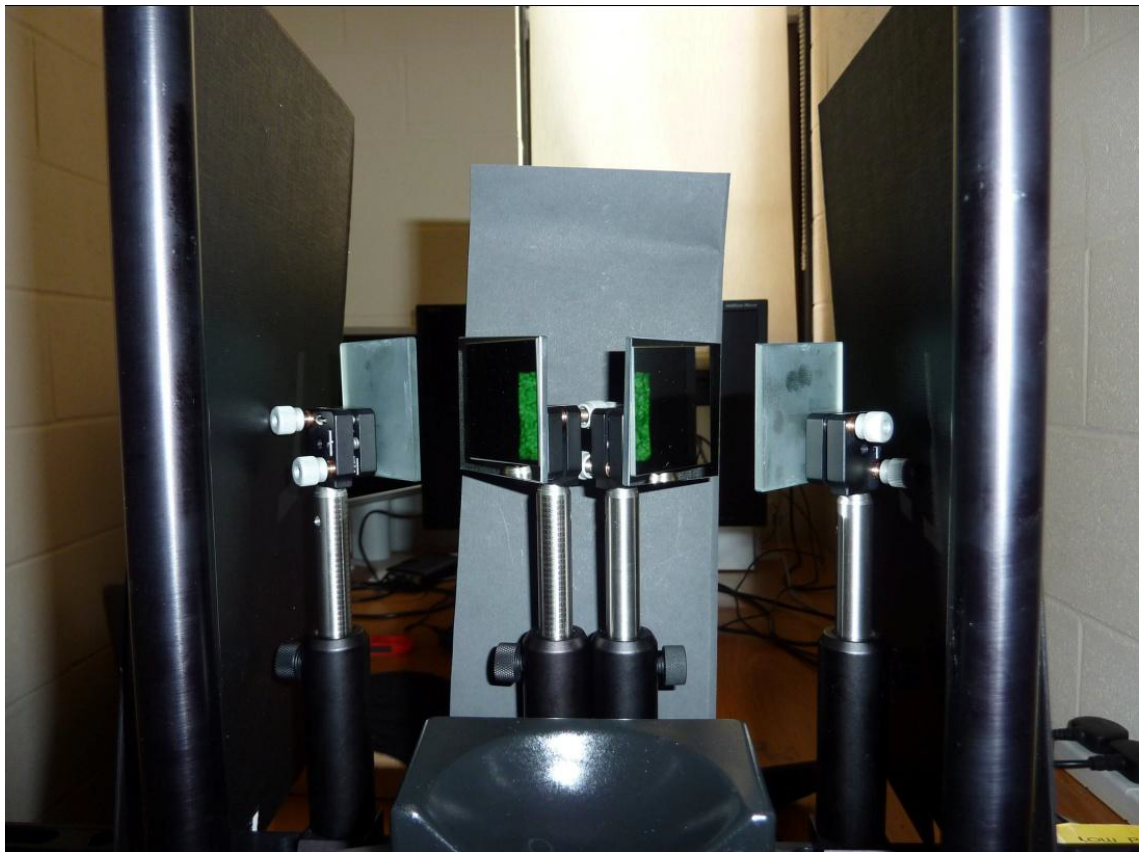
Figure 8.1 - Stereoscope setup used in the highlight disparity experiment. N.B. This is situated in our experimental space, but the window is not covered in this picture to improve the lighting conditions for the photograph.

This setup meant that if the two halves of the screen were exactly mirrored (i.e. the same surface shown to the left and right eye) then a participant would see a flat picture as if they were viewing it on a standard monitor. If instead, two slightly different images were show rendered from two correct viewpoints, the picture would take on depth due to the disparity in the images thanks to binocular vision.

The combination of the optical rail and adjustable mirror mounts allowed us to calibrate to different interocular distances as required. This calibration was done for every

participant in the experiment and used a custom program where the mirrors were adjusted to line up two targets, one visible in each eye. Participants were kept centred throughout the experiment using a chin rest to make sure their viewing position remained constant.

The monitor was positioned 80cm away from the viewer. This meant that the viewing distance for each eye was approximately 85cm when the diversion through the mirrors was taken into account. This distance to the monitor depended slightly on the participants' interocular distance.

During testing of the stereoscope it was noticed that the white walls of the room being used made it difficult to fuse stereoscopic pairs as the eye was drawn to the brighter walls. In addition, it was possible to see the monitor between the mirrors. To combat these problems, three pieces of black card were positioned to block the view of both the walls and the monitor.

Finally, a standard keyboard was positioned on the desk in front of the participants, allowing them easy access to the cursor and space keys. This was for participants to interact with the software used to present the stimuli to them, as is described in the next section.

8.4   Stimuli

As discussed in *Section 6.3*, for each of our five roughness levels we had rendered four different random phase surfaces. These surfaces each had statistically identical RMS roughness and had been rendered at 13 different glossiness levels along a linear scale. An example set is available in *Table C.2* (gloss level 1 to 13).

As we discovered in *Chapter 7* not all of the surfaces we created were perceived to be the same glossiness as each other. We did, however, find two pairs of surfaces for each roughness which statistically similar. Therefore in the experiment reported in this chapter, we used these pairs of surfaces. In each trial (which will be discussed in more detail in the next section) we presented two different surfaces to the participant and these two surfaces were always selected from the similar pairs shown in *Table 7.11*.

There were two ways in which the stimuli used in this experiment were different to those shown for the experiments in the previous chapter. Firstly, as we were introducing the concept of showing the stimuli as three-dimensional surfaces via binocular viewing

through a stereoscope, each stimulus could be seen either as a flat image or with realistic depth. This meant that each surface was rendered from three perspectives: one from the left eye position, one from the right and one from a midpoint which was shown to both eyes to give a flat image. Secondly, unlike the experiments discussed in the previous chapter we were allowing participants to adjust the glossiness of one surface. This means that we were using the entire scale of 13 different gloss levels we originally rendered, rather than just the two used previously.

Therefore, instead of referring to a single image as a stimuli, in this chapter, a 'stimulus' is a single surface which can be at any of 13 gloss levels and seen with or without disparity. When considered like this, we therefore have twenty adjustable stimuli overall, or four at each roughness level. These two pairs of surfaces correspond to those pairs which were selected in *Table 7.11*.

8.5   Design

At this stage, it is important to clarify the experimental design before continuing further. In this experiment, we are using a Method of Adjustment design, where we are asking a participant to adjust stimuli to match target stimuli. This is complicated by the results from our previous chapter, however, as well as the nature of our quantised stimuli and the comparison between monocular and binocular conditions.

As we showed in *Chapter 7*, only certain pairs of our rendered surfaces appear to be the same apparent gloss. To this end, we decided only to present comparisons between two surfaces we've shown participants see as similar. To explain the design, we will start with a single comparison (that between the surfaces labelled 1 and 2 in *Table C.1*) at a single target gloss level, before building this up into the other required conditions.

In this experiment, we have two binary conditions we want to balance to avoid ordering effects. First, we can change which of the pair of surfaces is the target and which is the adjustable surface. Second, as we always want the participant to compare between a monocular and binocular condition, we can change whether the target or adjustable surface is presented binocularly. This means for a single condition in this experiment, we have four possible ways of presenting it, as shown in *Figure 8.2*.

These four, balanced, conditions are then extrapolated outwards. First, to stop the participants simply learning the correct answer in the experiment there were three different target gloss levels. In our number scheme these were at levels '4', '7' and '10'.

Examples of the different gloss levels are available in *Table C.2*. These levels were chosen as they represented the entirety of the gloss levels, but left enough levels above and below them for participants to have freedom in their answers. This gave us 12 possible conditions.

Monocular		Binocular	Binocular		Monocular
Target Surface **1**	←Compared→	Adjustable Surface **2**	Target Surface **1**	←Compared→	Adjustable Surface **2**
		Condition 1	Condition 2		
		Condition 3	Condition 4		
Target Surface **2**	←Compared→	Adjustable Surface **1**	Target Surface **2**	←Compared→	Adjustable Surface **1**
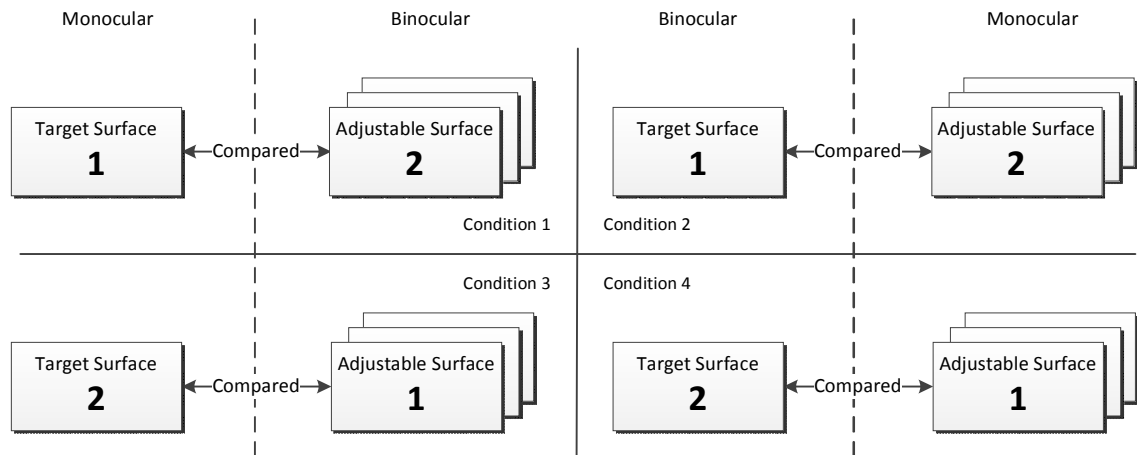
Figure 8.2 – Simplified Experimental Design for a single pair of surfaces and a single target gloss

In addition, we had a second pair of surfaces for participants to compare (labelled 2 and 3 in *Table C.1*), giving us 24 possible conditions at each roughness level. Finally, we repeated these 24 conditions at five different gloss levels, giving us 120 conditions for our participants in each experiment.

For every participant, the order in which these 120 conditions were presented was randomised to avoid possible learning effects.

8.6   Procedure

The experiment reported in this chapter was presented as a self-contained application which participants were asked to use. They were told to only view the monitor via the stereoscope described in *Section 8.3* and never to look at the monitor without it. After a short introduction from the author, the application was run and they were left alone to perform the experiment. After being asked to input their initials, an instruction screen was shown to the participants in white text on an otherwise blank, black screen. This was to ensure each participant received the same instructions as in *Figure 8.3*.

Once a participant pressed the space bar, the experiment began.

> *In this experiment, you will be shown pairs of surfaces in 3D.*
>
> *You will see one 'target' surface, then a second 'adjustable' surface.*
>
> *You will then be asked to adjust the 'adjustable' surface so*
>
> *that it appears to be the same glossiness as the 'target' surface.*
>
> *To view the 'target' surface, press <LEFT>, for 'adjustable' press <RIGHT>.*
>
> *To alter the glossiness level of the 'adjustable' surface use <UP> and <DOWN>.*
>
> *When you are happy with your answer, press <SPACE> to move on.*
>
> *If you need to take a break, you can before answering.*
>
> *When you're ready, PRESS THE SPACE BAR to start.*

Figure 8.3 - Experiment instructions show to each participant

For this experiment we used a method of adjustment experimental design. As stated in the instructions, participants were shown two different surfaces and asked to adjust one of these surfaces to appear the same glossiness as the other. The only change the participant could affect on the 'adjustable' surface was to increase or decrease the strength of the gloss, as described in *Section 8.4*. Participants were allowed to change between viewing the 'target' and 'adjustable' surface as often as they liked until happy with their answer.

Whenever a participant changed between the 'target' and 'adjustable' surface a grey masking image (as shown in *Figure 8.4*) was displayed for 500ms. When a participant submitted their answer, a masking image was shown for 2000ms before a new trial began.

The two surfaces shown as the 'target' and 'adjustable' surface were always different; the pairings used being the ones which were shown to be statistically similar for gloss in *Chapter 7*. As discussed in that chapter, this was to force participants to make judgements based on the nature of the entire surface, rather than single pixels or features. In addition to this, one of the surfaces was always displayed without surface disparity and one with. In effect, a participant would see one of the surfaces with depth and one without in each trial.

This method of comparing a flat image to one with depth is the main crux of this experiment. We were interested in whether the presence of disparity in an image makes the gloss appear stronger as shown in previous experiments [9, 11, 40]. As such, this experiment was designed to record how 'accurate' a participant was between the two

conditions. With this accuracy information we could see if the perceived gloss was significantly different from what the statistical measures would suggest.
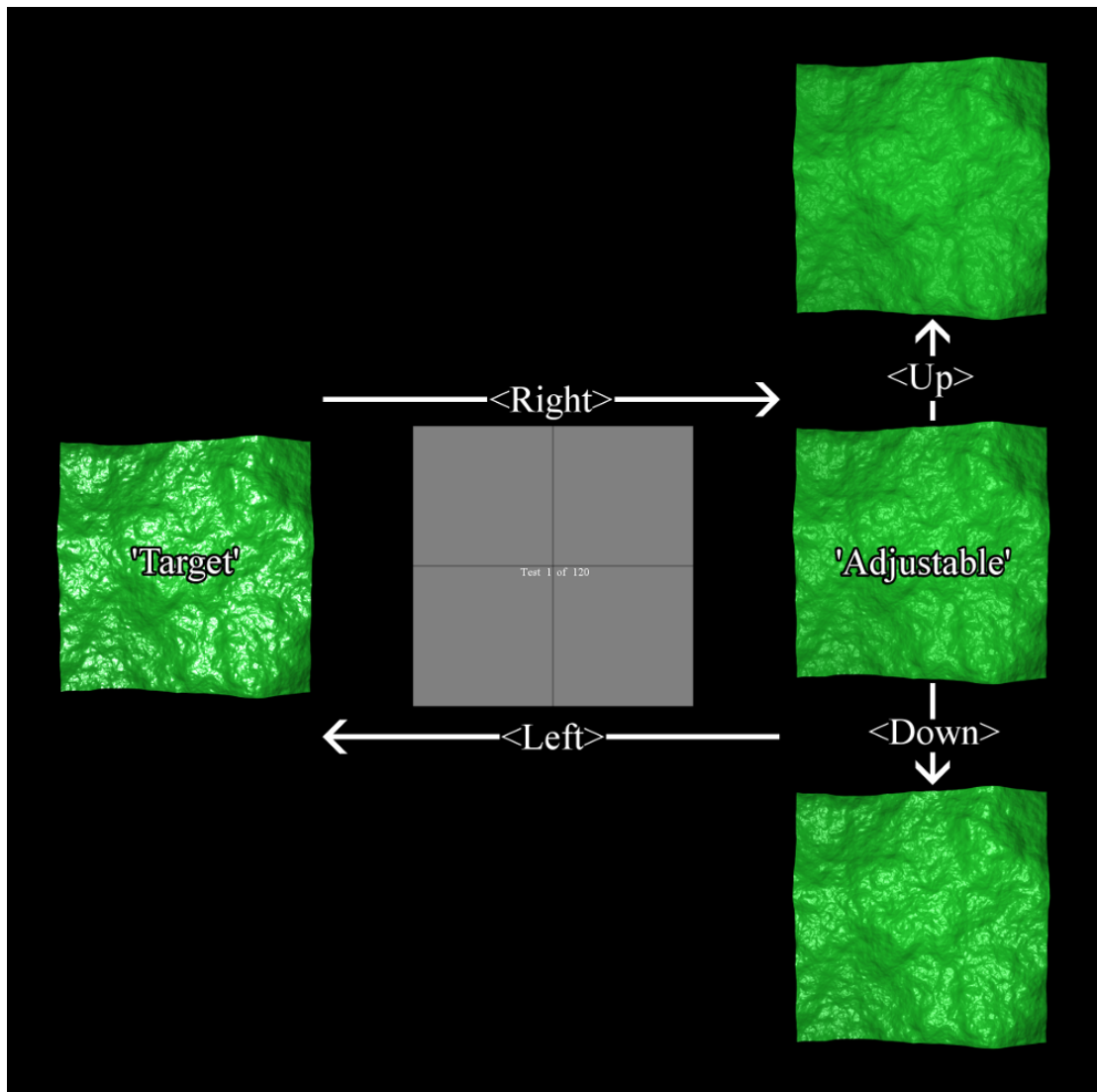
Figure 8.4 - Experiment Design for a single highlight disparity trial. A participant would press the space bar to move onto the next trial when happy with their answer. N.B. The directions refer to cursor keys on a standard desktop keyboard. The arrows and words were not visible to participants.

As described in more detail in *Section 8.5*, the presentation of the stimuli was perfectly balanced. This meant that for this experiment there were 120 different, unique trials for each participant, giving us four repeats for each pair of surfaces. Due to the increased length of time this experiment took over a 2AFC style, participants were not asked to repeat any of the trials. For every participant, the order in which the trials were presented was randomised to avoid possible learning effects.

8.7 Results

As discussed in *Section 8.5*, each participant gave four responses for each pair of surfaces at each roughness with three different target gloss levels. As we had two different pairs at each roughness, this meant that for each data point we had 8 responses per participant. For each participant, the arithmetic mean of these 8 responses was taken to give a single value in each of the 15 different conditions (5 roughness levels, 3 target gloss levels).

These raw results are available in *Table F.1* and will be analysed in the rest of this section.

In addition, the mean accuracy ratings of each participant selected (a total of 28 participants) are presented graphically. These results are presented twice to show the results across the two major conditions (Target Gloss and Roughness) and are presented in *Figure F.1* and *Figure F.2*, respectively.

8.7.1 Two-Way Repeated-Measures ANOVA

One of the questions we wanted to answer was whether changing the roughness of the surfaces used in this experiment had a significant effect on how accurately participants were able to estimate gloss values between a static image and the more realistically presented images with disparity.

To investigate this we used a two-way repeated-measures ANOVA, similar to the one reported in *Section 7.4.3*. In this case, our two conditions were roughness (5 levels) and target gloss (3 levels).

Before reporting the results, we calculated the sphericity of our data, the results of which are shown in *Table 8.1*. Mauchly's test indicated that the assumption of sphericity had not been violated for the effect of target gloss $\chi^2(2) = 4.76$, $p = 0.093$ or for the effect of roughness $\chi^2(9) = 12.65$, $p = 0.180$. Therefore we assumed sphericity and did not use a correction for the degrees of freedom. For the combined effect of both, however, Mauchly's test indicated that the assumption of sphericity had been violated $\chi^2(35) = 74.151$, $p < 0.001$. Therefore the degrees of freedom for the combined effect were corrected using the Greenhouse-Geisser estimates of sphericity ($\varepsilon = 0.66$). We have used the Greenhouse-Geisser correction as $\varepsilon < 0.75$ as recommended by Field and Hole ([123], p187).

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
Target Gloss	.833	4.759	2	.093	.857	.909	.500
Roughness	.608	12.646	9	.180	.805	.927	.250
Target Gloss * Roughness	.046	74.151	35	.000	.658	.835	.125

Table 8.1 - Mauchly's test of sphericity. For both target gloss and roughness we assumed sphericity, as neither result was significant. For the combination of both we cannot assume sphericity and therefore the degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity.

The within-subjects effects showed that the accuracy ratings of participants when comparing a static image and an image with disparity was affected by which roughness the participants were asked to compare $F(4, 108) = 3.25$, $p = 0.015$, in other words, at least one pair was significantly different to the others. Conversely, neither the change in target gloss $F(2, 54) = 2.07$, $p = 0.136$, nor the combined effect of both variables $F(5.26, 142.04) = 0.317$, $p = 0.910$ had any significant effect on how accurate participants were. This implies there was no interaction between the two variables in our experiment and as such they can be treated independently.

An abridged version of the within-subjects effects is presented here in *Table 8.2* with the full version available in *Table F.2*.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Target Gloss	Sphericity Assumed	1.520	2	.760	2.071	.136	.071	4.141	.408
Roughness	Sphericity Assumed	4.958	4	1.240	3.254	.015	.108	13.017	.819
Target Gloss * Roughness	Greenhouse-Geisser	.882	5.261	.168	.317	.910	.012	1.666	.130

a. Computed using alpha = .05

Table 8.2 - Abridged within-subjects effects showing that roughness does have a significant effect upon the accuracy of gloss perception $p = 0.015$, but target gloss and the combined effect of both does not $p = 0.136$, $p = 0.910$.

As we found a significant difference in one of our factors, we then proceeded to look at the pairwise comparisons to see whether we could detect any specific difference between different roughness levels. The p-values have been corrected for multiple

comparisons using the Bonferroni correction as described in *Section 7.3.2*. This abridged table is shown in *Table 8.3* below with a full version available in *Table F.3*.

(I) Roughness	(J) Roughness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1.8	2.0	-.229*	.072	.035	-.448	-.010
	2.2	-.150	.078	.630	-.387	.087
	2.4	-.295*	.093	.038	-.579	-.010
	2.6	-.284	.117	.221	-.642	.074

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Table 8.3 - Abridged pairwise comparisons showing two significantly different pairs. The two pairs are indicated with a shaded background and all p-values have been corrected using the Bonferroni correction method.

From this we can see that one roughness level (β = 1.8) was significantly different to two others (β = 2.0, 2.4). The raw mean values for each of the 15 data points from the 28 participants have been graphed in *Figure 8.5*, below. This, when considered with the ANOVA allows us to see that it is likely that whether highlight disparity strengthens gloss perception is reliant on what roughness level we are considering.

Figure 8.5 - A graph of the gloss similarity of surfaces with and without disparity. This graph shows average participant accuracy, so the closer to zero each point is, the more similar the two surfaces were perceived as. This graph shows the three different target gloss conditions vs. the five roughness levels. N.B. 'Target Gloss' refers to what gloss level the reference surface was which the participant was adjusting towards. This was to stop participants learning the correct answer.

8.7.2   Marginal Means

Although we now knew that roughness had an effect upon gloss strength between images with and without disparity, we weren't able to show whether there was a significant strengthening of perceived gloss from binocular viewing. To investigate this point, we used marginal means.

From to the ANOVA results in the previous section (*Table 8.2*) we saw there was no significant effect on our participants' results from which target gloss they were asked to consider. In addition, there was no combined effect between the two variables present. Owing to these two facts we could treat the three different target gloss levels as similar and calculate their marginal means. Simply put, we reduced our number of categories from 15 to 5 by considering the roughness level as our only axis.

At this stage, we wanted to see whether gloss was significantly strengthened by disparity for any or all of our roughness levels. To do this, we analysed the marginal means data directly. As the design of this experiment dictated, the closer our results were to zero, the more similar participants saw the strength of gloss between surfaces with disparity and without. Negative marginal means would show that disparity strengthened gloss; positive would mean disparity weakened apparent gloss.

The main consideration was whether or not the marginal mean values were significant or just due to noise in our experiment. To solve this, we used a technique similar to the binominal distribution analysis performed in *Chapter 7*. As described in *Section 7.3.2*, binominal distribution analysis simply looks to see whether results fall close enough to the expected result as to be due to noise, or whether the skew is down to some external influence. While we couldn't use this technique directly as there is no single 'success' and 'failure' state as required by that method, we can use the same principle here with our marginal mean accuracy scores.

To do this, we looked at our marginal mean values and compared them to our zero point. This was done by considering a 95% confidence interval that we have been using throughout this thesis. As we now had five marginal means, we corrected our confidence interval using the Bonferroni correction to ensure we didn't introduce any more Type I errors, as follows:

$$\propto = 1 - (\propto_{uncorrected} / n) \qquad \propto_{uncorrected} = 0.05$$

$$n = 5$$

This gave us a confidence interval of 99% which we used for our comparisons. It is important to note that although our error bars will have a 99% confidence interval, the overall confidence is still 95%.

