Visual discrimination of optical material properties: A large-scale study

Masataka Sawayama
Inria, Bordeaux, France
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa, Japan

Yoshinori Dobashi
Information Media Environment Laboratory, Hokkaido University, Hokkaido, Japan
Prometech CG Research, Tokyo, Japan

Makoto Okabe
Department of Mathematical and Systems Engineering, Graduate School of Engineering, Shizuoka University, Shizuoka, Japan

Kenchi Hosokawa
Advanced Comprehensive Research Organization, Teikyo University, Tokyo, Japan
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa, Japan

Takuya Koumura
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa, Japan

Toni P. Saarela
Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Maria Olkkonen
Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

Shin'ya Nishida
Cognitive Informatics Lab, Graduate School of informatics, Kyoto University, Kyoto, Japan
NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa, Japan

Complex visual processing involved in perceiving the object materials can be better elucidated by taking a variety of research approaches. Sharing stimulus and response data is an effective strategy to make the results of different studies directly comparable and can assist researchers with different backgrounds to jump into the field. Here, we constructed a database containing several sets of material images annotated with visual discrimination performance. We created the material images using physically based computer graphics techniques and conducted psychophysical experiments with them in both laboratory and crowdsourcing settings. The observer's task was to discriminate materials on one of six dimensions (gloss contrast, gloss

Citation: Sawayama, M., Dobashi, Y., Okabe, M., Hosokawa, K., Koumura, T., Saarela, T. P., Olkkonen, M., & Nishida, S. (2022). Visual discrimination of optical material properties: A large-scale study. *Journal of Vision*, *22*(2):17, 1–24, https://doi.org/10.1167/jov.22.2.17.

https://doi.org/10.1167/jov.22.2.17 Received December 1, 2019; published February 23, 2022 ISSN 1534-7362

distinctness of image, translucent vs. opaque, metal vs. plastic, metal vs. glass, and glossy vs. painted). The illumination consistency and object geometry were also varied. We used a nonverbal procedure (an oddity task) applicable for diverse use cases, such as cross-cultural, cross-species, clinical, or developmental studies. Results showed that the material discrimination depended on the illuminations and geometries and that the ability to discriminate the spatial consistency of specular highlights in glossiness perception showed larger individual differences than in other tasks. In addition, analysis of visual features showed that the parameters of higher order color texture statistics can partially, but not completely, explain task performance. The results obtained through crowdsourcing were highly correlated with those obtained in the laboratory, suggesting that our database can be used even when the experimental conditions are not strictly controlled in the laboratory. Several projects using our dataset are underway.

Introduction

Humans can visually recognize a variety of material properties of the objects they encounter daily. Although material properties, such as glossiness and wetness, substantially contribute to recognition, the contributions of value-based decision making, motor control, and computational and neural mechanisms underlying material perception had been overlooked until relatively recently—for a long time, vision science mainly used simple artificial stimuli to elucidate the underlying brain mechanisms. In the last two decades, however, along with advancements in computer graphics and machine vision, material perception has become a major topic in vision science (Adelson, 2001; Fleming, 2017; Nishida, 2019).

Visual material perception can be considered to be an estimation of material-related properties from an object image. For example, the perception of gloss and matte entails a visual computation of the specular and diffuse reflections of the surface, respectively. However, psychophysical studies have shown that human gloss perception does not have robust constancy against changes in surface geometry and illumination (e.g., Nishida & Shinya, 1998; Fleming, Dror, & Adelson, 2003), the other two main factors of image formation. Such estimation errors have provided useful information as to what kind of image cues humans use to estimate gloss. A significant number of psychophysical studies have been carried out not only on gloss but also on other optical material properties (e.g., translucency, transparency, wetness) (Fleming & Bulthoff, 2005; Motoyoshi, 2010; Xiao et al., 2014; Sawayama, Adelson, & Nishida, 2017a, Liao, Sawayama, & Xiao, 2022) and mechanical material properties (e.g., viscosity, elasticity) (Kawabe, Maruya, Fleming, & Nishida, 2015; Paulun et al., 2017; van Assen, Barla & Fleming, 2018; van Assen, Nishida, & Fleming, 2020). Neurophysiological and neuroimaging studies have revealed various neural mechanisms underlying material perception (Kentridge, Thomson, & Heywood, 2012; Nishio, Goda, & Komatsu, 2012; Nishio, Shimokawa, Goda, & Komatsu, 2014; Miyakawa et al., 2017). Some recent studies have also focused on developmental, environmental, and clinical factors of material processing (Yang, Kanazawa, Yamaguchi, & Motoyoshi, 2015; Goda, Yokoi, Tachibana, Minamimoto, & Komatsu, 2016; Ohishi et al. 2018). For example, Goda et al. (2016) showed in their monkey functional magnetic resonance imaging study that the visuohaptic experience of material objects alters the visual cortical representation. In addition, large individual differences in the perception of colors and materials depicted in one photo (#TheDress) have attracted much interest and provoked intense discussions (Brainard & Hurlbert, 2015; Gegenfurtner, Bloj, & Toscani, 2015).

