Asymmetric perceptual confounds between canonical lightings and materials

Fan Zhang

Perceptual Intelligence Laboratory,
Industrial Design Engineering,
Delft University of Technology, The Netherlands

Huib de Ridder

Perceptual Intelligence Laboratory,
Industrial Design Engineering,
Delft University of Technology, The Netherlands

Sylvia C. Pont

Perceptual Intelligence Laboratory,
Industrial Design Engineering,
Delft University of Technology, The Netherlands

To better understand the interactions between material perception and light perception, we further developed our material probe MatMix 1.0 into MixIM 1.0, which allows optical mixing of canonical lighting modes. We selected three canonical lighting modes (ambient, focus, and brilliance) and created scenes to represent the three illuminations. Together with four canonical material modes (matte, velvety, specular, glittery), this resulted in 12 basis images (the "bird set"). These images were optically mixed in our probing method. Three experiments were conducted with different groups of observers. In Experiment 1, observers were instructed to manipulate MixIM 1.0 and match optically mixed lighting modes while discounting the materials. In Experiment 2, observers were shown a pair of stimuli and instructed to simultaneously judge whether the materials and lightings were the same or different in a four-category discrimination task. In Experiment 3, observers performed both the matching and discrimination tasks in which only the ambient and focus light were implemented. Overall, the matching and discrimination results were comparable as (a) robust asymmetric perceptual confounds were found and confirmed in both types of tasks, (b) performances were consistent and all above chance levels, and (c) observers had higher sensitivities to our canonical materials than to our canonical lightings. The latter result may be explained in terms of a generic insensitivity for naturally occurring variations in light conditions. Our findings suggest that midlevel image features are more robust across different materials than across different lightings and, thus, more diagnostic for materials than for lightings, causing the asymmetric perceptual confounds.

Introduction

The appearance of an illuminated object is determined by its surface geometry (shape), its surface reflectance characteristics (material), and the illumination (lighting). With arbitrary combinations of material, shape, and lighting, the outcomes are difficult to predict. In computer graphics, given models for the shape, illumination, and material and enough computational power, an object can be precisely rendered by calculating the amount of illumination received by the hypothetical camera ("forward optics"). One classic approach that explains how the human visual system estimates physical properties is called "running physics in reverse" or "inverse optics" (Marr, 1982; Pizlo, 2001; Poggio & Koch, 1985; Poggio, Torre, & Koch, 1985). For material perception, using such an approach, the visual system would need to discount the lighting and shape while estimating the material. To do so, the visual system also would need to discount the material before it could estimate the lighting or the shape. Thus, this is a "chicken and egg" problem. Instead, we take as a given that shape, material, and lighting perception are perceptually confounded. Separate studies have been done on how humans visually perceive shapes, materials, or lightings, yet little is known about the interactions between shape, material, and lighting perception. Varying one of the three elements could result in systematic changes of appearance and, thus, could trigger systematic changes of light, material, *and* shape perceptions, and varying two or three of the elements simultaneously could result in similar ap-

pearances and, thus, trigger ambiguities (Dror, Adelson, & Willsky 2001; Morgenstern, Murray, & Harris, 2011; Pont & te Pas, 2006; te Pas & Pont, 2005; Zhang, de Ridder, & Pont, 2015). In this study, we focus on the interactions between lighting perception and material perception. In order to simplify the problem, we kept the shape of our stimuli constant, limited the study to opaque materials, and systematically varied materials and lightings.

Canonical lighting modes

Unlike in physics, light in space and the visual perception of its properties have not been intensively studied in psychophysics (Schirillo, 2013). Koenderink, Pont, van Doorn, Kappers, and Todd (2007) introduced a light probe to measure light perception. They placed a gauge object into a scene and asked observers to adjust the appearance of the probe such that it visually fit into the scene. Ever since then, progress has been made in measuring how humans estimate illumination properties, such as the relative intensity, direction, diffuseness, and color (Kartashova, de Ridder, te Pas, Schoemaker, & Pont, 2015; Kartashova, Sekulovski, de Ridder, te Pas, & Pont, 2016; Koenderink et al., 2007; Morgenstern et al., 2011; Toscani, Gegenfurtner, & Doerschner, 2017; Xia, Pont, & Heynderickx, 2013, 2014). Another approach is to use images of shaded objects as stimuli to investigate the perception of illumination properties, such as direction and diffuseness (Morgenstern, Geisler, & Murray, 2014, 2015; Pont & Koenderink, 2007; Xia et al., 2014), position of the light source (Schütt, Baier, & Fleming, 2016), complex 2-D light fields (van Doorn, Koenderink, & Wagemans, 2011), and complex natural 3-D light fields (Kartashova et al., 2016). Numerous studies implemented variation of illumination for measuring shape or material perception (e.g., Doerschner, Boyaci, & Maloney, 2010; Dror, Willsky, & Adelson, 2004; Fleming, Dror, & Adelson, 2003; Ho, Landy, & Maloney, 2006, 2008; M. Kim, Wilcox, & Murray, 2016; Marlow, Kim, & Anderson, 2012; Motoyoshi & Matoba 2012; Olkkonen & Brainard, 2010; Pont & te Pas, 2006; Wijntjes & Pont, 2010; Zhang et al., 2015) and found out that illumination influenced the judgments of shape and materials. Yet whether or not observers could perceive the changes of illumination remained unknown. In addition, the lightings involved in the abovementioned studies were mostly arbitrary complex natural luminance maps.

Mathematically, a light field can be described by five parameters $\{\theta, \varphi, x, y, z\}$ that describe the luminance for all directions and throughout the space (note that we neglect color and time for simplification). For a given position (knowing $\{x, y, z\}$), the local light field can be defined by just two parameters $\{\theta, \varphi\}$ that define the directions. Thus, the local light field can be defined as a spherical function and reconstructed by the sum of its spherical harmonics (SH): $f(\theta, \varphi) = \sum_{l=0}^{\infty} SH_l$, where l is the order of the angular mode (Mury, Pont, & Koenderink, 2007; Xia, Pont, & Heynderickx, 2016). The zeroth-order SH component (SH_0) is known as the "light density," and the first order SH component (SH_1) is known as the "light vector" (Mury et al., 2007). The diffuseness of a local light field can be calculated by subtracting the ratio of the powers of light vector SH_1 and light density SH_0 from one (Xia's diffuseness metric; see Xia, Pont, & Heynderickx, 2017a, 2017b). It ranges from zero, the most directed light, to one, the most diffuse light. In architectural perception-based lighting design, many designers build up their light plans in three canonical modes (Ganslandt, & Hofmann, 1992; Kelly, 1952), namely ambient, focus, and brilliance light. Phenomenologically, these modes correspond to the zeroth-, first-, and higher (than second) order components of the SH decompositions of the local light fields in physics (Mury, 2009). In this study, we implemented three canonical lighting modes by creating scenes representing the three abovementioned illuminations. The second order of the SH component of the physical light field is known as the "squash tensor," which we did not recreate in our laboratory environment. We ignored this component here because, in lighting architecture, it is not "designed" or addressed explicitly, probably because this component mostly comes from inter-reflections in natural scenes (Mury et al., 2007).

Canonical material modes

In material-perception studies, we are trying to understand to what extent and how we are able to recognize what things are made of (material categories, such as fabric, paper, plastic, etc.) or to make subjective judgments about the physical characteristics (material qualities, such as soft, smooth, glossy, etc.) or to attribute concepts to certain materials (material meanings, such as aggressive, nostalgic, industrial, etc.). In the material-perception literature, most often, computer graphic renderings are being used as stimuli, especially for materials within the glossy–matte variation. Computer graphics allows users to manipulate a large number of parameters to vary the geometry and surface reflectance of a 3-D object as well as the illumination to create stimuli sets. Using parametric models, it is calculated how incident light scatters from surfaces, resulting in a certain appearance of the rendered objects. It allows systematic control over the changes in the stimuli and, thus, often gives results that can be easily interpreted, but yet it consumes quite an

amount of computational power and sometimes generates images that appear unnatural or unrealistic. Because existing models (Blinn, 1977; Cook & Torrence, 1982; Ward, 1992) simulate glossy materials well, perceived glossiness has been studied intensively (Anderson & Kim, 2009; Fleming et al., 2003; Ho et al., 2006; J. Kim, Marlow, & Anderson, 2011; Marlow et al., 2012; Motoyoshi et al., 2007; Nishida & Shinya, 1998; Pellacini, Ferwerda, & Greenberg, 2000; Vangorp, Laurijssen, & Dutré, 2007). There are also some studies addressing how we perceive other (opaque) material qualities, such as velvetiness (Koenderink & Pont, 2003; Nishida, Sawayama, & Shimokawa, 2015). Other approaches include using real and photographed objects for glossiness perception (Hansmann-Roth, Pont, & Mamassian, 2017; van Assen, Wijntjes, & Pont, 2016), material categorization (Fleming, Wiebel, & Gegenfurtner, 2013; Sharan, Rosenholtz, & Adelson, 2009, 2014), or meaning attribution (Karana, Hekkert, & Kandachar, 2009).

We previously developed a material probe, MatMix 1.0, and found that it provided a perceptually intuitive measuring tool (Zhang, de Ridder, Fleming, & Pont, 2016). It was integrated in an interface for matching tasks, which allowed measurements of material perception in a purely visual and quantitative way. The probe implements optical mixing of four canonical material modes, namely matte, velvety, specular, and glittery. Each of them represents a very different surface scattering mode, and altogether they span a large part of the bidirectional reflectance distribution function (BRDF) space. In a previous study implementing MatMix 1.0, observers were asked to adjust the material probe and match the material to that of the stimuli, which were optical mixtures of photographs taken under one of three canonical lighting modes (Zhang et al., 2015). Results showed systematic, material-dependent influences of lighting on material perception, which was confirmed in an extra experiment using computer-rendered birds. In the current study, we implemented the same set of photographed basis images, the "bird set" (Figure 1), and conducted light-matching experiments by adjusting the probe to allow optical mixing of canonical lighting modes, i.e., by optically mixing the basis images per material instead of per lighting.