Our marginal means with corrected confidence intervals have been plotted in *Figure 8.6*, below.
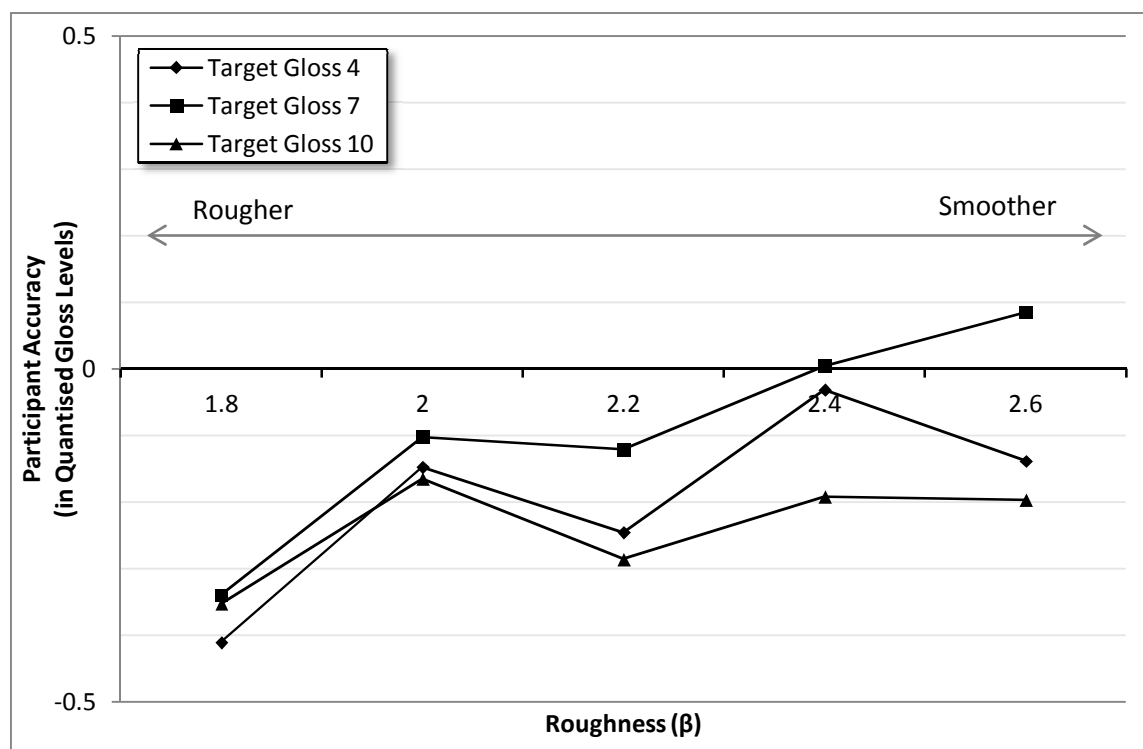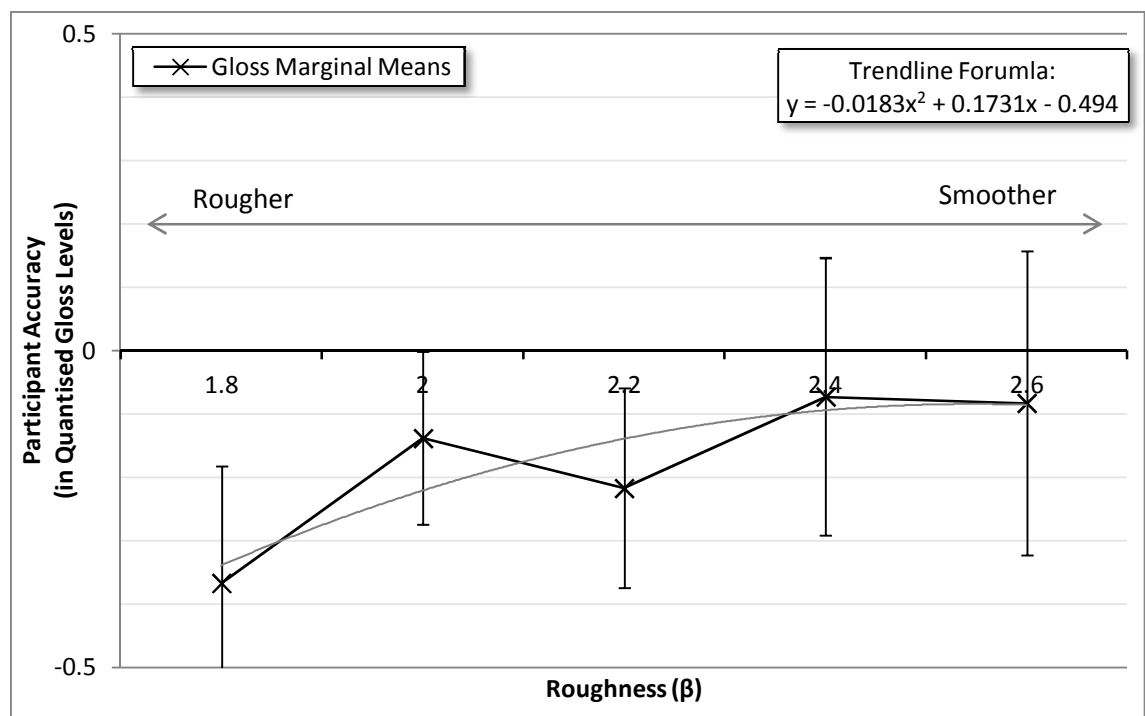

Figure 8.6 - A graph of the gloss similarity of surfaces with and without disparity. This graph shows average participant accuracy, so the closer to zero each point is, the more similar the two surfaces were perceived as. The error bars are 95% confidence intervals which have been corrected with the Bonferroni correction (0.05 / 5 = 0.01 or 99%) so as to ensure no additional Type I errors. This shows that gloss was significantly strengthened at roughness levels β = 1.8, 2.0 and 2.2.

From this graph we can conclude that for three of our roughness levels (β = 1.8, 2.0 and 2.2) the presence of disparity did strengthen participants' perception of gloss. For the two smoother roughness levels (β = 2.4 and 2.6), however, we cannot draw the same conclusion as the negative values might be due to noise.

Finally, with so few points it is difficult to determine what fitting function would best represent this data as there is the temptation to over fit the data by using high-order polynomial functions. For this reason, to offer an indication of a possible trend we have fitted a second order polynomial function (r^2 value of 0.7601) to the data as it was found

by Padilla to be the most accurate function when dealing with perceived roughness using surfaces similar to ours [7].

8.8   Conclusions

In this chapter, we aimed answer the question: does disparity strengthen gloss on rough surfaces? We have shown that there is no simple yes or no answer to this question; rather it is conditional on what roughness of surface the participant is looking at.

The experiment presented in this chapter asked participants to adjust a surface so that the two shown (one with disparity and one without) were the same apparent glossiness. This design is described in detail in *Section 8.5*. With 28 participants' results, we were able to calculate mean accuracy ratings for each of our 15 conditions. These 15 conditions were due to us having two variables in our experiment. First, we had five different roughness levels which the participants performed the experiment on. Second, we changed how glossy the surface which the participants were asked to match was. There were 3 different 'target gloss' levels.

In *Section 8.7.1* we then described how we used a two-way repeated-measures ANOVA to discover that while roughness did have a significant effect on how accurate participants were, the different target gloss levels did not. This showed us that our initial question would be dependent on the roughness of the surface being viewed.

Using the knowledge that different target gloss levels didn't alter a participant's accuracy, we analysed the marginal means for each roughness directly. This is presented in *Section 8.7.2*. We found that by using the marginal means and corrected 95% confidence intervals we could show that the marginal means for three of our roughness levels (β = 1.8, 2.0 and 2.2) were significantly below zero. Due to the design of our experiment, this meant that for those three roughness levels participants were seeing the surfaces with disparity as having significantly stronger gloss. This answers the question presented in this chapter and is the main output from this chapter and thesis.

As this was the last major experiment which will be reported in this thesis, the next and final chapter will summarise the major contribution and output of this thesis and present future work which is informed by the work presented in this document.

Chapter 9

Summary and Conclusions

In this final chapter, we will discuss and summarise this thesis as a whole and present our conclusions. First, in *Section 9.1* we will summarise the argument presented, showing the mostly linear progression of the argument. In *Section 9.3* we present possible future work, based on questions presented by the results reported in this thesis. Finally, in *Section 9.4* we present the main contributions of our work.

9.1   Summary

Our initial aim of this thesis was to investigate the interaction of multiple important properties which were easily understood by naïve participants. We were unaware, however, what properties were important to these non-expert observers. To this end, in *Chapter 2* we investigated what work had already been performed and whether there was a common list or taxonomy of perceptual properties which were widely used.

We began this investigation by discussing early work into texture features which formed the foundation of much of the following research. We discussed how this led to experiments investigating high-level features using free grouping techniques. In particular, we discussed two important studies which offered lists of possible terms and interesting techniques for gathering participant feedback. Rao and Lohse presented a technique of free grouping and hierarchical clustering analysis for discovering how participants related different properties. From Picard et al. we discovered a labelling technique useful for creating taxonomies. We concluded that the closest to a non-expert taxonomy available was offered by Atkinson, who incorporated the results of 15 different fabric studies together into a single list of terms. We decided, however, that this list was unsuitable for our purpose as some of the terms had been sourced from experts or from other languages.

As we concluded there was no suitable taxonomy available, *Chapter 3* was dedicated to creating a simple taxonomy of property terms which could be used to narrow the focus

of our search. In this chapter, we conducted a three-step process (inspired by the work of Rao and Lohse and Picard et al.) to create a two-level taxonomy which was understandable by English speaking non-experts, representative of the whole property space, concise, and well distributed.

We began by examining the list of property terms provided by Atkinson to ensure they were understandable. Each word was sorted into one of three groups by 30 participants, based on how well used and understood each was. The terms were then ordered by their respective scores, allowing us to remove rarely used words from the original list. Then, using data provided by Orzechowski, we updated the list with common terms which were missing from the original set to produce a non-expert property list. Finally, 20 participants were asked to group these 79 words in a free-grouping experiment and pick a representative word for each group. These results were used to cluster the properties into 11 groups, each with a representative word, providing a two-level taxonomy.

In *Chapter 4*, as we had decided to present our stimuli via digital methods, we investigated which tools would be suitable for representing the different property groups. We identified three main categories which represented these property groups: tactile, physically simulated, and visual properties.

To represent tactile properties, we investigated haptic pens, but found that the devices available suffered from a lack of motor strength, making them unsuitable for our experiments. Furthermore, we decided that due to a lack of experience, inability to measure real fabric properties, and a desire to avoid real-time rendering we would not investigate physically-simulated properties.

This led us to investigate how best to represent visual stimuli in a more realistic way than using a single, static image. There were two areas we investigated: head tracking and stereoscopic display methods. While ideally we would have used both, it was infeasible in our case. The only stereoscopic display method we investigated which did not suffer from crosstalk was the mirror stereoscope. As this requires participants' heads to be kept still we could not use both systems together. Therefore, as we believed it was more suitable for our purposes, we decided to use a mirror stereoscope to display our stimuli.

The final outcome of *Chapter 4* was the discovery of a large body of research into both *Gloss* and *Roughness* using stereoscopes. As both of these terms were present in our

property taxonomy, we decided that these two properties would be fruitful to investigate further.

Therefore, *Chapter 5* presented our in depth literature review of the gloss and roughness literature, using this knowledge to decide how best to create stimuli which would enable us to investigate these two properties effectively. We began this literature review by examining work already performed using real glossy samples, but decided against this path for our own work as we desired more control over the creation of our stimuli.

As we were not using real stimuli, we investigated methods of generating synthetic surfaces instead. We investigated eight different surface generation techniques, before deciding to use $1/f^\beta$ noise surfaces due to their isotropic, naturalistic appearance and parsimonious parameter set. After selecting a method of creating synthetic surfaces, we investigated what work had already been performed using rough, glossy surfaces. We found that while there has been research performed on the interaction between surface shape, gloss and stereoscopic vision, little has been done to investigate the perception of rough surfaces with physically accurate lighting and reflectance models.

Therefore, we then investigated the different reflectance models which were common in the literature. We found that there was a wide range of models of varying physical accuracy and computational complexity, providing different compromises between accuracy and rendering time. We decided to use the Schlick model due to its good compromise between accuracy, accounting for the physical properties of light, and computational complexity. When this model was combined with multiple bounce path tracing to address the issues of interreflection and soft, self-shadowing (also discussed in *Chapter 5*), it would account for most light phenomena. A brief discussion of the light source used follows, with the conclusion being that an area light source would be used.

Finally, due to the lack of research in this area, we decided to investigate the following question: Does highlight disparity strengthen perceived gloss on rough surfaces?

In *Chapter 6* we detailed exactly how we constructed our stereoscope and our stimuli. We began by deciding on the design of our stereoscope, choosing to build a single-plane mirror stereoscope. We then proceeded to discuss in detail the construction of our stereoscope so the reader can reconstruct the system if necessary. We paid close attention to the types of mirrors used, as they are of paramount importance to a

successful mirror stereoscope. This was presented first as the stereoscope design had bearing on the maximum size of stimuli we could use.

We then discussed the stimuli we created for our experiments. Our stimuli were $1/f^\beta$ noise surfaces, created at five different mesoscale roughness levels. These surfaces were rendered as static images, using a path tracer and the Schlick reflectance model, from three different angles. This allowed the surfaces to be displayed as stereogram pairs with virtual depth, or as simple monocular images. Our stimuli were high detail and thanks to the use of path tracing, considering 10 bounces for each light probe and the use of an area light, took into account effects of interreflections and soft shadowing. We go into detail about how we constructed and rendered these surfaces allowing the reader to recreate similar stimuli as required.

The surfaces created were each generated from a different random phase spectra to prevent participants performing per-pixel comparisons. While previous experiments have shown that random phase has no affect on perception of roughness, to our knowledge no-one had confirmed that there is no effect on gloss perception.

To this end, in *Chapter 7* we presented two psychophysical experiments designed to investigate whether random phase alters a participant's perception of how glossy a surface is. First, we investigated a single roughness level to see if there was any effect on gloss perception from changing the random phase of a surface. Using both binomial distribution analysis and a one-way repeated measures ANOVA we showed that there was a significant difference in the perceived strength of gloss between surfaces of the same roughness, but different random phase.

As we discovered that random phase did alter participants' perception of gloss, a second experiment was carried out to see whether these results generalised to our four other roughness levels. A two-way repeated measures ANOVA showed, as before, that there was a significant difference in the perceived strength of gloss due to random phase.

As these results showed that stimuli of the same global roughness were perceived at different gloss strengths to others, we decided to select pairs which participants had selected as being most similar. We normalised the results from both previous experiments, so perceptually similar surfaces from each of the five roughness levels we used could be selected. The conclusion of this chapter identified two pairs of surfaces for each of our five roughness levels which participants judged to be similar strengths of gloss.

These pairs of surfaces were used in *Chapter 8* where we answered the question posed earlier: Does highlight disparity strengthen perceived gloss on rough surfaces? In this chapter, we showed that there is no simple yes or no answer to this question; rather it is conditional on what roughness of surface the participant is judging.

The experiment presented in this chapter asked participants to adjust the strength of gloss of a surface so that the two shown (one with disparity and one without) were the same apparent glossiness. We used five different roughness levels (specifically the pairs of surfaces selected in *Chapter 7*) and three different 'target gloss' levels, giving us 15 conditions. With 28 participants' results, we were able to calculate mean accuracy ratings for each of these conditions.

We performed a two-way repeated-measures ANOVA on these accuracy ratings to discover that while roughness did have a significant effect on how accurate participants were, the different target gloss levels did not. This showed us that our initial question would be dependent on the roughness of the surface being viewed.

Using the knowledge that different target gloss levels didn't alter a participant's accuracy, we analysed the marginal means for each roughness directly. We found that participants' accuracy for three of our roughness levels (β = 1.8, 2.0 and 2.2) was significantly skewed from the result expected if there was no experimental effect. Due to the design of our experiment, this showed that for those three roughness levels, participants were seeing the surfaces with disparity as having significantly stronger perceptual gloss.

9.2 Discussion

At this stage, we can relate the findings of this thesis back to our initial motivations presented in *Chapter 1* and ask how these results inform us about digital and online presentation of items. While our findings that the random phase of fractal surfaces can alter participant's perception of gloss are important for research purposes, we believe it is our finding that monocular images significantly under represent a surface's true gloss which has wider significance.

This result, of course, only applies to items which are highly reflective and those with distinct, rough, surface texture. There are, however, many cases where we believe this knowledge can be used to improve the digital representation of products and items with little additional cost.

For example, as 3D display systems become more prevalent and accepted, it will allow manufactures of high-value items to present images of their products in a more realistic way over the internet, or through television and film advertising. Items which rely on the quality of their materials as a selling point, such as the finished metal on high-end smartphones, can better communicate their gloss and shine.

In addition, many museums over the past decade have begun digitising their more delicate works, or those works they don't have space to display to the public. Items such as marble or bronze sculptures could easily be captured with disparity and displayed to the public via 3D kiosks, allowing visitors to experience a more accurate vision of the items at their best.

As a final example, during the design stage of a product where the final look is important, designers could easily (with a little additional time) render the product so it could be presented in 3D. This would allow a more accurate representation of what the finished product's reflectance properties would be like, before having to prototype it with the desired materials.

In addition, there are now affordable devices on the consumer market (such as the FinePix REAL 3D camera which is available at time of writing for around £150) which allow people to record 3D images or video which can easily be played back through 3D display systems.

In conclusion then, we believe this thesis has shown that for little additional monetary or time cost, glossy items can be represented digitally more accurately and realistically by simply recording them with a 3D capture device, or rendering them with the correct stereo disparity.

9.3   Future Work

At this stage, we would like to briefly discuss two possible opportunities for future work which are suggested by the results presented in this thesis.

While in this thesis we have shown that random phase spectra in $1/f^{\beta}$ noise surfaces can alter the perceived strength of gloss on rough surfaces, we did not investigate in detail to discover why this might be. This result is interesting, as work by Padilla showed that random phase did not alter participants' perception of roughness. As we ensured that all our surfaces had the same global, root mean square (RMS) roughness, it would suggest that for gloss judgements, local surface statistics take precedent over global ones.

Another interesting question worthy of future research is why the presence of highlight disparity only altered perceived gloss strength on the roughest three of our surfaces. Even accounting for experimental noise, it seems apparent that the difference between ratings with disparity and without tends towards zero as the surface get smoother. Why then, is highlight disparity so important for rough surfaces, but not so for smooth? We believe this would be an interesting question to investigate in the future, as many contrasting explanations have been forwarded.

9.4　Final Conclusions

In this thesis, we have shown that the presence of highlight disparity on glossy surfaces increases the perceived strength of gloss, at least for rough surfaces. In other words, when a rough, glossy surface is presented without correct stereoscopic disparity, participants are unable to accurately estimate how glossy it is. We consider this to be the main contribution of this thesis as although there has been previous work investigating gloss and disparity (such as Blake and Bülthoff's early work [9] or the more recent work of Wendt et al. [10, 11] which all used relatively smooth, unnatural surfaces) to the author's knowledge no one has used the same combination of naturalistic, rough surfaces and physically accurate lighting before.

In showing this, however, we have also presented two other results which we believe are also novel. First, we present a two-level taxonomy which was determined by non-expert participants' decisions. From an original list presented by an expert in fabrics, we offered a non-expert list of fabric properties, confirmed by naïve participants. Using different naïve participants, these properties were clustered into 11 groups, each with a representative word, giving us a two-level taxonomy readily understandable by non-expert, English speakers. We believe this taxonomy to be novel as it is the only one known to the author to be understandable by naïve participants and to cover such a wide range of different fabric types.

The final piece of novel work is that we have shown, perhaps surprisingly considering earlier work, that the random phase of $1/f^{\beta}$ noise surfaces can alter the perceived strength of gloss on the surface. That is to say, two surfaces of the same global root mean square roughness, made of the same material can be perceptually different in terms of gloss strength. Although Padilla showed that roughness is unchanged by

changing the phase spectrum [7] to the author's knowledge no one has shown that phase can alter perceived gloss strength before.

We believe, therefore, that this thesis offers interesting, novel contributions to the research in this area. In particular, the work on phase, highlight disparity and strength of gloss presented in *Chapter 7* and *Chapter 8* advances the research in these areas with new, novel results.

Appendix A

Example Consent Form

Gloss Adjustment Experiment Consent Form

Thank you for agreeing to take part in our research. Today, I will ask you to perform an experiment which will take around 1 hour. In this experiment, you will be shown a series of surfaces displayed in 3D and asked to adjust similar surfaces to be the same 'glossiness'. You should be aware that due to the nature of this experiment you will be asked to use a chin rest throughout, but can take as many breaks as you like. In addition, you will go through a short eye-test beforehand to make sure you have working depth perception.

We will keep your information confidential and will anonymise all of your written results. All your results will be recorded under whatever initials you give at the beginning of this experiment and in any future publications or discussions these initials will be used so your results can never be linked to you personally.

All data will be collected and stored in accordance with the Data Protection Act 1998.

You have the right to withdraw at anytime without giving a reason.

Participant's Name:

- May we keep an anonymous copy of any results you provide?

  ☐ Yes   ☐ No

- May we use these results in the future for conferences, publications or presentations?

  ☐ Yes   ☐ No

- May we report (without identifying you) any comments/feedback on this experiment you provide?

  ☐ Yes   ☐ No

I have been fully informed as to what this experiment will entail and am aware of my right to withdraw at anytime. I hereby fully and freely consent to participation in the study, which has been fully explained to me.

Signed _____

Date _____

Figure A.1 - Example consent form as used in the experiment in *Chapter 8*. All the consent forms used in the PhD followed the same basic template with the title and first paragraph changed to reflect the study being performed.
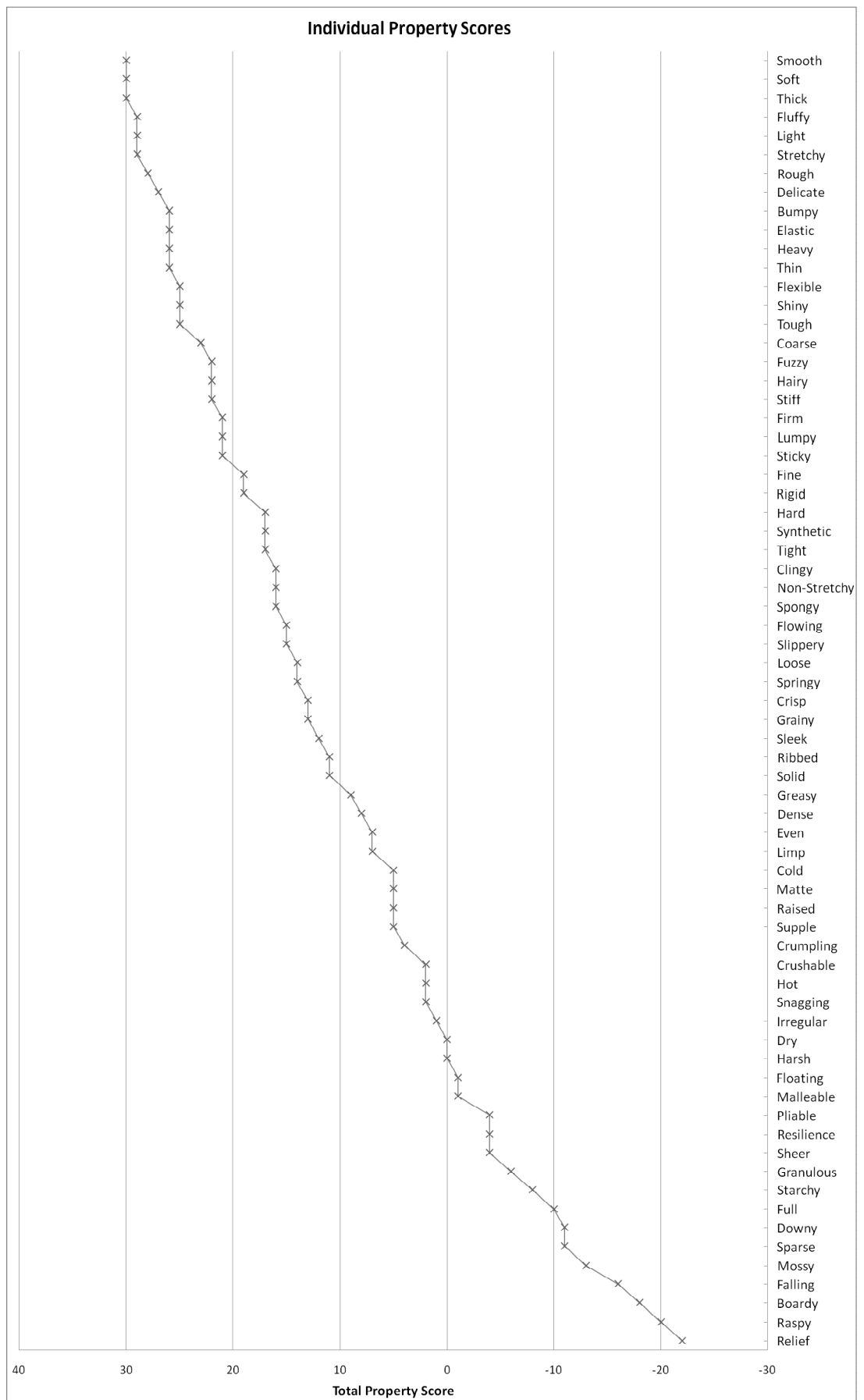
Detailed Word Sorting Results