A promising strategy for a more global understanding of material perception is to promote multidisciplinary studies comparing behavioral and physiological responses of humans and animals obtained under a variety of developmental, environmental, cultural, and clinical conditions. There are two problems, however. One lies in the high degree of freedom in selecting experimental stimulus parameters and task procedures. Because the appearance of a material depends not only on reflectance parameters but also on geometry and illumination, all of which are high dimensional, the use of different stimuli (and different tasks) in different studies could impose serious limitations on direct data comparisons. The other problem is the technical expertise necessary for rendering realistic images, which could discourage researchers unfamiliar with graphics from beginning material perception studies.

Aiming at removing these obstacles, we attempted to build a database that can be shared among multidisciplinary material studies. We rendered several sets of material images. The images in each set were changed in one of material dimensions in addition to illumination and viewing conditions. We then measured the behavioral performance for those image sets using a large number of "standard" observers. We used a simple task that can be used in a variety of human, animal, and computational studies. By using our database, one would be able to efficiently begin a new study, shortening the time for stimulus preparation, as well as time for control data collection with standard human observers.

Specifically, we selected six dimensions of material property (Figure 1). These dimensions have been extensively studied in the past material perception studies. Most of them can be unambiguously manipulated by changing the corresponding rendering parameters. Although we attempted to cover a wide range of optical material topics, we do not believe that

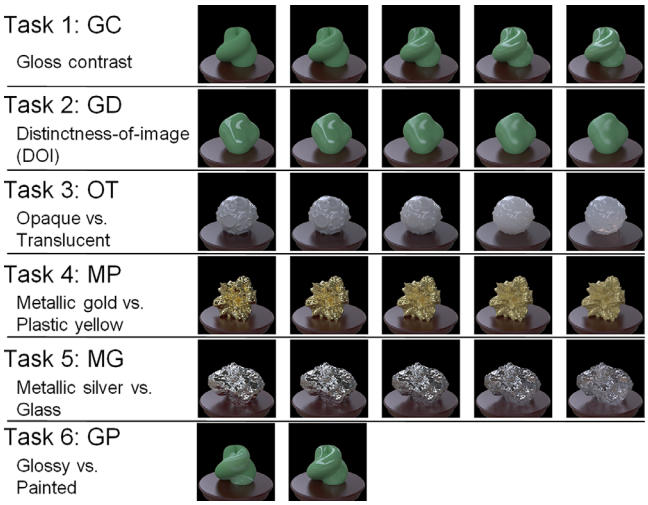

Figure 1. Schematic overview of six tasks recorded in the database.

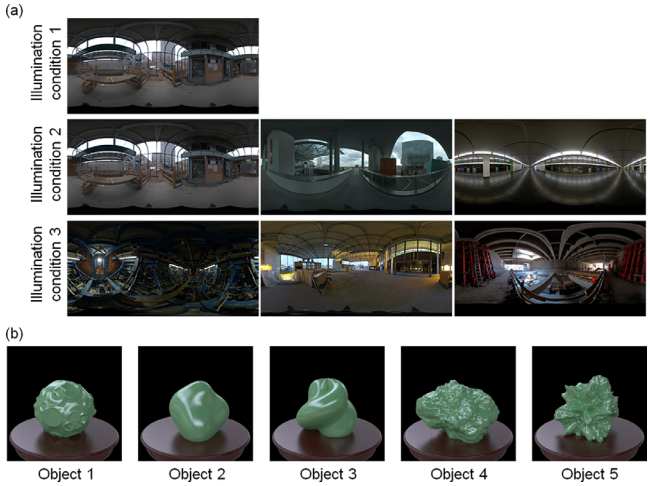

Figure 2. (a) Illumination conditions. Object images were rendered with six global illumination environments and were presented to observers under three illumination conditions. Under illumination condition 1, a stimulus display consisted of four objects (same shape, different poses) rendered with the same illumination environment. Under illumination condition 2, a stimulus display consisted of three objects (same shape, same pose) rendered with slightly different (in terms of their pixel histograms) light probes. Under illumination condition 3, a stimulus display consisted of three objects (same shape, same pose) rendered with largely different illumination environments. (b) Geometrical conditions. We used five different object shapes for each material task under each illumination condition. The stimulus condition is also summarized in Table 1.