To first answer to what extent observers can discount material while matching optically mixed canonical lighting modes, we conducted Experiment 1, in which observers were asked to mix and match the lighting modes of the probe to a mixed illumination in the stimulus. The material modes in the stimulus and the probe could be either the same or different. Observers could only manipulate the illumination of the probe in this task, not its material. In Experiment 2, using a four-category discrimination task and a different group
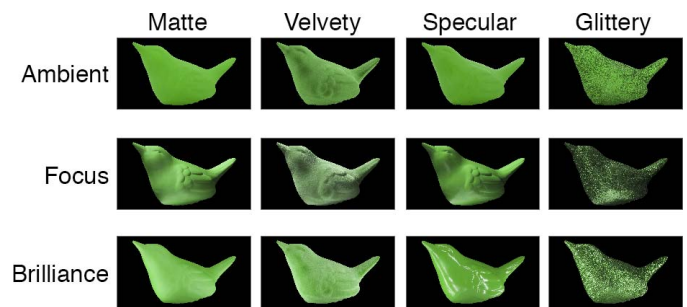

Figure 1. The 12 basis images combining three canonical lighting modes and four canonical material modes, i.e., the "bird set". From left to right, each column represents a canonical material mode (matte, velvety, specular, and glittery). From top to bottom, each row represents a canonical lighting mode (ambient, focus, brilliance). In the matching experiments of the previous work, we optically mixed basis images per row such that materials were optically mixed (Zhang et al., 2015). In the current study, we optically mixed the basis images per column, such that lighting was optically mixed in the stimuli and the probe.

of observers, we tested to what extent observers can simultaneously discriminate materials and lightings. They were shown a pair of basis images selected from the 12 basis images shown in Figure 1 and asked to make simultaneous judgments about whether the materials were the same or not and whether the illuminations were the same or not. In Experiment 3, we compared the matching and four-category discrimination tasks for a reduced stimulus set. A third group of observers was asked to first finish a reduced version of the matching experiment and then, after a short break, a reduced version of the four-category discrimination experiment. The reduction concerned removing the brilliance light stimuli and keeping those of the ambient and focus light, i.e., only using the images in the first two rows in Figure 1.

Experiment 1: Can people discount materials while matching lighting?

Methods

The MixIM 1.0 interface

In previous work, we found that even inexperienced observers performed well above chance in matching optically mixed materials using our MatMix 1.0 interface (Zhang et al., 2016). In this study, MatMix 1.0 was adjusted to MixIM 1.0 (mix illuminations and materials) to allow light mixing and study whether people can match optically mixed canonical lighting modes for objects that are made of the same material or
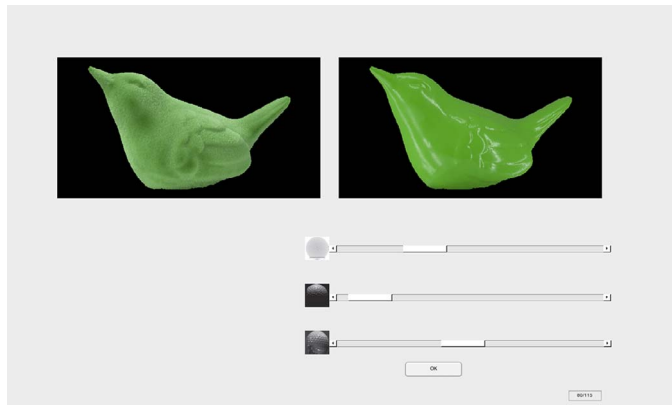
Figure 2. The interface of Experiment 1. Left: A stimulus image. Right: The probe image. The material of stimulus and probe could be the same or different (here they are different). The three sliders represent the three canonical lighting modes. The icon next to each slider visualizes the corresponding lighting mode. The position of each slider bar represents a weight value, ranging from zero to 1.2. The task of the observers was to move the sliders to match the illumination of the probe image with that of the stimulus image. In this figure, the illumination of the probe image does not match the illumination of the stimulus image.

No.	$w_{ambient}$	w_{focus}	$w_{brilliance}$
1	1	0	0
2	0	1	0
3	0	0	1
4	0.5	0.5	0
5	0.5	0	0.5
6	0	0.5	0.5
7	0.33	0.33	0.33

Table 1. Weight of each canonical lighting mode in the stimuli for Experiment 1.

different ones. In contradistinction to optically mixing materials, mixing canonical lighting modes is actually physically realistic. In the MixIM 1.0 interface (Figure 2), three sliders below the right image (probe) represent the three canonical lighting modes, namely ambient, focus, and brilliance light, respectively. How a golf ball appeared under the corresponding light was shown next to each slider to give observers a purely visual reference about what each slider represents. The use of a golf ball as a light probe (Kartashova et al., 2015; Pont & Koenderink, 2007) was chosen because the texture gradients due to the surface structure of the golf ball helps to disambiguate the diffuseness and direction of the light (Xia et al., 2014). In each matching trial, a stimulus image (at left) and the probe image (at right) were presented to observers in corresponding image windows for comparison and matching. The interface was developed using the graphic user interfaces features in MATLAB R2014a (MathWorks, Natick, MA) and presented to the observers on a linearly calibrated Apple, Inc., 15-in. retina display.

Basis images

In our laboratory, we simulated the three canonical lighting modes and took photographs of each canonical material mode under each lighting mode (Zhang et al., 2015) as already shown in Figure 1. For the ambient light, we placed both the camera and the object into a white photo tent and then took the photographs for each canonical material mode. For the focus light, we illuminated the object from the left upper side with a halogen spotlight. For the brilliance light, we hung an LED-strip (150 LEDs) surrounding the object. Note that, in order to register the basis images when performing optical mixing, it was important to keep the same relative position between the objects and the camera. This was done by attaching a horizontal, 1-m-long camera slider on a tripod on wheels. The camera was fixed on one side of the camera slider and the object on the other side. The whole setup could then be moved from one scene to another. The photograph was calibrated by adjusting the white balance of the raw images to set the highlights to be white. Then, to avoid color interaction, we set the hue value to 0.33 (green) for all images using MATLAB. The influence of the hue transformation was negligible as the birds were pure green (RAL 6018, except the glittery bird for which the color was matched visually).

Stimuli

For Experiment 1, we designed seven weight combinations of the three lighting modes as shown in Table 1. Basis images in each column in Figure 1 were linearly superimposed by implementing Equation 1 as shown below, per material mode:

$$I_{stimulus\_material} = w_{ambient} \cdot I_{ambient\_material} + w_{focus} \cdot I_{focus\_material} + w_{brilliance} \cdot I_{brilliance\_material}, \quad (1)$$

where $\{w_{ambient}, w_{focus}, w_{brilliance}\}$ are the weights of the lighting modes (Table 1) and $\{I_{ambient\_material}, I_{focus\_material}, I_{brilliance\_material}\}$ are the basis images shown in Figure 1 with material denoting one of the four canonical material modes: either matte, velvety, specular, or glittery. No linear combinations of materials were used; i.e., the optical mixing of three lighting modes were performed per material. As a result, the linearly mixed stimulus image $\{I_{stimulus\_material}\}$ presents matte, velvety, specular, or glittery material in a combination of ambient, focus, and brilliance light. In Figure 2, the top left image gives an example of

stimulus no. 7 for velvety material; i.e., the weights for all basis images of the velvety bird were equal to 0.33.

Probe

In Experiment 1, observers could manipulate the appearance of the probe image by moving the sliders and, thus, perform the matching accordingly. The probe image was also a linearly superimposed optical mixing result of the basis images per material mode. The mixing process can be illustrated by Equation 2:

$$I_{probe\_material} = w'_{ambient} \cdot I'_{ambient\_material} + w'_{focus} \cdot I'_{focus\_material} + w'_{brilliance} \cdot I'_{brilliance\_material}, \quad (2)$$

where $\{w'_{ambient}, w'_{focus}, w'_{brilliance}\}$ are the weight values corresponding to the positions of the slider bars in the corresponding sliders (see Figure 2: the interface) and $\{I'_{ambient\_material}, I'_{focus\_material}, I'_{brilliance\_material}\}$ are the basis images shown in Figure 1 per material mode, which could be either the same or a different material mode than the material mode used in the stimulus image. No linear combinations of materials were used in the probe either. The linearly mixed probe $\{I_{probe\_material}\}$ allows real-time dynamic and interactive variation of a visual presentation of canonical lighting modes through adjustments of the slider bars.

Procedure

The positions of the slider bars were randomly initialized in each trial. The trials were presented in pseudorandom order. At the start of the experiment, observers were instructed that their task was to move the sliders to adjust the appearance of the bird in the top right window (probe) until it appeared to be in the same illumination as the bird in the top left window (stimulus). They were told that the materials could be the same or different, so the task was not to match the images themselves, but the illumination of the birds. Three trials were performed as practice trials before the first session started. In the practice trials, participants were told that they could move the slider bars by dragging the mouse or pressing the left and right arrow keys on the keyboard. Moving the slider bars by dragging the mouse resulted in bigger steps, and pressing the arrow keys resulted in smaller steps and more gradual changes in the probe. In the actual experiment, four material modes in the probe image were combined with four material modes in the stimulus, resulting in 16 material combinations. Together with seven weight combinations for the stimuli lighting in the optical mixture (Table 1) per material combination, there were 112 trials in total for each

observer. It took around 60 min to finish the experiment.

Observers

We recruited four unpaid observers who had participated in at least five psychophysical experiments, and 11 paid inexperienced observers participated in Experiment 1. The four unpaid observers are grouped as "experienced" as they had participated in former experiments working with the experimental interface. All 15 participants had normal or corrected-to-normal vision. Participants read and signed a consent form before the experiments. The experiments were approved by the human research ethics committee of Delft University of Technology and conducted in accordance with the declaration of Helsinki and Dutch law.

Analysis and results

Least squares fit

The matching performance using the MixIM 1.0 interface can be evaluated by solving the linear factor matrix X of Equation 3 using least squares fitting:

$$[P]_{3 \times (112 \times N)} = [X]_{3 \times 3} \cdot [S]_{3 \times (112 \times N)} + [E]_{3 \times (112 \times N)}, \quad (3)$$

$[S] = \begin{bmatrix} w_{ambient} \\ w_{focus} \\ w_{brilliance} \end{bmatrix}$, the weights of the stimuli,

$[P] = \begin{bmatrix} w'_{ambient} \\ w'_{focus} \\ w'_{brilliance} \end{bmatrix}$, the weights of the probe,

and the residuals $[e] = \begin{bmatrix} e_{ambient} \\ e_{focus} \\ e_{brilliance} \end{bmatrix}$.

In Equation 3, each row represents a canonical lighting mode, specifically the ambient, focus, and brilliance lighting mode from top to bottom. Per observer, there were 112 trials, and together with the number of participants N, there were in total $(112 \times N)$ columns in matrix S, matrix P, and matrix E. Each column in matrix S represents the weights of the three canonical lighting modes in the stimulus image, and the corresponding column in matrix P represents the weights of the three canonical lighting modes in the probe image, i.e., the values represented by the positions of the three sliders set by the observers. The 3 × 3 linear factor matrix X was solved using a least squares fit in MATLAB, and then matrix E was the subtraction between P and $X \cdot S$. If the matching would be veridical, X would be a 3 × 3 identity matrix, and the

$r = 0.56$	$w'_{ambient}$	w'_{focus}	$w'_{brilliance}$
$w_{ambient}$	0.63	0.22	0.31
w_{focus}	0.29	0.66	0.29
$w_{brilliance}$	0.20	0.22	0.62

Table 2. The linear factor matrix X in Equation 3 solved using the least square method ($N = 15$).

matrix E would be a zero matrix. The ratio r between the sum of the diagonal values in X and the sum of X, i.e., $r = \sum diag(X) / \sum(X)$, can be used to evaluate the performance, ranging from zero (only possible mathematically) to one (veridical) with 0.33 being the chance level.