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	Score
Smooth	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30
Soft	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30
Thick	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	30
Fluffy	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	29
Light	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	29
Stretchy	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	29
Rough	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	28
Delicate	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	27
Bumpy	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	0	26
Elastic	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0	1	1	0	1	1	1	1	1	26
Heavy	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1	1	1	1	1	0	1	1	1	1	1	26
Thin	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	26
Flexible	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0	0	1	1	1	0	1	0	1	1	1	1	1	1	1	1	25
Shiny	1	1	1	1	1	1	1	1	1	0	1	1	0	1	1	0	-1	1	1	1	1	1	1	1	1	1	1	1	1	1	25
Tough	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	-1	25
Coarse	0	1	1	1	1	1	1	1	1	1	1	1	-1	0	1	0	0	1	1	1	1	1	1	0	1	1	1	1	1	1	23
Fuzzy	1	1	1	0	1	1	1	0	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	0	-1	1	1	1	1	1	22
Hairy	1	1	1	0	1	0	0	1	1	0	1	1	0	1	1	1	1	0	-1	1	1	1	1	1	1	1	1	1	1	1	22
Stiff	1	1	1	0	1	1	1	0	1	1	0	0	1	1	1	0	1	1	-1	1	1	1	1	1	0	1	1	1	1	1	22
Firm	0	0	1	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	0	-1	1	1	1	1	1	1	1	1	1	0	21
Lumpy	1	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	-1	1	1	1	1	1	0	1	1	1	1	0	21
Sticky	1	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	-1	1	1	1	1	0	21
Fine	1	1	1	0	1	1	0	1	1	0	1	0	0	0	1	0	0	1	1	1	0	1	1	1	0	1	1	0	1	1	19
Rigid	0	1	0	1	1	1	1	0	1	1	1	0	0	0	-1	0	1	-1	1	1	1	1	1	1	1	1	1	1	1	1	19
Hard	1	0	1	1	0	0	0	0	1	1	-1	0	1	1	1	1	1	1	1	-1	-1	1	1	1	1	1	1	1	1	0	17
Synthetic	0	1	1	0	1	1	1	1	1	1	1	1	-1	1	1	1	-1	0	1	1	1	0	1	-1	1	0	1	0	0	1	17
Tight	1	1	1	0	1	1	0	1	0	0	0	1	0	1	1	-1	1	1	0	1	1	0	1	1	-1	0	1	1	1	1	17
Clingy	1	1	1	-1	1	1	1	1	1	0	-1	1	1	1	1	0	1	1	1	0	0	-1	1	0	0	0	0	1	1	1	16
Non-Stretchy	0	0	1	1	0	0	1	0	1	0	1	0	0	1	1	1	0	1	0	0	0	1	1	1	0	0	0	1	1	1	16
Spongy	1	1	1	0	1	0	0	0	1	1	1	0	1	1	0	0	0	1	1	1	1	1	1	0	-1	0	1	1	1	-1	16
Flowing	1	1	0	-1	1	1	1	0	1	1	1	1	0	-1	1	1	1	1	1	-1	0	0	1	1	1	-1	0	1	1	1	15
Slippery	1	1	1	0	1	0	0	0	0	1	1	1	1	0	-1	0	0	0	1	0	1	1	1	1	-1	1	1	1	1	-1	15
Loose	1	1	1	0	1	1	1	0	1	1	-1	1	-1	1	1	1	-1	1	1	-1	0	1	0	1	0	-1	1	0	1	1	14
Springy	1	0	1	0	1	0	1	0	1	0	1	-1	1	0	1	0	0	1	0	0	0	0	1	0	0	1	1	1	1	1	14
Crisp	1	0	0	0	1	1	1	0	1	0	0	0	0	1	1	-1	0	1	1	-1	1	0	1	1	0	0	0	1	1	1	13
Grainy	0	1	1	1	1	0	0	0	1	0	0	0	0	1	1	1	1	1	0	0	-1	-1	1	0	1	0	0	1	1	0	13
Sleek	1	0	1	-1	1	1	1	0	1	0	1	1	0	0	1	0	0	1	1	-1	0	0	0	1	0	0	0	0	1	1	12
Ribbed	-1	1	1	-1	1	1	-1	0	1	1	0	1	0	1	0	1	0	1	1	0	0	0	1	0	0	0	0	0	1	1	11
Solid	1	0	1	0	0	1	0	0	0	1	0	0	0	1	-1	-1	0	1	0	-1	-1	1	1	1	1	1	1	1	1	0	11
Greasy	1	1	0	-1	0	0	0	0	1	1	1	0	0	0	1	0	1	1	0	-1	-1	1	1	-1	1	0	0	1	1	0	9
Dense	0	1	1	0	0	0	0	0	1	1	0	1	1	0	1	0	1	0	0	-1	0	0	0	-1	0	0	1	1	1	-1	8
Even	1	0	1	-1	1	0	1	0	1	1	0	0	0	0	1	1	0	1	-1	-1	1	-1	1	0	-1	0	1	0	-1	1	7
Limp	1	0	0	-1	0	0	1	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	1	1	0	0	1	1	0	0	7
Cold	1	0	0	1	0	0	0	0	0	1	1	-1	1	1	-1	1	0	0	1	-1	0	1	0	1	-1	0	0	-1	1	-1	5
Matte	-1	0	1	-1	1	1	0	0	1	1	0	1	-1	1	-1	-1	1	0	0	0	0	1	0	0	0	0	0	0	1	0	5
Raised	0	1	0	-1	0	0	0	0	1	0	1	1	1	0	0	0	-1	0	-1	0	0	1	1	0	-1	1	1	0	1	-1	5
Supple	0	0	0	0	0	1	1	0	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	1	0	5
Crumpling	1	0	1	1	0	1	-1	0	1	-1	-1	0	0	0	1	0	-1	-1	1	1	-1	0	0	0	0	-1	0	1	1	1	4
Crushable	0	0	0	0	0	0	-1	0	1	0	0	0	0	1	0	-1	-1	0	-1	0	0	0	1	0	0	1	0	0	1	1	2
Hot	1	0	0	0	-1	0	0	0	0	1	1	1	-1	1	1	-1	-1	0	0	1	-1	0	1	0	1	-1	0	0	0	-1	2
Snagging	0	1	0	-1	0	1	-1	1	1	1	-1	0	1	1	0	-1	1	1	0	0	-1	0	0	-1	-1	0	0	0	1	1	2
Irregular	0	1	0	0	1	0	0	0	0	1	0	0	1	-1	-1	0	0	1	0	0	-1	0	-1	1	0	0	0	0	1	-1	1
Dry	1	0	0	1	0	0	0	0	0	0	-1	0	1	1	1	0	-1	1	1	-1	-1	0	0	-1	-1	-1	1	0	0	-1	0
Harsh	0	0	0	1	0	1	0	0	0	0	-1	0	0	-1	1	0	-1	-1	0	-1	0	1	1	1	-1	0	0	0	1	1	0
Floating	1	1	0	-1	0	0	0	1	-1	0	1	-1	0	0	-1	-1	1	0	-1	-1	1	-1	1	-1	0	1	1	-1	0	1	-1
Malleable	-1	1	0	0	-1	1	0	1	1	0	0	0	-1	-1	0	0	0	1	-1	0	-1	-1	-1	1	0	0	0	-1	1	-1	-1
Pliable	-1	0	0	-1	1	1	0	1	1	1	-1	0	-1	0	-1	-1	0	1	1	-1	0	-1	-1	-1	1	0	0	0	-1	-1	-4
Resilience	-1	0	0	-1	1	1	1	0	0	0	1	1	-1	-1	0	0	-1	-1	-1	1	1	-1	0	1	1	0	0	-1	1	1	-4
Sheer	0	0	1	-1	1	1	0	1	0	-1	-1	0	0	0	0	-1	-1	-1	0	-1	1	1	0	0	-1	-1	1	0	1	-1	-4
Granulous	-1	0	0	-1	1	0	-1	0	0	-1	0	0	-1	-1	1	1	0	0	0	-1	0	1	0	1	1	0	-1	1	1	-1	-6
Starchy	0	0	0	-1	1	1	0	-1	1	0	-1	-1	0	0	-1	-1	-1	0	0	1	-1	1	-1	-1	0	0	-1	1	1	-1	-8
Full	1	0	0	-1	0	1	1	0	0	0	-1	-1	-1	0	-1	0	-1	0	0	-1	0	-1	-1	-1	-1	0	-1	0	0	0	-10
Downy	0	0	0	-1	0	0	-1	0	1	-1	-1	0	0	-1	0	-1	0	0	0	-1	0	-1	-1	-1	-1	0	-1	0	0	-1	-11
Sparse	0	0	1	-1	0	1	0	0	-1	0	-1	-1	0	-1	0	0	-1	-1	0	0	1	-1	-1	-1	-1	0	0	-1	-1	-1	-11
Mossy	0	0	0	-1	0	0	-1	-1	1	0	-1	-1	0	0	-1	0	-1	-1	0	1	0	-1	-1	-1	-1	-1	0	0	0	-1	-13
Falling	1	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	0	-1	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-16
Boardy	0	-1	0	-1	0	-1	0	-1	0	-1	-1	0	1	0	0	-1	-1	-1	-1	0	-1	-1	0	-1	-1	-1	-1	0	-1	-1	-18
Raspy	0	0	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	-1	0	-1	0	1	-1	-1	-1	-1	-1	-1	-1	0	-20
Relief	1	0	0	-1	-1	0	0	-1	-1	-1	-1	-1	-1	-1	-1	0	-1	-1	0	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-22

Figure B.1 - Detailed results for the properties study described in *Section 3.3.3* with individual word scores for each of the 30 participants.

Figure B.2 - A graph of the total scores for every word used in the Properties Experiment described in *Section 3.3.3*.

137

Figure B.3 - Full dissimilarity matrix of all properties

	Bendable	Breezy	Brittle	Bumpy	Clingy	Coarse	Cold	Creasable	Crinkly	Crisp	Crumpling	Crunchy	Crushable	Delicate
Bendable	0	20	16	19	19	19	19	16	20	19	16	18	14	19
Breezy	20	0	20	19	16	19	16	16	16	20	18	19	18	13
Brittle	16	20	0	19	20	15	18	19	18	14	18	14	15	17
Bumpy	19	19	19	0	20	15	20	19	15	20	18	18	18	20
Clingy	19	16	20	20	0	20	18	17	19	20	19	19	19	19
Coarse	19	19	15	15	20	0	20	19	17	17	19	14	17	19
Cold	19	16	18	20	18	20	0	19	19	18	18	20	19	19
Creasable	16	16	19	19	17	19	19	0	12	16	5	16	6	18
Crinkly	20	16	18	15	19	17	19	12	0	16	10	15	13	18
Crisp	19	20	14	20	20	17	18	16	16	0	15	11	14	20
Crumpling	16	18	18	18	19	19	18	5	10	15	0	16	7	19
Crunchy	18	19	14	18	19	14	20	16	15	11	16	0	15	17
Crushable	14	18	15	18	19	17	19	6	13	14	7	15	0	19
Delicate	19	13	17	20	19	19	19	18	18	20	19	17	19	0

Figure B.4 - Detailed view of the dissimilarity matrix, showing the first 15 properties with the highlighted elements showing example dissimilarities of 20 and 5.
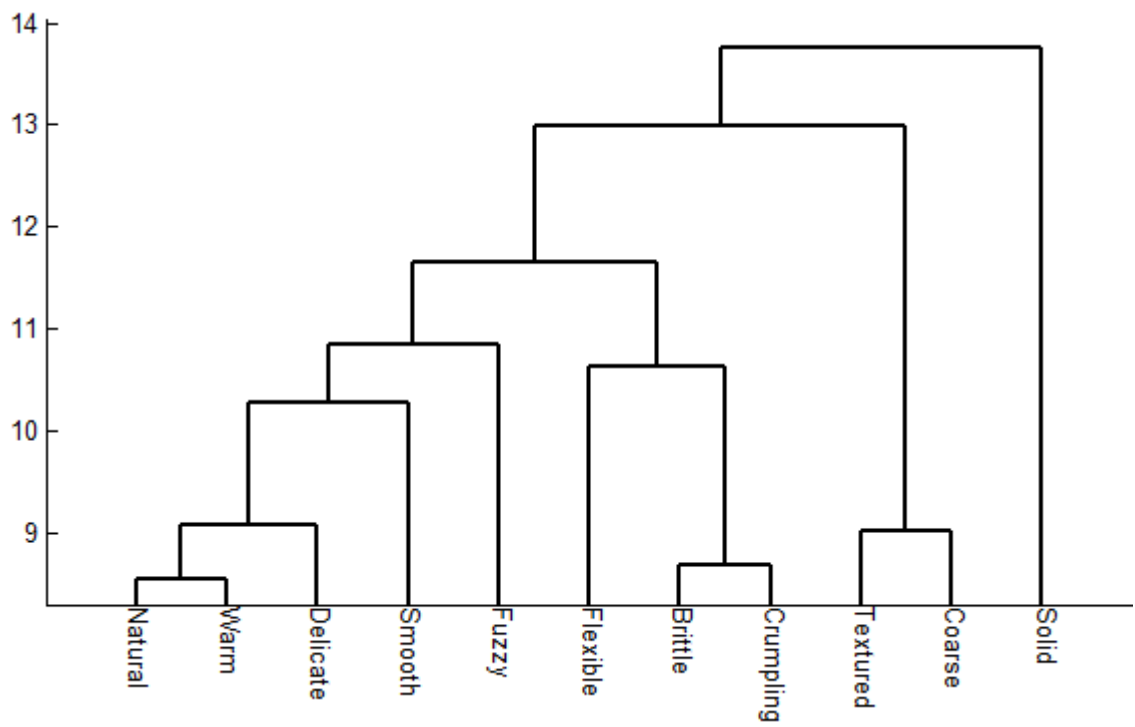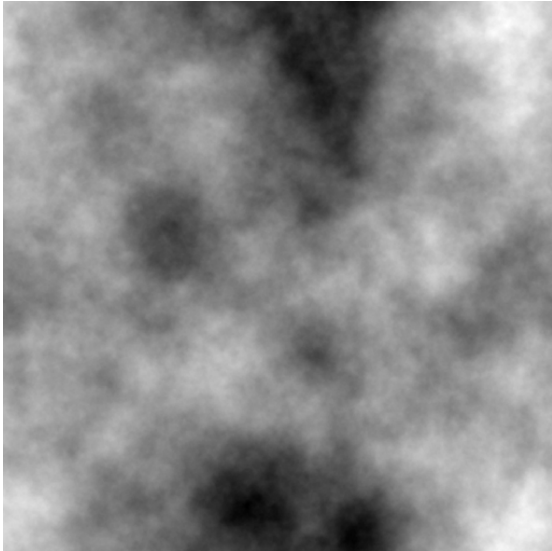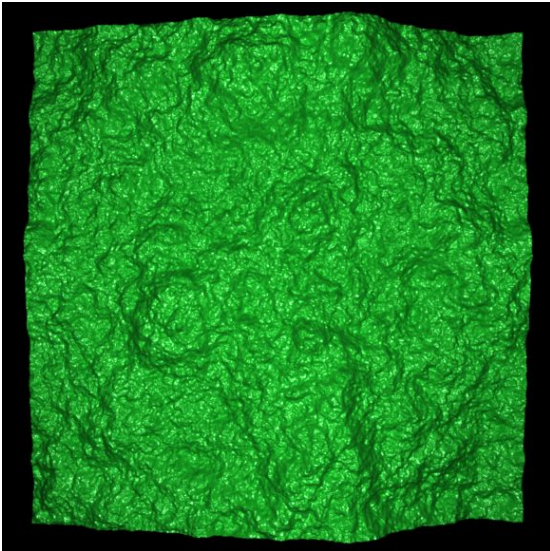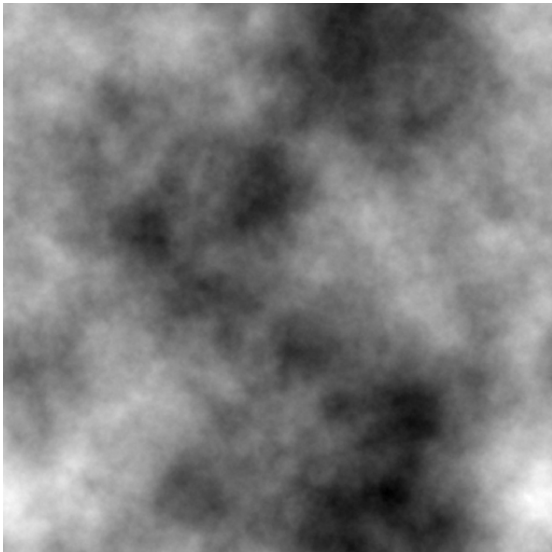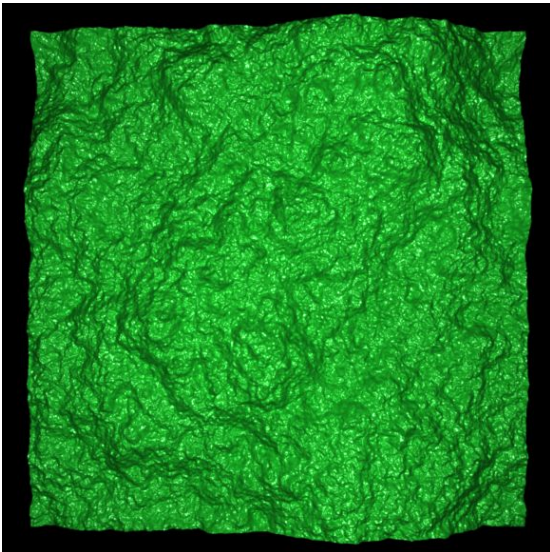
Figure B.5 - A simplified version of the dendrogram showing eleven groups denoted by their representative properties. Note that the vertical axis still displays the dissimilarity but now starts at the 'cutting' point shown in *Figure 3.6*.

Appendix C

1/f$^\beta$ Height Maps and Surfaces

#	Height map	Rendered Surface (at Gloss Level 7)
1	'beta18set1.png'	'mono18Set1-07.bmp'
2	'beta18set2.png'	'mono18Set2-07.bmp'

3
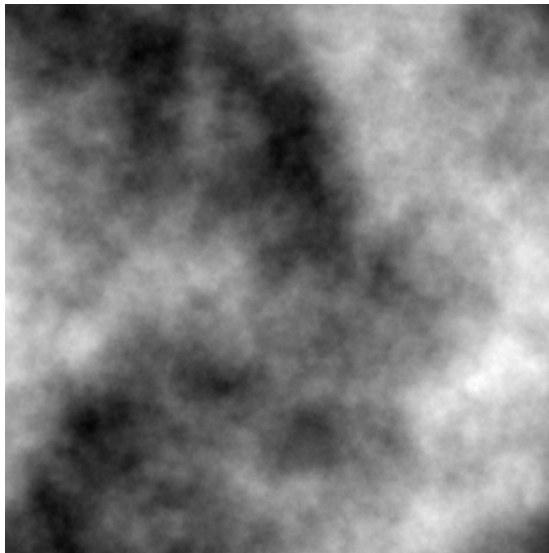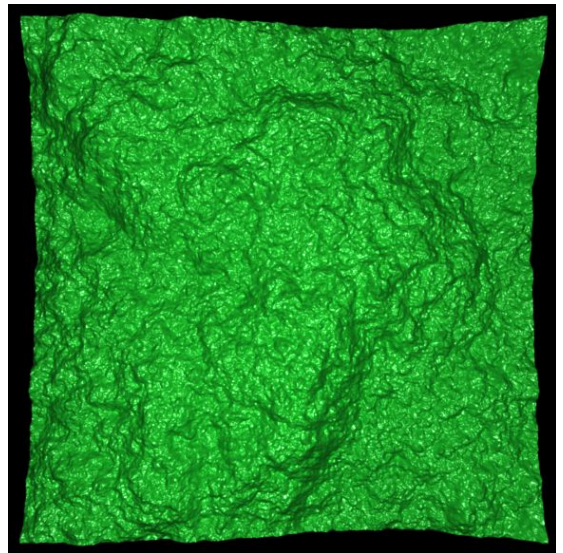
'beta18set3.png'   'mono18Set3-07.bmp'

4
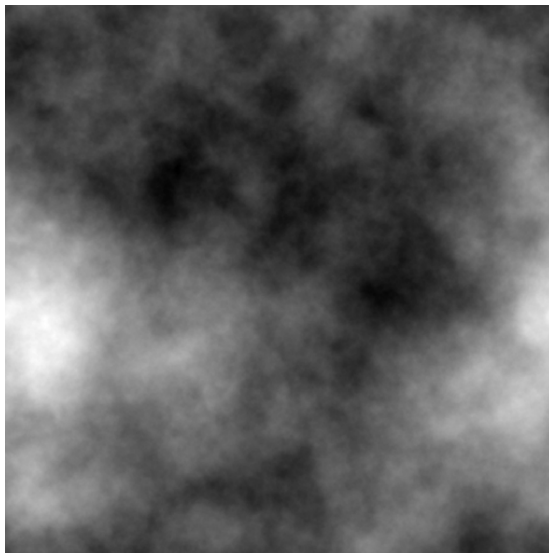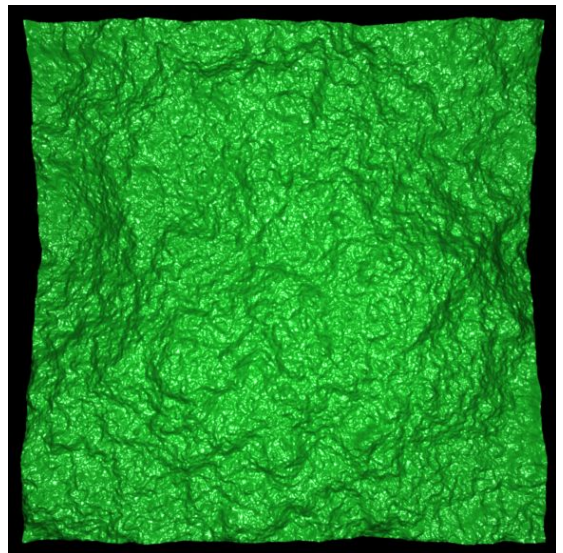
'beta18set4.png'   'mono18Set4-07.bmp'

5
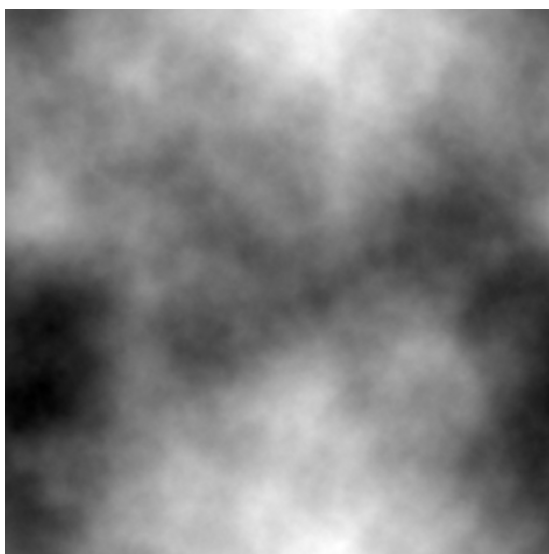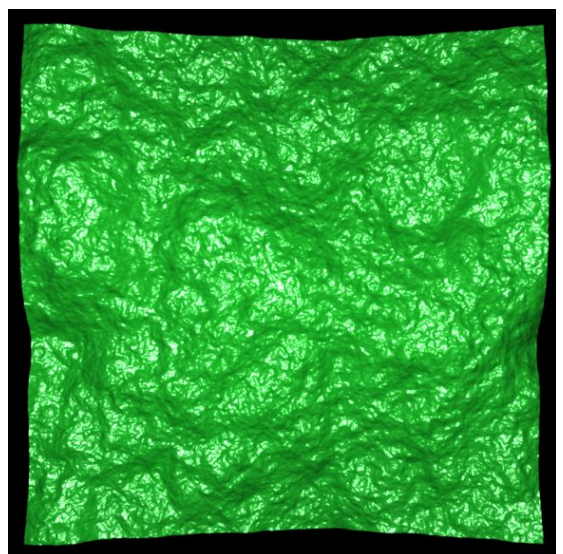
'beta20set1.png'   'mono20Set1-07.bmp'

