# Brightness Perception of Surfaces with Mesoscale Structures

Michael Ludwig[1, a)] and Gary Meyer[1, b)]
*University of Minnesota*

Surface geometry can play an important role in our ability to understand and interpret material appearance and properties. This property ranges from large-scale shape changes impacting our identification of reflections to visible surface roughness affecting gloss perception. In this work we present a user study that examines numerous surface geometries that are defined at the mesoscale: small enough to be considered indicative of the material and not object geometry, but large enough to be visible from a distance with the naked eye. Models of perceived brightness were compared against sparsely collected brightness judgments from the study and used to densely compare many generated mesoscale surface patterns. Averaging incoming luminance over a spatially-varying surface proved effective at modeling brightness judgments. The effects of the mesoscale structure on perceived brightness were not directly correlated to parameters such as shape, size, or depth of the bumpy texture elements.

Keywords: Perceived brightness, Mesoscale structure, Spatially-varying appearance

## INTRODUCTION

In this work we examine numerous surface geometry patterns and how subjects perceive their brightness under different viewing and lighting conditions. The surface structures are generated at the mesoscale, a scale above the microscopic that is well described by microfacet distributions and below the macroscopic scale that fundamentally influences the overall shape of an object. Mesoscale surface patterns are interesting because they provide small, but visible, cues as to lighting and shading while still frequently being identified as part of a material (e.g. the ridges of stucco, coarseness of brick, and grains of certain woods). The patterns may be random, produced naturally, or follow very regular forms based on artistic design decisions. This last example is important because people can manufacture mesoscale patterns with much more ease than new paints or coatings, and such patterns are frequently employed in product design to create signature looks or improved haptic feel while maintaining appearance standards. Indeed our work is inspired by our interactions with the automotive industry where they design surface patterns at this scale for the interior of vehicles.

The mesoscale is somewhat subjective and dependent on the distance to the viewer. For example, when examined up close, concrete and other building materials can have visible non-smooth surfaces but at a distance it can be accurately described as a plane with a reflectance model such as Oren-Nayar's BRDF.[1] The rule of thumb that was applied to the surfaces generated in this study is that the mesoscale is just large enough to produce visible patterns from shading but not so large that it significantly alters the silhouette of the object. Unfortunately, the parameter space for surfaces patterns at this scale is infinite; besides the shape elements in a pattern, the spatial dimensions and resolution of the pattern can be increased infinitely. To help constrain the scope of this study, we restrict the surface geometries to a set of parameterizable generators that produce distinct pattern families. Additionally, we have chosen to evaluate the perceived brightness of these surfaces. There are many aspects of appearance, ranging from brightness, lightness, color, glossiness, texture, apparent tactile roughness, and material makeup that can be inferred by the visual system and are fundamental to daily interactions with the world.[2] We begin with brightness for its relative simplicity and because it acts as a good foundation before moving to more advanced appearance attributes.

In order to evaluate the effects of surface structure on brightness, two user studies were performed to collect brightness judgments on a subset of the generate surfaces. This data was compared against several hypothesized perceived brightness models, the most accurate of which turned out to be simply averaging incoming luminance. This model was then used to calculate similarities between all generated surface patterns and embed them in a metric space via multidimensional scaling. Several geometric variables of the surface patterns were then correlated against the embedding's dimensions to see if any were tied to brightness perception.

The next section describes background work examining brightness and lightness perception in humans, material perception, and other related work from the psychophysics and computer graphics communities. After that, the subsequent section describes the process for generating multiple families of parameterized mesoscale surfaces and the experimental setup for the two user studies. This is followed by a results section covering consistency and basic properties of the collected brightness judgments. Then, the analysis section compares the brightness models to the collected data to determine the most successful, and applies that model to understanding the geometric influence on perceived brightness. Lastly, we conclude with a summary of our work and a discussion on implications for future research into spatially-varying appearance.

[a)] Electronic mail: mludwig@cs.umn.edu
[b)] Electronic mail: meyer@cs.umn.edu

## BACKGROUND

Brightness has often been studied alongside lightness. In the fields of psychophysics and perception, brightness is defined as the perceived luminance of an object and lightness is the apparent reflectance of the object.[3] Within the color science field, these terms have slightly different meanings, where brightness is the attribute of visual sensation where a region exhibits more or less light and lightness is the perceived brightness relative to the brightness of a similarly illuminated "white".[4,5] The two definitions of brightness are compatible with one another, while the lightness definitions are not. However, since we are only concerned with brightness, there will be no need to provide subjects with a reference white point or make judgments of reflectance. Brightness and lightness have been studied in very synthetic scenarios with 2D elements arranged as concentric annuli,[6] complex rectangular patterns,[7] and designed to evoke depth relations.[8] Additionally, past research has focused on flat or smooth patterns when judging brightness and this work is the first to our knowledge to approach the spatially-varying problem.

We are interested in exploring brightness when the stimuli is a much more physically accurate simulation of a 3D surface. Research has shown that realistic lighting can have an impact on subjects' abilities to identify gloss.[9] Other work has shown that perceived shape and depth can affect our interpretation of color.[10] Given this, it is necessary to include the supporting scenery that assists in identifying lighting and shading from texture.

Elements of this study have been inspired by the computer graphics and psychophysics experiments into the perception of glossiness. This body of work has demonstrated the effectiveness of user studies comparing and matching rendered images,[11] and the impact that surface geometry variations can have on the perception of surface properties.[12–14] Although the discovery that mesoscale surface structure can impact how humans perceive glossiness is significant, the geometry patterns have not been analyzed in a systematic way. Several studies that have approached the subject have chosen arbitrary and distinct structures and patterns ($\frac{1}{f^\beta}$ noise or overlapping bubbles) leading to difficulties comparing their results. Additionally one can argue that given the complexity of introducing visible surface variations into a stimuli, a property such as gloss is perhaps too far-reaching before understanding simpler traits.

To that end, we report our study designed to evaluate how humans interpret the brightness of surfaces with visible roughness or structure that is still small enough to be considered part of appearance and not geometry. We target brightness as it is one of the most primitive of perceived quantities when viewing a surface. We examine numerous surface patterns to identify commonalities in how humans interpret surface geometry at this scale. A goal is to provide insight into future studies that consider perceived spatially-varying appearances, as well as guidelines for interpreting past research involving mesoscale surface patterns.

## EXPERIMENT

Our experiments were designed to serve two purposes: first, to see how consistently people judge the brightness of a surface with mesoscale patterns of shading; and second, to see if there are trends across varieties of surface geometries. During the experiment, a subject is presented with two copies of a scene side-by-side, as shown in Figure 1. On the left presentation, the disc in the middle of the pillar displays a ray-traced surface with mesoscale texture. The right presentation's disc's brightness is controlled by a slider. Subjects were tasked with adjusting the brightness until it best matched the overall brightness of the complex surface on the left. A matching adjustment task was chosen to avoid the pair-wise explosion that would occur if subjects were to compare patterns. Given the size of surface pattern space considered, a comparison task was deemed inappropriate.

Once matched, as reported by the subject, the screen was cleared briefly before advancing to another trial with a different scene configuration. The matching task was time-limited to 15 seconds. If this time was exceeded the trial was advanced automatically. This short time period encouraged measuring brightness of the overall pattern. When longer or unlimited periods of time were given during pilot studies, subjects frequently attempted to match the unshadowed portion of the stimuli or tried to discount lighting effects in other ways. The specific time window was chosen, based on the pilot study performance, to be a short time that allowed subjects to complete the task while also forcing them to perform our simpler task. Periodically subjects were given a short break to relieve fatigue. Prior to the experiment, all subjects were given a demonstration of mesoscale surfaces in the real world using a molded plastic plaque from an automotive company and then trained with the user interface.