Overall results

The overall results of all observers in Experiment 1 is expressed as the linear factor matrix X, solved by least squares fitting, and is shown in Table 2 ($N = 15$). In the matrix, the diagonal values are 0.63, 0.66, and 0.62 for ambient, focus, and brilliance light, respectively, and the nondiagonal values are all between 0.20 and 0.31, so the matrix is dissimilar from an identity matrix. The ratio r is 0.56, which is far above chance level ($r = 0.33$, see individual performance).

The performance per material combination in stimulus and probe for all observers can be seen in Figure 3. The plot shows the ratio r calculated per material combination with the colors of the bars coding the materials of the probe. Each subplot shows results for one material of the stimulus (matte, velvety, specular, and glittery from left to right with labels on the x-axis coded in corresponding colors). When the materials were the same in the stimulus and the probe, the performances were closest to veridical ($r = 1$) in each subplot. When the materials were different in the stimulus and the probe, the performances were still above chance but less close to veridical than when materials were the same. When the velvety material mode was presented, irrespective of whether it was in the probe or in the stimulus, the results were the least veridical. This shows that material differences decreased the performance of matching optically mixed lighting modes. Thus, for our very diverse material and lighting modes, there were strong perceptual interactions between materials and lightings.

Individual performance

The individual matching results (the histogram of the ratios r for all observers) can be seen in Figure 4 ($Mean = 0.57$, $SD = 0.14$). It clearly shows that four out of 15 observers performed just above chance level (0.33), and the other 11 observers performed well above
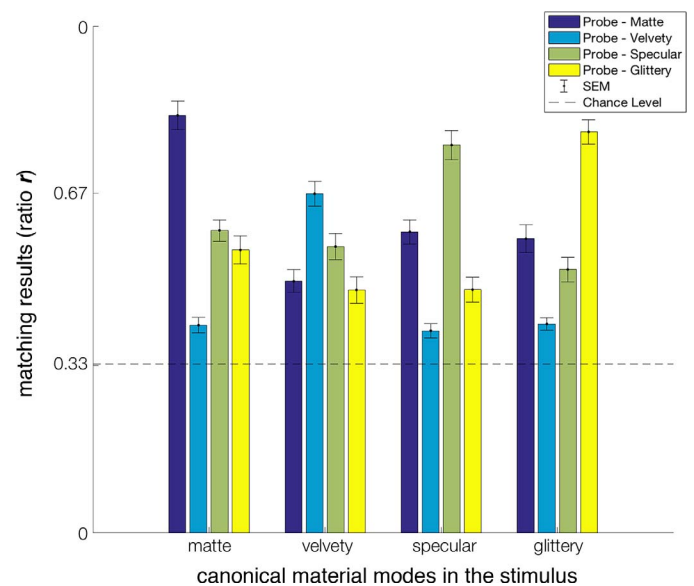

Figure 3. Ratio r calculated per material combination of the stimulus and the probe. The four subplots show the results for matte, velvety, specular, or glittery stimuli from left to right, respectively. The material of the probe is color-coded; see legend. The y-axis represents the ratio r. Each ratio was calculated over all data of the 15 observers per material combination. The error bars depict one standard error of the mean.

chance; i.e., most of the observers were able to match the optically mixed canonical lighting modes. The four observers who performed just above chance level were all inexperienced observers (colored in blue).
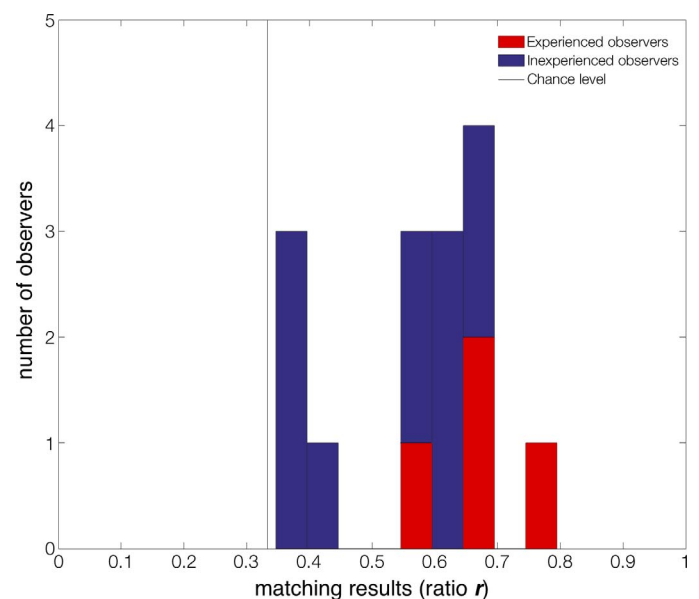

Figure 4. Histogram of number of observers for the performance ratio r. The red-colored bars are the results of four observers who are experienced in psychophysical experiments. The blue-colored bars are the results of 11 observers who had no experience in psychophysical experiments at all.
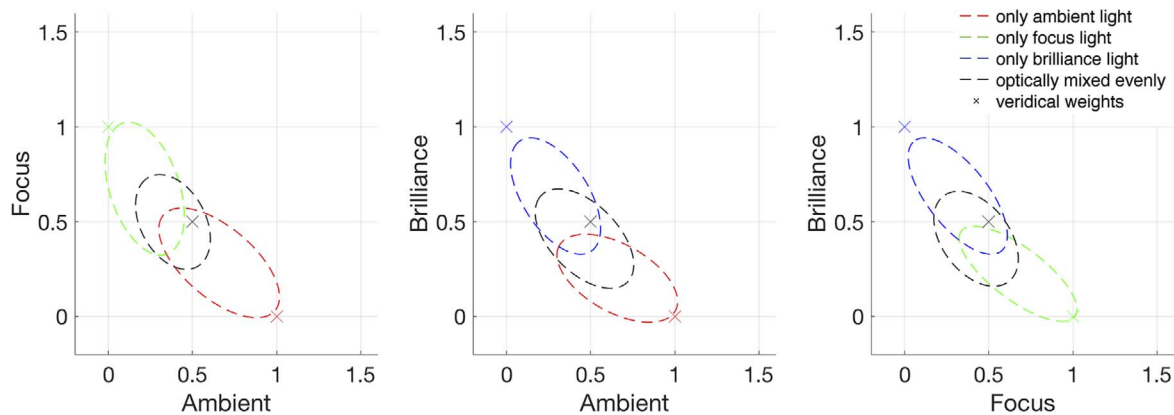
Figure 5. Bivariation plots for each combination of two lighting modes for all observers. The three subplots are results for different lighting combinations. Different colors correspond to different lighting-weight combinations in the stimuli, which are depicted by the crosses (the veridical weights). Specifically, the red color corresponds to the stimuli in which only ambient light was presented, the green color corresponds to the stimuli in which only focus light was presented, the blue color corresponds to the stimuli in which only brilliance light was presented, and the black color corresponds to the stimuli when two lighting modes were optically mixed (each 50% in the mixture). The ellipses represent one standard deviation of bivariate normal distributions fitted to the data.

Perceptual interactions between canonical lighting modes: Bivariation plot

Another way of interpreting the data from our matching experiment is to visualize the interactions between the basis modes in the mixtures. The interactions between each combination of two lighting modes were visualized by means of ellipses representing one standard deviation values of bivariate normal distributions fitted to the data for all observers for the 16 material combinations (four materials in the stimulus by four materials in the probe). The fitted ellipses are shown per lighting combination in Figure 5 and for different groups of observers in Figure 6. Every data point represents the settings of two of the three sliders in the probe in one trial. For clarity of presentation, the data points themselves were rendered invisible in the plots. Each subplot contains three ellipses, which depict the results for three different weight combinations in the stimuli. The coordinates of the crosses depict the corresponding weight combinations of the stimuli (see Table 1). This provides a means to visualize the extent to which participants confuse the lighting modes. In general, if there is less overlap between ellipses, if the ellipses are centered closer to the crosses, and if the ellipses are smaller, then the lighting modes interact less. The general results can be seen in Figure 5. In the plots, the red color corresponds to the stimuli in which only ambient light was present, the green color corresponds to the stimuli in which only focus light was present, the blue color corresponds to the stimuli in which only brilliance light was present, and the black color corresponds to the stimuli when two lighting modes were optically mixed (each 50% in the mixture). We find that the ellipses are in the right order but tend to shift toward each other in the center. Blue ellipses

shifted away from the blue crosses the most, showing that the responses for mixtures containing the brilliance light were the least veridical.

To further analyze the interactions between materials and lightings, we looked into the results per material combination as shown in Supplementary Figure S1. The rows of Supplementary Figure S1, containing three subplots, show the matching results per material combination of the stimuli and the probe under different lightings, corresponding to the results (one of the 16 ratios r) shown in Figure 3. For symmetric matching, if the materials in the stimuli and the probe were the same, we found that the crosses (the stimulus centers) fell into the ellipses (one standard deviation of bivariate normal distribution fitting). The only exception happened if velvety was presented in the stimuli and the probe, for which the probing results of the ambient and brilliance light deviated more than one standard deviation. For asymmetric matching, Supplementary Figure S1 shows that when velvety was presented in the probe, the ellipses tended to shift toward the green cross representing focus lighting or to the origin for conditions without focus lighting. This explains why the results were less veridical when the velvety mode was present as shown in Figure 3.