6
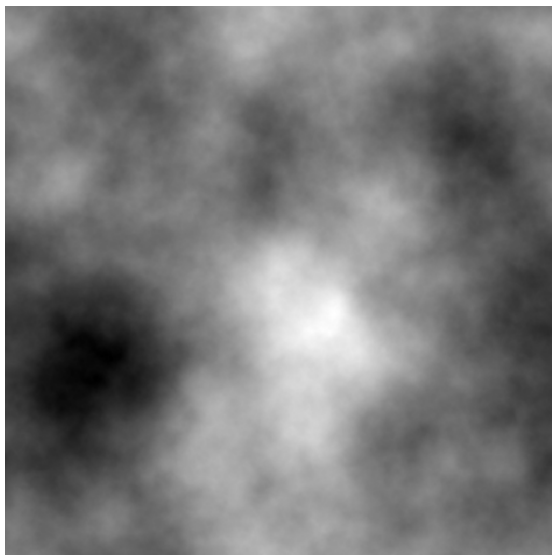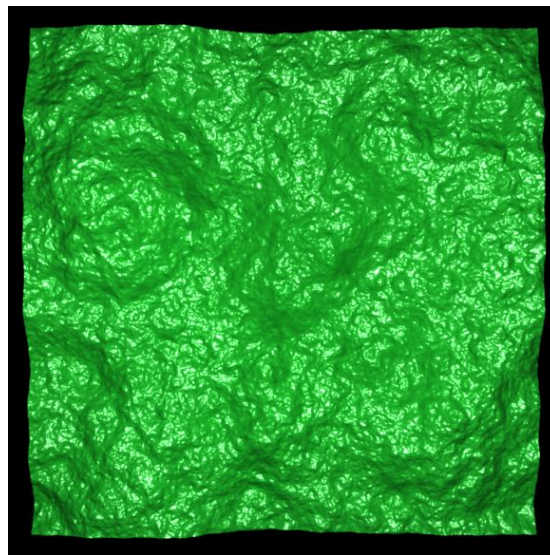
'beta20set2.png'　　　　　'mono20Set2-07.bmp'

7
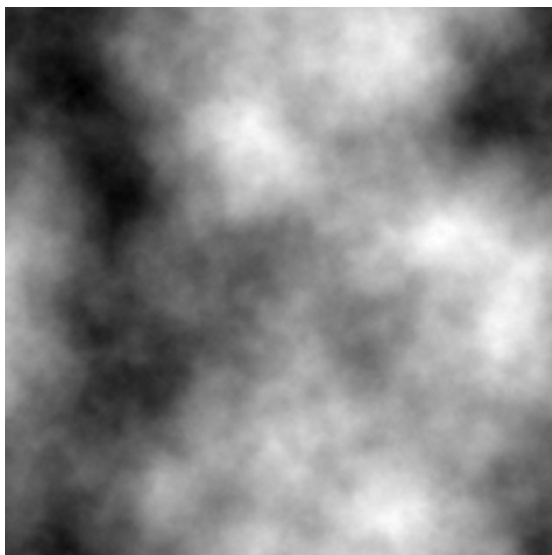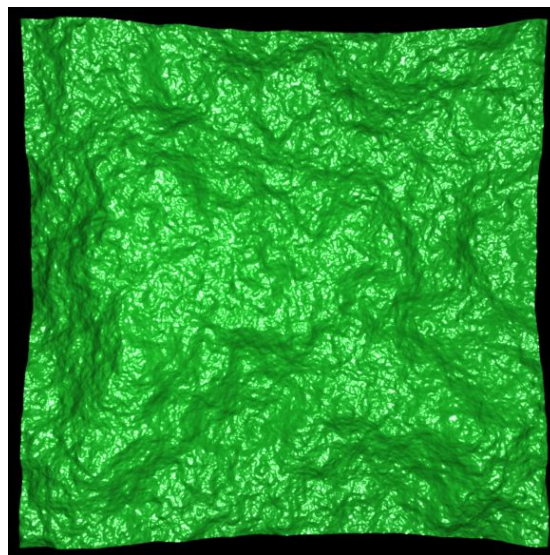
'beta20set3.png'　　　　　'mono20Set3-07.bmp'

8
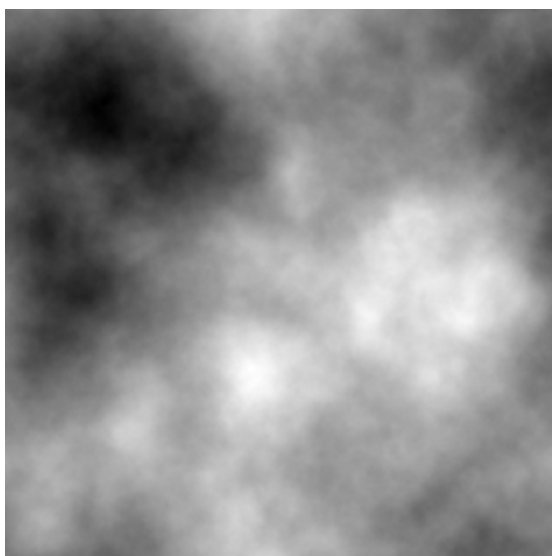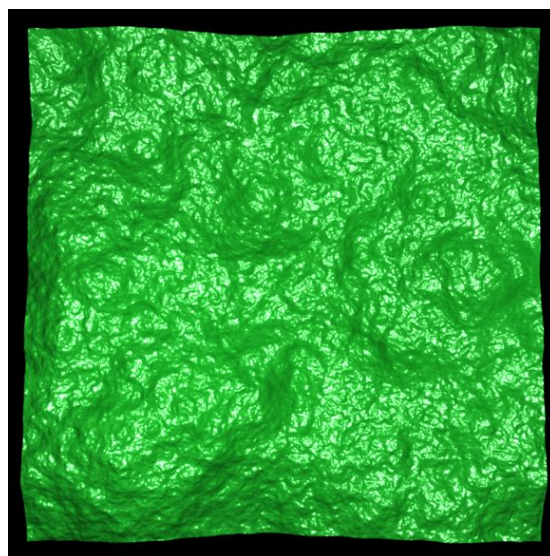
'beta20set4.png'　　　　　'mono20Set4-07.bmp'
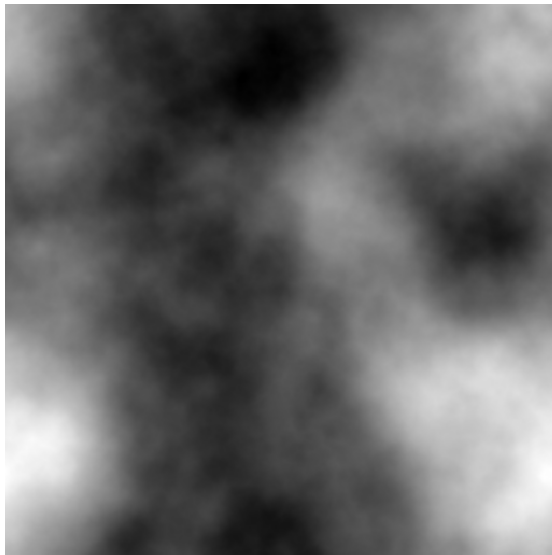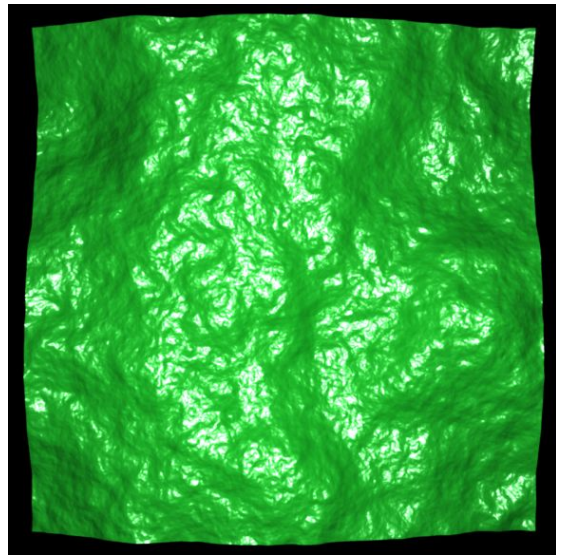
9	'beta22set1.png'	'mono22Set1-07.bmp'
10	'beta22set2.png'	'mono22Set2-07.bmp'
11	'beta22set3.png'	'mono22Set3-07.bmp'
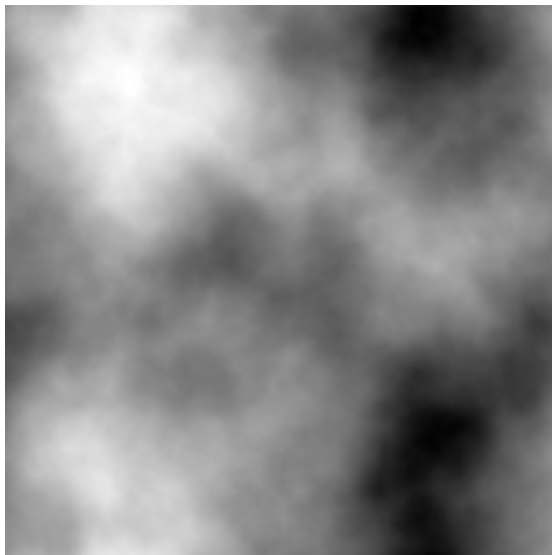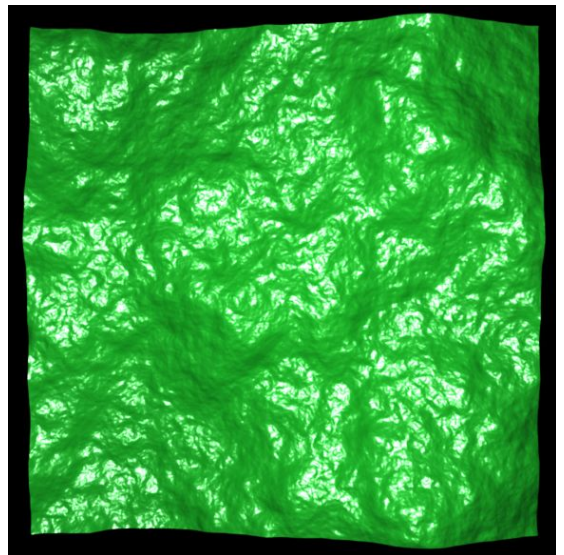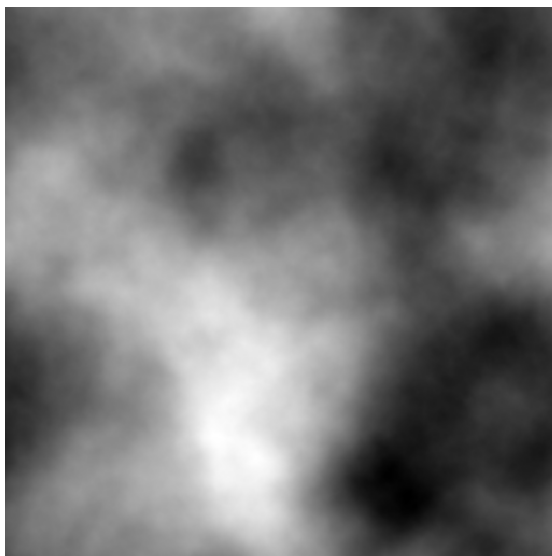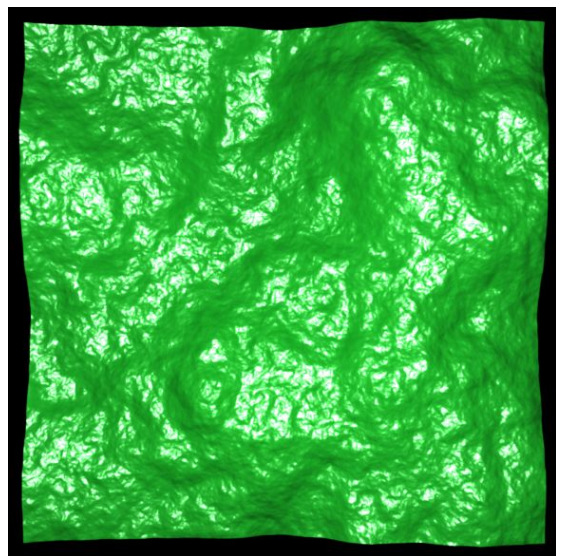
12
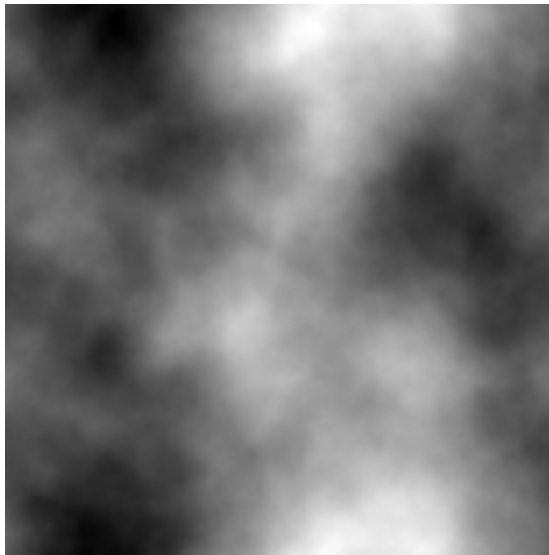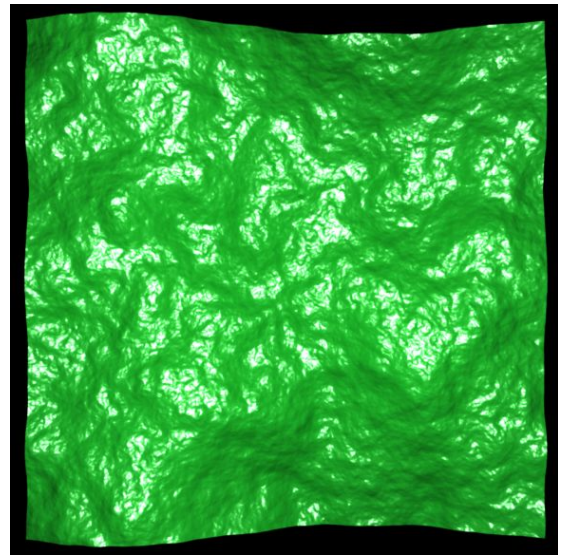
'beta22set4.png' 'mono22Set4-07.bmp'

13
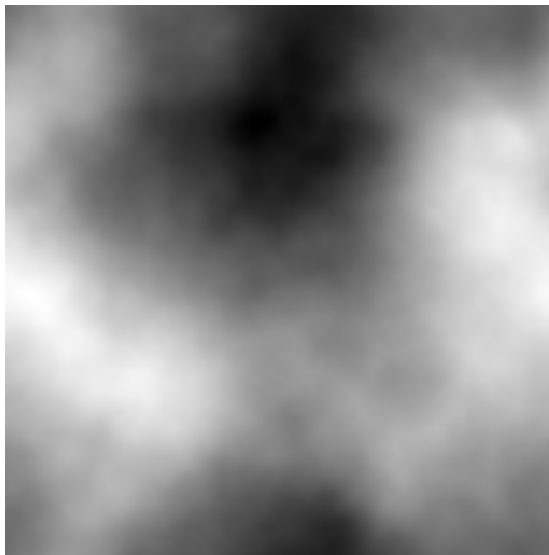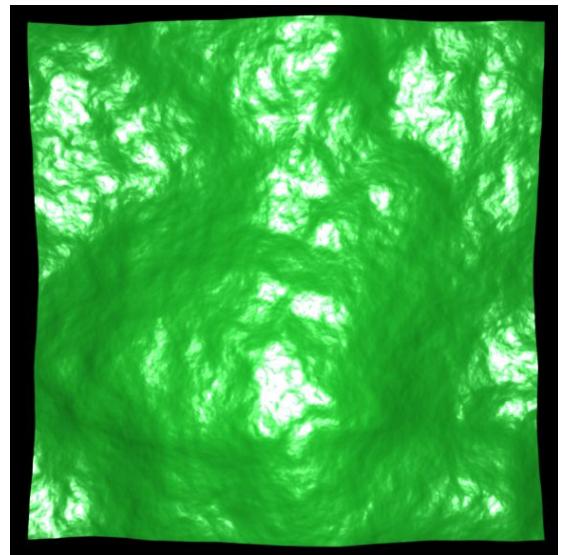
'beta24set1.png' 'mono24Set1-07.bmp'

14
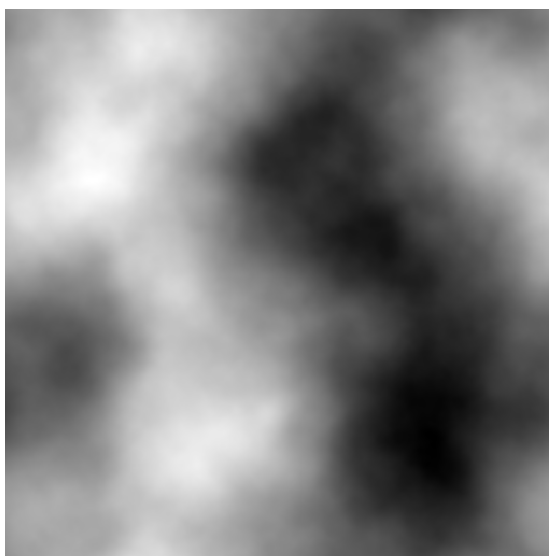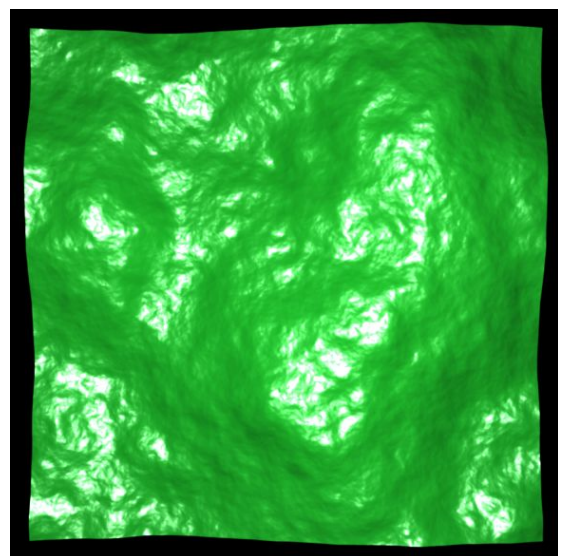
'beta24set2.png' 'mono24Set2-07.bmp'

145

15
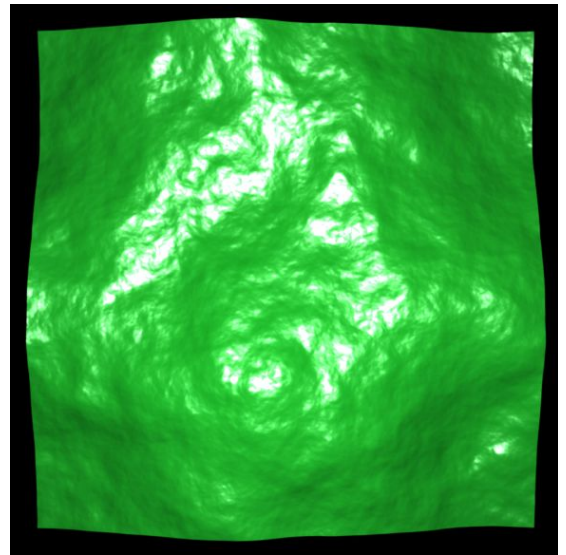
'beta24set3.png'

'mono24Set3-07.bmp'

16
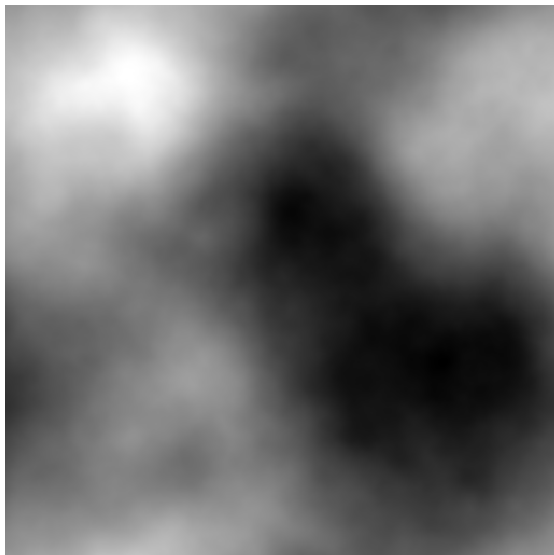
'beta24set4.png'
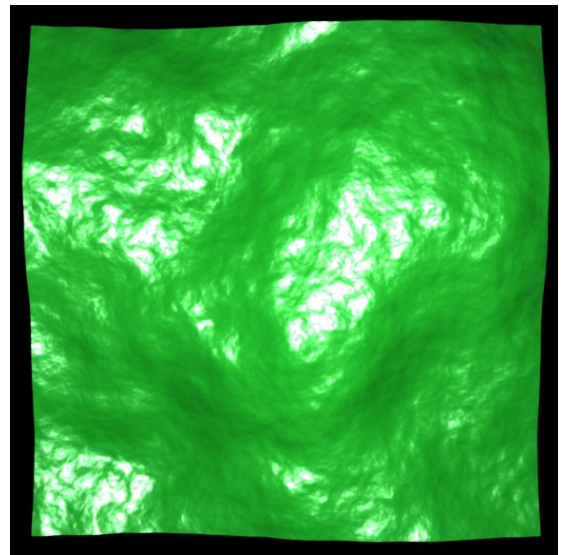
'mono24Set4-07.bmp'

17
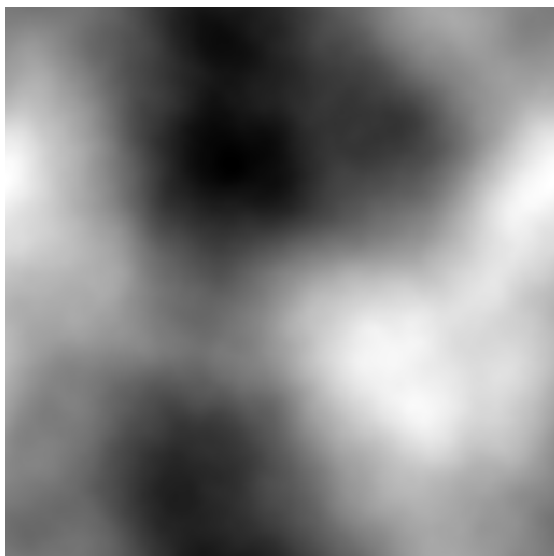
'beta26set1.png'