Training consisted of performing the same adjustment task, but on surface patterns not included in the actual study. A window of reasonable values was selected by the authors, and visual feedback was provided if the subject's matching attempt fell outside of the window. All subjects consistently fell within the acceptable window by the end of five training trials, many even on their first trial. Several had issues at the very beginning of training while they became accustomed the 15-second time window.

Two user studies were performed; the studies were identical except for the selected stimuli as described in the later section on stimuli sampling. Each study had twelve participants with no subject participating in both studies. Both studies had 8 female and 4 male subjects each. Subjects ranged in age from 18 to mid-60's with normal or corrected-to-normal vision and no reported visual impairments. The study was conducted in a dark-

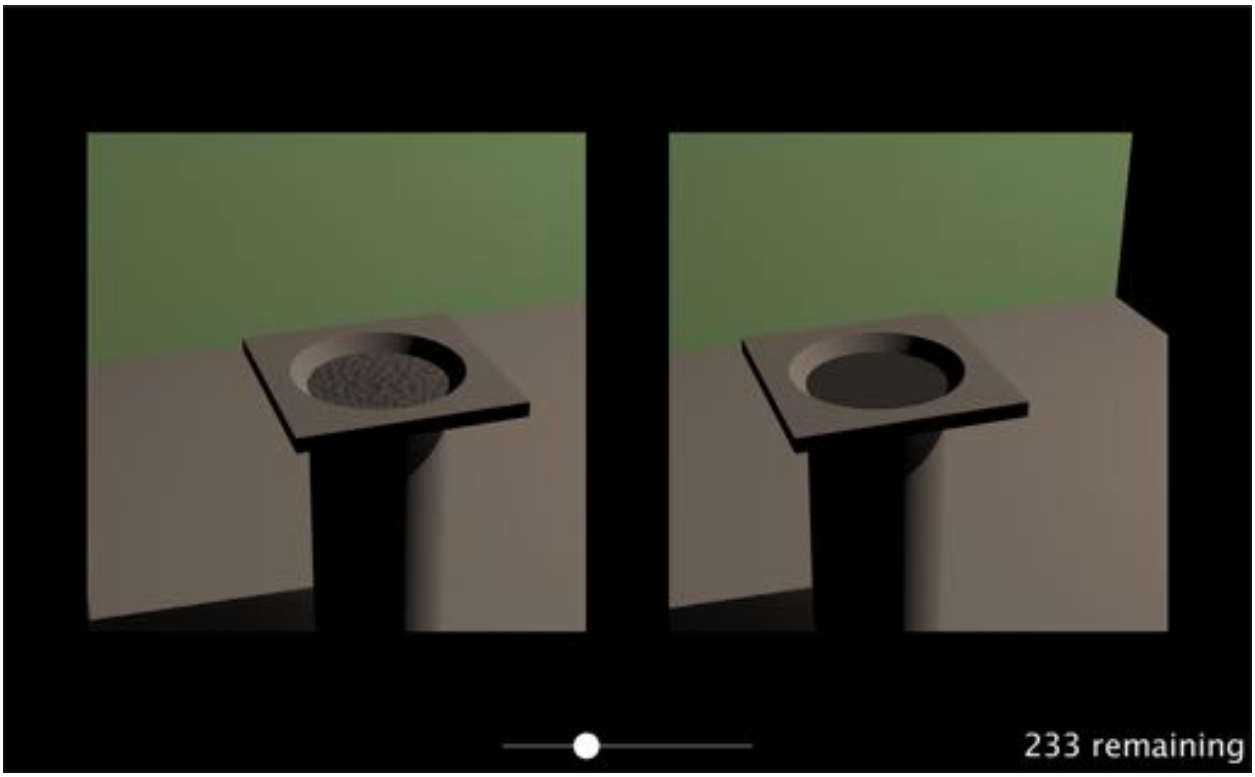



Figure 1. The experiment's user interface presented to subjects. The slider along the bottom is controlled by mouse movement and adjusts the brightness of the right pedestal.

ened room on a single 24-inch Dell monitor in full screen with a resolution of $1920 \times 1080$; the monitor was viewed at a distance of 60cm. The black point of the monitor was $0.0165\,\mu W\,cm^{-2}$ and its white point was $25.00\,\mu W\,cm^{-2}$; the gamma was measured and calibrated but no color calibration was performed due to the monochromatic nature of the tests. The viewing distance was chosen to correspond well to both the distance of the camera in the simulations and the ad-hoc, arms' length approach designers frequently used when viewing physical samples. The stimuli, as described below, was tonemapped to the display using Reinhard's photographic operator.[15]

We next describe our process for producing numerous parameterizable mesoscale surfaces, followed by a discussion of the overall stimulus presented to subjects, and lastly our strategy for sparsely sampling the large number of scenes.

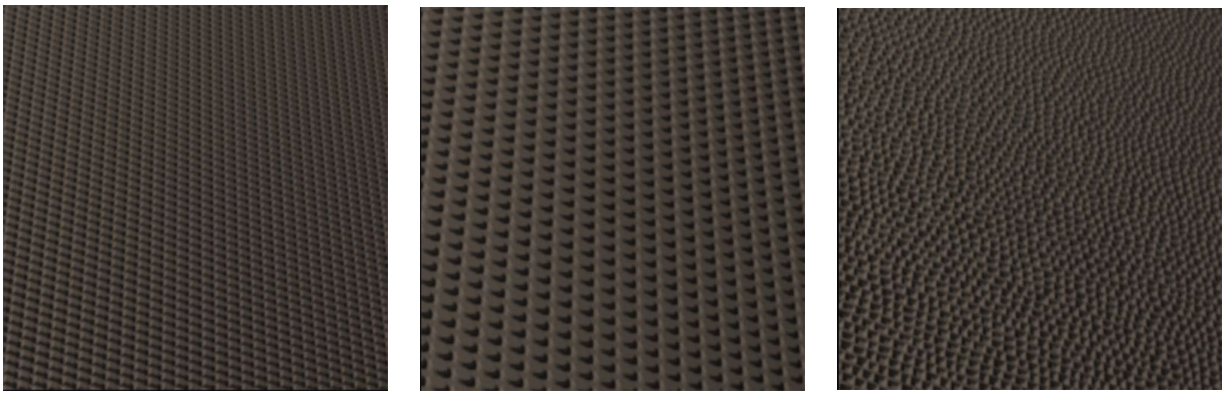
### Mesoscale Surface Generation

Surface patterns were automatically generated using a variety of processes to achieve a spread of patterns that ranged from completely regular patterns that might be manufactured to stochastic or natural patterns. This range includes the classes of surface examined by past studies on the perception of gloss. Each surface generator was parameterized so that many variations could be produced while still having a cohesive structure. The patterns created are described in Figure 2. The patterns shown in Figure 2e and Figure 2f feature irregular structure formed from Perlin noise.[16]

Overall, 136 total surface patterns were generated by varying their available parameters to get a range of stipple sizes, elevations, and other properties. The size of elements within the surfaces ranged from 1mm to 8mm, which given the scene stimuli parameters, covered the targeted characteristic sizes of mesosurfaces.