To further analyze the individual results, we separated the group of four observers who performed just above chance level from the group of inexperienced observers that performed better, according to both the results from the least squares fitting method (Figure 4) and their individual bivariation plots (as shown in Supplementary Figures S2 through S4). In addition, the four experienced observers were separated as one group (colored in red in Figure 4). In Figure 6, results of the three observer groups can be seen in the rows. In each row, on the left, it shows the 3×3 linear factor matrix
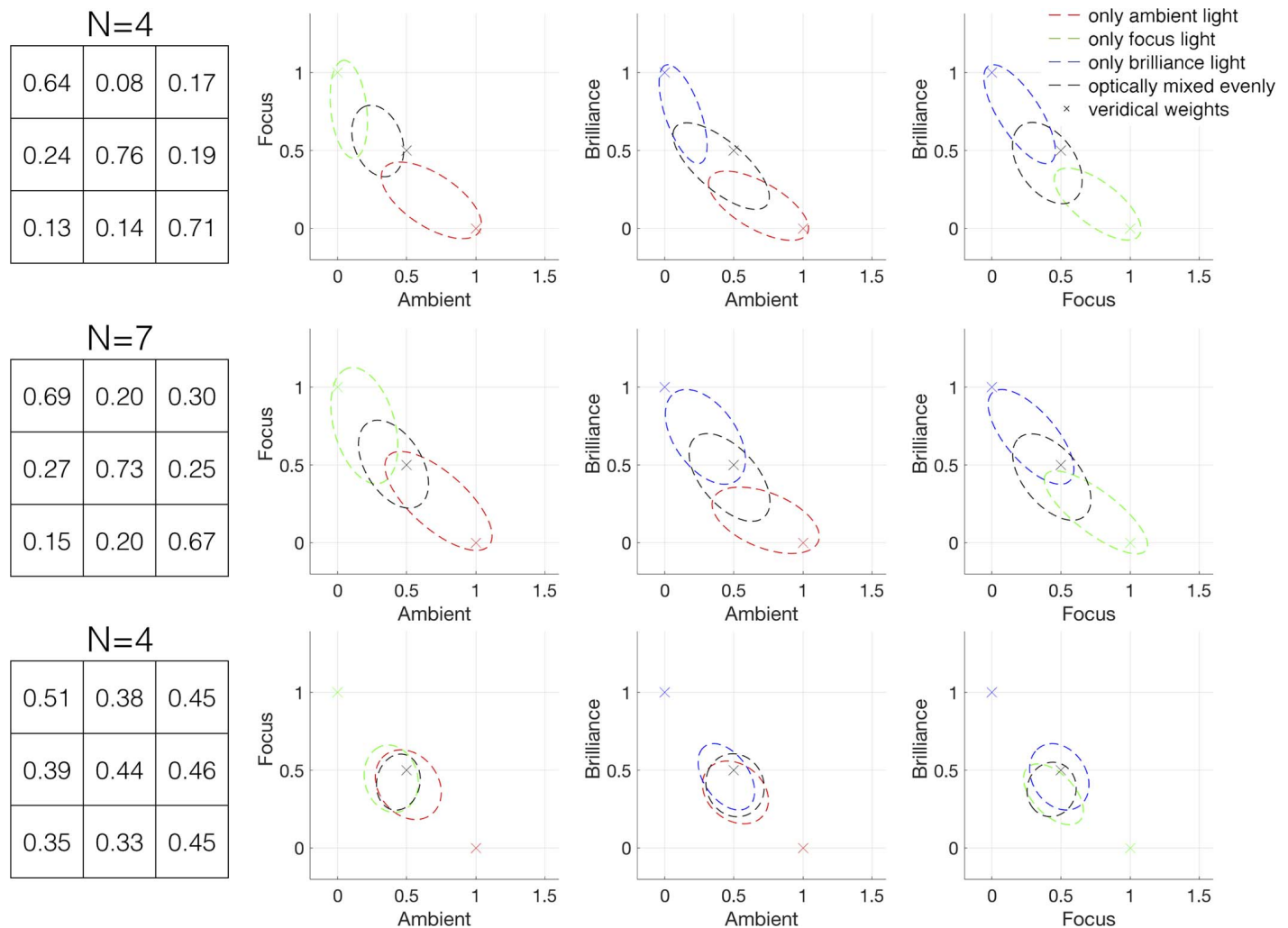
Figure 6. Left: Linear factor matrices that were fitted using the least squares method, per group, in the same format as in Table 2. Right: Bivariation plots for each combination of two lighting modes (in the columns) for three groups of observers (in the rows). Top: Results of the four experienced observers. Middle: Results of the seven inexperienced observers who performed far above chance. Bottom: Results of the four inexperienced observers who performed just above chance. Different colors correspond to different lighting-weight combinations in the stimuli, which are depicted by the crosses (the veridical weights). Specifically, the red color corresponds to the stimuli in which only ambient light was presented, the green color corresponds to the stimuli in which only focus light was presented, the blue color corresponds to the stimuli in which only brilliance light was presented, and the black color corresponds to the stimuli when two lighting modes were optically mixed (each 50% in the mixture). The ellipses represent one standard deviation of bivariate normal distributions fitted to the data.

X that was calculated per group (the same format as Table 2). On the right, each subplot shows a combination of two lighting modes (in colors). The first row shows the data for the group of the four experienced observers; note that all of them performed well above chance ($r = 0.60, 0.69, 0.69, 0.79$). The second row shows the data for the group of the seven well-performing inexperienced observers ($r = 0.55, 0.55, 0.57, 0.61, 0.62, 0.66, 0.68$). The third row shows the data for the group of the four inexperienced observers that performed just above chance ($r = 0.35, 0.35, 0.39, 0.40$). The ellipses for the experienced observers (the first row) show less overlap

than those for the inexperienced observers (the second row) and certainly than those for the just-above-chance performers. The crosses, depicting the veridical settings, were all within the ellipses for the experienced observers, and the blue crosses (brilliance light) were outside the blue ellipses for the well-performing inexperienced observers; i.e., the veridical weights of the brilliance lighting mode differed more from the mean probing results for this group of inexperienced observers. The results of the observers who performed just above chance level according to the least square fitting analysis, as shown in the third row, cluster in the center. Overall, the ellipses tend to shift to the center of

the plots. Apparently, the participants always use at least two sliders even when only one slider is required for a perfect match. This is especially obvious with the inexperienced observers, but it is also apparent for the experienced participants.

Intermediate discussion

In Experiment 1, we asked observers to match optically mixed lightings in two conditions: symmetric matching (same materials in the stimulus and the probe) and asymmetric matching (different materials in the stimulus and the probe). The goal was to test whether observers could match the mixture of canonical lighting modes while discounting materials. In general, observers were above chance level in the light-matching tasks. Individual differences were found as four out of 15 observers tended to mix all lightings no matter if they were presented in the stimulus, which led to their less-veridical performances. We also found that when velvety was presented in the probe or in the stimulus, the overall performance was significantly less veridical. To conclude, using our optical mixing interface, we found that observers were able to either match lightings while discounting materials (Experiment 1) or match materials while discounting lightings (Zhang et al., 2016). To further investigate the confounds between our canonical material and lighting modes, we designed Experiment 2 to test whether observers could *simultaneously* discriminate materials and lightings and Experiment 3 to relate the results of the two types of tasks.

Experiment 2: Can people simultaneously discriminate material and lighting?

Methods

This experiment was to test whether observers can discriminate our canonical material and lighting modes simultaneously and to what extent material and lighting perceptions are confounded. The task was similar to a previous study in which observers were asked to judge materials and illuminations separately for a series of spherical objects (te Pas and Pont, 2005). Here, we asked observers to make discrimination judgments for a more systematic set, that is, our canonical material and lighting modes, and observers had to judge materials and lightings simultaneously. In each trial, observers were shown a pair of stimulus images and asked to choose from four response

categories—"same materials same lightings," "same materials different lightings," "different materials same lightings," and "different materials different lightings"—based on the appearance of two birds (Figure 7). The aim of the experiment was to test whether (and for which modes) observers can judge if differences in appearance are due to material and/or lighting variations for systematically chosen modes that strongly differ optically and together span much of the reflectance and lighting spaces. The interface was developed with the Psychophysics Toolbox extensions (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997) in MATLAB and presented to the observers on a linearly calibrated Apple, Inc., 15-in. retina display.

Stimuli

Only the 12 basis images were used as stimuli in Experiment 2 (Figure 1); i.e., no optical mixing was performed in Experiment 2.

Observers

Eight paid inexperienced observers participated in Experiment 2. All participants had normal or corrected-to-normal vision. Participants read and signed the consent form before the experiments. The experiments were approved by the human research ethics committee at Delft University of Technology and conducted in accordance with the declaration of Helsinki and Dutch law.

Procedure

Because all observers were inexperienced and did not participate in Experiment 1, they were instructed to browse through all stimulus images in pseudorandom order before the actual experiment started to give them a brief idea of how similar or different the images could be. Each stimulus image was repeated twice and displayed for at least 0.5 s before the observer could click a button to display the next one. They were told that there were four different material types and three lighting types and every image would be one of the four materials in one of the three types of lighting. They were also told that, in the actual experiment, their task would be to compare two of the images and answer whether the materials are the same or different and whether the lightings are the same or different.

With 12 basis images as stimuli, there were 78 possible combinations, 12 of "same materials same lightings," 12 of "same materials different lightings," 18 of "different materials same lightings," and 36 of "different materials different lightings." In order to balance the number of trials for each stimulus category, they were repeated six, six, four, and two times per
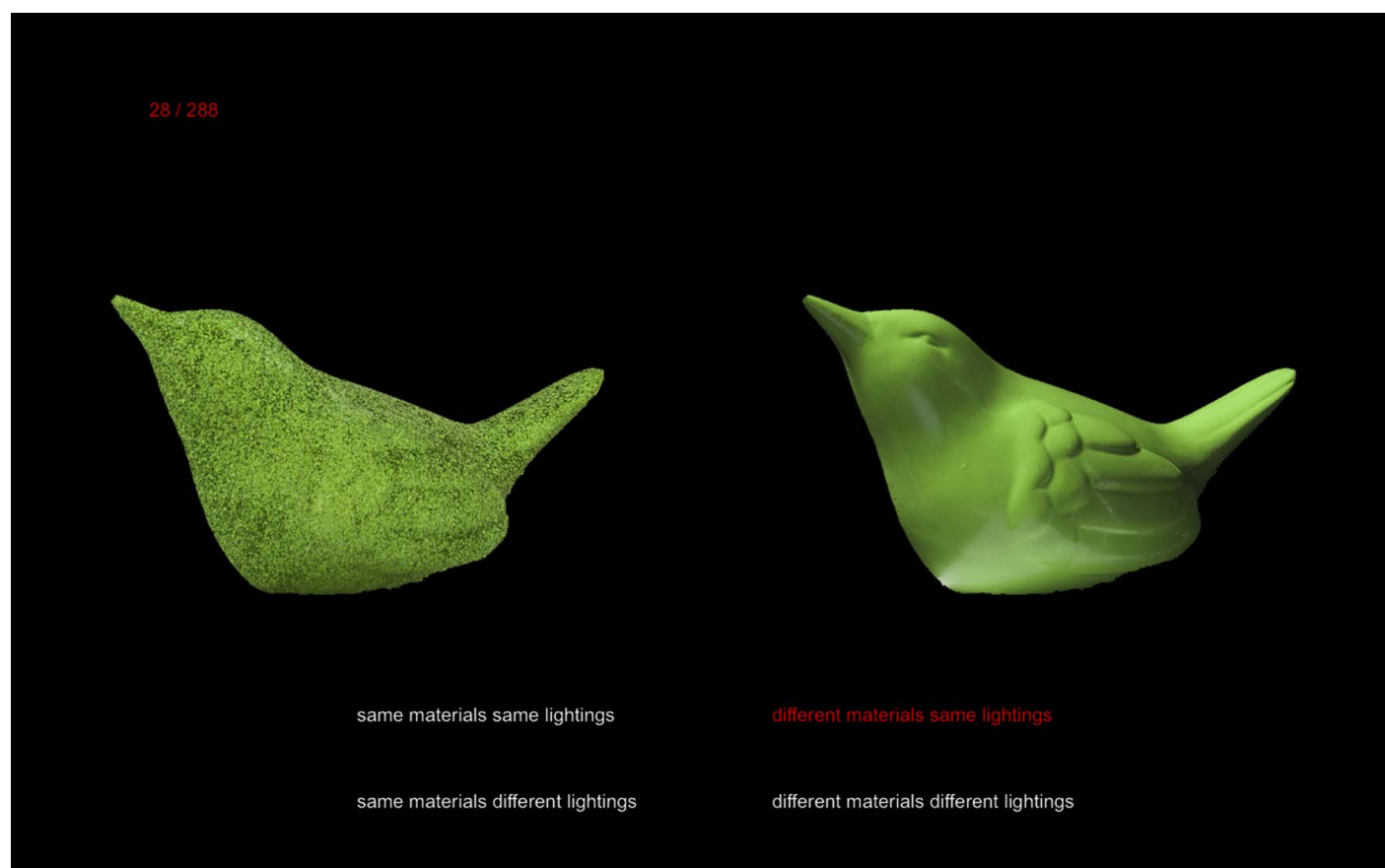
Figure 7. The interface of Experiment 2. Left: Glittery material under ambient light. Right: Specular material under focus light. The four response options are listed below the images. The selected option is marked red. The number in the top left corner indicates the progress (number of trials done as a ratio of the total number of trials). Here, the selected option is not correct.

category, respectively, so that we got 72 trials per stimulus category, i.e., 288 trials per observer. Without time limits for the task, it took around an hour to finish the experiment.