'mono26Set1-07.bmp'
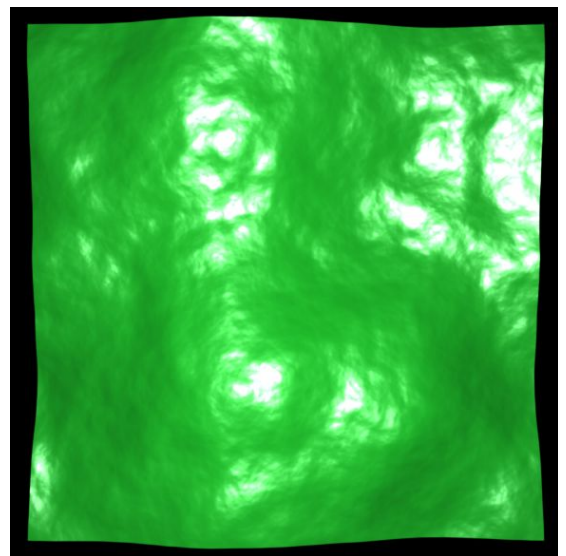
18		
	'beta26set2.png'	'mono26Set2-07.bmp'
19		
	'beta26set3.png'	'mono26Set3-07.bmp'
20		
	'beta26set4.png'	'mono26Set4-07.bmp'
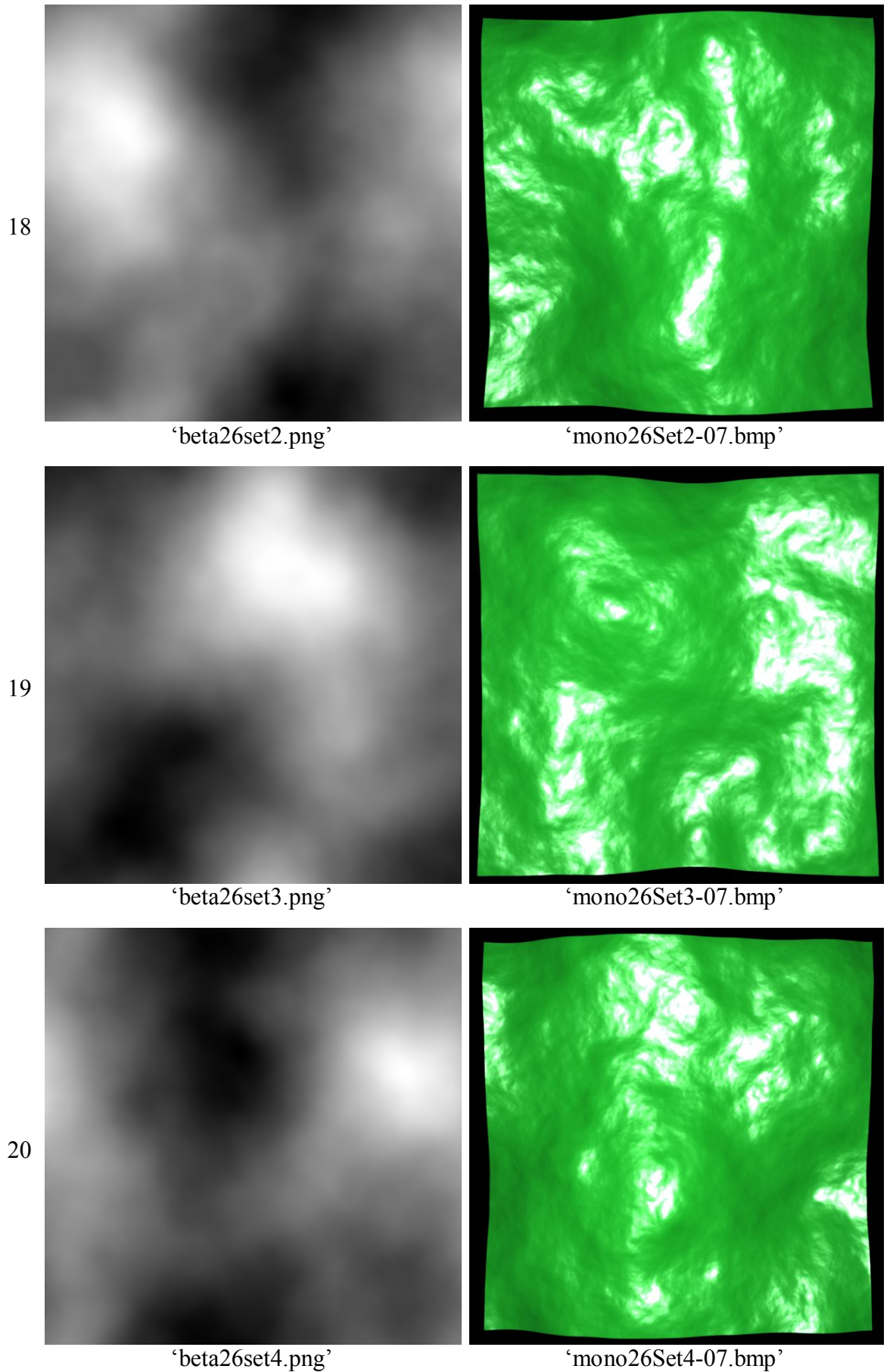
Table C.1 - $1/f^\beta$ height maps (Left) and finished rendered surface (Right). Please note, we only show each surface at a single gloss level here. For an example of the possible range of gloss levels, see *Table C.2.*
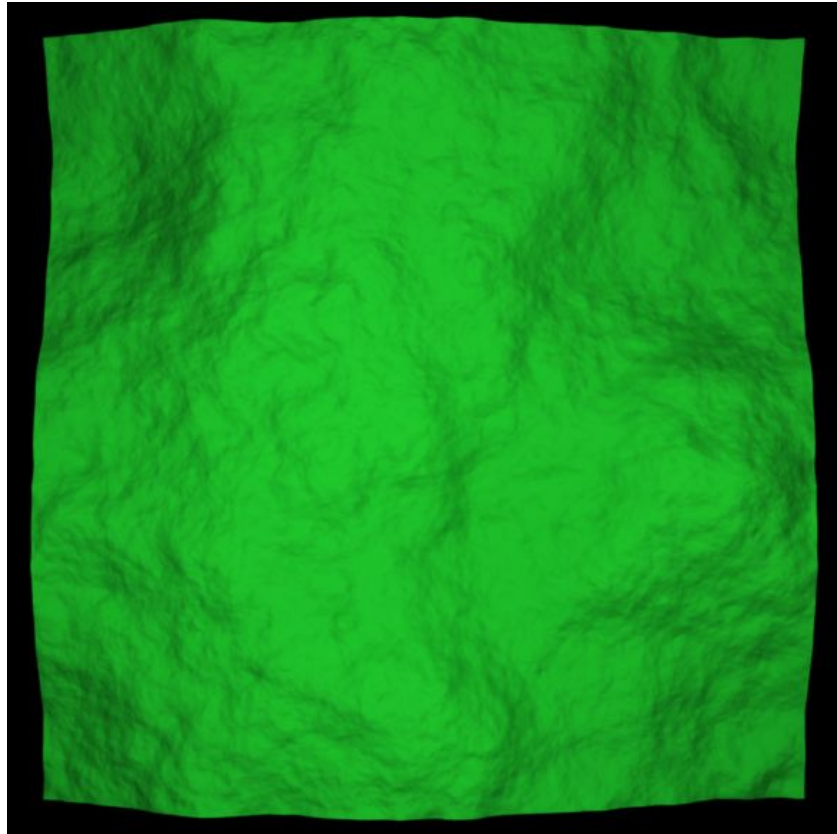
'mono22Set1-00.bmp'
N.B. Perfectly diffuse surfaces were never used in our
experiments, but are provided here to give the full possible range.
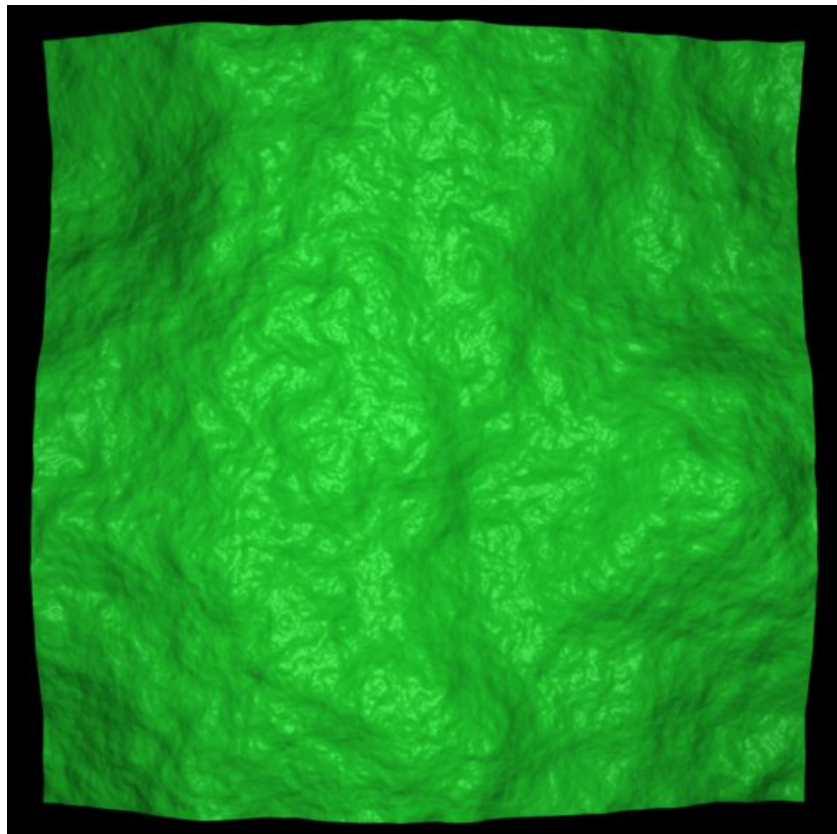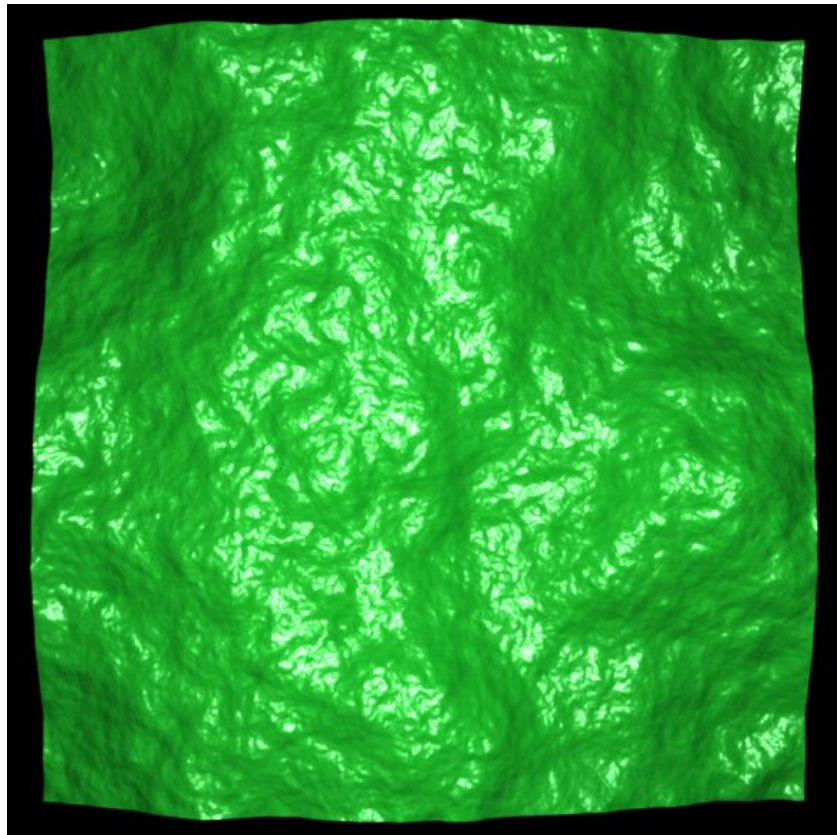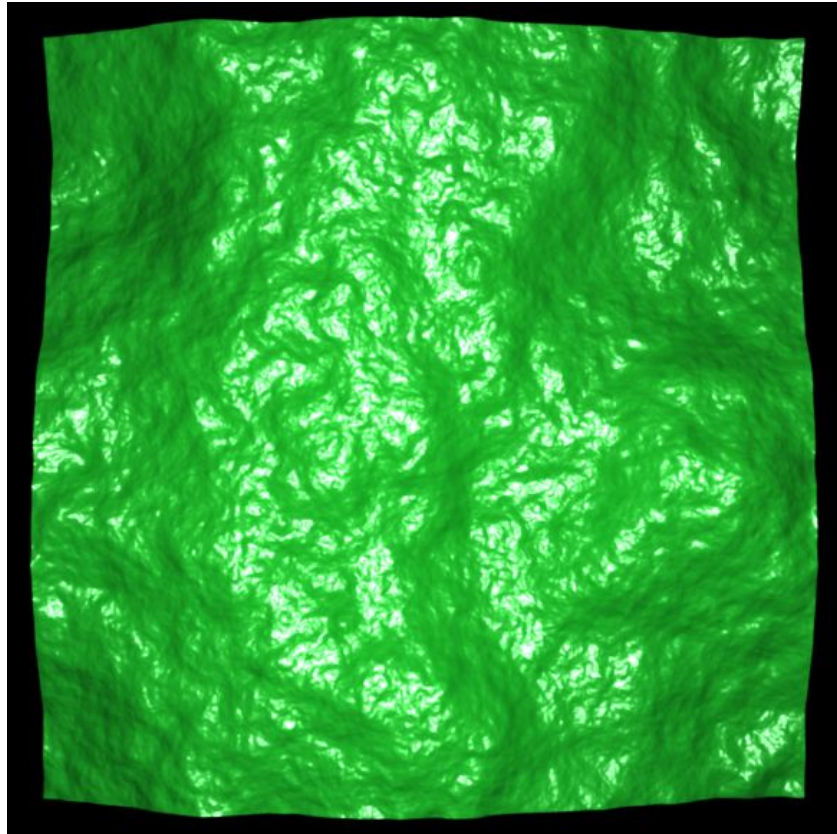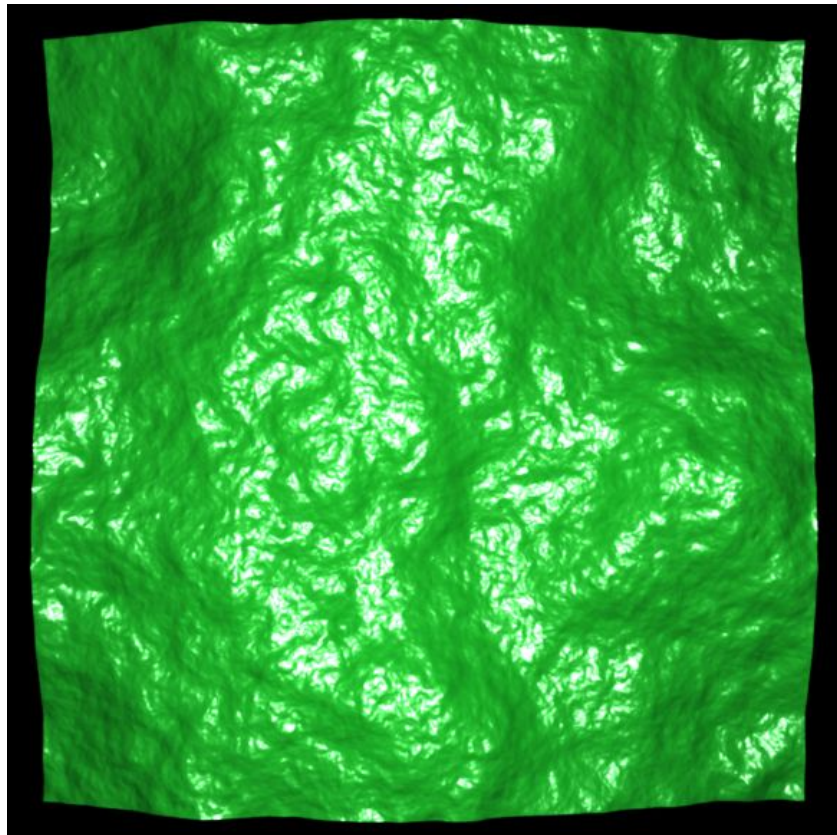

'mono22Set1-01.bmp'
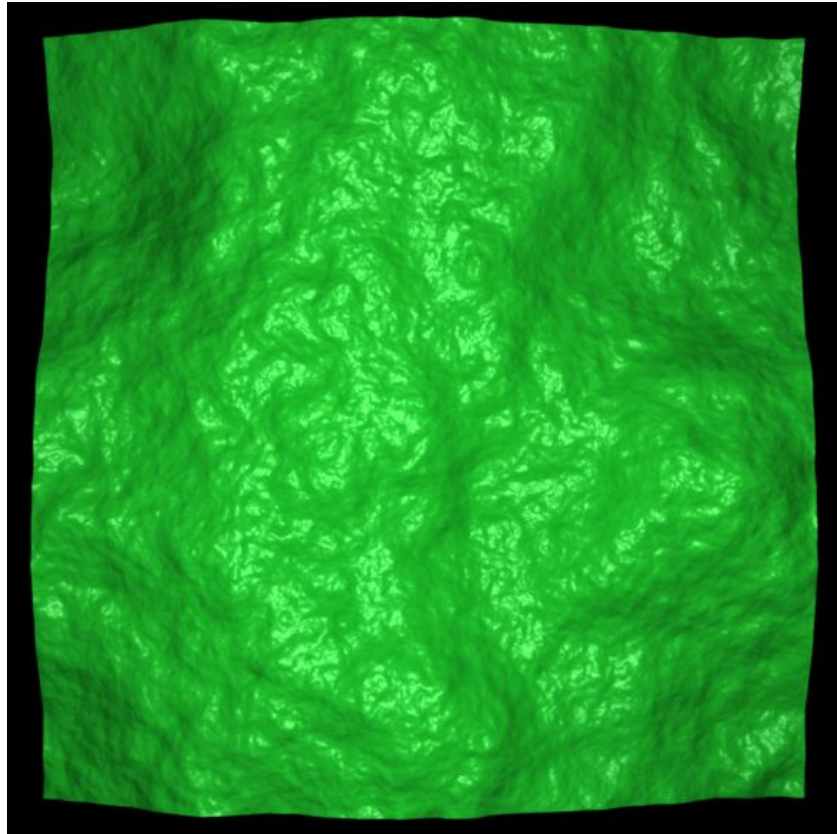
'mono22Set1-02.bmp'

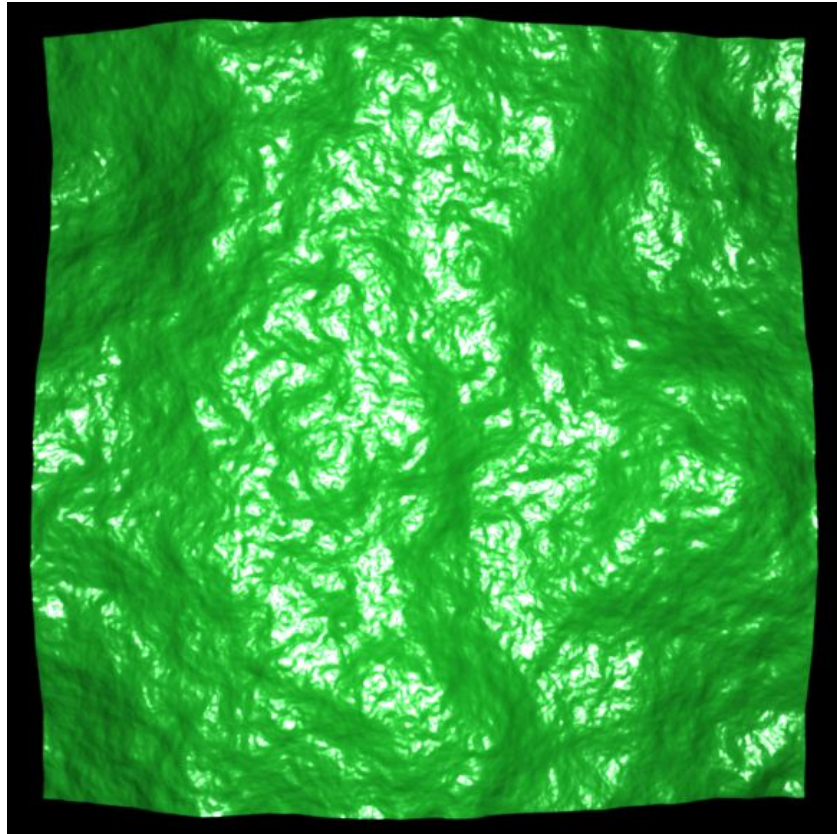
'mono22Set1-03.bmp'

149
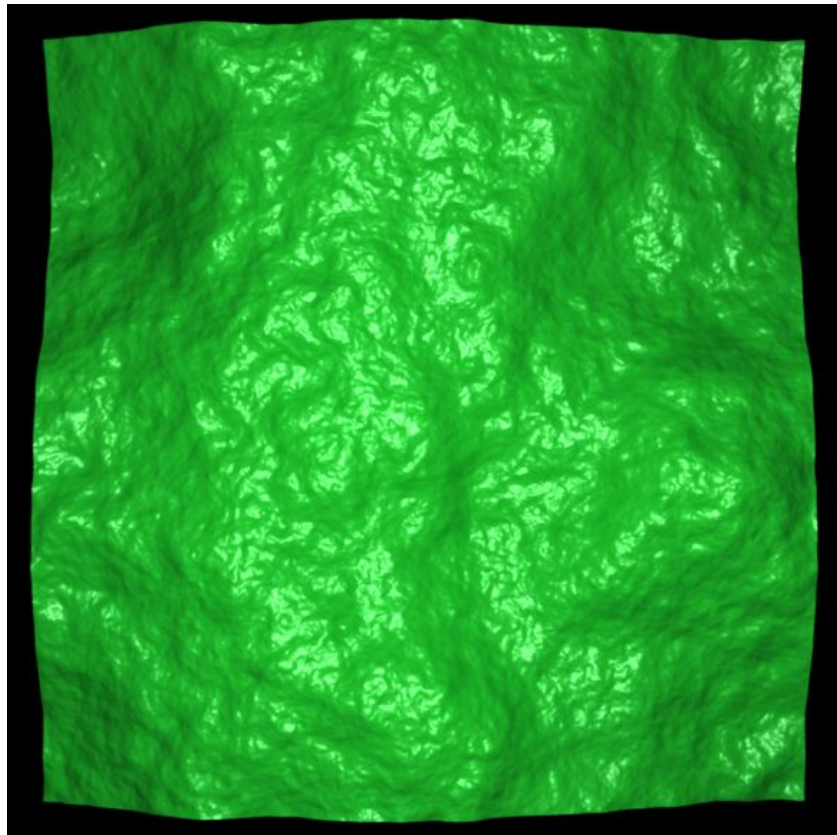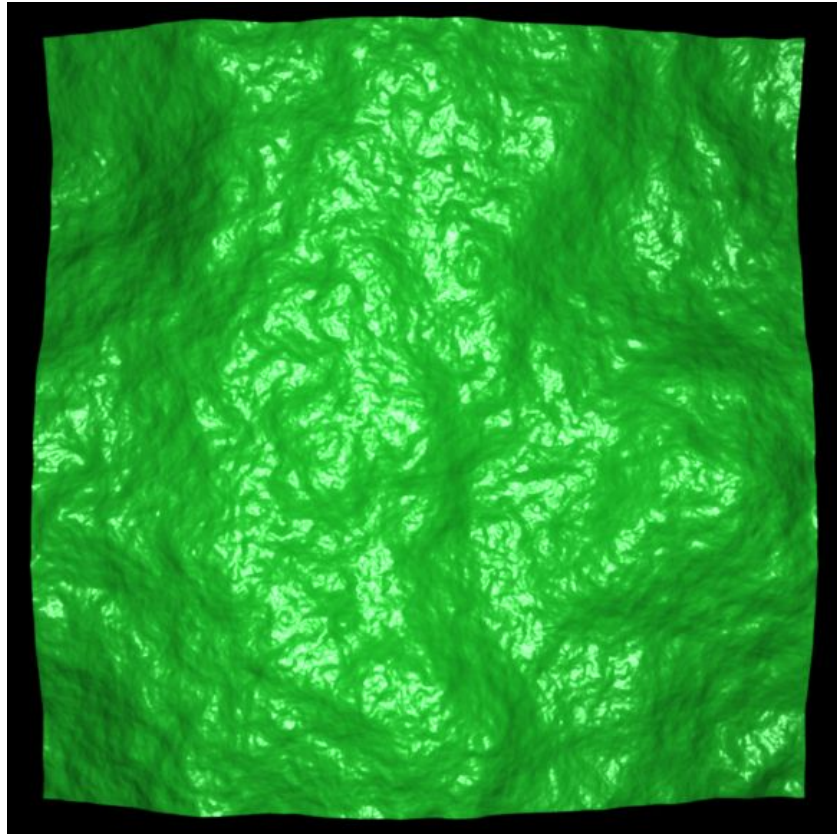
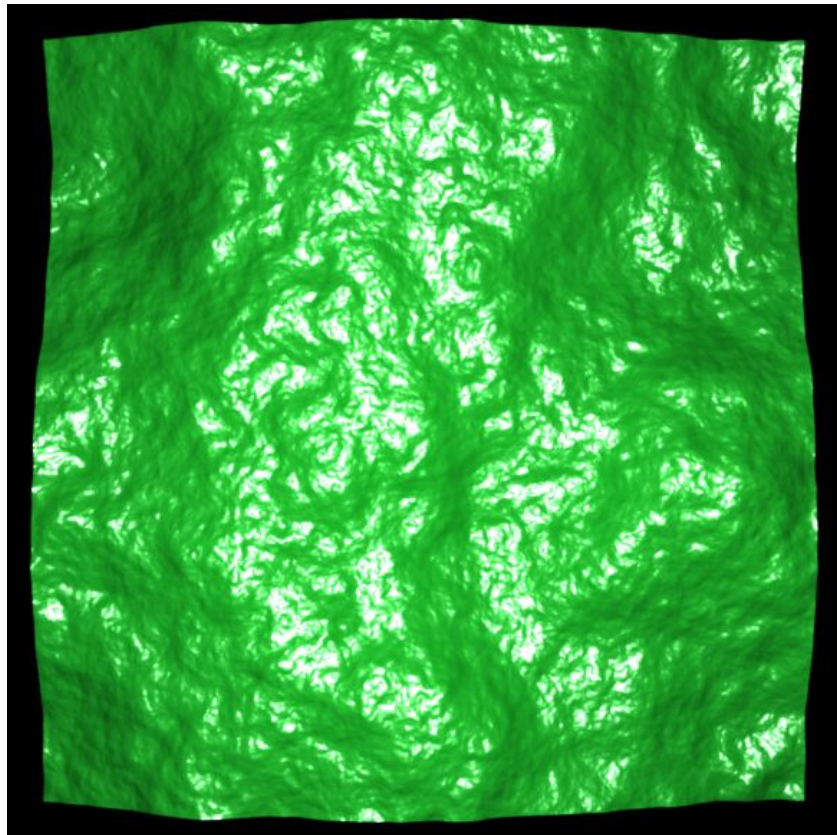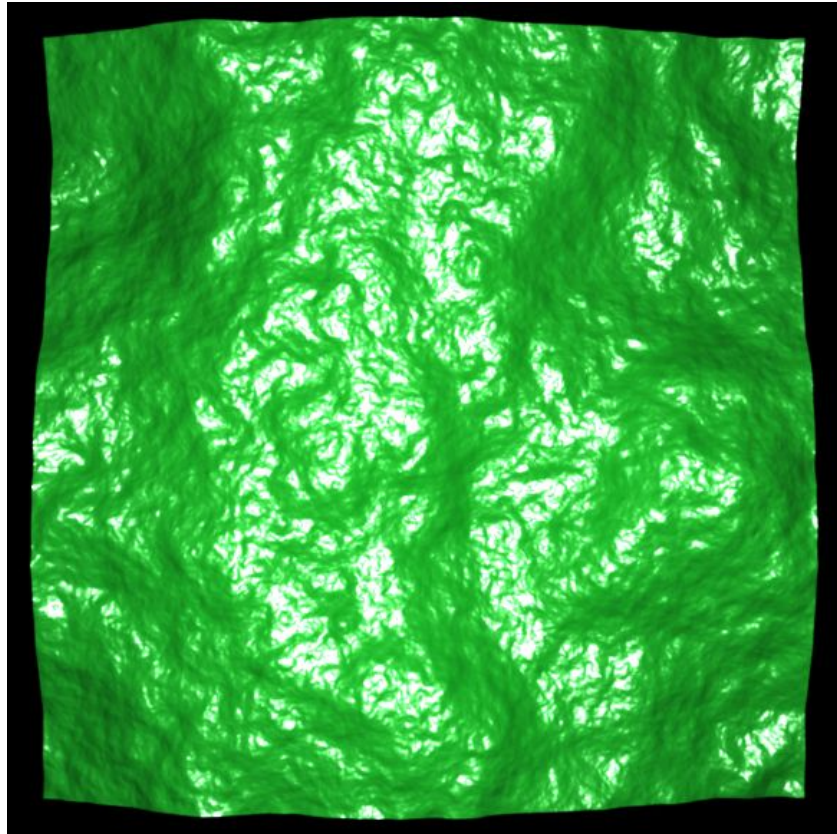'mono22Set1-04.bmp'