### Stimuli Design

Every generated surface pattern was rendered from sixteen view points and sixteen lighting directions, for a total of $256 \times 136 = 34,816$ images. Surface patterns were lit and viewed from multiple directions so that any view or light direction dependence on the perceived brightness could be detected. The height elevation of each surface was applied to a plane. This was chosen over a more complex macroscale geometry to remove any confounding factors caused by the shading gradients of the macro surface. The view and lighting positions were distributed evenly over the hemisphere. The sixteen positions were equivalent for views and lighting, the particular number chosen experimentally to adequately sample the reflectance behavior of the mostly diffuse panel while



(a) Rectangular stipples, with parameterized size, spacing, and depth.

(b) Dotted stipples, with parameterized radius, spacing, bubbliness, and depth.

(c) Randomized rectangular stipples, as rectangular stipples but with random variation per stipple.

(d) Randomized dotted stipples, as dotted but with random variation per stipple.

(e) Thresholded Perlin noise to evoke semi-natural ridges.

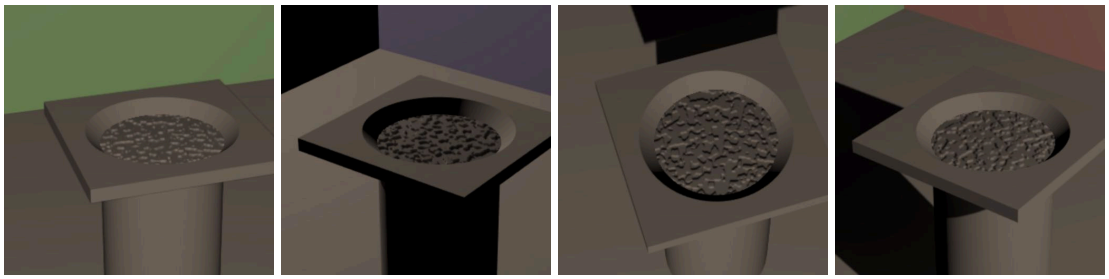
(f) Perlin noise for a natural roughness.

Figure 2. Surface pattern classes generated as part of this study into mesoscale surface appearance.

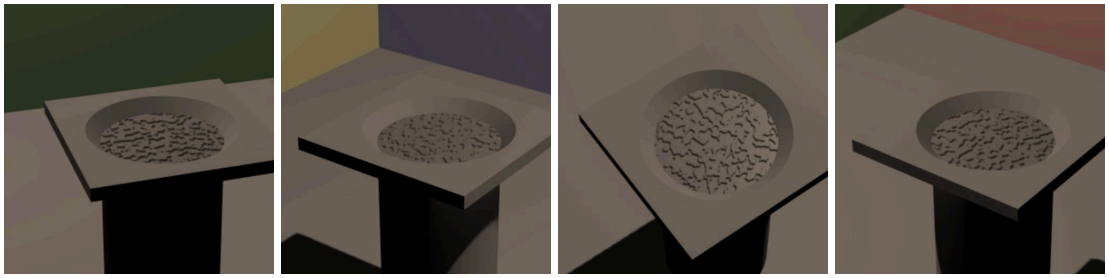
avoiding an unnecessary amount of simulation and stimuli image generation.

These images are presented in a relatively complex scene to provide improved depth and lighting cues. This helps remove inversions in the interpretation of the bump patterns and misinterpretations of the shadowing pattern as an albedo texture. Figure 1 displays the scene the surfaces are placed within. The pedestal provides shading gradients and casts a strong shadow to help the subject infer the light direction. The pedestal stands in the center of a room with four differently colored walls, which alleviates the sense of the object floating in space and helps the subject track where they are viewing from in each trial. The walls and floor are modeled with a perfectly diffuse material, while the mesoscale surface is a plastic material modeled with the GGX distribution[17] and parameters chosen to be similar to plastic sample plaques we have studied from industrial designers. Specifically, the diffuse reflection coefficient was 0.55 and the width of the microfacet distribution function, which models roughness, was 0.153. The viewing position variable affects the entire stimuli scene, which means each presentation can display the pedestal or walls from a different point of view. Figure 3 shows close-up examples of the stimuli, without the paired pedestal that displays the subject-controlled surface.

The light within the room is a 5500K temperature sphere approximately the size of a light bulb and is placed according to the trial's lighting condition. This relatively simple lighting scenario allows changing the direction of the light to have a meaningful impact on the surface appearance while remaining a plausible real-world configuration, such as an indoor room with a bare light bulb. Although there is evidence that real-world, complex environments help perceive glossiness of a material, because the chosen material of our sample is not significantly specular this is less critical. By using a simpler light, we are able to measure the baseline performance on the brightness judgment task before advancing to more complex lighting scenarios in future work.



(a) Four different views of a threshold noise surface under fixed lighting.



(b) The same surface and four views, but under a different fixed lighting condition.

Figure 3. Closeup of stimuli scene presented to subjects, demonstrating shading, strong cast shadows, and walls for context. The two rows show the same surface under different views and lighting conditions.

### Scene Selection for Subjects

Even with the constraints imposed on pattern generation and limiting the scenes to sixteen views and lighting conditions, our dataset of rendered mesoscale surfaces consists of 34,816 images. This is far too many to present to subjects in a reasonable time frame. Instead we opt to do a semi-randomized sparse sampling of the surface patterns to maximize the number of geometries seen while ensuring reasonable repetition across subjects. To that end, the two user studies conducted used different scene selection criteria while otherwise following the exact same experimental procedure. The number of stimuli presented to each subject was constrained so they were guaranteed to complete in one hour, based on the maximum time per presentation previously chosen.

In the first study, each subject was assigned a random subset of the generated surface patterns. For each assigned pattern a random view or light was chosen and the sixteen images matching that condition are included in the trial set for the subject. Additionally, a noise-patterned surface (Figure 2f) was evaluated in the fixed-view and fixed-lighting conditions by every subject, for an additional 32 trials per subject. The noise pattern was chosen for viewing by every subject because it has frequently been used in glossiness perception studies. All selected trials for a subject were shuffled together to avoid ordering adaptation. The shuffled block of trials was repeated three times to collected repeated measures to test whether subjects were significantly changing their responses over time, and to form a better estimate of their matched brightness. This first study captures data for a single surface viewed by many subjects, as well as a sparse sampling of other surfaces viewed by a single subject, all from multiple view and lighting directions.

The second study's selection criteria was designed to complement the data acquired from the first study. Half of the trials considered by a subject in the second study were chosen from conditions previously seen by only a single subject. These conditions were drawn randomly but were weighted towards view and light poses that had a higher perceived brightness variance. Preliminary analysis of the data showed that this higher variance within and between subjects' measured brightness occurred when the light was oriented away from the normal of the stimuli plane and when the viewing direction approached specular. The second half of trials for a subject relied on the same variance-based sampling to choose a view and light pose, but the surface pattern was drawn from the set of patterns not previously seen in the first study. Like before all trials were repeated three times and shuffled. Unlike the first study, each selected set of trials was presented to multiple subjects. This second study provides additional data to validate the responses from the first study's subjects and broadens the number of viewed surface patterns.

In the next section we present and discuss the data gathered from these two studies.

## RESULTS

The goal of this experiment was to determine if the mesoscale surface pattern has a significant impact on our brightness judgments of the object. Prior to making any

substantive claims it is necessary to measure reliability and consistency of the data. The studies were designed to have redundancy within a subject's responses and across multiple subjects. The results and analysis in the following sections consider three subsets of the collected data.

The first subset are the responses to the trials shared by all subjects in the first study. The surface pattern shown in these responses was a rough noise surface of the class described by Figure 2f. Subjects evaluated the surface for all sixteen light directions (with a single fixed view) and all sixteen viewing directions (with a single fixed light) for a total of 31 poses. One light and view pose was present in both conditions. The second subset of data contains the first and consists of all brightness judgments of stimuli presented to at least three subjects. This includes trials from both the first and second studies. The final data set is the totality of brightness judgments.