In the actual experiment, a pair of stimuli was displayed and one of the four options was randomly initialized. For the images in each stimuli pair, being left or right was also randomized. Observers were instructed that they could press up, down, left, and right arrow keys on the keyboard to select their answer. The selected one was marked red. Then observers could press the spacebar to finish the current trial and start the next one. The numbers on the top left corner of the interface indicated the progress of the experiment.

Results

Overall performance

In Figure 8, the fractions of responses per stimulus category are shown. Each square shows the fraction represented as a gray level with the number showing the exact value, calculated by dividing the total counts of the responses by the number of trials per stimulus category

(i.e., 72 in this task). Each row represents one stimulus category, and each column represents an answering option. Note that, for each row, the fractions of the four answers add up to one, and the diagonals show the fractions of the correct answers, i.e., the discrimination accuracy. Also note that chance level is 0.25 for this four-category discrimination experiment. As expected, when the materials and lightings were both the same in the stimuli image pair, observers got the highest accuracy (0.97). When the materials were the same and lightings were different, the accuracy somewhat decreased (0.78). But when the materials were different, the accuracy strongly decreased to be just above 0.5 independent of whether the lightings were the same (0.58) or different (0.54). Off-diagonal values are negligible except for two cases (0.27 and 0.33). The responses were found to be significantly associated with the stimulus categories ($\chi^2(9) = 3247.2, p < 0.001$). They also showed that, when materials were different, observers would indeed perceive the materials to be different but then be less accurate about whether the lightings were the same or different. In Supplementary Figure S5, we present the stimulus image pairs that resulted in the least and best performances in Experi-
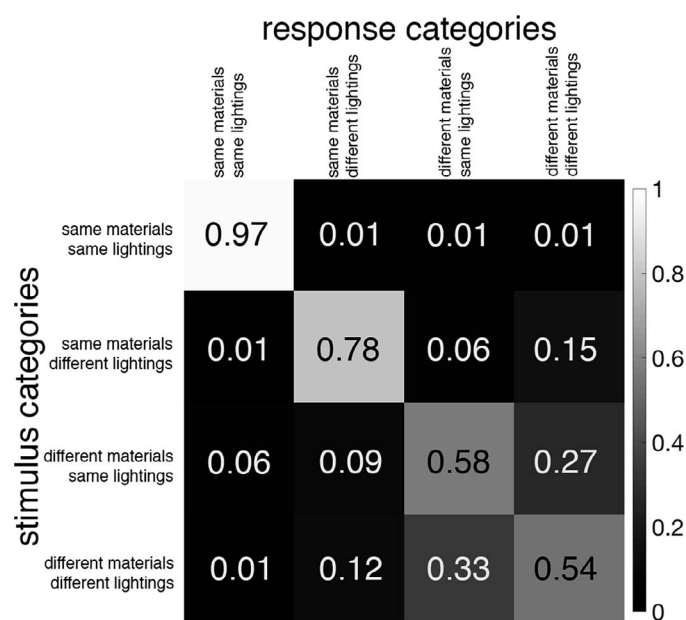
Figure 8. The fractions of responses per stimulus category. Each row represents a stimulus category, and each column represents a response category. The squares on the diagonal are the fractions of answering correctly, i.e., the discrimination accuracies.

ment 2 (only for the "different materials different lightings" category). To conclude, both the material and lighting differences caused the accuracy to decrease, but material differences caused the accuracy to decrease more. For different materials, the observers had much difficulty in judging whether the lightings were the same or not but still performed well above chance.

In order to further analyze the results, we implemented signal-detection theory by considering the four-category discrimination task as two yes-or-no questions: (a) "Are the materials the same?" and (b) "are the lightings the same?" Explicitly, when analyzing materials, lighting was not considered and vice versa. For example, stimulus (or response) categories "same materials same lightings" and "same materials different lightings" were combined as one stimulus (or response) category for materials ("the same"). Answering "the same" when the stimuli were the same constitutes a "hit," and answering "the same" when the stimuli were actually different constitutes a "false alarm." The hits and false alarms could be converted to z scores $z(Hit)$ and $z(Fa)$, respectively (Macmillan & Creelman, 2005). From $z(Hit)$ and $z(Fa)$, one can derive the sensitivity d', where $d' = z(Hit) - z(Fa)$, and the response bias c, where $c = -[z(Hit) + z(Fa)]/2$. The former refers to the ability to successfully indicate whether two stimuli are the same or different. The latter refers to the tendency to answer "same" independent of the type of stimulus pair (same or different). It turns out that all participants were sensitive to differences in materials as

well as in lightings (see Supplementary Table S1 presenting the resulting d' and c values per participant). On average, they were significantly more sensitive to the material differences ($d' = 2.36\pm0.10$) than to the lighting differences ($d' = 1.82\pm0.15$). This was confirmed in a paired t test: $t(7) = 3.86, p = 0.006$. Because we found a significant difference between the averaged hit rates, paired t test, $t(7) = 3.20, p = 0.015$, but not between the averaged false alarms, paired t test, $t(7) = -1.27, p = 0.25$, the higher sensitivities for materials may be attributed to higher hit rates for materials. The average response biases for materials ($c = 0.08\pm0.06$) and for lightings ($c = 0.01\pm0.12$) were negligible and not significantly different as confirmed in a paired t test: $t(7) = -0.84, p = 0.43$. This is consistent with the observation that the usage of the four types of responses was almost equal: the sums of the columns in Figure 8 are 1.05, 1.00, 0.98, and 0.97. Finally, the largest range of individual values happened with $z(Fa)$ for lighting ($SEM = 0.18$; see Supplementary Table S1), confirming that there are individual differences comparable to those found for the performance measure in Experiment 1. In Figure 9, sensitivity d' and response bias c for materials and lightings in Experiment 2 are plotted in red. It is clear that observers had higher sensitivity for materials than for lightings. Note that, in this figure, we also show the results from Experiment 3 in blue plots.

Experiment 3: Are matching performances and discrimination accuracies within observers comparable?

Because we found similar effects and idiosyncratic differences in Experiments 1 and 2, we wanted to further investigate the relationship between the matching and discrimination performances. Thus, we conducted a third experiment consisting of two sessions, one with the matching task and the other with the category discrimination task. A different group of observers was recruited and asked to participate in both sessions in order to be able to directly compare the results of the two tasks. Both tasks were simplified by removing the brilliance lighting mode and keeping ambient and focus lighting modes only in the stimuli.

Methods

Observers

Ten inexperienced observers participated in both sessions of Experiment 3. All participants had normal
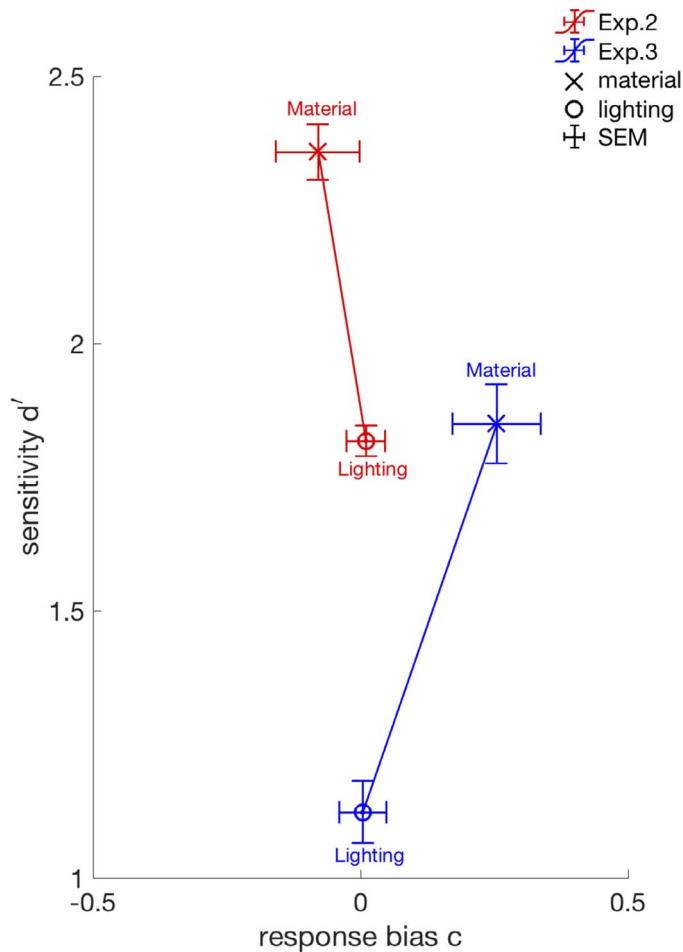
Figure 9. Sensitivity *d'* and response bias *c* for materials and lighting in Experiments 2 and 3. Red-colored plots show results from Experiment 2, and blue-colored plots show results from Experiment 3. Crosses depict results for materials; circles depict results for lighting. Each error bar depicts the corresponding standard error of the mean for both axes.

or corrected-to-normal vision. Participants read and signed the consent form before the experiments. The experiments were approved by the human research ethics committee at Delft University of Technology and conducted in accordance with the declaration of Helsinki and Dutch law.

Session 1: Simplified version of the matching task

Because the brilliance light was removed from the MixIM 1.0 interface (Figure 10), the basis images used in the mixing process are only the top two rows in Figure 1. For mixing only ambient and focus light, the mixing process for the stimuli was simply adjusted to Equation 4 with the weights as in Table 3:

$$I_{stimulus\_material} = w_{ambient} \cdot I_{ambient\_material} + w_{focus} \cdot I_{focus\_material}. \quad (4)$$
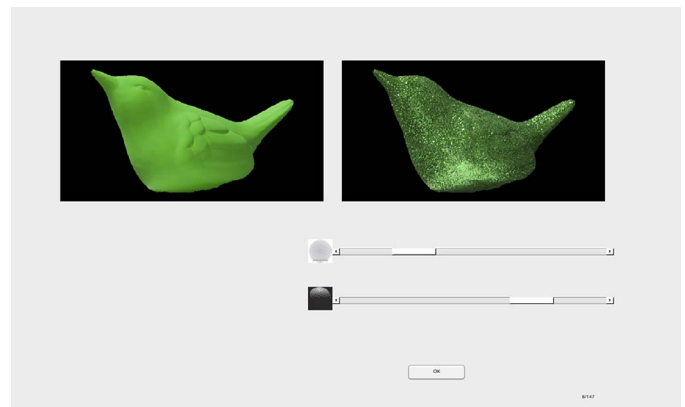

Figure 10. The interface for the first session in Experiment 3. Left: A stimulus image, consisting of a mixture of matte material in 50% ambient light and 50% focus light. Right: A probe image (glittery material mode). Top slider represents the contribution of ambient light. Bottom slider represents the contribution of focus light. In this figure, the illumination of the probe image does not match the illumination of the stimulus image.