we have assembled an exclusive list of critical material properties that vision science should challenge. Our intention is not to build the standard database for all material recognition research, but to establish one primitive test set that promotes further examination of the previous findings on material recognition in more diverse research contexts (see Discussion).

Three of these dimensions are related to gloss (Figure 1, tasks 1, 2, and 6), the most widely investigated material attribute (Pellacini, Ferwerda, & Greenberg, 2000; Fleming et al., 2003; Motoyoshi, Nishida, Sharan, & Adelson, 2007; Olkkonen & Brainard, 2010; Doerschner, Fleming, Yilmaz, Schrater, Hartung, & Kersten, 2011; Kim, Marlow, & Anderson, 2011; Marlow, Kim, & Anderson, 2011; Kentridge et al., 2012; Marlow et al., 2012; Nishio et al., 2014; Sun, Ban, Di Luca, & Welchman, 2015; Adams, Kerrigan, & Graf, 2016; Miyakawa, Banno, Abe, Tani, Suzuki, & Ichinohe, 2017; Schmid, Barla, & Doerschner, 2021; Storrs, Anderson, & Fleming, 2021). We controlled the contrast gloss and gloss distinctness of image (DOI) as in previous studies (Pellacini et al., 2000; Fleming et al., 2003; Nishio et al., 2014). For example, Nishio et al. (2014) found neurons in the inferior temporal cortex of monkeys that selectively and parametrically respond to gloss changes in these two dimensions. We also controlled the spatial consistency of specular highlights, which is another stimulus manipulation of gloss perception (Figure 1, task 6). By breaking the spatial consistency, some highlights look like albedo changes by white paint (Beck & Prazdny, 1981; Kim et al., 2011; Marlow et al., 2011; Sawayama & Nishida, 2018). In addition to gloss perception, translucency perception has also been widely investigated (Fleming & Bülthoff, 2005; Motoyoshi, 2010; Gkioulekas, Xiao, Zhao, Adelson, Zickler, & Bala, 2013; Nagai, Ono, Tani, Koida, Kitazaki, & Nakauchi, 2013; Xiao et al., 2014; Chadwick, Cox, Smithson, & Kentridge, 2018; Gigilashvili, Thomas, Hardeberg, & Pedersen, 2021). We adopted the task of discriminating opaque from translucent objects by controlling the thickness of the translucent media (Figure 1, task 3). Furthermore, we adopted the task of plastic-yellow/gold discrimination (task 4, MP) (Okazawa, Koida, & Komatsu, 2011) and glass/silver discrimination (task 5, MG) (Kim & Marlow, 2016; Tamura, Prokott, & Fleming, 2019).

We used an oddity task (Figure 3) to evaluate the capability of discriminating each material dimension. We chose this task because it requires neither complex verbal instruction nor verbal responses by the observer. Therefore, it can be applied to a wide variety of observers including infants, animals, and machine vision algorithms, and their task performances can be directly compared. Indeed, several research projects using our dataset are underway (see the Discussion section).

To control the task difficulty, we varied the value of the parameter of each material dimension. In addition, we manipulated the stimulus in two ways that affected the task difficulty. First, we set three illumination conditions. One set of stimuli included images of different poses taken in identical illumination environments (Figure 2a, illumination condition 1); the

Illumination	Task 1 (GC)	Task 2 (GD)	Task 3 (OT)	Task 4 (MP)	Task 5 (MG)	Task 6 (GP)
1	Object (5), illumination (1), pose (5)	Object (5), illumination (1), pose (5)	Object (5), illumination (1), pose (5)	Object (5), illumination (1), pose (5)	Object (5), illumination (1), pose (5)	Object (5), illumination (1), pose (5)
2	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	—
3	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	Object (5), illumination (3), pose (1)	Object (5), illumination, (3), pose (1)	—

Table 1. Summary of stimulus conditions. Digits in parentheses indicate the number of each condition.