The next section presents analysis confirming that subjects consistently measured brightness over multiple presentations. The within-subjects analysis uses the first subset of data. The subsequent section analyzes the variability between subjects responses. The between-subjects analysis uses the second set of data. The last results section analyzes the distribution similarity between stimuli viewed or lit at the same angle as one another, and also relies on the second set of data. The third set of data is used solely for the model comparison in the analysis section.

#### Within Subjects

To test whether or not subjects' responses changed over time from the repeated measures, an rANOVA–a one-way ANOVA grouped by the repeated measures–was performed for the 31 shared scenarios viewed by 12 subjects. The null hypothesis of an rANOVA is that there is no significant effect on response over time, i.e. the samples come from the same distribution. The probability $p$-value of a rANOVA test is used to reject this null hypothesis when that value is below some fixed threshold, such as $p < 0.05$. If rejected, there is some detectable, statistically significant change in response over repeated trials. Each of the 31 shared scenarios were analyzed with rANOVA separately–each having 3 repeated trials, with 12 samples in each–which provides a range of $p$-values. The minimum $p$-value is 0.104 and the maximum value is 0.997. After completing the second user study, we extended the rANOVA analysis to all trials that had been seen by at least three subjects (from the pool of 24 subjects). This amounted to 369 unique surface pattern and light/view combinations. Only 16 of these had rANOVA $p$-values less than 0.05 but since they were viewed by only a few subjects it is likely noise from outliers. The median $p$-value over these 369 scenarios is 0.4833. Given this, we cannot reject the null hypothesis that subjects' responses are unchanged over time. Or more simply, there was no detectable change over in subject responses over time. Thus, we assume the responses are from the same distribution and average them together to get a more accurate estimate for each subjects' reported brightness, which we use in the remaining analysis.

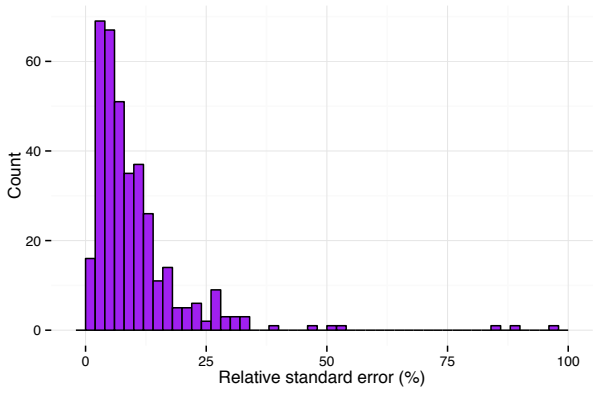



Figure 4. Histogram of the relative standard error between subject responses.

#### Between Subjects

Relative standard error (as a percentage) is used to quantify the consistency between subjects. Only the 369 trial scenarios viewed by at least three subjects were considered. Relative standard error is used so that errors can be compared across the different scenarios. The error was calculated over each scenario's subject responses, after averaging over each subject's repeated measures, and measured against the sample mean. The distribution of error is shown in Figure 4 and is heavily skewed towards the lower end, with a peak around 5%. This is a strong indication that subjects evaluate brightness in a consistent manner. Anecdotally it was reported that more trouble was had evaluating surfaces that presented a mixture of very bright highlights combined with dark shadows. This is verified by the increased variance in responses for scenes at specular with a glancing light angle. Figure 5 shows a view of one surface pattern that presents such a challenge.

#### Stimuli Distribution Similarity

This section looks at the collected results in terms of the similarity between brightness judgment distributions of a surface's stimuli imagery. Each surface has a total of 256 stimuli images associated with it, from all combinations of 16 viewing and lighting directions. However, there is substantial redundancy of these stimuli in terms of viewing and lighting angle, if the orientation of the surface pattern is ignored. These can be described by $N \cdot H$ and $N \cdot L$, where $N$ refers to the geometric normal

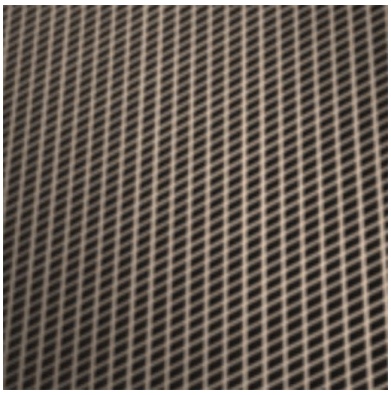


Figure 5. A stimuli featuring strong shadows and bright reflections that make providing a single brightness judgment difficult.

(stimuli plane in this case), $L$ refers to the direction to the light, and $H$ is the half vector between $L$ and the view direction. We will refer to this as the *relative pose* for a stimuli. For each surface, the subject response distributions of its stimuli images can be grouped by their relative pose. Within each relative pose group, it may not necessarily be the case that subjects perceive brightness the same. This would be the case if the orientation of the surface geometry with respect to $L$ or $H$ influenced the perceived brightness.

The inspiration for the relative pose comes from the field of computer graphics where the two quantities, $N \cdot H$ and $N \cdot L$, are often used to describe material reflectance models. For lit regions, the value of $N \cdot H$ ranges from 0 to 1 where values close to 1 signify perfect specular alignment; it is a function of both light and view position. The value of $N \cdot L$ is a function only of light position and similarly ranges from 0 to 1, where larger values correspond to more direct illumination.

Figure 6 helps illustrate the relative pose data for a single surface and the grouping of subject response distributions. Figure 6a shows the average perceived brightness for each relative pose of an example surface. Figure 6b breaks the single value at each coordinate into the average values for the stimuli images of the coordinate. Figure 6c further decomposes it into the full distribution of measured brightnesses from subjects for each stimuli image at each relative pose coordinate. Note how the sizes of the dots, which represent brightness, are very similar to each other within a relative pose coordinate. This suggests that it is acceptable to collapse all measured data into the lower-dimensional relative pose space.

To confirm if this is the case, the Kolmogorov-Smirnov test was used for each pair of subject brightness distributions within a bin. The KS test can be used to determine if two distributions are distinct for low $p$-values. Our approach forms a matrix of $p$-values for the pairwise comparisons, with $p = 1$ along the diagonal. If any element of this matrix has $p < 0.05$ we consider that relative pose coordinate of the scene to be inconsistent. Less than 2% of these bins exhibited inconsistencies and there was no trend amongst those for a particular surface pattern or scene, so they are likely due to outliers. The absolute orientation of surface geometry does not significantly impact brightness judgments after factoring out relative orientation effects. This dimensionality reduction from the 256 absolute poses to relative poses is used later in the analysis section to help form the metric embedding for all surface patterns.

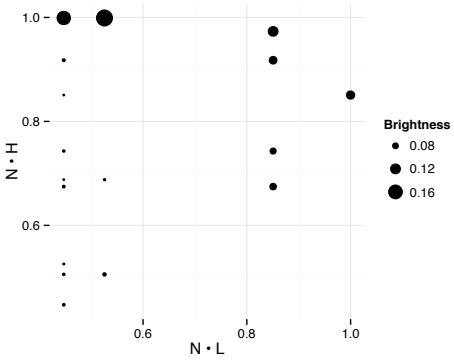
## ANALYSIS

The experimental design previously developed only provided a sparse sampling of brightness evaluations over the set of all stimuli images rendered. In order to evaluate the entirety of generated surfaces, three model estimators of perceived brightness were developed that could be used to produce missing brightness values. The first model, *mean normal*, calculates the average surface normal of the surface and estimates brightness based on that normal. The second model, *flat surface*, estimates perceived brightness as if the subject disregarded shadowing effects. Both of these two models would imply some level of geometric understanding of the surface. The third model, *mean luminance*, simply averages the incoming luminance of the stimuli, which represents a local operator on the image and does not require a higher-level understanding of the surface mesoscale. These models will be evaluated against the collected subject data in the following section.