And similarly, the mixing process for the probe becomes

$$I_{probe\_material} = w'_{ambient} \cdot I'_{ambient\_material} + w'_{focus} \cdot I'_{focus\_material}. \quad (5)$$

In this session, the four material modes in the stimuli and the four material modes in the probe images were combined with three weight combinations for the light modes, which resulted in 48 trials per run. With three repetitions plus three practice trials, there were 147 trials per observer, which resulted in a session lasting between 30 and 60 min.

Session 2: Simplified version of the four-category discrimination task

After observers finished the first session, they did a second session: the four-category discrimination task using the same interface as in Experiment 2 (Figure 7). Unlike in Experiment 2, before the actual experiment started, observers did not browse through all stimuli images. Instead, they were told all stimuli images they were about to see had appeared in the previous session. They were also told that all stimuli images in this session would be images of one of the four material

No.	$w_{ambient}$	w_{focus}
1	1	0
2	0	1
3	0.5	0.5

Table 3. Weight of each canonical lighting mode in the stimuli for Experiment 3.

modes in one of the two lighting modes, which they just manipulated by moving the sliders in the first session.

For each observer, with eight basis images as stimuli, there were 36 possible combinations, including eight "same materials same lightings," four "same materials different lightings," 12 "different materials same lightings," and 12 "different materials different lightings." To create the same number of stimuli per category, these combinations were repeated three, six, two, and two times, respectively. The resulting total number of trials was 24 per category, in total 96 trials per observer, which took approximately half an hour to finish.

Results

Matching

Because the brilliance lighting mode was removed in Experiment 3, the linear factor matrix X solved by least square fitting changed accordingly as in Equation 6:

$$[P]_{2 \times (96 \times N)} = [X]_{2 \times 2} \cdot [S]_{2 \times (96 \times N)} + [E]_{2 \times (96 \times N)}, \quad (6)$$

$$[S] = \begin{bmatrix} w_{ambient} \\ w_{focus} \end{bmatrix}, \text{ the weights of the stimuli,}$$

$$[P] = \begin{bmatrix} w'_{ambient} \\ w'_{focus} \end{bmatrix}, \text{ the weights of the probe,}$$

and the residuals $[E] = \begin{bmatrix} e_{ambient} \\ e_{focus} \end{bmatrix}$.

For N participants, there were in total $(96 \times N)$ columns in matrix S, matrix P, and matrix E when solving the Equation 6. The linear factor matrix X became 2×2, and the matrix E was again the subtraction between P and $X \cdot S$. If the matching would be veridical, S would be a 2×2 identity matrix, and matrix E would be a $2 \times (96 \times N)$ zero matrix. The matching performance could be evaluated in the same manner as in Experiment 1, i.e., taking the ratio between the sum of the diagonal values and the sum of X, i.e., $r = \sum diag(X)/\sum(X)$, which ranges from zero (only possible mathematically) to one (veridical). Note that, in Experiment 3, the chance level is 0.5, which is higher than the chance level (0.33) in Experiment 1.

The overall matching results of all observers in Experiment 3 ($N = 10$) are

$$X = \begin{bmatrix} 0.72 & 0.35 \\ 0.41 & 0.73 \end{bmatrix}, \ r = 0.66.$$

Because the chance level is 0.50, the ratio r being 0.66 shows that, overall, observers performed above chance in the matching session in Experiment 3. The bivariation plot of all observers is shown in Figure 11 in the same
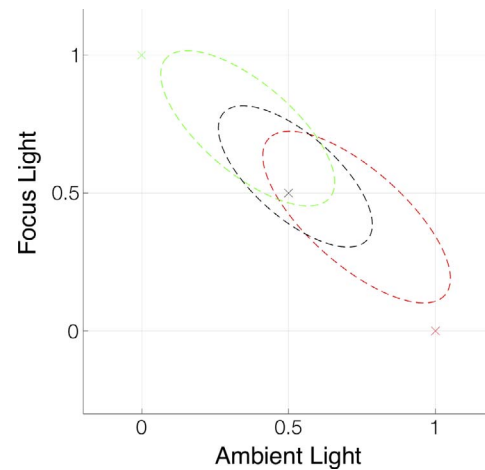

Figure 11. The bivariation plot of the overall matching results in Experiment 3 ($N = 10$). Different colors correspond to weight combinations (Table 3) of ambient light (x-axis) and focus light (y-axis) in the stimuli, which are depicted by the crosses. Specifically, the red color corresponds to the stimuli in which only ambient light was presented, the green color corresponds to the stimuli in which only focus light was presented, and the black color corresponds to the stimuli when two lighting modes were optically mixed (each 50% in the mixture). The ellipses represent one standard deviation of bivariate normal distributions fitted to the data.

format as in Figures 5 and 6 for Experiment 1. Each ellipse represents one standard deviation of bivariate normal distribution fitted to 16 data points (rendered invisible for clarity of presentation). The coordinates of the crosses depict the corresponding weight combinations of the stimuli as shown in Table 3, corresponding to the color of the ellipses. Specifically, the red color corresponds to the stimuli in which only ambient light was presented, the green color corresponds to the stimuli in which only focus light was presented, the black color corresponds to the stimuli in which both lighting modes were optically mixed 50% each in the mixture. Similar to what can be seen in Figure 5, the ellipses show a shift from the veridical values toward the center but are still in the correct order. Check Supplementary Figure S6 for individual results.

Discrimination

The results of the four-category discrimination task are shown in Figure 12. Similar to the results of Experiment 2 (Figure 8), when the materials and lightings were both the same in the stimuli image pair, observers got the highest accuracy, being 0.89. When the materials were the same and lightings were different, the accuracy decreased to 0.48. The accuracy was 0.43 when the materials were different and the lightings were the same, and 0.57 when both the materials and the lightings were different. Note that
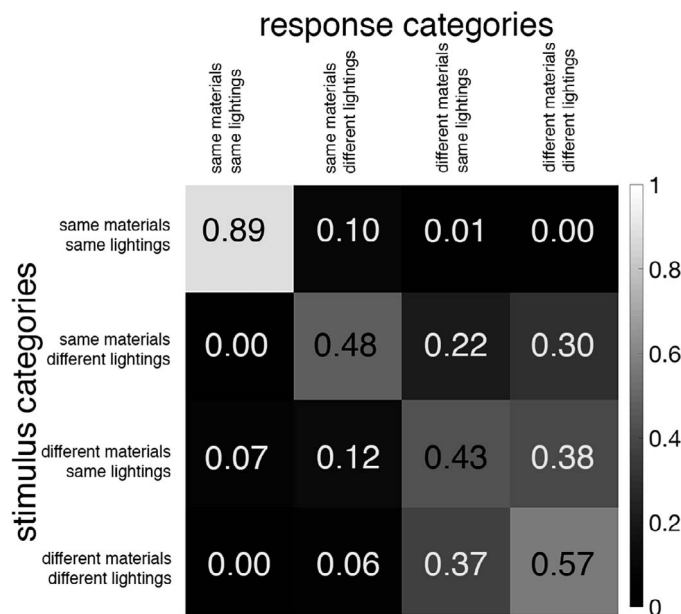
response categories

Figure 12. The fractions of responses per stimulus category of the four-category discrimination task in Experiment 3. Each row represents a stimulus, and each column represents a response category. The squares on the diagonals are the fractions of answering correctly, i.e., the accuracies.

here the chance level is 0.25, the same as in Experiment 2. We again found a strong association between the responses and the stimulus categories ($\chi^2(9) = 926.32, p < 0.001$).

Again, we implemented signal-detection theory by considering the four-category discrimination task as two yes-or-no questions: (a) "Are the materials the same?" and (b) "are the lightings the same?" As in Experiment 2, observers were all found to be sensitive to the differences in both the materials and the lightings (the resulting values of sensitivity d' and response bias c are listed in Supplementary Table S2). They were also significantly more sensitive to material differences ($d' = 1.85 \pm 0.16$) than to lighting differences ($d' = 1.12 \pm 0.16$), confirmed in a paired t test, $t(9) = 6.832, p < 0.001$. The average response bias for materials ($c = 0.25 \pm 0.07$) was not significantly different from that for lightings ($c = 0.00 \pm 0.09$), confirmed in a paired t test, $t(9) = 2.151, p = 0.06$. Unlike in Experiment 2, there was no significant difference between the averaged hit rates, paired t test, $t(9) = 0.97, p < 0.36$, but now there was one between the averaged false alarms, paired t test, $t(9) = -4.45, p = 0.002$, suggesting that the higher sensitivities for materials may be attributed to lower false alarm rates.

Comparison

To directly compare the performances of the matching task (session 1) and the four-category
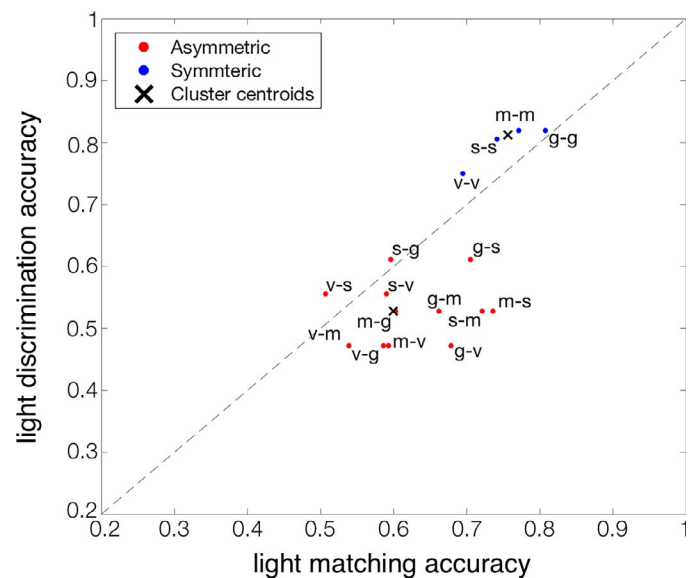

Figure 13. Comparison between the lighting-matching and the discrimination results in Experiment 3. The data points depict different material combinations with "m", "v", "s", and "g" denoting matte, velvety, specular, and glittery, respectively (e.g., "m-m" means the materials in the trial were both matte; "s-g" means the materials in the trial were specular and glittery). Colors were assigned using a *k*-means clustering algorithm for two clusters with the crosses depicting the cluster centroids. The dashed line depicts the identity line.

discrimination task (session 2), we first tested at a global level by correlating the individual light-matching accuracies with the corresponding sensitivities d' and response biases c for both materials and lighting (see Supplementary Table S2). The light-matching accuracy was found to be significantly correlated with one variable only, namely response bias c for light discrimination (negatively correlated, $r^2 = 0.40, p = 0.049$).