'mono22Set1-05.bmp'

150

'mono22Set1-06.bmp'
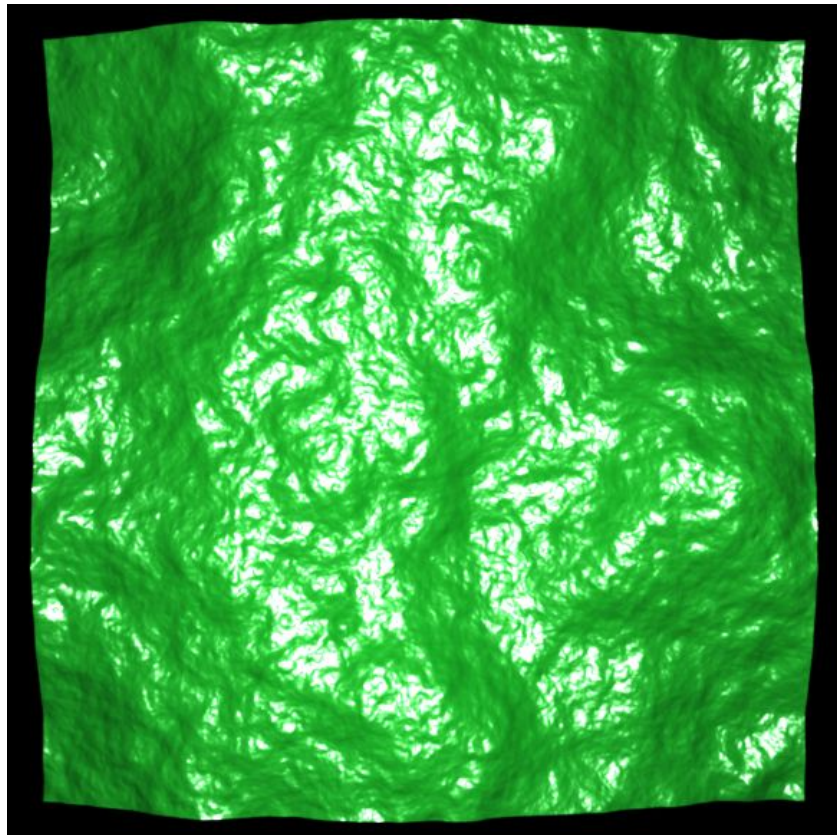

'mono22Set1-07.bmp'
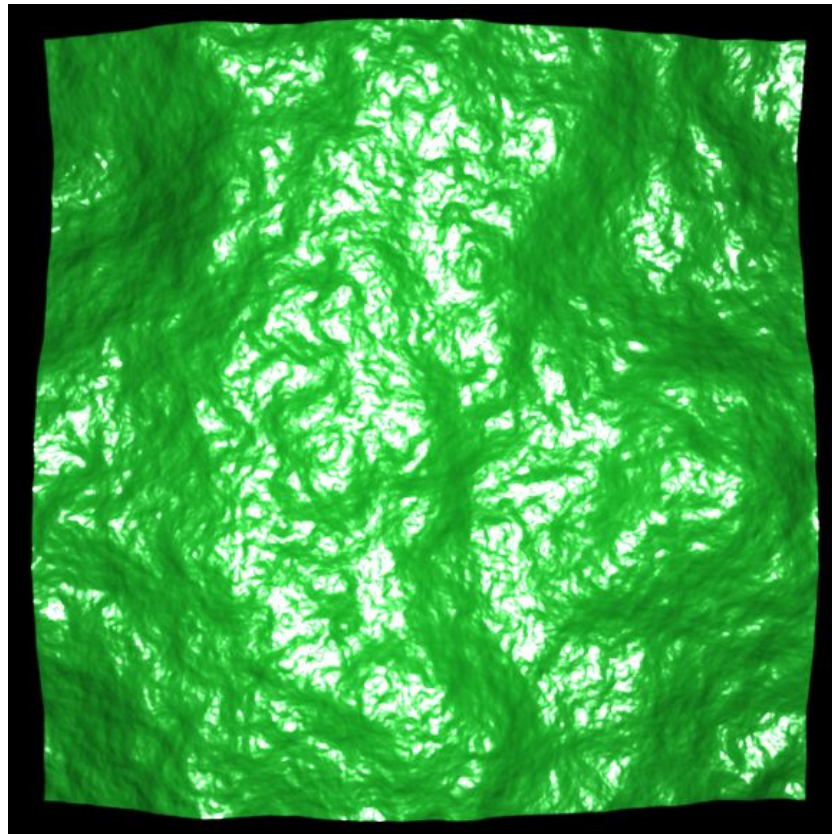
151

'mono22Set1-08.bmp'

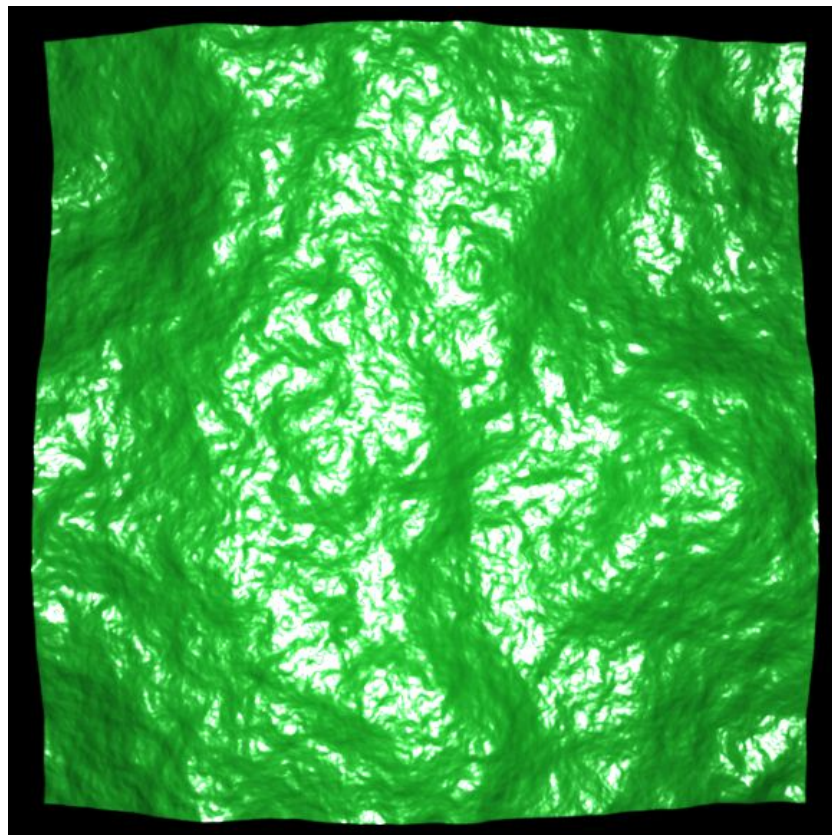'mono22Set1-09.bmp'

152

'mono22Set1-10.bmp'

'mono22Set1-11.bmp'

'mono22Set1-12.bmp'

'mono22Set1-13.bmp'

Table C.2 - An example full range of quantised gloss levels for a single surface. All the 13 possible gloss levels for a single stimulus (β = 2.2), labelled #9 in our numbering scheme.

Appendix D

Detailed Phase Results

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Pairs	Sphericity Assumed	.830	5	.166	9.316	.000	.509	46.580	1.000
	Greenhouse-Geisser	.830	2.758	.301	9.316	.000	.509	25.693	.987
	Huynh-Feldt	.830	4.098	.203	9.316	.000	.509	38.177	.999
	Lower-bound	.830	1.000	.830	9.316	.014	.509	9.316	.775
Error(Pairs)	Sphericity Assumed	.802	45	.018					
	Greenhouse-Geisser	.802	24.822	.032					
	Huynh-Feldt	.802	36.882	.022					
	Lower-bound	.802	9.000	.089					

a. Computed using alpha = .05

Table D.1 - Full within-subjects effects for β = 2.2 showing that phase does have a significant effect upon the perception of gloss. As we are assuming sphericity, the important row is the shaded one where $p < 0.001$ (Actual: 0.000004).

(I) Pairs	(J) Pairs	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
9 vs. 10	9 vs. 11	.165	.056	.251	-.058	.388
	9 vs. 12	.015	.061	1.000	-.228	.258
	10 vs. 11	.160	.078	1.000	-.149	.469
	10 vs. 12	.005	.088	1.000	-.344	.354
	11 vs. 12	-.185	.050	.074	-.383	.013
9 vs. 11	9 vs. 10	-.165	.056	.251	-.388	.058
	9 vs. 12	-.150	.039	.063	-.306	.006
	10 vs. 11	-.005	.047	1.000	-.192	.182
	10 vs. 12	-.160	.069	.670	-.431	.111
	11 vs. 12	-.350[*]	.056	.002	-.571	-.129
9 vs. 12	9 vs. 10	-.015	.061	1.000	-.258	.228
	9 vs. 11	.150	.039	.063	-.006	.306
	10 vs. 11	.145	.053	.343	-.064	.354
	10 vs. 12	-.010	.066	1.000	-.272	.252
	11 vs. 12	-.200	.055	.084	-.419	.019
10 vs. 11	9 vs. 10	-.160	.078	1.000	-.469	.149
	9 vs. 11	.005	.047	1.000	-.182	.192
	9 vs. 12	-.145	.053	.343	-.354	.064
	10 vs. 12	-.155	.042	.073	-.320	.010
	11 vs. 12	-.345[*]	.048	.001	-.535	-.155
10 vs. 12	9 vs. 10	-.005	.088	1.000	-.354	.344
	9 vs. 11	.160	.069	.670	-.111	.431
	9 vs. 12	.010	.066	1.000	-.252	.272
	10 vs. 11	.155	.042	.073	-.010	.320
	11 vs. 12	-.190	.065	.252	-.447	.067
11 vs. 12	9 vs. 10	.185	.050	.074	-.013	.383
	9 vs. 11	.350[*]	.056	.002	.129	.571
	9 vs. 12	.200	.055	.084	-.019	.419
	10 vs. 11	.345[*]	.048	.001	.155	.535
	10 vs. 12	.190	.065	.252	-.067	.447

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

*. The mean difference is significant at the .05 level.

Table D.2 - Full pairwise comparisons for β = 2.2, showing two significantly different pairs. The pairs are indicated with a shaded background and have been corrected using the Bonferroni correction method.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
Roughness	Sphericity Assumed	.336	3	.112	3.241	.038	.265	9.724	.679
	Greenhouse-Geisser	.336	2.077	.162	3.241	.060	.265	6.733	.555
	Huynh-Feldt	.336	2.711	.124	3.241	.043	.265	8.789	.644
Error Roughness	Sphericity Assumed	.934	27	.035					
	Greenhouse-Geisser	.934	18.694	.050					
	Huynh-Feldt	.934	24.403	.038					
	Lower-bound	.934	9.000	.104					
Pairs	Sphericity Assumed	.311	5	.062	2.964	.021	.248	14.821	.810
	Greenhouse-Geisser	.311	3.117	.100	2.964	.047	.248	9.240	.649
	Huynh-Feldt	.311	4.959	.063	2.964	.022	.248	14.698	.808
Error Pairs	Sphericity Assumed	.944	45	.021					
	Greenhouse-Geisser	.944	28.054	.034					
	Huynh-Feldt	.944	44.628	.021					
Roughness * Pairs	Sphericity Assumed	.982	15	.065	2.012	.019	.183	30.180	.948
	Greenhouse-Geisser	.982	4.610	.213	2.012	.102	.183	9.274	.591
	Huynh-Feldt	.982	10.043	.098	2.012	.041	.183	20.207	.854
Error Roughness* Pairs	Sphericity Assumed	4.392	135	.033					
	Greenhouse-Geisser	4.392	41.486	.106					
	Huynh-Feldt	4.392	90.390	.049					

a. Computed using alpha = .05

Table D.3 - Full within-subjects effects for β = 1.8, 2.0, 2.4 and 2.6 showing that phase does have a significant effect upon the perception of gloss. The shaded rows show which rows we are using when taking into account sphericity. For roughness we are using the Greenhouse-Geisser correction for degrees of freedom, for pair type we are assuming sphericity.

| (I) Pairs	(J) Pairs	Mean Difference (I-J)	Std. Error	Sig.a	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1	2	.069	.036	1.000	-.075	.213
	3	.081	.030	.371	-.038	.201
	4	.100	.037	.364	-.046	.246
	5	.106	.045	.623	-.071	.283
	6	.046	.024	1.000	-.050	.142
2	1	-.069	.036	1.000	-.213	.075
	3	.013	.029	1.000	-.102	.127
	4	.031	.021	1.000	-.051	.114
	5	.037	.036	1.000	-.103	.178
	6	-.023	.036	1.000	-.166	.120
3	1	-.081	.030	.371	-.201	.038
	2	-.013	.029	1.000	-.127	.102
	4	.019	.031	1.000	-.105	.142
	5	.025	.023	1.000	-.068	.118
	6	-.035	.028	1.000	-.146	.075
4	1	-.100	.037	.364	-.246	.046
	2	-.031	.021	1.000	-.114	.051
	3	-.019	.031	1.000	-.142	.105
	5	.006	.034	1.000	-.127	.139
	6	-.054	.033	1.000	-.184	.076
5	1	-.106	.045	.623	-.283	.071
	2	-.037	.036	1.000	-.178	.103
	3	-.025	.023	1.000	-.118	.068
	4	-.006	.034	1.000	-.139	.127
	6	-.060	.034	1.000	-.196	.075
6	1	-.046	.024	1.000	-.142	.050
	2	.023	.036	1.000	-.120	.166
	3	.035	.028	1.000	-.075	.146
	4	.054	.033	1.000	-.076	.184
	5	.060	.034	1.000	-.075	.196

Based on estimated marginal means

a. Adjustment for multiple comparisons: Bonferroni.

Table D.4 - Full pairwise comparisons for β = 1.8, 2.0, 2.4 and 2.6. This shows there are no significantly different pairs when the confidence interval has been altered using the Bonferroni correction method.
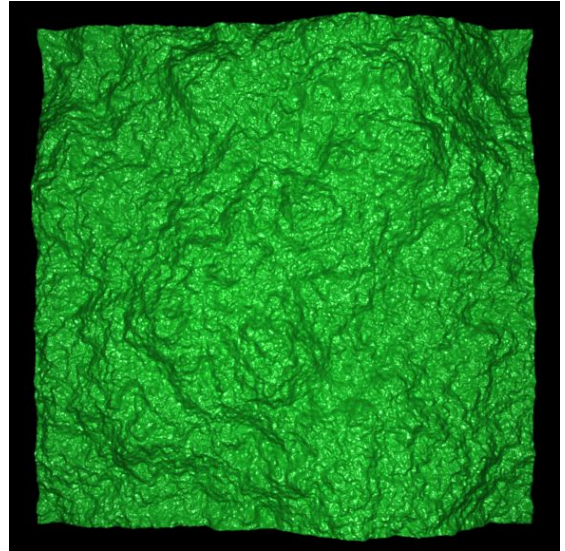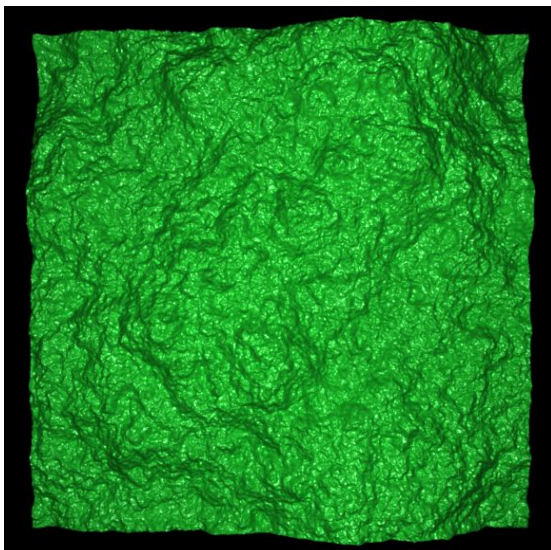
Appendix E

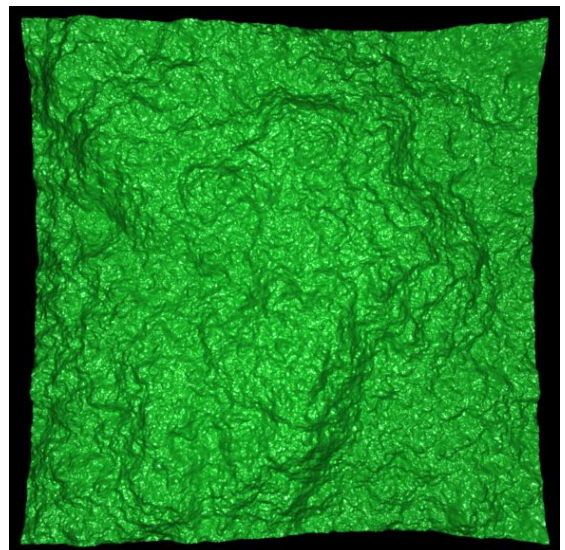## Similar Gloss Surface Pairs

$\beta = 1.8$
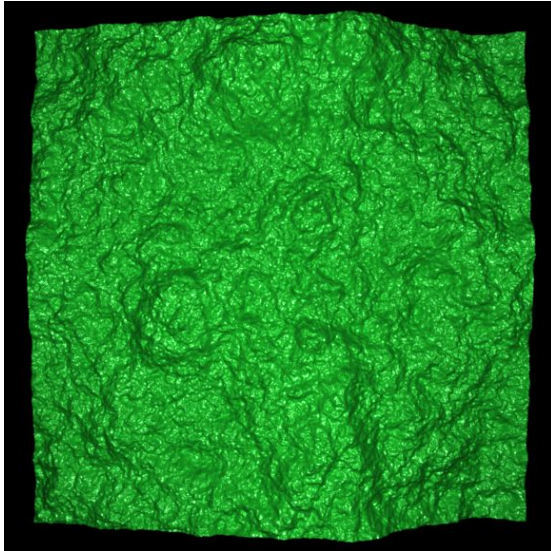
'mono18Set1-07.bmp'

'mono18Set2-07.bmp'
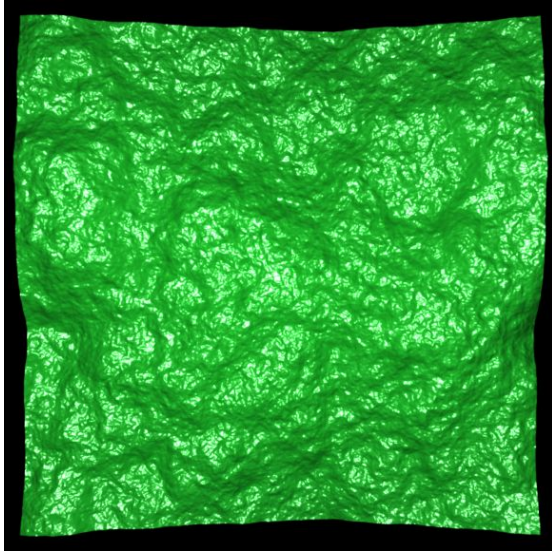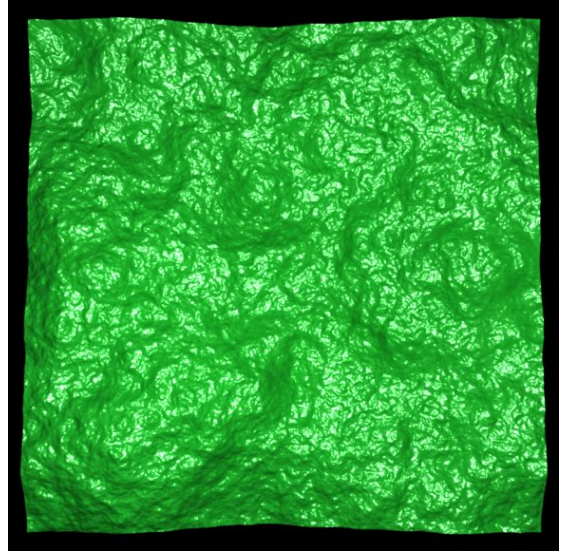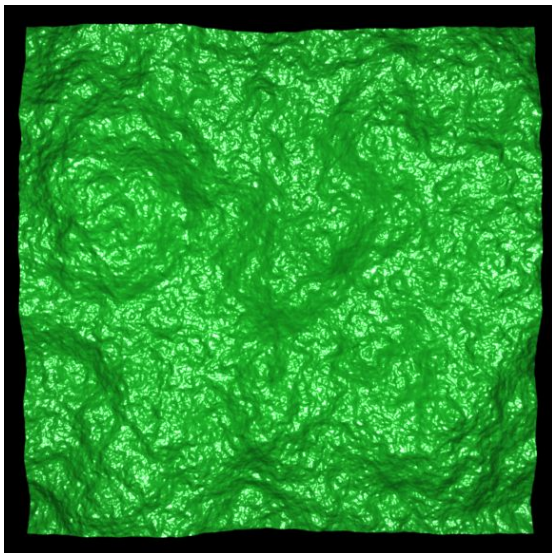

'mono18Set2-07.bmp'