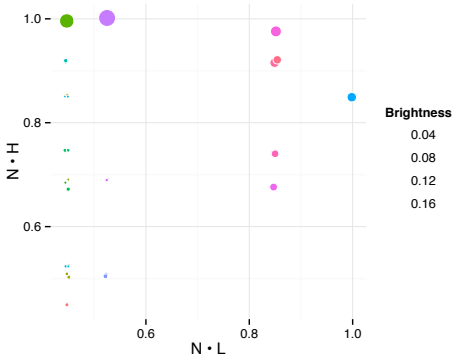
Following the analysis of each model, we embed the surfaces in a metric space based on their brightness profiles from the simulated view and lighting directions. This embedding can be used to explore correlations with geometric properties of the surface texture. High correlations would suggest that the mesoscale surface structure has a strong relationship to perceived brightness, although we find that there is little such influence. Although nominally a negative result, this is a useful for future spatially-varying appearance studies and will be discussed in the conclusion.

### Model Comparison

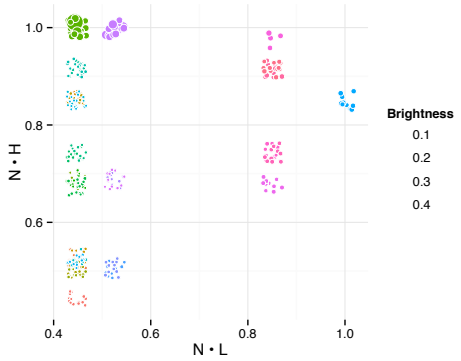
The mean normal model is easily rejected as an independent model because it is not significantly different from the flat surface model. The mean normal model was considered initially because it represents the limit of what happens in real-time rendering as a normal map is filtered and down-sampled. Figure 7 shows the projected $x$ and $y$ coordinates of the mean normal vectors of all surfaces. This is a useful visualization of how the normals deviate from the $z$-axis, the origin, which represents the normal vector of a perfectly flat surface. It is important to note the scales of the axis: no mean normal deviates from the $z$-axis by more than a thousandth of a unit.




(a) Averaged across all responses in each pose group.




(b) Averaged across stimuli conditions in each pose group. Each color represents one of the 256 viewing/lighting images.




(c) Full subject response distribution, colored by their stimuli image.

Figure 6. Distribution breakdown by relative pose coordinates for the rough noise surface used in the first study. For plots showing multiple points per distribution, the positions of each sample are jittered to aid in display.

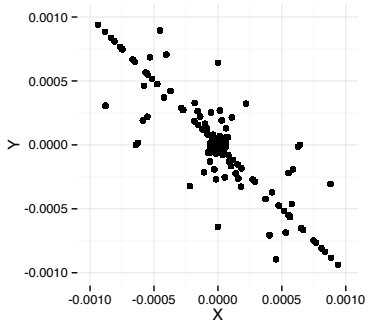



Figure 7. Projection of mean normal vectors onto the $XY$-plane, showing little deviation from the $z$-axis.

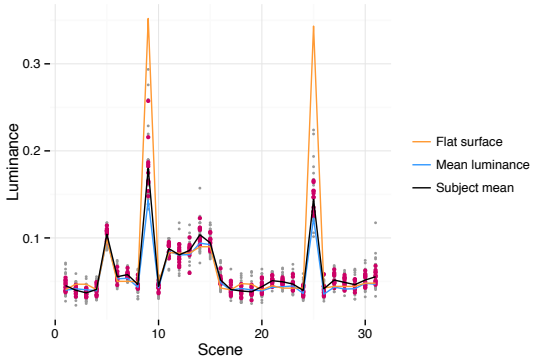



Figure 8. Perceived brightness for the rough-noise surface presented to all subjects in the first study.

Thus the mean normal model provides no significant improvements or differences to the flat surface model. For the remainder of the analysis, the mean normal model is ignored in lieu of the flat surface model.

Next, the flat surface and mean luminance models are evaluated using the first data subset containing responses on a single surface from the first 12 subjects. The responses to the 31 poses are shown in Figure 8 where the

light and view poses are arranged arbitrarily along the horizontal axis. The 36 separate responses (12 subjects $\times$ 3 repeated measures) are shown as transparent black points. Each subject's repeated measures are averaged and color-coded per-user across all poses. The average over every subject is shown as a purple trend line, alongside the two models: the average luminance of the image, and the luminance of a flat surface.

The accuracy of the models can be measured by the probability of their prediction being the population mean of perceived brightnesses. Performing the Student's $t$-test for each scenario shows that depending on the viewing and lighting condition, subject responses are significantly different from either model, although the differences are not substantial. The flat surface model $t$-test produced $p$-values ranging from $2.853 \times 10^{-12}$ to 0.882 and the surface mean $p$-values range from $1.464 \times 10^{-5}$ to 0.818. Given that these distributions were based on only twelve subjects it is hard to rule out the models based on this test alone. The $p$-values for the mean luminance model were higher and more frequently above a significance test of 0.05 compared to the flat surface model. For certain poses, the mean luminance model does represent the population's perceived brightness. However, the at-specular scenarios in Figure 8 show a distinct separation of the flat surface model from both the subject average and mean luminance model. Interestingly though, the subject responses are frequently brighter than the surface mean model when at specular even if they are not as bright as the flat surface model predicts. When off specular, both models regularly fall amongst the subject distribution for the scene.

It is possible that a specular-dependent effect occurs in our perception of brightness over a complex mesosurface. Alternatively, the at-specular stimuli images have a higher probability of producing high contrast images like the example from Figure 6 . The additional amount of reflected light creates more contrast with the shadowed regions of the mesoscale pattern. Increased difficulty was reported verbally by several subjects when they encountered the scenarios shown. This disparity makes it potentially more difficult to calculate the average, or perhaps the subject is biased more towards the brighter reflection. New perceptual tasks and questions will need to be designed to answer this hypothesis. While the data shown in Figure 8 suggests a reasonable fit for the mean luminance model to subject data, it is only for a single surface pattern. To compare the mean luminance and flat surface models to the subject data across many surface patterns, non-metric multidimensional scaling ($MDS$)[18] is used to embed surfaces into metric spaces. These metric spaces of the models are compared against that induced by subject data.

In order to build an embedding, a distance relation between the surface patterns must be defined that is based on the perceived brightness of the surface. We model correlation between surface patterns as the dot product between vectors containing the perceived brightness for each relative pose of the surface. Relative pose refers to the compressed representation based on $N \cdot L$ and $N \cdot H$ of a stimuli image previously described in the results section. The dot product gracefully handles comparisons between surfaces that have disparate samplings of relative poses. Evaluating this correlation function for each pair of surfaces creates a distance matrix that can be used with MDS and other dimension or principle component analysis algorithms. The embedded points of each surface will respect, to the best degree possible, their distances or similarities defined in the distance matrix.