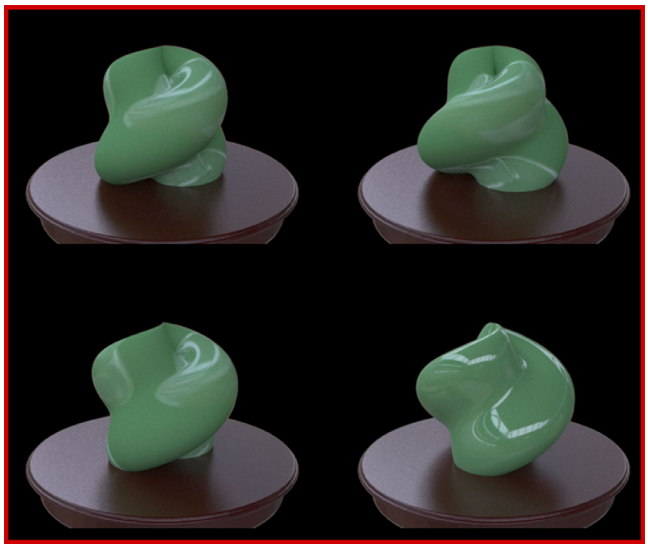

Figure 3. Example of a four-object oddity task (illumination condition 1) used for collecting standard observer data. The observers were asked to select which image was the odd one out in the four images. We did not tell the observer that the experiment was on material recognition. We conducted experiments both in the laboratory and through crowdsourcing.

second set contained stimuli of identical poses taken in slightly different illumination environments (Figure 2a, illumination condition 2); and the third set contained identical poses taken in largely different illumination environments (Figure 2a, illumination condition 3). Then, we used the five different object geometries for each task (Figure 2b).

We wished to collect data from a large number of observers. A laboratory experiment affords control over the stimulus presentation environment but is unsuited to collecting a large amount of data from numerous participants. In contrast, one can collect a lot of data through crowdsourcing, at the expense of reliable stimulus control. To overcome this trade-off, we conducted identical psychophysical experiments both in the laboratory and through crowdsourcing. This enabled us to evaluate individual difference distributions along with the effects of environmental factors on task performance.

In sum, we made a large set of image stimuli for evaluations of visual discrimination performance on six material dimensions (gloss contrast, gloss DOI, translucency–opaque, plastic–gold, glass–silver, and glossy–painted) and measured a large number of adult human observers performing oddity tasks in the laboratory and through crowdsourcing. The tasks had three illumination conditions and five object geometries. Although the original motivation of this project was to make a standard stimulus–response dataset of material recognition for promotion of multidisciplinary studies, it also has its own scientific value as, to the best of our knowledge, it is the first systematic comparison of the effects of illumination condition and object geometry, as well as of individual variations across a variety of material dimensions. Our data include several novel findings, as reported below.

Methods

We evaluated the observers' performance of six material recognition tasks. We selected tasks that had been used in previous material studies: (1) gloss contrast discrimination (GC); (2) gloss DOI discrimination (GD); (3) opaque versus translucent (OT); (4) metallic gold versus plastic yellow (MP); (5) metallic silver versus glass (MG); and (6) glossy versus painted (GP). For each task, we used five geometry models. We used six global illuminations for tasks 1 to 5 and one for task 6. We conducted behavioral experiments using an oddity task, which can be used even with human babies, animals, and brain-injured participants, because it does not entail complex verbal instructions. In the experiment, the observers were asked to select the stimulus that represented an oddity among three or four object stimuli. They were not given any feedback about whether or not their responses

were correct. We controlled the task difficulty by changing the illumination and material parameters. To test the generality of the resultant database, we conducted identical experiments in the laboratory and through crowdsourcing. Our dataset is available at https://github.com/mswym/material_dataset.

Image generation for making standard image database

We utilized the physically based rendering software called Mitsuba (Jakob, 2010) to make images of objects consisting of different materials, and we controlled six different material dimensions.

Task 1 (GC) and task 2 (GD) conditions

To control the material property of the gloss discrimination tasks, we used the perceptual light reflection model proposed by Pellacini et al. (2000). They constructed a model based on the results of psychophysical experiments using stimuli rendered by the Ward reflection model (Ward, 1992) and rewrote the Ward model parameters in perceptual terms. The model of Pellacini et al. has two parameters, d and c, which roughly correspond to the DOI gloss and the contrast gloss, respectively, of Hunter (1937). The difficulty of our two gloss discrimination tasks was controlled by separately modulating these two parameters.