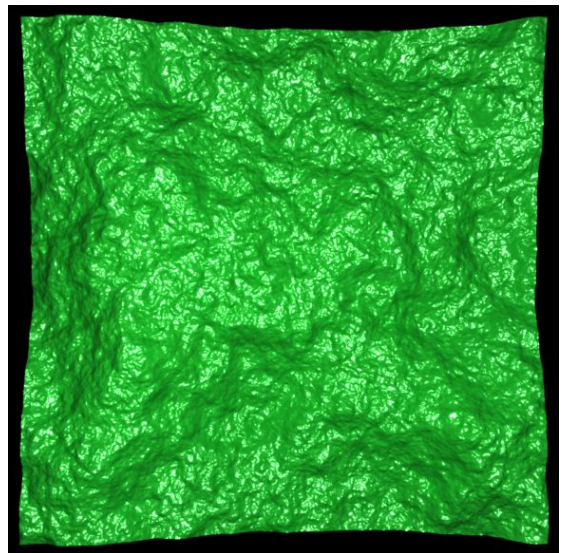
'mono18Set3-07.bmp'
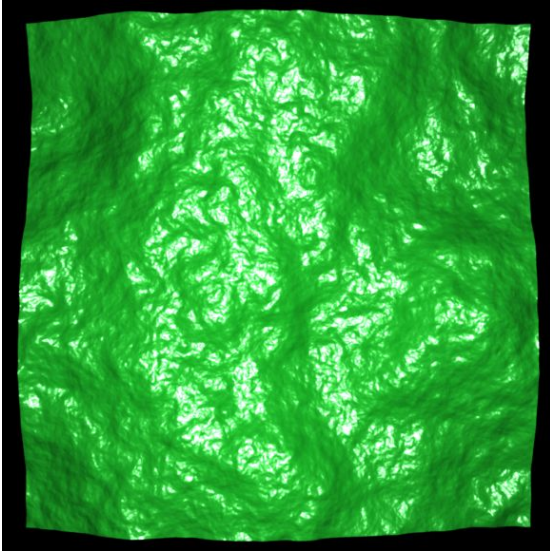
$\beta = 2.0$

'mono20Set1-07.bmp'

'mono20Set4-07.bmp'

'mono20Set2-07.bmp'

'mono20Set3-07.bmp'

$\beta = 2.2$

'mono22Set1-07.bmp'
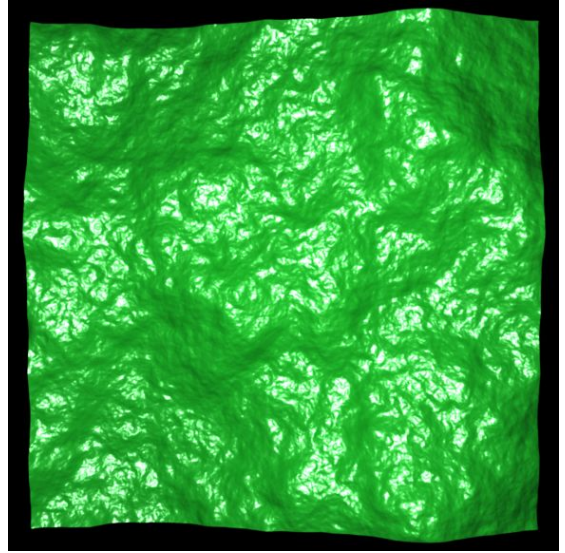
'mono22Set2-07.bmp'
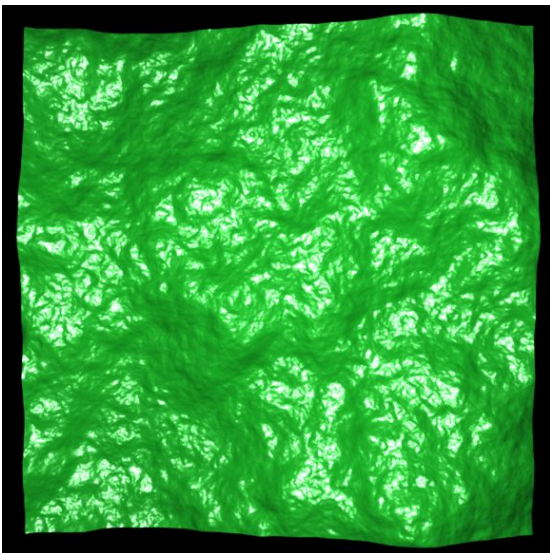
'mono22Set2-07.bmp'
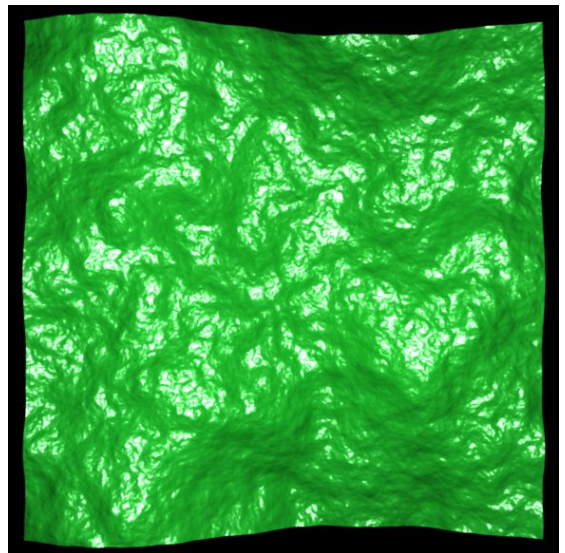
'mono22Set4-07.bmp'

$\beta = 2.4$
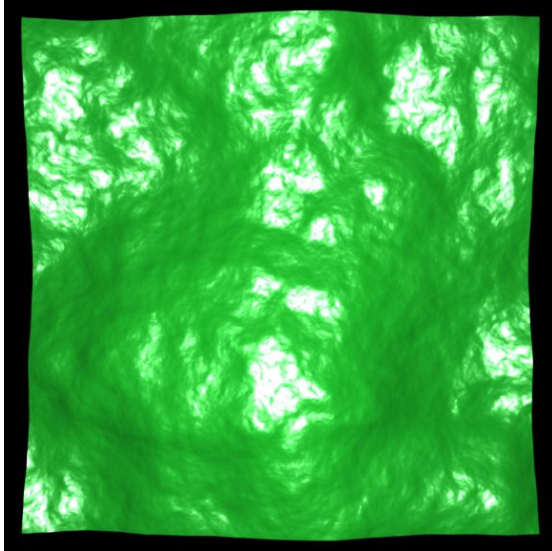
'mono24Set1-07.bmp'
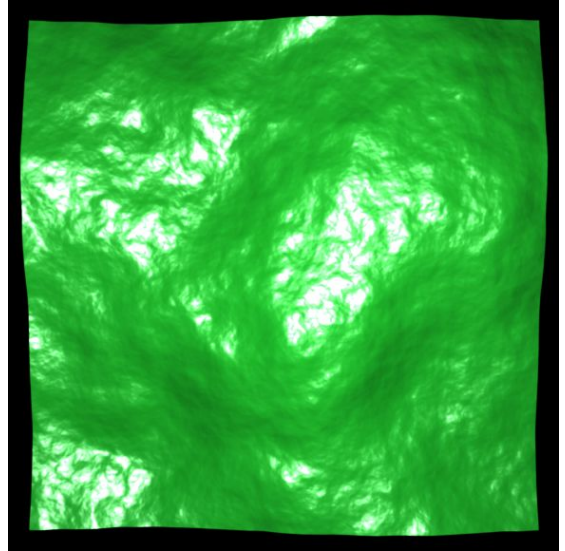
'mono24Set4-07.bmp'
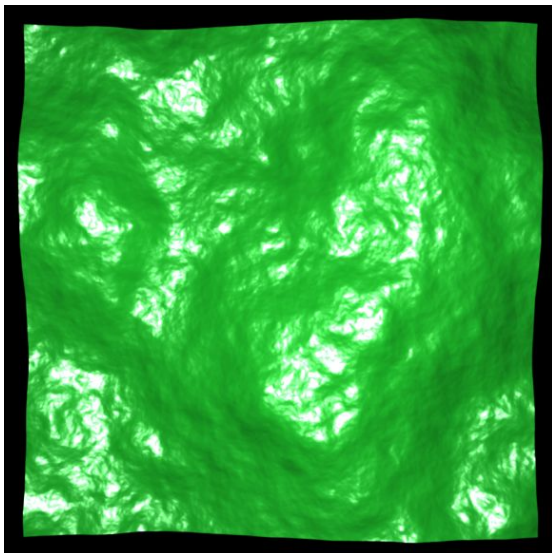
'mono24Set2-07.bmp'
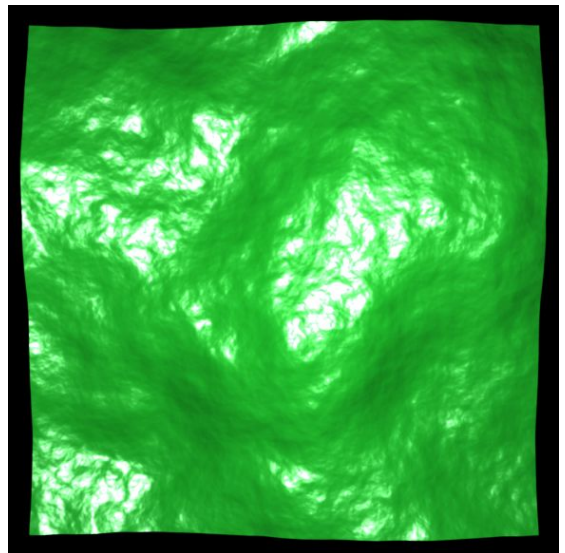
'mono24Set4-07.bmp'

$\beta = 2.6$
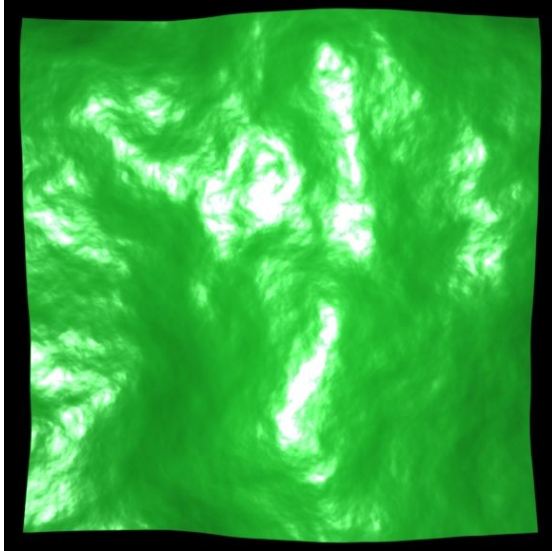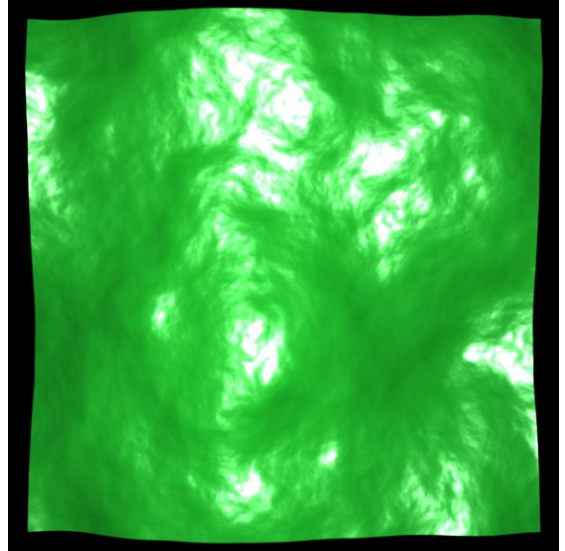
'mono26Set2-07.bmp'   'mono26Set4-07.bmp'
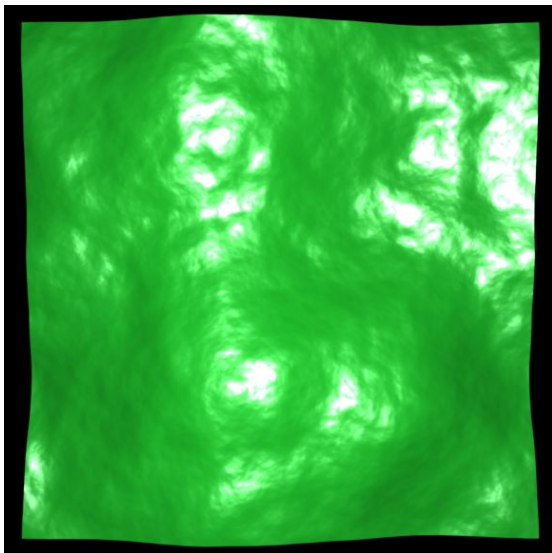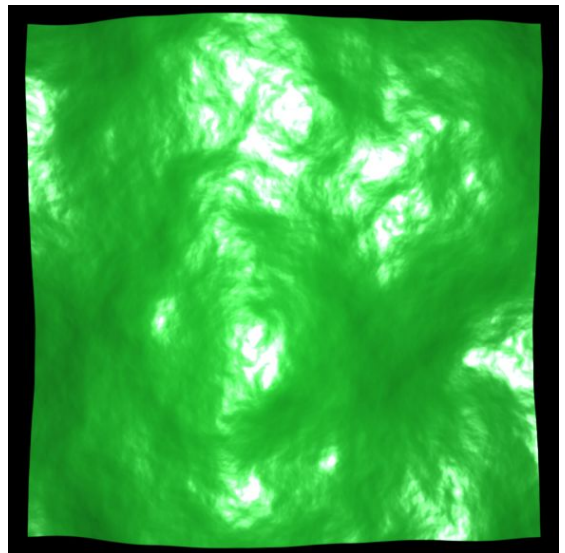
'mono26Set1-07.bmp'   'mono26Set4-07.bmp'

Table E.1 - Selected pairs of perceptually similar glossy surfaces. Each page shows the two pairs of surfaces chosen for each roughness level. N.B. The pairs are represented by rows, so only surfaces on the same row were compared to each other in *Chapter 8*. Only gloss level 7 is shown here, in the experiments to select these pairs, gloss level 6 was also used.

Appendix F

Detailed Disparity Results

(a) Participant	Target Gloss 4				
	$\beta = 1.8$	$\beta = 2$	$\beta = 2.2$	$\beta = 2.4$	$\beta = 2.6$
1	0.25	-0.5	-0.5	0.25	-0.25
2	-1.125	-0.25	0	-0.625	0
3	-0.5	-0.25	-0.25	0	-0.25
4	-0.875	0.75	0	0	0.125
5	-0.25	0.375	0	0	-0.375
6	-0.75	0	-0.5	-0.5	-0.5
7	-1.25	-0.125	0.375	-0.625	-0.125
8	-0.25	-0.125	0	0.125	0
9	-0.875	0.125	0	0.125	-0.25
10	0	-0.125	-0.125	0.25	0
11	-0.5	-0.125	-0.25	0	0.125
12	-0.125	-1	-0.25	0.625	0.5
13	0.25	-0.375	0.25	-0.625	0.125
14	-0.625	-0.5	-0.5	-0.25	-0.625
15	0	-0.25	-0.75	0	0
16	0	0.25	-0.125	-0.25	-0.25
17	0.375	-0.125	0	-0.5	-1
18	-0.625	0.125	0	-0.125	0.375
19	-1.375	-1	-2	1.625	-1.5
20	0	0	-0.375	0	0.5
21	-0.75	-0.125	-0.125	0.125	0.125
22	-0.25	0.125	-0.25	-0.125	-0.125
23	-0.625	-0.125	-0.25	0.125	-0.125
24	-0.625	-0.625	-0.75	-0.5	0.125
25	0.125	-0.375	-0.375	0	0
26	0.25	-0.25	0	-0.125	-0.375
27	-1	0.25	-0.25	0.25	0.125
28	-0.375	0.125	0.125	-0.125	-0.25

(b) Participant	Target Gloss 7				
	$\beta = 1.8$	$\beta = 2$	$\beta = 2.2$	$\beta = 2.4$	$\beta = 2.6$
1	-0.875	0	0.875	-0.125	-0.875
2	-0.75	-1.125	-0.25	-0.25	1.375
3	-0.125	-0.375	-0.375	-0.625	0.25
4	-1.125	0.25	-1	0.25	-0.25
5	0.125	0.25	-0.375	-0.5	-0.25
6	-0.5	-0.75	-0.875	-0.75	0.5
7	-0.625	-0.25	0	0.75	0.875
8	0.25	-0.125	0.25	0.375	0.375
9	-1.375	0.25	-0.375	-0.25	0.375
10	-1	-0.25	0.375	-0.375	0.125
11	0.625	-0.5	-0.125	0.875	-1
12	0.75	0.25	-0.375	0.625	1.5
13	0.25	0	1.125	-0.125	0.125
14	-0.125	0.75	-0.125	0.875	1.5
15	-0.625	0.375	-0.625	-2.125	0.125
16	1	-0.375	-0.375	-0.25	1.125
17	0	-0.625	0.5	0.5	-0.125
18	-1	-0.25	0.375	-0.125	0.125
19	-1.125	1.25	0.25	0.25	-1
20	-1.75	0	-0.75	-0.125	-0.75
21	0	-0.75	0.125	0.25	0.375
22	-0.625	-0.5	-0.25	0.75	-1.25
23	0.125	0.375	0.25	0.75	-0.125
24	-0.875	0.125	-0.5	-0.5	-0.125
25	-0.625	0.25	-0.5	-1	-0.25
26	0	-0.375	0	-0.875	-0.75
27	0.75	-0.5	-0.375	1.625	0.25
28	-0.25	-0.25	-0.25	0.25	0.125

(c) Participant	Target Gloss 10				
	$\beta = 1.8$	$\beta = 2$	$\beta = 2.2$	$\beta = 2.4$	$\beta = 2.6$
1	-0.5	0.75	-0.5	0.75	-0.125
2	-1.125	0.375	0	-0.875	1.125
3	-1.125	-0.625	-0.125	-0.25	0.25
4	-0.5	-0.625	-0.75	0.25	0.125
5	-0.125	0	0.125	-0.625	-0.5
6	-0.75	-0.625	-0.25	0.375	0.25
7	0.75	-0.125	0.125	-1.125	-1.25
8	0.125	0	-0.375	0.375	-0.375
9	0.375	0.125	0.5	-0.5	-0.25
10	-0.5	0	0.125	0.75	0.875
11	-0.125	-1.125	-0.25	0.625	-0.5
12	-1.5	-1	0.75	-0.25	0.75
13	-0.875	-0.625	0.125	-0.875	-0.875
14	0	0	-0.5	0	0.375
15	0.125	0.625	-1.25	-0.25	0.125
16	-1.25	0.625	0.125	-1.125	-0.125
17	0	0.875	-0.25	0.625	-1
18	0.125	0.125	-0.875	0.375	-0.25
19	-0.25	-0.875	-0.625	-3.375	-1.375
20	-1.625	-0.875	-0.375	1.25	-0.75
21	-0.5	0.5	-0.375	-0.75	1.25
22	-0.125	-0.875	-0.625	-0.5	0.75
23	-0.125	0.375	-0.625	0.75	-0.125
24	0	-0.625	0.125	-0.5	-0.5
25	-0.375	0.125	0.125	0.625	-0.25
26	-0.375	0	-0.5	-1.5	-0.625
27	1.5	0.375	-1.25	0.25	-1.75
28	-1.125	-1.5	-0.625	0.125	-0.75

Table F.1 - The raw data from the highlight disparity experiment. Table *(a)*, *(b)* and *(c)* show the results for the three different target gloss levels used in the experiment. N.B. the results are in 'quantised gloss levels' and are an arithmetic mean of 8 different judgements for each data point.

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared	Noncent. Parameter	Observed Power[a]
TargetGloss	Sphericity Assumed	1.520	2	.760	2.071	.136	.071	4.141	.408
	Greenhouse-Geisser	1.520	1.713	.887	2.071	.144	.071	3.548	.375
	Huynh-Feldt	1.520	1.818	.836	2.071	.141	.071	3.765	.387
Error (TargetGloss)	Sphericity Assumed	19.823	54	.367					
	Greenhouse-Geisser	19.823	46.262	.429					
	Huynh-Feldt	19.823	49.091	.404					
Roughness	Sphericity Assumed	4.958	4	1.240	3.254	.015	.108	13.017	.819
	Greenhouse-Geisser	4.958	3.221	1.540	3.254	.023	.108	10.481	.750
	Huynh-Feldt	4.958	3.708	1.337	3.254	.017	.108	12.067	.796
Error (Roughness)	Sphericity Assumed	41.136	108	.381					
	Greenhouse-Geisser	41.136	86.954	.473					
	Huynh-Feldt	41.136	100.116	.411					
TargetGloss * Roughness	Sphericity Assumed	.882	8	.110	.317	.959	.012	2.534	.153
	Greenhouse-Geisser	.882	5.261	.168	.317	.910	.012	1.666	.130
	Huynh-Feldt	.882	6.683	.132	.317	.941	.012	2.117	.142
Error (TargetGloss * Roughness)	Sphericity Assumed	75.181	216	.348					
	Greenhouse-Geisser	75.181	142.035	.529					
	Huynh-Feldt	75.181	180.454	.417					

a. Computed using alpha = .05

Table F.2 - Full within-subjects effects for the highlight disparity experiment showing that roughness does have an effect on participant's accuracy matching gloss between stimuli with and without disparity. The shaded rows show which rows we are using when taking into account sphericity. For target gloss and roughness we are assuming sphericity, for the combined effect of both we are using the Greenhouse-Geisser correction.

(I) Roughness	(J) Roughness	Mean Difference (I-J)	Std. Error	Sig.[a]	95% Confidence Interval for Difference[a]	
					Lower Bound	Upper Bound
1.8	2.0	-.229*	.072	.035	-.448	-.010
	2.2	-.150	.078	.630	-.387	.087
	2.4	-.295*	.093	.038	-.579	-.010
	2.6	-.284	.117	.221	-.642	.074
2.0	1.8	.229*	.072	.035	.010	.448
	2.2	.079	.079	1.000	-.162	.319
	2.4	-.065	.093	1.000	-.351	.220
	2.6	-.055	.103	1.000	-.370	.260
2.2	1.8	.150	.078	.630	-.087	.387
	2.0	-.079	.079	1.000	-.319	.162
	2.4	-.144	.097	1.000	-.441	.152
	2.6	-.134	.092	1.000	-.414	.146
2.4	1.8	.295*	.093	.038	.010	.579
	2.0	.065	.093	1.000	-.220	.351
	2.2	.144	.097	1.000	-.152	.441
	2.6	.010	.117	1.000	-.349	.369
2.6	1.8	.284	.117	.221	-.074	.642
	2.0	.055	.103	1.000	-.260	.370
	2.2	.134	.092	1.000	-.146	.414
	2.4	-.010	.117	1.000	-.369	.349

*. The mean difference is significant at the .05 level.

a. Adjustment for multiple comparisons: Bonferroni.