Figure 9 shows the results of applying multidimensional scaling to the set of surfaces evaluated by at least one subject; each surface is drawn as a large point. The six hue blocks correspond to the six pattern classes from Figure 7 for the surfaces. The flat surface model is included in the distance matrix used by MDS, where its brightness is evaluated for all necessary relative pose coordinates. Because the surface geometry does not affect the flat surface model, only a single black point is displayed. The mean luminance model was evaluated on the same stimuli seen by subjects and its corresponding MDS-generated arrangement is shown as small points. A line connects the locations of surfaces between the subject data arrangement and that from the mean luminance model. The embeddings of each space are aligned by use of a Procrustes affine transformation that translates their centroids to the origin, normalizes scale to have unit root mean squared distance to the origin, and calculates a rotation that minimizes distances between paired points. It is valid to apply such an affine transformation because MDS defines the space up to an affine transformation, i.e. only relative distances between points are enforced.

Figure 10 plots the normalized stress of the MDS projections to various dimensions. Low stress values indicate that the distance matrix can still be accurately embedded within that particular dimension. Two plots are shown, one for MDS based on the subject data brightness profiles and one based on the mean luminance modeled profiles. They are almost identical and show an elbow in their curves at four dimensions. However, a stress of under 0.3 represents an acceptable error in the projection, making the arrangements shown in Figure 9 still valid.

While the mean luminance model does not perfectly align with the space formed from subject data, it is quite similar. Many of the paired points are very closely aligned, with only several outliers. These outliers are can be attributed to the sparsity of the distance matrix used in the MDS calculations. While the stress is acceptable when projected to two dimensions, it is not as robust a fit compared to what is shown in Figure 11, which is based on a dense distance matrix. This allows outliers in actual subject responses to significantly affect one of two things: the final projection, or the Procrustes alignment between the model and subject response MDS plots. The pair-wise distances are almost always less than the distance to the flat surface model point, which indicates that the mean luminance model provides the best

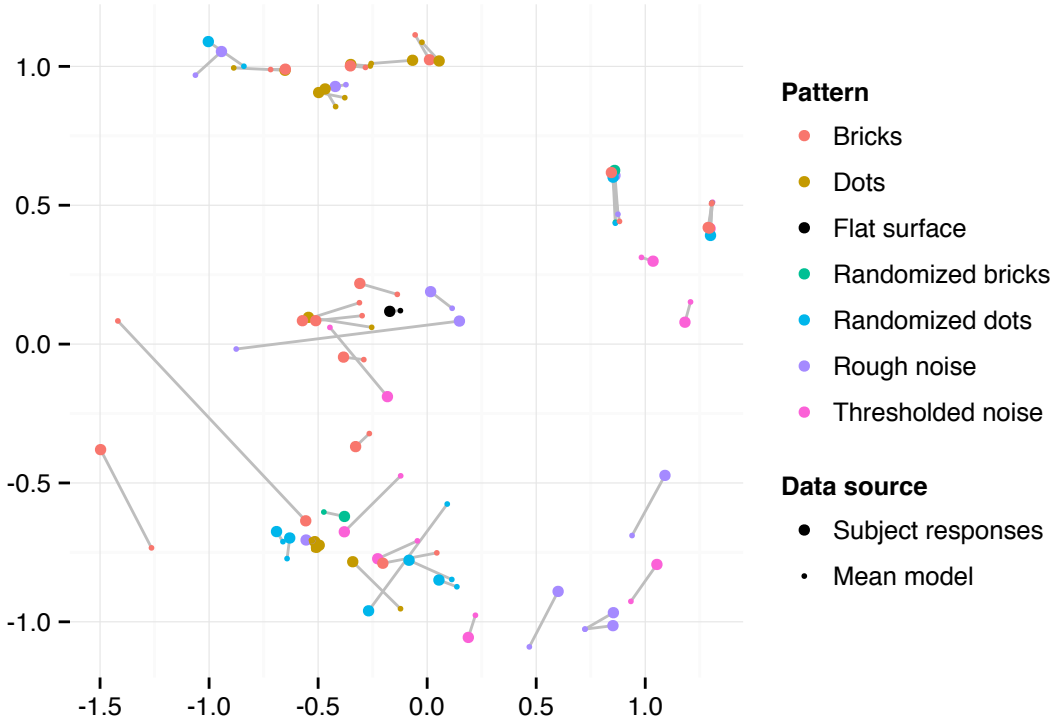



Figure 9. MDS embedding comparison between subject data (large points), mean luminance model (small points), and flat surface model (black point).

approximation to perceived brightness of the models considered. The independence of brightness with respect to surface orientation, which was determined in the previous results section, is further indication that a low-level operator like averaging is consistent with human behavior. If brightness of a complex surface was judged with a high-level understanding of the surface geometry then that surface's orientation to the viewer would likely have an impact on the brightness, which was not found to be the case.

The next section extends the MDS analysis described here to the full set of generated surfaces, and uses these results to examine how geometric properties of the surface relate to perceived brightness.

#### Geometric Correlations with Brightness

The following analysis is inspired by the approach described in Wills' work on identifying perceptual dimensions of gloss.[19] In their study, once a two-dimensional metric space had been found for simulated gloss images, the physical parameters of their gloss model were plotted as a third dimension and multiple correlation analysis was performed to identify any trends. This provided both a correlation coefficient indicating the strength of the correlation across the two spatial dimensions of the embedding as well as the vector direction.

For this analysis, various geometric quantities of the generated surface patterns are correlated against the two dimensional embedding of surfaces using MDS. Given its effectiveness, the mean luminance model is used to calculate a fully-specified, dense distance matrix for every generated surface. Figure 10 shows the two dimensional space calculated using MDS, as well as the stress curve for this modeled data set. This stress curve indicates that two dimensions provides a very accurate reconstruction of the relative distances between surfaces. The density of brightness values and inclusion of additional surface patterns is responsible for the different arrangement compared to what was shown in Figure 9.

While the perceptual gloss work of Wills et al. investigated physical parameters from their gloss model, this work considers physical properties of the mesoscale surfaces. Four physical variables were considered, many of which depend on the concept of a *feature* on the surface. A feature is defined as a contiguous region of the surface that is protruding up from the surface. In the case of surfaces produced from the stipple-driven pattern generators, each raised stipple is a distinct feature. When the stipples are inverted and from concavities on the surface, the contiguous upper flat plane is considered to be a single feature. This was intentionally chosen so that inversions of other patterns have distinct feature descriptors. Features were automatically extracted from the generated surface patterns by utilizing a depth threshold and

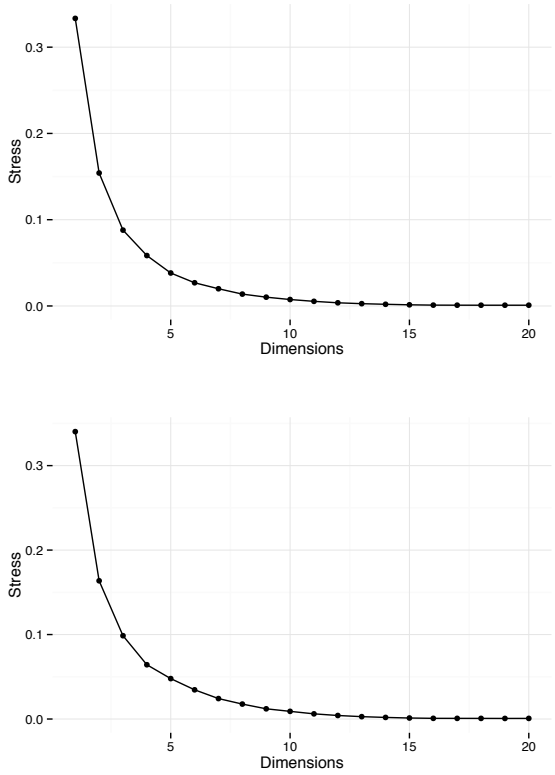



Figure 10. Stress of the MDS projection by dimension, for subject data (first) and mean luminance model (second).