The parameter space of the Ward reflection model can be described as follows:

$$\rho\left(\theta_i, \phi_i, \theta_o, \phi_o\right) = \frac{\rho_d}{\pi} + \rho_s \frac{\exp\left[-\tan^2\delta/\alpha^2\right]}{4\pi\alpha^2\sqrt{\cos\theta_i \cos\theta_o}},$$

where $\rho(\theta_i, \phi_i, \theta_o, \phi_o)$ is the surface reflection model, and θ_i, ϕ_i and θ_o, ϕ_o are the incoming and outgoing directions, respectively. The model has three parameters; ρ_d is the diffuse reflectance of a surface, ρ_s is the energy of its specular component, and α is the spread of the specular lobe. As noted earlier, Pellacini et al. (2000) defined two perceptual dimensions, c and d, on the basis of the Ward model parameters, such that d corresponds to DOI gloss and is calculated from α, and c corresponds to perceptual glossiness contrast and is calculated from ρ_s and ρ_d using the following formula:

$$d = 1 - \alpha$$
$$c = \sqrt[3]{\rho_s + \frac{\rho_d}{2}} - \sqrt[3]{\frac{\rho_d}{2}}.$$

Although more physically feasible bidirectional reflectance distribution function (BRDF) models than the Ward model have been proposed for gloss simulation (Ashikmin, Premože, & Shirley, 2000; Walter, Marschner, Li, & Torrance, 2007), we based ours on the Ward model because it has been used in many previous psychophysics and neuroscience studies (Nishio et al., 2014).

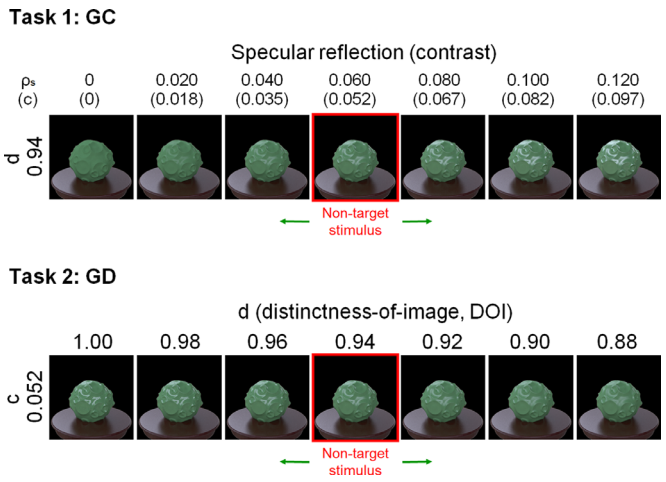

Figure 4. Material examples of tasks 1 (GC) and 2 (GD). For task 1 (GC), the specular reflectance of the odd target stimulus was varied from 0.00 to 0.12. The non-target stimuli that were presented as the context objects in each task had specular reflectance of 0.06. For task 2 (GD), the DOI parameter of the target specular reflection was varied from 1.00 to 0.88, and that of the non-target stimuli was 0.94.

For the task of gloss discrimination in the contrast dimension, the specular reflectance (ρ_s) was varied in a range from 0.00 to 0.12 in steps of 0.02 while keeping the diffuse reflectance (ρ_d) constant (0.416), resulting in the contrast parameters 0, 0.018, 0.035, 0.052, 0.067, 0.082, and 0.097. The DOI (d) was the fixed value (0.94) (Figure 4, task 1). As c approaches 0, the object appears to have a matte surface. The specular reflectance (ρ_s) of the non-target stimulus in the task was 0.06.

For the task of gloss discrimination in the DOI dimension, the parameter d was varied from 0.88 to 1.00 in 0.02 steps while keeping ρ_s constant (0.06) (Figure 4, task 2). As d approaches 1.00, the highlights of the object appear sharper. The DOI parameter (d) of the non-target stimuli was 0.94.