Table F.3 - Full pairwise comparisons for β = 1.8, 2.0, 2.2, 2.4 and 2.6. This shows there are two significantly different pairs when the confidence interval has been altered using the Bonferroni correction method. These are shown with the shaded background.
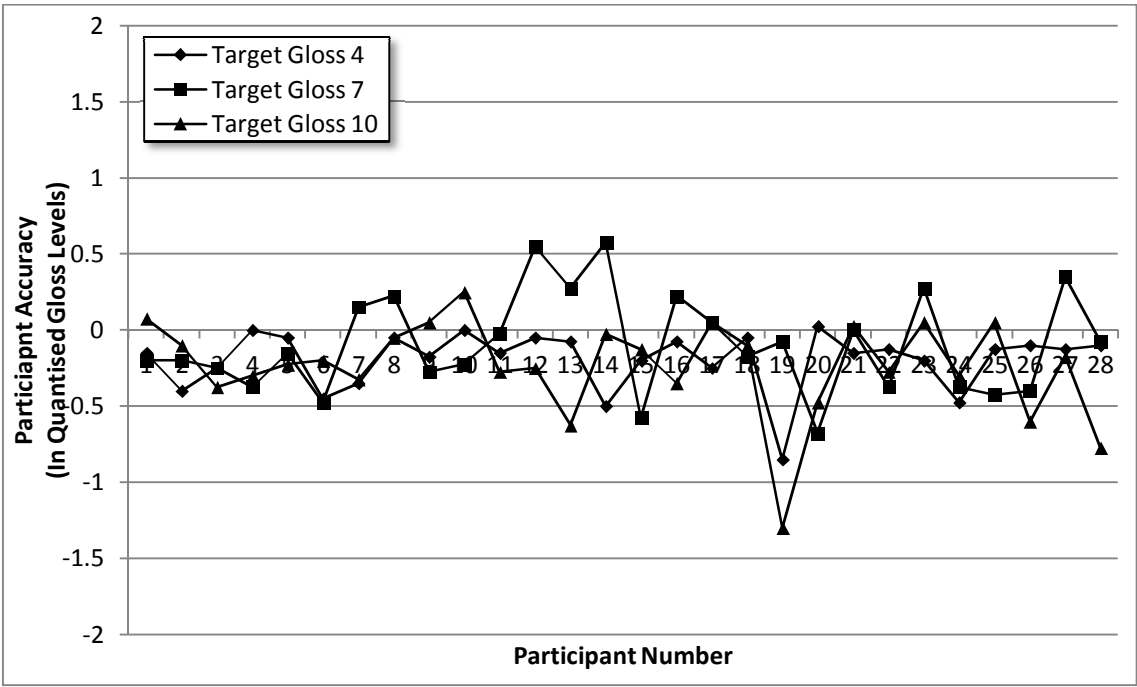
Figure F.1 – Participant accuracy data sorted by target gloss level. All 28 participants whose results have been shown, with the accuracy between the monocular and binocular conditions displayed in quantised gloss levels.
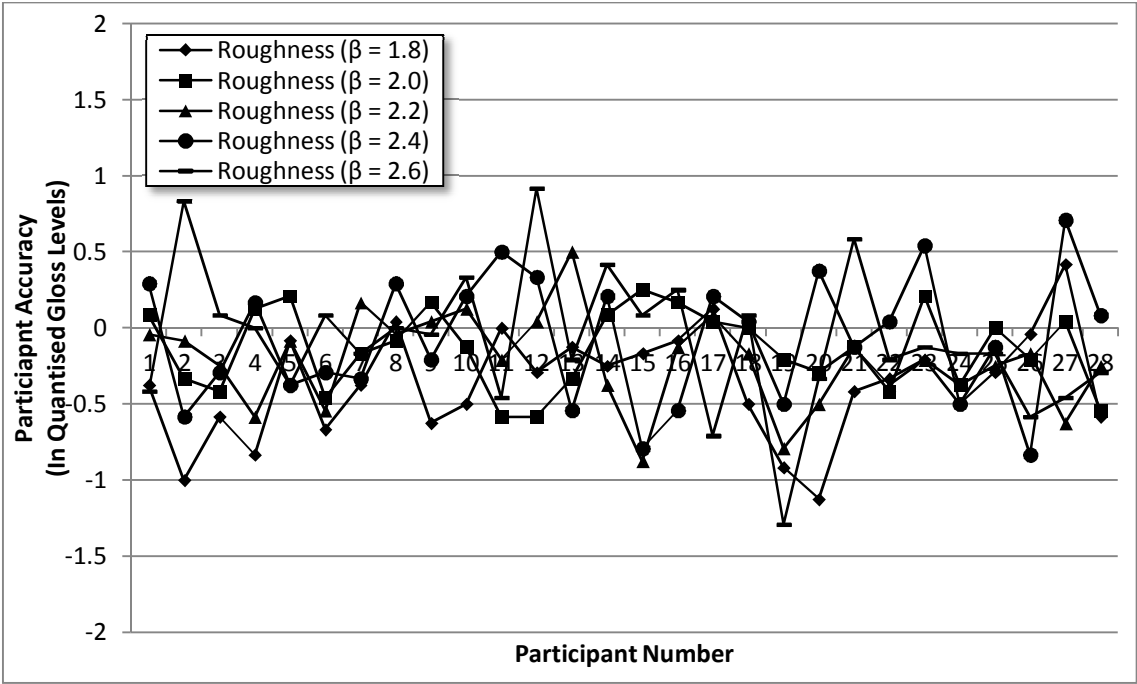

Figure F.2 - Participant accuracy data sorted by target roughness level. All 28 participants whose results were accepted have been shown, with the accuracy between the monocular and binocular conditions displayed in quantised gloss levels.

Bibliography

1.   Ho, Y.X., M.S. Landy, and L.T. Maloney, *How direction of illumination affects visually perceived surface roughness.* Journal of Vision, 2006. 6(5): p. 634-648.

2.   Ho, Y.X., L.T. Maloney, and M.S. Landy, *The effect of viewpoint on perceived visual roughness.* Journal of Vision, 2007. 7(1).

3.   Rao, A.R. and G.L. Lohse, *Identifying high level features of texture perception.* CVGIP: Graphical Models and Image Processing, 1993. 55(3): p. 218-233.

4.   Picard, D., et al., *Perceptual dimensions of tactile textures.* Acta Psychologica, 2003. 114(2): p. 165-184.

5.   Petrou, M., A. Talebpour, and A. Kadyrov, *Reverse engineering the way humans rank textures.* Pattern Analysis and Applications, 2007. 10(2): p. 101-114.

6.   Padilla, S., et al., *Perceived roughness of 1/f[beta] noise surfaces.* Vision Research, 2008. 48(17): p. 1791-1797.

7.   Padilla, S., *Mathematical Models for Perceived Roughness of Three-Dimensional Surface Textures*, 2008, Heriot-Watt University.

8.   Qi, L., et al., *How Mesoscale and Microscale Roughness Affect Perceived Gloss.* 2011. 2011.

9.   Blake, A. and H. Bulthoff, *Does the brain know the physics of specular reflection[quest].* Nature, 1990. 343(6254): p. 165-168.

10.   Wendt, G., et al., *Disparity, motion, and color information improve gloss constancy performance.* Journal of Vision, 2010. 10(9).

11.   Wendt, G., F. Faul, and R. Mausfeld, *Highlight disparity contributes to the authenticity and strength of perceived glossiness.* Journal of Vision, 2008. 8(1).

12.   Methven, T.S., et al. *A Comparison of Crowd-Sourcing vs. Traditional Techniques for Deriving Consumer Terms.* in *Digital Engagement 2011.* 2011.

13.   Methven, T.S. and M.J. Chantler, *Problems of Perceiving Gloss on Complex Surfaces*. 2012.

14.   Joh, A.S., et al., *Why walkers slip: Shine is not a reliable cue for slippery ground.* Perception & Psychophysics, 2006. 68(3): p. 339-352.

15.   Bajcsy, R., *Computer description of textured surfaces*, in *Proceedings of the 3rd international joint conference on Artificial intelligence*1973, Morgan Kaufmann Publishers Inc.: Stanford, USA. p. 572-579.

16.   Loh, H.H., J.G. Leu, and R.C. Luo, *The analysis of natural textures using run length features.* Industrial Electronics, IEEE Transactions on, 1988. 35(2): p. 323-328.

17.   Galloway, M.M., *Texture analysis using gray level run lengths.* Computer Graphics and Image Processing, 1975. 4(2): p. 172-179.

18.   Weszka, J.S., C.R. Dyer, and A. Rosenfeld, *A Comparative Study of Texture Measures for Terrain Classification.* Systems, Man and Cybernetics, IEEE Transactions on, 1976. SMC-6(4): p. 269-285.

19.   Haralick, R.M., Shanmuga.K, and I. Dinstein, *Textural Features for Image Classification.* Ieee Transactions on Systems Man and Cybernetics, 1973. Smc3(6): p. 610-621.

20.   Tamura, H., S. Mori, and T. Yamawaki, *Textural Features Corresponding to Visual-Perception.* Ieee Transactions on Systems Man and Cybernetics, 1978. 8(6): p. 460-473.

21.   Amadasun, M. and R. King, *Textural Features Corresponding to Textural Properties.* Ieee Transactions on Systems Man and Cybernetics, 1989. 19(5): p. 1264-1274.

22.   Brodatz, P., *Textures: a photographic album for artists and designers*. Vol. 66. 1966: Dover New York.

23.   Long, H. and W.K. Leow. *Perceptual texture space improves perceptual consistency of computational features*. in *INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE*. 2001. LAWRENCE ERLBAUM ASSOCIATES LTD.

24. Gurnsey, R. and D.J. Fleet, *Texture space.* Vision Research, 2001. 41(6): p. 745-757.

25. Vanrell, M., J. Vitria, and X. Roca, *A multidimensional scaling approach to explore the behavior of a texture perception algorithm.* Machine Vision and Applications, 1997. 9(5-6): p. 262-271.

26. Long, H.Z., C.W. Tan, and W.K. Leow, *Invariant and perceptually consistent texture mapping for content-based image retrieval.* 2001 International Conference on Image Processing, Vol Ii, Proceedings, 2001: p. 117-120.

27. Heaps, C. and S. Handel, *Similarity and features of natural textures.* Journal of Experimental Psychology-Human Perception and Performance, 1999. 25(2): p. 299-320.

28. Halley, F., *Perceptually Relevant Browsing Environments for Large Texture Databases*, in *School of Mathematical and Computer Sciences* 2012, Heriot-Watt University.

29. Emrith, K., *Perceptual Dimensions for Surface Texture Retrieval*, in *School of Mathematical and Computer Sciences* 2008, Heriot-Watt University.

30. Chantler, M.J., *Why Illuminant Direction Is Fundamental to Texture Analysis.* Iee Proceedings-Vision Image and Signal Processing, 1995. 142(4): p. 199-206.

31. Chantler, M.J., *The effect of variation in illuminant direction on texture classification*, 1994, Heriot-Watt University.

32. Picard, D., *Partial perceptual equivalence between vision and touch for texture information.* Acta Psychologica, 2006. 121(3): p. 227-248.

33. Atkinson, D. and D. Sensoria, *Private Communication interal to Digital Sensoria RCUK Project regarding initial word survey*, 2009.

34. Orzechowski, P.M., et al., *Interactivity to Enhance Perception: Does increased interactivity in mobile visual presentation tools facilitate more accurate rating of textile properties?* MobileHCI - In Print, 2011.

35.   Ding, C. and X.F. He, *Cluster merging and splitting in hierarchical clustering algorithms.* 2002 Ieee International Conference on Data Mining, Proceedings, 2002: p. 139-146.

36.   Gordon, A.D., *A Review of Hierarchical-Classification.* Journal of the Royal Statistical Society Series a-Statistics in Society, 1987. 150: p. 119-137.

37.   Lederman, S.J., J.M. Loomis, and D.A. Williams, *The role of vibration in the tactual perception of roughness.* Perception & Psychophysics, 1982. 32(2): p. 109-116.

38.   totallytextures Ltd. *SmarterSwatch*. Last Accessed: May 14th, 2013; Available from: http://www.totallytextures.com/products/SmarterSwatch/.

39.   Geomagic. *Phantom Omni® Haptic Device Specifications*. Last Accessed: May 14th, 2013; Available from: http://www.geomagic.com/en/products/phantom-omni/specifications/.

40.   Sakano, Y. and H. Ando, *Effects of head motion and stereo viewing on perceived glossiness.* Journal of Vision, 2010. 10(9).

41.   Wikipedia. *Nintendo Wii Sensor Bar (Licensed under Creative Commons Attribution-Share Alike 3.0 Unported)*. Last Accessed: 15th May, 2013; Available from: http://en.wikipedia.org/wiki/File:Nintendo_Wii_Sensor_Bar.jpg.

42.   Lee, J.C. *Head Tracking for Desktop VR Displays using the WiiRemote* Last Accessed: 15th May, 2013; Available from: http://youtu.be/Jd3-eiid-Uw.

43.   NVIDIA. *NVIDIA 3D Vision*. Last Accessed: 16th May, 2013; Available from: http://www.nvidia.com/object/3d-vision-main.html.

44.   Qi, L., et al., *How mesoscale and microscale roughness affect perceived gloss.* Perception, 2012. 41(3): p. 375-375.

45.   Hunter, R.S. and D.B. Judd, *Development of a method of classifying paints according to gloss.* ASTM Bulletin, 1939. 97: p. 11.

46.   Billmeyer Jr, F.W. and F.X.D. O'Donnell, *Visual gloss scaling and multidimensional scaling analysis of painted specimens.* Color Research & Application, 1987. 12(6): p. 315-326.

47.   Harrison, V. and S. Poulter, *Gloss measurement of papers-the effect of luminance factor.* British Journal of Applied Physics, 1951. 2: p. 92.

48.   Qi, L., *Measuring Perceived Gloss of Rough Surfaces*, in *School of Mathematical and Computer Sciences*2012: Heriot-Watt University.

49.   Leloup, F.B., et al., *Luminance-based specular gloss characterization.* Journal of the Optical Society of America a-Optics Image Science and Vision, 2011. 28(6): p. 1322-1330.

50.   Ji, W., et al., *Gloss as an aspect of the measurement of appearance.* Journal of the Optical Society of America a-Optics Image Science and Vision, 2006. 23(1): p. 22-33.

51.   Obein, G., K. Knoblauch, and F. Vienot, *Difference scaling of gloss: Nonlinearity, binocularity, and constancy.* Journal of Vision, 2004. 4(9): p. 711-720.

52.   O'Donnell, F.X.D. and F.W. Billmeyer, Jr., *Psychometric Scaling of Gloss.* Review and Evaluation of Appearance: Methods and Techniques pp. 14 - 32, ed. J.J. Rennilson and W. N. Hale. Vol. ASTM STP 914. 1986, Philadelphia: American Society for Testing and Materials.

53.   ASTM International, *D523 - 08: Standard Test Method for Specular Gloss*, 2008.

54.   International Organization for Standardization, *ISO 2813: Paints and varnishes - Determination of specular gloss of non-metallic paint films at 20 degrees, 60 degrees and 85 degrees*, 1994.

55.   Deutsches Institut für Normung e. V., *DIN 67530: Reflectometer as a means for gloss assessment of plane surfaces of paint coatings and plastics*, 1982.

56.   Hunter, R. and R.W. Harold, *The measurement of appearance.* 1987: Wiley-Interscience.

57.   Perlin, K. *A Unified Texture/Reflectance Model, Course on Advanced Image Synthesis*. in *ACM SIGGRAPH* 1984.

58.   Perlin, K., *An image synthesizer*, in *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*1985, ACM. p. 287-296.

59.   Wikipedia. *Perlin Noise (Left Image: Creative Commons Attribution-Share Alike 3.0 Unported, Right Image: Public Domain)*. Last Accessed: 23rd April, 2013; Available from: http://en.wikipedia.org/wiki/Perlin_noise.

60.   Perlin, K., *Improving noise.* Acm Transactions on Graphics, 2002. 21(3): p. 681-682.

61.   Perlin, K. and E.M. Hoffert, *Hypertexture*, in *Proceedings of the 16th annual conference on Computer graphics and interactive techniques*1989, ACM. p. 253-262.

62.   Ramanarayanan, G., et al., *Visual equivalence: Towards a new standard for image fidelity.* Acm Transactions on Graphics, 2007. 26(3).

63.   Perlin, K., *Chapter 2: Noise Hardware*, in *Real-Time Shading Languages SIGGRAPH Course Notes*. 2002, Marc Olano: ACM SIGGRAPH.

64.   Turing, A.M., *The chemical basis of morphogenesis.* Philosophical Transactions of the Royal Society of London, 1952. Series B, No. 641, Vol. 237, 14th August 1952: p. 37 - 72.

65.   Meinhardt, H. and H. Meinhardt, *Models of biological pattern formation*. Vol. 6. 1982: Academic Press London.

66.   Bard, J.B.L., *A model for generating aspects of zebra and other mammalian coat patterns.* Journal of Theoretical Biology, 1981. 93(2): p. 363-385.

67.   Turk, G., *Generating Textures on Arbitrary Surfaces Using Reaction-Diffusion.* Siggraph 91 Conference Proceedings, 1991. 25: p. 289-298.

68.   Meinhardt, H., *Models for positional signalling, the threefold subdivision of segments and the pigmentation pattern of molluscs.* Journal of embryology and experimental morphology, 1984. 83(Supplement): p. 289-311.

69.   Munafo, R.P. *Reaction-Diffusion by the Gray-Scott Model: Pearson's Parameterization*. 2013 [cited 2013 24th April].

70.   Clarke, A.D.F., *Modelling visual search for surface defects*, 2010, Department of Computer Science, School of Mathematics and Computer Science, Heriot-Watt University, Edinburgh.

71.   Liu, Y., W.-C. Lin, and J. Hays. *Near-regular texture analysis and manipulation*. in *ACM Transactions on Graphics (TOG)*. 2004. ACM.

72.   Liu, Y. and W.-C. Lin, *Deformable texture: the irregular-regular-irregular cycle*. 2003.

73.   Liu, Y., Y. Tsin, and W.-C. Lin, *The promise and perils of near-regular texture*. International Journal of Computer Vision, 2005. 62(1-2): p. 145-159.

74.   Nicoll, A., et al. *Fractional Fourier Texture Masks: Guiding Near-Regular Texture Synthesis*. in *Computer Graphics Forum*. 2005. Wiley Online Library.

75.   Ho, Y.X., M.S. Landy, and L.T. Maloney, *Conjoint measurement of gloss and surface texture*. Psychological Science, 2008. 19(2): p. 196-204.

76.   Mandelbrot, B.B., *The fractal geometry of nature/Revised and enlarged edition*. New York, WH Freeman and Co., 1983, 495 p., 1983. 1.

77.   Saupe, D., *Algorithms for random fractals*. The science of fractal images, 1988: p. 71-136.

78.   Brown, R., *A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies*. The Philosophical Magazine, or Annals of Chemistry, Mathematics, Astronomy, Natural History and General Science, 1828. 4(21): p. 161-173.

79.   van der Schaaf, A. and J.H. van Hateren, *Modelling the power spectra of natural images: statistics and information*. Vision Research, 1996. 36(17): p. 2759-2770.

80.   Wong, P., J. Howard, and J.S. Lin, *Surface roughening and the fractal nature of rocks*. Physical review letters, 1986. 57(5): p. 637-640.

81.   Shah, P., et al. *Perceived directionality of 1/f β noise surfaces*. in *Proceedings of the 5th symposium on Applied perception in graphics and visualization*. 2008. ACM.

82.   Padilla, S., et al., *Perceived roughness of 1/fβ noise surfaces*. Vision Research, 2008. 48(17): p. 1791-1797.

83.   Linnett, L., *Multi-texture image segmentation*. 1991.

84.   McGunnigle, G. and M.J. Chantler, *Comparison of three rough surface classifiers.* Iee Proceedings-Vision Image and Signal Processing, 2002. 149(5): p. 263-271.

85.   Sayles, R.S. and T.R. Thomas, *Surface topography as a nonstationary random process.* Nature, 1978. 271(5644): p. 431-434.

86.   Mulvaney, D.J., D.E. Newland, and K.F. Gill, *A complete description of surface texture profiles.* Wear, 1989. 132(1): p. 173-182.

87.   Ogilvy, J.A., *Theory of wave scattering from random rough surfaces.* 1991: Adam Hilger.

88.   McGunnigle, G., *The classification of textured surfaces under varying illuminant direction*, 1998, Heriot-Watt University.

89.   Fleming, R.W., R.O. Dror, and E.H. Adelson, *Real-world illumination and the perception of surface reflectance properties.* Journal of Vision, 2003. 3(5): p. 347-368.

90.   Kerrigan, I.S. and W.J. Adams, *Highlights, disparity, and perceived gloss with convex and concave surfaces.* Journal of Vision, 2013. 13(1).

91.   Nishida, S. and M. Shinya, *Use of image-based information in judgments of surface-reflectance properties.* Journal of the Optical Society of America a-Optics Image Science and Vision, 1998. 15(12): p. 2951-2965.

92.   Hartung, B. and D. Kersten, *Distinguishing shiny from matte.* Journal of Vision, 2002. 2(7): p. 551-551.

93.   Hartung, B. and D. Kersten, *How does the perception of shape interact with the perception of shiny material?* Journal of Vision, 2003. 3(9): p. 59-59.

94.   Muryy, A.A., R.W. Fleming, and A.E. Welchman, *Binocular cues for glossiness.* Journal of Vision, 2012. 12(9): p. 869-869.

95.   Phong, B.T., *Illumination for Computer Generated Pictures.* Communications of the Acm, 1975. 18(6): p. 311-317.

96.   Vangorp, P., J. Laurijssen, and P. Dutré. *The influence of shape on the perception of material reflectance.* in *ACM Transactions on Graphics (TOG).* 2007. ACM.

97.   Ward, G.J., *Measuring and Modeling Anisotropic Reflection.* Siggraph 92 : Conference Proceedings, 1992. 26: p. 265-272.

98.   Ward, G.J. *The RADIANCE lighting simulation and rendering system.* in *Proceedings of the 21st annual conference on Computer graphics and interactive techniques.* 1994. ACM.

99.   Pellacini, F., J.A. Ferwerda, and D.P. Greenberg. *Toward a psychophysically-based light reflection model for image synthesis.* in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques.* 2000. ACM Press/Addison-Wesley Publishing Co.

100.   Ferwerda, J.A., F. Pellacini, and D.P. Greenberg. *Psychophysically based model of surface gloss perception.* in *Photonics West 2001-Electronic Imaging.* 2001. International Society for Optics and Photonics.

101.   Veach, E. and L.J. Guibas. *Metropolis light transport.* in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques.* 1997. ACM Press/Addison-Wesley Publishing Co.

102.   Lambert, J.-H., *JH Lambert,... Photometria, sive de Mensura et gradibus luminis, colorum et umbrae.* 1760: sumptibus viduae E. Klett.

103.   Blinn, J.F., *Models of light reflection for computer synthesized pictures.* SIGGRAPH Comput. Graph., 1977. 11(2): p. 192-198.

104.   Ngan, A., F. Durand, and W. Matusik, *Experimental validation of analytical BRDF models*, in *ACM SIGGRAPH 2004 Sketches*2004, ACM: Los Angeles, California. p. 90.

105.   Matusik, W., *A data-driven reflectance model*, 2003, Citeseer.

106.   Cook, R.L. and K.E. Torrance, *A reflectance model for computer graphics.* SIGGRAPH Comput. Graph., 1981. 15(3): p. 307-316.

107.   Cook, R.L. and K.E. Torrance, *A Reflectance Model for Computer Graphics.* ACM Trans. Graph., 1982. 1(1): p. 7-24.

108.   Torrance, K.E. and E.M. Sparrow, *Theory for off-specular reflection from roughened surfaces.* JOSA, 1967. 57(9): p. 1105-1112.

109. He, X.D., et al. *A comprehensive physical model for light reflection*. in *ACM SIGGRAPH Computer Graphics*. 1991. ACM.

110. Schlick, C., *An Inexpensive BRDF Model for Physically-based Rendering*. Computer Graphics Forum, 1994. 13(3): p. 233-246.

111. Ashikhmin, M. and P. Shirley, *An anisotropic phong BRDF model*. Journal of graphics tools, 2000. 5(2): p. 25-32.

112. LuxRender. *LuxRender v0.7 - GPL Physically Based Renderer*. Released: July 9th, 2010; Available from: http://www.luxrender.net/.

113. Debevec, P., *Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography*, in *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*1998, ACM. p. 189-198.

114. Wheatstone, C., *Contributions to the physiology of vision.--Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision*. Philosophical transactions of the Royal Society of London, 1838. 128: p. 371-394.

115. ThorLabs. *ME2S-G01 - 2" Square Protected Aluminum Mirror, 3.2 mm Thick* Last Accessed: 10th April, 2013; Available from: http://www.thorlabs.de/thorProduct.cfm?partNumber=ME2S-G01.

116. ThorLabs. *Economy Front Surface Mirrors (See section entitled 'Square Protected Aluminum Mirrors')*. Last Accessed: 10th April, 2013; Available from: http://www.thorlabs.de/newgrouppage9.cfm?objectgroup_id=890.

117. NEC. *24" MultiSync PA241W Monitor - 24" Widescreen Professional Graphics Desktop Monitor*. Released: February 1st, 2010; Available from: http://www.necdisplay.com/p/desktop-monitors/pa241w-bk.

118. Blender Foundation. *Blender v2.49b - Blender is 3D creation for everyone, free to use for any purpose*. Released: September 16th, 2009; Available from: http://www.blender.org/.

119. Kelemen, C., et al. *A simple and robust mutation strategy for the metropolis light transport algorithm*. in *Computer Graphics Forum*. 2002. Wiley Online Library.

120.   Baker, S. *NEC PA241W Review*. TFT Central Written: April 8th, 2010, Last Accessed: 26th July, 2011; Available from: http://www.tftcentral.co.uk/reviews/nec_pa241w.htm.

121.   Roehrig, H., et al., *Why should you calibrate your display?* Penetrating Radiation Systems and Applications V, 2003. 5199: p. 181-192.

122.   Field, A.P., *Discovering statistics using SPSS : (and sex, drugs and rock 'n' roll)*. 2nd ed. ISM introducing statistical methods. 2005, London ; Thousand Oaks, Calif.: Sage Publications. xxxiv, 779 p.

123.   Field, A.P. and G. Hole, *How to design and report experiments*. 2003, London ; Thousand Oaks, Calif.: Sage publications Ltd. xii, 384 p.