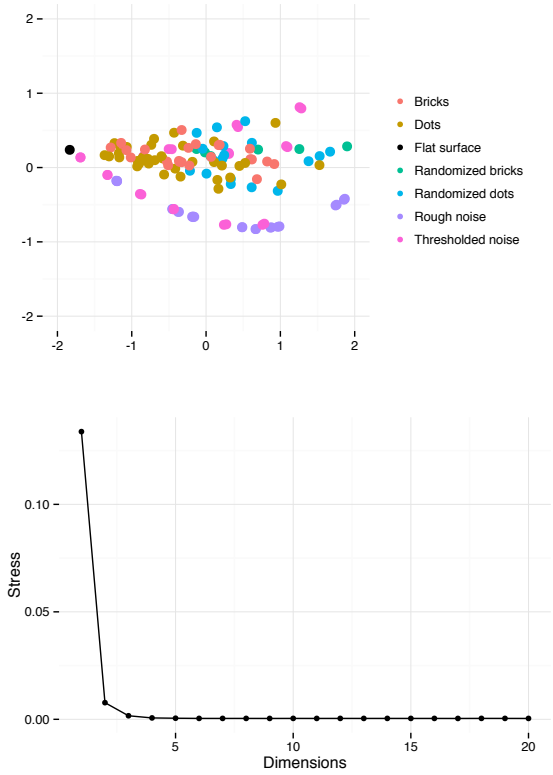



Figure 11. Densely evaluated MDS arrangement for all surfaces in two dimensions, along with its associated stress curve.

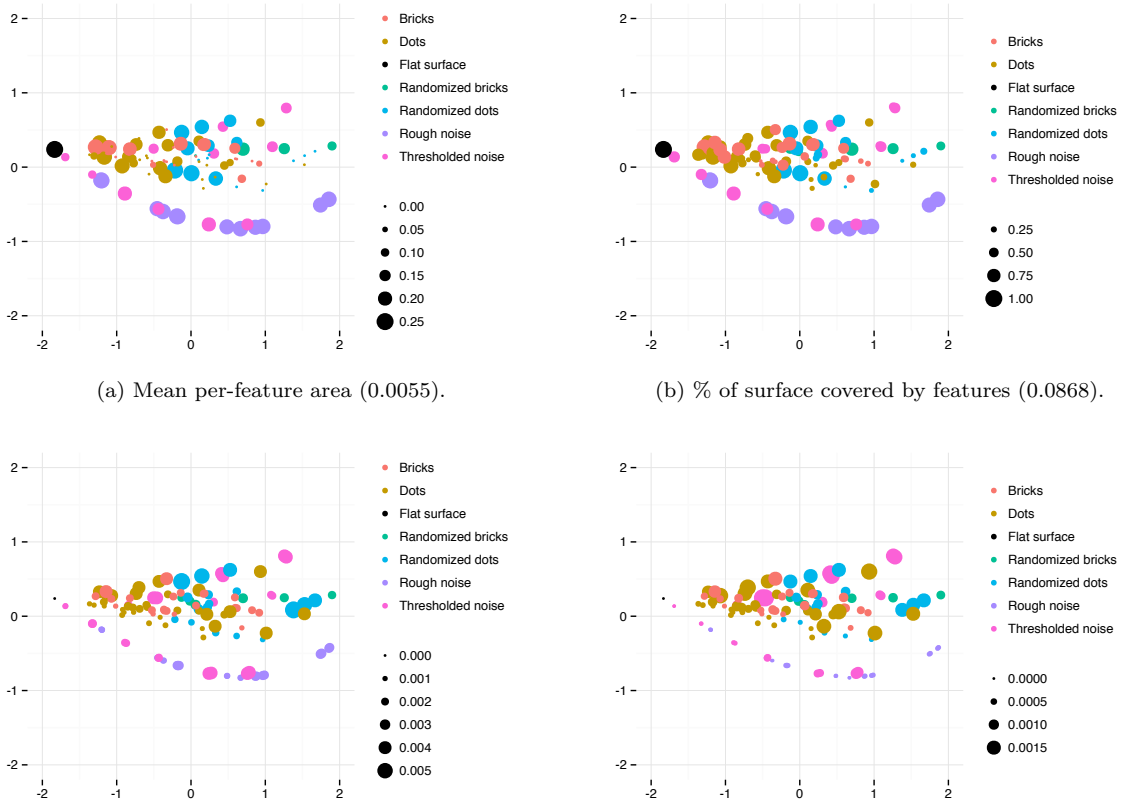
connected-component labeling.

Four geometric variables are considered, three of which are based on the features detected within a surface. The four plots of Figure 12 map the geometric variable to the area of each point in the dense metric embedding previously calculated. The first variable, shown in Figure 12a, is the average area of each feature within the surface pattern. This characterizes the two-dimensional size of the features on a surface while ignoring depth. The second variable, in Figure 12b, is the percentage of total surface area covered by features, which captures how densely packed the feature elements are. Figure 12c presents the third variable: maximum feature depth, which specifies the amount of protrusion or indentation from the flat plane. Finally, Figure 12d displays the standard deviation of depth values throughout the mesoscale surface.

Reviewing these visualizations, no clear trends or structured arrangement of points correlate with the geometric variables. The coefficient of multiple correlation is shown for each of these variables in the caption for each plot. No geometric variable has a strong correlation with the two dimensions of the metric embedding. The largest correlation, for the standard deviation of depth values, is 0.4322, which is not significant. Additionally, there is no pattern or grouping of the surface families specified as the color of each point.

There is obviously some effect on perceived brightness due to a surface's mesoscale structure, however, it does not appear to be related to simple geometric properties of the mesoscale pattern. Dot stipples, rounded bubbled stipples, rectangular brick stipples, and stochastic patterns all are capable of producing similar brightness profiles given appropriate generation parameters. The depth, size, and variation of mesoscale features do not seem directly relevant to the brightness judgment task, given the specific lighting and material parameters chosen for this study. The next section discusses possible impacts of changing these variables.

At face value this seems like a negative result. However, we feel that the lack of a consistent geometric effect on perceived brightness actually makes future research into perception of spatially-varying appearances easier. If each family of pattern produced distinct, non-overlapping clusters of embedded points, it would suggest that the choice of mesoscale structure for a stimuli could have a unique and isolated effect on whatever perceptual attribute is being studied. Given that this is not the case, the choice of pattern is less critical to a study's design and its applicability to untested surfaces.




(a) Mean per-feature area (0.0055).

(b) % of surface covered by features (0.0868).

(c) Maximum feature depth (0.2287).

(d) Feature depth standard deviation (0.4322).

Figure 12. Multiple correlation of embedding against several variables of the surface geometry. The correlation coefficient is given in parentheses for each plot. The direction of maximum correlation is not shown because there are no strong correlations.

## CONCLUSION

The results presented are both promising and challenging. It is a good indication that there is significant agreement between subjects and that brightness judgments are consistent over time. Additionally the subject responses were robust enough to find that surface orientation of the mesoscale pattern did not influence the perceived brightness. The subject evaluations were sparsely distributed over the set of mesoscale patterns generated and used to validate and compare several models of perceived brightness. It was found that simply averaging the incoming spatially-varying luminance provided the best estimate. This model was then applied to the entirety of generated surfaces and multidimensional scaling was utilized to try and discover any relations between surfaces' brightness profiles and geometric properties.