Task 3 (OT) conditions

To make translucent materials, we used the function of homogeneous participating medium implemented in the Mitsuba renderer. In this function, a flexible homogeneous participating medium is embedded in each object model. The intensity of the light that travels in the medium is decreased by scattering and absorption and is increased by nearby scattering. The parameters of the absorption and scattering coefficients of the medium describe how the light is decreased. We used the parameters of the "whole milk" measured by Jensen, Marschner, Levoy, M., and Hanrahan (2001). The parameter of the phase function describes the directional scattering properties of the medium. We used an isotropic phase function. To control the task difficulty, we modulated the scale parameter of the

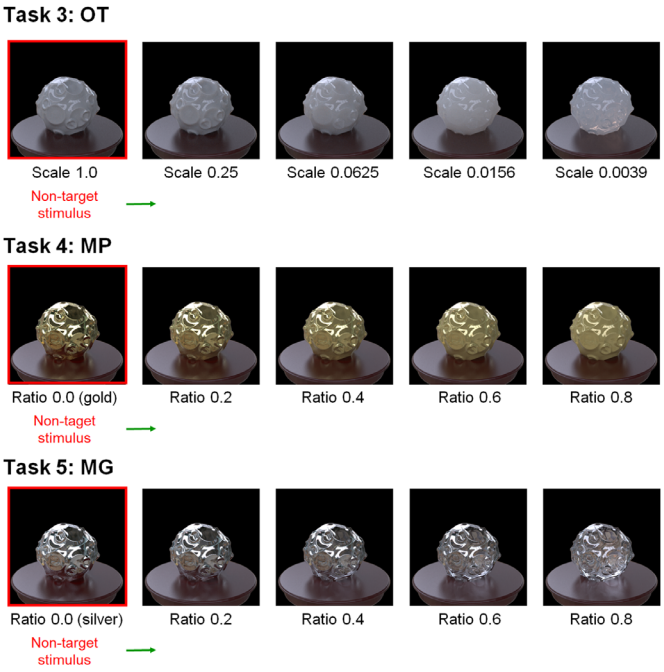

Figure 5. Material examples of tasks 3 (OT), 4 (MP), and 5 (MG). For task 3 (OT), the scale of the volume media that consisted of milk was varied from 1.0 to 0.0039. For task 4 (MP) and 5 (MG), the blending ratio of the two materials was varied from 0.0 to 0.8. The non-target stimuli in the tasks were as shown in the legend.

scattering and absorption coefficients. The parameter describes the density of the medium. The smaller the scale parameter is, the more translucent the medium becomes. The scale parameter was varied as follows: 0.0039, 0.0156, 0.0625, 0.25, and 1.00 (Figure 5, task 3). The scale parameter of the non-target stimulus in the task was 1.00. In addition, the surface of the object was modeled as a smooth dielectric material to produce strong specular highlights, as in previous studies (Gkioulekas et al., 2013; Xiao et al., 2014). That is, non-target objects were always opaque, and the degree of transparency of the target object was changed.

Task 4 (MP) conditions

To morph the material between gold and plastic yellow, we utilized a linear combination of gold and plastic BRDFs, implemented in the Mitsuba renderer. By changing the weight of the combination, the appearance of a material (e.g., gold) can be modulated toward that of the other material (e.g., plastic yellow). In this task, the weight was varied in a range from 0.00 to 0.80 in steps of 0.20 (Figure 5, task 4). The parameter of the non-target stimulus was 0, at which the material appeared to be pure gold.

Task 5 (MG) conditions

Similar to task 4, we utilized a linear combination of dielectric glass and silver materials, also implemented in the Mitsuba renderer. The weight of the combination was varied from 0.00 to 0.80. The parameter of the non-target stimulus was 0, at which the material appeared to be pure silver (Figure 5, task 5).

As noted above, for tasks 3, 4, and 5 in which the parameters of the target stimulus were varied between two material states (i.e., opaque vs. transparent, metallic vs. plastic, and metallic vs. glass), we placed the non-target objects at one end (i.e., one of two material states). If we placed the non-target stimuli in the middle of the stimulus variable, as in tasks 1 and 2, and when the difference between the target and non-target stimuli was small, the display contained only ambiguous material objects. In such cases, the observers might not pay attention to the material dimension relevant to the task. By placing the non-target at one extreme value, we could make the stimulus display always contain the object images in a specific material state, helping participants focus on the task relevant material dimension.