Subsequently, we further tested the correlation between the light-matching accuracy (the ratio *r*) and the light-discrimination accuracy (the fraction of correctly answering "same lighting") per material combination (Figure 13). Overall, a significant correlation between the light-matching and light-discrimination accuracy was found in Experiment 3 ($r^2 = 0.45, p < 0.01$). More specifically, some observations are listed below:

- For the symmetric cluster in which materials were the same (blue data points), we observed that the material combinations including velvet tend to produce lower performances in both the discrimination and matching tasks.
- For the asymmetric cluster in which materials were different (red data points), we observed that when specular material was involved, the discrimination accuracy (0.57 ± 0.04) was significantly higher

than when specular material was not involved (0.49 ± 0.03).

- The combinations with specular and glittery resulted in the highest performance, showing that those two modes interacted least of all asymmetric combinations.
- The combinations with velvety and matte gave the lowest performance among all cases, showing that these modes interacted most of all our material modes.

Discussion

In this paper, we present three experiments. In Experiment 1, we asked observers to optically mix three canonical lighting modes (ambient, focus, and brilliance) while discounting our canonical material modes (matte, velvety, specular, and glittery) in a matching task. Eleven out of 15 observers' performance levels were well above chance, and the remaining four observers performed just above chance (Figure 4). In Experiment 2, we asked observers to simultaneously discriminate materials and lightings in a four-category discrimination task and found that observers were more sensitive in discriminating our material modes than our lighting modes; i.e., they were better at judging the material modes than the lighting modes to be the same or not. In Experiment 3, we implemented a simplified version of both the matching and four-category discrimination tasks by removing the brilliance light and then asked observers to first perform the matching task and then the four-category discrimination task. Results from Experiment 3 showed that the matching and discrimination results were comparable and confirmed the asymmetric perceptual confounds between materials and lightings that we observed in Experiments 1 and 2. Across these experiments, observers were found to be more sensitive to material differences than to lightings differences.

For the matching task, an interface inspired by audio-mixing desks was tested in a previous study (Zhang et al., 2016) and further developed in this study. Here, the number of sliders in the interface was reduced from four for the material mixing in the previous study to three in Experiment 1 and two in Experiment 3 of this study for lighting matching. This actually reduced the level of complexity of manipulating the interface and increased the level of chance performance from one of four to one of three for Experiment 1 and one of two for Experiment 3 (if calculated as the ratio *r* using least squares fitting). However, the general performance of the light-matching task in this study was lower than the performance of the material-matching task in our former studies. So observers were better at discounting

our lightings in matching the optically mixed canonical material modes than discounting our materials in matching the optically mixed canonical lighting modes. This again confirms the asymmetric perceptual confounds we found in Experiments 2 and 3.

One possible cause of this asymmetric perceptual confound might be that we showed the appearance of the objects without a context. In our experiments, observers had to make judgments based purely on the objects' appearances. If observers would have access to other information about the light, such as from the background or the appearance of other objects, it might be easier for them to make more accurate judgments. Indeed, light is usually inferred by looking at the appearance of (the objects in) a scene.

Ecologically, this asymmetric confound makes sense as human beings have to recognize and interact with materials under different illumination in our daily lives. Yet most of us (except for instance lighting professionals) do not normally have the necessity to recognize or interact with different types of lightings. In fact, we may simply be used to changes of illumination in natural environments without realizing it, especially for those changes that occur naturally, which is the case for the variations and modes that we used.

It should be realized, however, that we are comparing apples and oranges (lighting and materials) and that there is no obvious physical basis to compare the magnitudes of the differences between materials and lightings. In this study, we approached this by selecting canonical modes, which are optically very different from each other and altogether span much of the reflectance (BRDF) space and descriptions of natural light fields. The limitations of our conclusions are obviously set by this choice of modes and their representations via the bird photographs. Detailed characteristics of the modes, such as lighting direction, beam width, the statistical characteristics of the brilliance lighting, and microscattering properties of the glittery flakes or velvet hairs, are expected to have an influence on the results. However, considering the coarse characteristics of the modes and especially how wide apart they are in the spaces of possible reflectance and lighting types, we reasoned that the asymmetric confound in this study suggests a more generic phenomenon with an ecologically plausible basis.

This connects to how our visual systems represent materials and lightings. In material-perception studies, instead of the "inverse optics" and the "image statistics" approaches, the "statistical appearance models" approach represents an alternative, for instance, for the study of gloss perception (Fleming, 2014). Similarly, in our studies, we presented "a painterly approach" (Zhang et al., 2016), i.e., optical mixing of canonical material or lighting modes, that allows observers to intuitively manipulate the midlevel
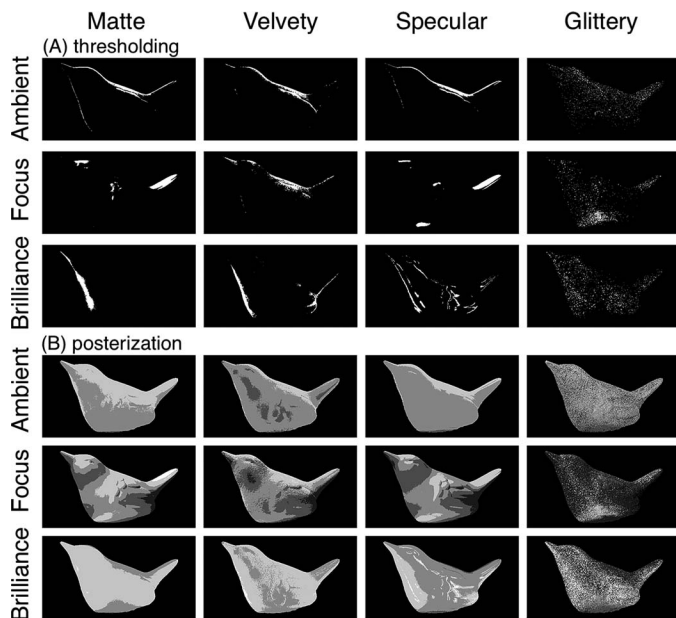
Figure 14. Examples of image analyses of the basis images (in the same format as in Figure 1). Top: The red-channel thresholding showing the upper 5% percentile of brightest pixels. Note that the thresholding level varies per image. Bottom: The green-channel of the basis images after posterization from 255 to four levels.

image cues in a weighted-mixture manner. From the results of our earlier material-matching experiment, we argued that these key midlevel image features form the triggers for material perceptions, such as the smooth shading along the surface of the matte mode, the bright contours for the velvety mode, the highlights for the specular mode, and the bright speckles all over the surfaces of the glittery mode (Zhang et al., 2016). Here, we argue that midlevel image features could also be the triggers for our lighting perceptions: the overall brightness and lack of gradients for the ambient or mathematical zeroth-order component of the light; the contrast, main highlight, and the shading gradient direction for the focus or first-order component of the light; and the contrast and spatial patterns of the glint for the brilliance or higher order components of the light (Ganslandt & Hofmann, 1992; Kelly, 1952). Close observation of our photographs in Figure 1 plus their mixtures and computer-rendered simulations may suggest that these features are, overall, less robust for variations of material than for variations of lighting (Figure 14). In Figure 14A, we show the top 5% brightest pixels in each basis image by applying thresholding to the red channel of the images. In Figure 14B, we show the shading patterns by posterizing the green channel of the basis images from 255 to four levels. The last column of the thresholded images shows that the images of glittery material are clearly dominated by the spread of the dots, i.e., the glints, that

result in the glittery appearance regardless of illumination. The images of the matte, velvety, and specular materials show otherwise spatially varied patterns. Specifically, we observed smooth shading gradients for matte mode; smooth shading gradients, bright contours, and fine-grained textures that might trigger the velvetiness in the velvety image; and the specular highlight regions spread along the curvature of the surface for specular mode (except for specular under the ambient lighting, which caused interactions with matte mode). One may argue that, in ambient lighting, the bright contours, which we suggested trigger velvetiness, can be observed in the thresholded images for matte, velvety, and specular material, too. However, by closely looking at the spread of the pixels on those bright contours in velvety images, combining the patterns of their shadings, we could discriminate velvety from matte or specular (not quantitatively though). In natural scenes with arbitrary materials and light, this difference in feature robustness would make it harder to judge the lighting than the material. Similarly, these midlevel image features varied differently for matte, specular, and glittery materials under the canonical lighting modes, being more diagnostic for our canonical materials than our canonical lightings, causing the asymmetric perceptual confounds. Simple image statistics (such as comparing the image histograms, then calculating the difference between each two images, and the correlation between each two images) could not explain the asymmetric confounds. In order to better understand what and how midlevel image features account for material and lighting perception, novel quantitative metrics are required for image analysis, such as separating specific features from object color (Klinker, Shafer, & Kanade 1987).

Conclusion

In this study, we implemented two types of tasks, namely a light-matching task and a four-category discrimination task for our canonical material and lighting modes. From the results of the light-matching tasks in Experiments 1 and 3, we found that most of our observers could match optically mixed canonical lighting modes while discounting materials although a small portion of the observers tended to use only a narrow range around the center of the possible slider positions. In particular, observers performed better when the materials in the stimulus and the probe were the same than when they were different. From the results of the four-category discrimination tasks in Experiments 2 and 3, we found that observers could discriminate our material modes better than our lighting modes. Their sensitivities for the material

discrimination were found to be higher than those for the lighting discrimination. Observers also found it difficult to discriminate lighting modes when the materials were different. Moreover, in Experiment 3, by conducting a simplified version of both matching and discrimination tasks with the same group of observers, we found that the performances of matching and discrimination task were indeed comparable.

To conclude, in all three experiments and across all observers, the sensitivities for judging the differences between our canonical material modes are higher than those for the canonical lighting modes. If materials are different, it is harder to see whether or not the illuminations are different than if materials are the same. If lightings are different, it is almost as easy to see whether the materials are different or not as when the lightings are the same. Our findings suggest that midlevel image features are more robust across different materials than across different lightings and, thus, more diagnostic for our canonical materials than our canonical lightings, causing the asymmetric perceptual confounds.

Keywords: material perception, light perception, material–lighting confounds, matching, discrimination

Acknowledgments

Commercial relationships: none.
Corresponding author: Fan Zhang.
Email: f.zhang-2@tudelft.nl.
Address: Perceptual Intelligence Laboratory, Industrial Design Engineering, Delft University of Technology, The Netherlands.

References

Anderson, B. L., & Kim, J. (2009). Image statistics do not explain the perception of gloss and lightness. *Journal of Vision*, 9(11):10, 1–17, https://doi.org/10.1167/9.11.10. [PubMed] [Article]

Blinn, J. F. (1977). Models of light reflection for computer synthesized pictures. *ACM SIGGRAPH Computer Graphics*, 11(2), 192–198.

Brainard, D. H. (1997) The Psychophysics Toolbox, *Spatial Vision 10*: 433–436.