No such relationships or correlations were found for the variables that were considered. As noted earlier, this is desirable because if there had been strong influences from pattern family, size, or depth, then the sheer number of possible mesoscale surfaces would prove insurmountable to future spatially-varying appearance research, and this would bode ill for existing perceptual work on the impacts of mesoscale structure on perceived gloss. It is important to note that only simple geometric variables were considered in this analysis, and it is possible that a more complex geometric description would correlate well with the brightness profiles of each surface. One such model would be to apply Chubb et al.'s work on achromatic texture matching that suggests brightness judgments are dependent on the distribution and histogram of brightness values present in the stimuli.[20] Another possible geometric model would be to apply the more advanced normal map filtering techniques that are used in real-time rendering to better match the appearance of distant objects.[21]

This work is valuable as a foundation for pursuing more complex questions regarding mesosurfaces. While this study considered variations of surface pattern as well as viewing and lighting directions, many configurations that influence appearance were fixed. Only a single microfacet reflection model was used for the surface simulations. Changing the degree of specular and diffuse, or even using a different reflection model, could yield different results. Based on qualitative reports from subjects, stimuli with higher contrast proved harder to judge. Intuitively this makes sense for a local averaging operator; its error is not that significant when contrast is low. The

primary source of contrast in bumpy surfaces with a homogeneous underlying material is the mesoscale structure that creates sporadic shadows and highlights. Materials featuring lighter colors, or shinier materials that reflect additional light at specular, are more likely to exceed a threshold for mapping to a single brightness value. Additional studies can be run utilizing the same study design presented here, but include additional material variations to the set of stimuli. However, it is our expectation that materials that produce similar levels of contrast can be adequately modeled by the local averaging model proposed here.

The effects of contrast on the ability to make brightness judgments are an important topic of additional study. It is not just material parameters that can influence this: lighting conditions are also a significant factor in appearance perception. While different directions were evaluated in this work, it utilized fairly direct illumination with no ambient or other environmental contribution. This allows for potentially very dark shadows, thus increasing contrast. Increasing the intensity of the light source in this situation would also increase contrast and potentially cause the average model to break down, or for subject responses to deviate from those presented in this study. However, using more complex or natural environmental lighting will likely have the opposite effect and decrease contrast. These illumination conditions often have more ambient light so areas in shadow will still be partially lit. Any follow up work to explore lighting influence on mesoscale surface brightness perception should utilize an HDR monitor. This will more accurately capture the differences between intensities in complex environments and help remove any confounding effects of the tonemapping operator required for conventional displays. For this reason an HDR display would also be important when studying more-specular materials. It is clear that controlled lab studies such as this cannot easily scale to the magnitude required to tackle general spatially-varying appearance research. Instead, crowdsourced study methodologies must be developed that are trust worthy as they have the most potential to recruit the necessary number of subjects. Lab-based studies can then be performed to validate crowd-based results. This also provides an opportunity to perform equivalent tasks in the real world with manufactured or printed surfaces to confirm or detect any deviations. Although the rendering algorithms used here are physically correct, and the use of an HDR display would only further increase realism, there is a chance that behavior changes when moving from the real to the virtual.

In summary, we have presented a user study designed with a more complex and natural stimuli for the subject and have evaluated many different surface patterns. Our experimental design has allowed us to consider a wide variety of surface patterns. By densely sampling a small subset and then progressively sparser sets, the surface pattern domain was evaluated robustly. Additionally the experimental time for each subject was kept to a minimum. Analysis of subject responses shows that brightness is consistent over time and that there is little variation between subjects, although it increases along the specular direction. We have also developed an approach for correlating sparsely evaluated brightness profiles of two surfaces. This method was used to show the consistency between collected subject responses and the *mean luminance* model for perceived brightness. It was also used to show that the simple geometric variables that define surface gometry do not correlate with brightness. This is a positive "negative" result because it indicates that the exact choice of surface structure does not completely restrict the applicability of other spatially-varying appearance perception research.

## REFERENCES


[1] S. K. Nayar and M. Oren, "Visual Appearance of Matte Surfaces," Science **267**, 1153–1156 (1995).

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

[3] L. E. Arend, "Mesopic lightness, brightness, and brightness contrast." Perception & Psychophysics **54**, 469–476 (1993).

[4] R. W. G. Hunt and M. R. Pointer, *Measuring Colour*, 4th ed. (John Wiley & Sons, Ltd, Chichester, UK, 2011).

[5] G. Wyszecki and W. S. Stiles, "Uniform Color Scales," in *Color Science* (Wiley-Interscience, 1982) pp. 486–513.

[6] R. Shapley and R. C. Reid, "Contrast and assimilation in the perception of brightness." Proc. Natl. Acad. Sci. **82**, 5983–5986 (1985).

[7] L. E. Arend and B. Spehar, "Lightness, brightness, and brightness contrast," Perception & Psychophysics **54**, 446–456 (1993).

[8] J. Schirillo, A. Reeves, and L. Arend, "Perceived lightness, but not brightness, of achromatic surfaces depends on perceived depth information," Perception & Psychophysics **48**, 82–90 (1990).

[9] R. W. Fleming, R. O. Dror, and E. H. Adelson, "Real-world illumination and the perception of surface reflectance properties," Journal of Vision **3**, 347–368 (2003).

[10] M. G. Bloj, D. Kersten, and A. C. Hurlbert, "Perception of three-dimensional shape influences colour perception through mutual illumination." Nature **402**, 877–879 (1999).

[11] J. A. Ferwerda, F. Pellacini, and D. P. Greenberg, "A psychophysically based model of surface gloss perception," in *Proc SPIE*, edited by B. E. Rogowitz and T. N. Pappas (SPIE, 2001) pp. 1–11.

[12] S. Padilla, O. Drbohlav, P. R. Green, A. Spence, and M. J. Chantler, "Perceived roughness of $1/f^{\beta}$ noise surfaces," Vision Res. **48**, 1791–1797 (2008).

[13] P. J. Marlow, B. L. Anderson, and J. Kim, "The Perception and Misperception of Specular Surface Reflectance," Current Biology **22**, 1909–1913 (2012).

[14] L. Qi, M. J. Chantler, J. P. Siebert, and J. Dong, "The joint effect of mesoscale and microscale roughness on perceived gloss," Vision Res. **115**, 209–217 (2015).

[15] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic Tone Reproduction for Digital Images," ACM Trans Graph. **21**, 267–276 (2002).

[16] K. Perlin, "An Image Synthesizer," in *SIGGRAPH* (1985) pp. 287–296.

[17] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance, "Microfacet Models for Refraction through Rough Surfaces," in *EGSR* (Eurographics Association, 2007) pp. 195–206.

[18] L. Tsogo, M. H. Masson, and A. Bardot, "Multidimensional Scaling Methods for Many-Object Sets," Multivariate Behavioral Research **35**, 307–319 (2000).

[19] J. Wills, S. Agarwal, D. Kriegman, and S. Belongie, "Toward a Perceptual Space for Gloss," ACM Trans. Graph. **28**, 1–15 (2009).

[20] C. Chubb, M. S. Landy, and J. Econopouly, "A visual mechanism tuned to black," Vision Res. **44**, 3223–3232 (2004).

[21] E. Bruneton and F. Neyret, "A Survey of Nonlinear Prefiltering Methods for Efficient and Accurate Surface Shading," IEEE Trans. Visual. Comput. Graphics **18**, 242–260 (2012).