Task 6 (GP) conditions

The skewed intensity distribution due to specular highlights of an object image can be a diagnostic cue for gloss perception (Motoyoshi et al., 2007). However, when the specular highlights are inconsistent in terms of their position and/or orientation with respect to the diffuse shading component, they look more like white blobs produced by surface reflectance changes even if the intensity distribution is kept constant (Beck & Prazdny; 1981; Anderson & Kim, 2009; Kim et al., 2011; Marlow et al., 2011; Sawayama & Nishida, 2018). For our last task of glossy objects versus matte objects with white paint, we rendered the glossy objects on the basis of the model of Pellacini et al. (2000). The parameter c was set to 0.067, and the parameter d ranged from 0.88 to 1.00 in steps of 0.04 (Figure 6, lower). Considering material naturalness, these objects may not be typically encountered in the real world, but this task is theoretically important because it will provide insights into the underlying visual computation of material recognition.

To make object images with inconsistent highlights (white paints), we rendered each scene twice with different object materials with identical shapes. First, we rendered a glossy object image by setting the diffuse reflectance to 0 (i.e., the image including only specular highlights). The rendered image of specular highlights was a two-dimensional texture for the second rendering. We eliminated the brown table when rendering the first scene. Next, we rendered a diffuse object image (i.e., one without specular reflection with the texture of specular highlights). The object and illumination for the first and second renderings were the same. We mapped the specular image rendered in one object pose to the three-dimensional geometry by spherical mapping and repeating the image. Because the position

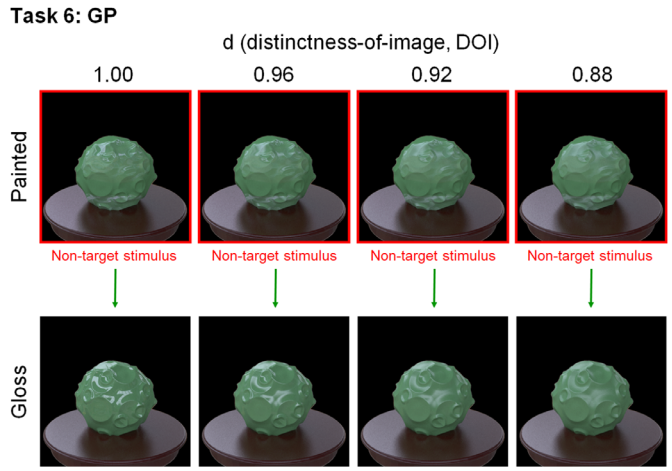

Figure 6. Material examples of task 6. The DOI of the specular reflection was varied from 1.00 to 0.88. This parameter was the same for the non-target painted objects and the target glossy object in each stimulus display.

of the texture mapping was randomly determined, the highlight texture positions were inconsistent with diffuse shadings. We varied the parameter *d* of the first rendering from 1.00 to 0.88 (Figure 6, lower). After we rendered the inconsistent-highlights image, the color histogram of the image was set to that of a consistent glossy object image by using a standard histogram matching method (Sawayama & Nishida, 2018).

We made task 6 only under illumination 1. This is because it was difficult to match the color distributions of the target and non-target stimuli for illuminations 2 and 3, where one stimulus set was rendered under different illuminations. If we matched the color histograms of the objects under these conditions, the colors of the objects could be incongruent with their background colors (i.e., the table and the shadow in this scene). This could produce another cue to find an outlier, thus making these conditions inappropriate for the task purpose.

Geometry

For each material, we rendered the object images by using five different abstract geometries (Figure 2b). These geometries were made from a sphere by modulating each surface normal direction with different kinds of noise using ShapeToolbox (Saarela & Olkkonen, 2016; Saarela, 2018). Specifically, Object_1 was made from modulations of low-spatial-frequency noise and crater-like patterns (the source code of this geometry is available on the web, http://saarela.github.io/ShapeToolbox/gallery-moon.html). Object_2 was a bumpy sphere modulated by lowpass bandpass noise. Object_3 was a bumpy sphere modulated by sine-wave noise. Object_4 and Object_5 were bumpy spheres modulated by Perlin noise. These objects were also rendered using ShapeToolbox.

Five samples were too small to systematically vary shape parameters. Instead, we handcrafted sphere-based abstract shapes in such a way expected to maximize the shape diversity. It is known that, even when rendering with the same reflectance function (BRDF), objects with smooth, low-frequency surface modulations and those with spiky, high-frequency surface modulations could have very different material appearance (Nishida & Shinya, 1998, Vangorp, Laurijssen, & Dutré, 2007). We therefore created five geometries with a variety of low and high spatial frequency surface modulations to see human material perception under widely different geometry conditions.