Cook, R. L., & Torrence, K. E. (1982). A reflectance model for computer graphics. *ACM Transactions on Graphics*, 1(1), 7–24.

Doerschner, K., Boyaci, H., & Maloney, L. T. (2010). Estimating the glossiness transfer function induced by illumination change and testing its transitivity. *Journal of Vision*, 10(4):8, 1–9, https://doi.org/10.1167/10.4.8. [PubMed] [Article]

Dror, R. O., Adelson, E. H., & Willsky, A. S. (2001). Estimating surface reflectance properties from images under unknown illumination. In *Human Vision and Electronic Imaging VI* (pp. 231–243). International Society for Optics and Photonics.

Dror, R. O., Willsky, A. S., & Adelson, E. H. (2004). Statistical characterization of real-world illumination. *Journal of Vision*, 4(9):11, 821–837, https://doi.org/10.1167/4.9.1. [PubMed] [Article]

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

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 3(5):3, 347–368, https://doi.org/10.1167/3.5.3. [PubMed] [Article]

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *Journal of Vision*, 13(8):9, 1–20, https://doi.org/10.1167/13.8.9. [PubMed] [Article]

Ganslandt, R., & Hofmann, H. (1992). *Handbook of lighting design*. Brauschweig/Wiesbaden, Germany: ERCO GmbH.

Hansmann-Roth, S., Pont, S. C., & Mamassian, P. (2017). Contextual effects on real bicolored glossy surfaces. *Journal of Vision*, 17(2):17, 1–13, https://doi.org/10.1167/17.2.17. [PubMed] [Article]

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2006). How direction of illumination affects visually perceived surface roughness. *Journal of Vision*, 6(5):8, 634–648, https://doi.org/10.1167/6.5.8. [PubMed] [Article]

Ho, Y. X., Landy, M. S., & Maloney, L. T. (2008). Conjoint measurement of gloss and surface texture. *Psychological Science*, 19(2), 196–204.

Karana, E., Hekkert, P., & Kandachar, P. (2009). Meanings of materials through sensorial properties and manufacturing processes. *Materials & Design*, 30(7), 2778–2784.

Kartashova, T., de Ridder, H., te Pas, S. F., Schoemaker, M., & Pont, S. C. (2015). The visual light field in paintings of Museum Prinsenhof: Com-

paring settings in empty space and on objects. In B. E. Rogowitz, T. N. Pappas, H. de Ridder (Eds.), *SPIE/IS&T Electronic Imaging* (pp. 93941M–93941M). International Society for Optics and Photonics.

Kartashova, T., Sekulovski, D., de Ridder, H., te Pas, S. F., & Pont, S. C. (2016). The global structure of the visual light field and its relation to the physical light field. *Journal of Vision*, 16(10):9, 1–16, https://doi.org/10.1167/16.10.9. [PubMed] [Article]

Kelly, R. (1952). Lighting as an integral part of architecture. *College Art Journal*, 12(1), 24–30.

Kim, J., Marlow, P., & Anderson, B. L. (2011). The perception of gloss depends on highlight congruence with surface shading. *Journal of Vision*, 11(9): 4, 1–19, https://doi.org/10.1167/11.9.4. [PubMed] [Article]

Kim, M., Wilcox, L. M., & Murray, R. F. (2016). Perceived three-dimensional shape toggles perceived glow. *Current Biology*, 26(9), R350–R351.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, 36(14), 1–16.

Klinker, G. J., Shafer, S. A., & Kanade, T. (1987). Using a color reflection model to separate highlights from object color. In *Proceedings - First international conference on computer vision. Vol. 87* (pp. 145–150). New York, NY: IEEE.

Koenderink, J., & Pont, S. (2003). The secret of velvety skin. *Machine Vision and Applications*, 14(4), 260–268.

Koenderink, J. J., Pont, S. C., van Doorn, A. J., Kappers, A. M., & Todd, J. T. (2007). The visual light field. *Perception*, 36(11), 1595–1610.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Marlow, P. J., Kim, J., & Anderson, B. L. (2012). The perception and misperception of specular surface reflectance. *Current Biology*, 22(20), 1909–1913.

Marr, D. (1982). *Vision*. San Francisco: W. H. H. Freeman.

Morgenstern, Y., Geisler, W. S., & Murray, R. F. (2014). Human vision is attuned to the diffuseness of natural light. *Journal of Vision*, 14(9):15, 1–18, https://doi.org/10.1167/14.9.15. [PubMed] [Article]

Morgenstern, Y., Geisler, W. S., & Murray, R. F. (2015, March). The role of natural lighting diffuseness in human visual perception. In *SPIE/IS&T Electronic Imaging* (pp. 93940P–93940P). International Society for Optics and Photonics, https://doi.org/10.1117/12.2086258.

Morgenstern, Y., Murray, R. F., & Harris, L. R. (2011). The human visual system's assumption that light comes from above is weak. *Proceedings of the National Academy of Sciences, USA*, 108(30), 12551–12553.

Motoyoshi, I., & Matoba, H. (2012). Variability in constancy of the perceived surface reflectance across different illumination statistics. *Vision Research*, 53(1), 30–39.

Motoyoshi, I., Nishida, S. Y., Sharan, L., & Adelson, E. H. (2007, May 10). Image statistics and the perception of surface qualities. *Nature*, 447(7141), 206–209.

Mury, A. A., Pont, S. C., & Koenderink, J. J. (2007). Light field constancy within natural scenes. *Applied Optics*, 46(29), 7308–7316.

Mury, A. A., Pont, S. C., & Koenderink, J. J. (2009). Structure of light fields in natural scenes. *Applied Optics*, 48(28), 5386–5395.

Nishida, S. Y., Sawayama, M., & Shimokawa, T. (2015). Material-dependent shape distortion by local intensity order reversal. *Journal of Vision*, 15(12):940, https://doi.org/10.1167/15.12.940. [Abstract]

Nishida, S. Y., & Shinya, M. (1998). Use of image-based information in judgments of surface-reflectance properties. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 15(12), 2951–2965.

Olkkonen, M., & Brainard, D. H. (2010). Perceived glossiness and lightness under real-world illumination. *Journal of Vision*, 10(9):5, 1–19, https://doi.org/10.1167/10.9.5. [PubMed] [Article]

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically based light reflection model for image synthesis. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 55–64). New York, NY: ACM Press/Addison-Wesley Publishing Co.

Pelli, D. G. (1997) The VideoToolbox software for visual psychophysics: Transforming numbers into movies, *Spatial Vision 10*, 437–442.

Pizlo, Z. (2001). Perception viewed as an inverse problem. *Vision Research*, 41(24), 3145–3161.

Poggio, T., & Koch, C. (1985). Ill-posed problems in early vision: From computational theory to analogue networks. *Proceedings of the Royal Society of London B: Biological Sciences*, 226(1244), 303–323.

Poggio, T., Torre, V., & Koch, C. (1985, Setpember 26). Computational vision and regularization theory. *Nature*, 317(6035), 314–319.

Pont, S. C., & Koenderink, J. J. (2007). Matching illumination of solid objects. *Attention, Perception, & Psychophysics*, 69(3), 459–468.

Pont, S. C., & te Pas, S. F. (2006). Material-illumination ambiguities and the perception of solid objects. *Perception*, 35(10), 1331–1350.

Schirillo, J. A. (2013). We infer light in space. *Psychonomic Bulletin & Review*, 20(5), 905–915.

Sharan, L., Rosenholtz, R., & Adelson, E. (2009). Material perception: What can you see in a brief glance? *Journal of Vision*, 9(8): 784, https://doi.org/10.1167/9.8.784. [Abstract]

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of Vision*, 14(9):12, 1–24, https://doi.org/10.1167/14.9.12. [PubMed] [Article]

Schütt, H. H., Baier, F., & Fleming, R. W. (2016). Perception of light source distance from shading patterns. *Journal of Vision*, 16(3):9, 1–20, https://doi.org/10.1167/16.3.9. [PubMed] [Article]

te Pas, S. F., & Pont, S. C. (2005). A comparison of material and illumination discrimination performance for real rough, real smooth and computer generated smooth spheres. In *Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization* (pp. 75–81). New York, NY: ACM.

Toscani, M., Gegenfurtner, K. R., & Doerschner, K. (2017). Differences in illumination estimation in #thedress. *Journal of Vision*, 17(1):22, 1–14, https://doi.org/10.1167/17.1.22. [PubMed] [Article]

Ward, G. J. (1992). Measuring and modeling anisotropic reflection. *ACM SIGGRAPH Computer Graphics*, 26(2), 265–272.

van Assen, J. J. R., Wijntjes, M. W., & Pont, S. C. (2016). Highlight shapes and perception of gloss for real and photographed objects. *Journal of Vision*, 16(6):6, 1–14, https://doi.org/10.1167/16.6.6. [PubMed] [Article]

van Doorn, A. J., Koenderink, J. J., & Wagemans, J. (2011). Light fields and shape from shading. *Journal of Vision*, 11(3):21, 1–21, https://doi.org/10.1167/11.3.21. [PubMed] [Article]

Vangorp, P., Laurijssen, J., & Dutré, P. (2007). The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics*, 26(3), 77.

Wijntjes, M. W., & Pont, S. C. (2010). Illusory gloss on Lambertian surfaces. *Journal of Vision*, 10(9):13, 1–12, https://doi.org/10.1167/10.9.13. [PubMed] [Article]

Xia, L., Pont, S. C., & Heynderickx, I. (2013). Probing light in real scenes using optical mixtures. In *Proceedings of the ACM Symposium on Applied Perception* (pp. 137–137). New York, NY: ACM.

Xia, L., Pont, S. C., & Heynderickx, I. (2014). The visual light field in real scenes. *i-Perception*, 5(7), 613–629.

Xia, L., Pont, S. C., & Heynderickx, I. (2016). Effects of scene content and layout on the perceived light direction in 3D spaces. *Journal of Vision*, 16(10):14, 1–13, https://doi.org/10.1167/16.10.14. [PubMed] [Article]

Xia, L., Pont, S. C., & Heynderickx, I. (2017a). Light diffuseness metric part 1: Theory. *Lighting Research & Technology*, 49(4), 411–427, https://doi.org/10.1177/1477153516631391.

Xia, L., Pont, S. C., & Heynderickx, I. (2017b). Light diffuseness metric, Part 2: Describing, measuring and visualising the light flow and diffuseness in three-dimensional spaces. *Lighting Research & Technology*, 49(4), 428–445, https://doi.org/10.1177/1477153516631392.

Zhang, F., de Ridder, H., & Pont, S. (2015). The influence of lighting on visual perception of material qualities. In *SPIE/IS&T Electronic Imaging* (pp. 93940Q–93940Q). International Society for Optics and Photonics.

Zhang, F., de Ridder, H., Fleming, R. W., & Pont, S. (2016). MatMix 1.0: Using optical mixing to probe visual material perception. *Journal of Vision*, 16(6): 11, 1–18, https://doi.org/10.1167/16.6.11. [PubMed] [Article]