Illumination and pose

We used six high-dynamic-range light-probe images as illuminations for rendering. These images were obtained from Bernhard Vogl's light probe database (http://dativ.at/lightprobes/). To vary the task difficulty, we used three illumination conditions (illumination conditions 1, 2, and 3) (Figure 2a). Under illumination condition 1, the observers selected one oddity from four images in a task. We rendered the images by using an identical light probe: "Overcast Day/Building Site (Metro Vienna)." We prepared five poses for each task of illumination condition 1 by rotating each object in 36° steps; four of them were randomly selected in each task.

Under illumination condition 2, the observers selected one oddity from three images in a task. We created the images by using slightly different (in terms of their pixel histograms) light probes, including "Overcast Day/Building Site (Metro Vienna)," "Overcast Day at Techgate Donaucity," and "Metro Station (Vienna Metro)." The task procedure of illumination condition 3 was the same as that of illumination condition 2.

For illumination condition 3, we created the three images by using light probes that were rather different from each other: "Inside Tunnel Machine," "Tungsten Light in the Evening (Metro Building Site Vienna)," and "Building Site Interior (Metro Vienna)." We computed the pixel histogram similarity for each illumination pair and used it as the distance for the multidimensional scaling analysis (MDS). We extracted three largely different light probes in the MDS space and used them for illumination condition 3. We also selected three similar light probes in the space and used them for illumination condition 2. The pose of each object in illumination conditions 2 and 3 was not changed. The stimulus condition is summarized in Table 1.

Rendering

To render the images, we used the integrator of the photon mapping method for tasks 1, 2, 4, 5, and 6 and used the integrator of the simple volumetric path tracer implemented in the Mitsuba renderer for task 3 (OT). The calculation was conducted using single-float

precision. Each rendered image was converted into sRGB format with a gamma of 2.2 and saved as an 8-bit PNG image.

Behavioral experiments

Laboratory experiment

Twenty paid volunteers participated in the laboratory experiment. Before starting the experiment, we confirmed that all of them had normal color vision by having them take the Farnsworth Munsell 100 Hue Test and that all had normal or corrected-to-normal vision by having them take a simple visual acuity test. The participants were naïve to the purpose and methods of the experiment. The experiment was approved by the ethical committees at NTT Communication Science Laboratories.

The generated stimuli were presented on a calibrated 30-inch color monitor (ColorEdge CG303W; EIZO, Hakusan, Ishikawa, Japan) controlled with a Quadro 600 video card (NVIDIA Corporation, Santa Clara, CA). Each participant viewed the stimuli in a dark room at a viewing distance of 86 cm, where a single pixel subtended 1 arcmin. Each object image of 512 × 512 pixels was presented at a size of 8.5° × 8.5°.

In each trial, four (illumination 1) or three (illuminations 2 and 3) object images chosen for each task were presented on the monitor (Figure 3). Measurements of different illumination conditions were conducted in different blocks. Under illumination condition 1, four different object images in different orientations were presented. Under illumination conditions 2 and 3, the three different object images had different illuminations. The order of illumination conditions 1, 2, and 3 was counterbalanced across observers. The observers were asked to report which of the object images looked odd by pushing one of the keys. The stimuli were presented until the observer made a response. The task instructions were simply to find the odd one with no further explanation about how it was different from the rest. The observers were not given any feedback about whether or not their response was correct. All made 10 judgments for each task of illumination condition 1. Seventeen observers made 10 judgments for each task of illumination condition 2, and three made only seven judgments due to the experiment's time limitation. Seventeen observers made 10 judgments for each task of illumination condition 3, and three made seven judgments due to the experiment's time limitation.

Crowdsourcing experiment

In the web experiment, 416, 411, and 405 paid volunteers participated in the tasks of illumination conditions 1, 2, and 3, respectively. We recruited these observers through a Japanese commercial crowdsourcing service. All who participated under illumination condition 3 also participated under illumination conditions 1 and 2. Moreover, all who participated in illumination condition 2 had also participated under illumination condition 1. The experiment was approved by the ethical committees at NTT Communication Science Laboratories.

All observers used the web browsers of their own personal computers or tablets to participate in the experiment. We asked them to watch the screen from a distance of about 60 cm. Each object image was shown on the screen at a size of 512 × 512 pixels. We did not strictly control the visual angle of the image participants observed.

The procedure was similar to that of the laboratory experiment. In each trial, four or three object images that had been chosen depending on the task were presented on the screen, as shown in Figure 3. The measurement was conducted under illumination condition 1 first, followed by one under illumination condition 2 and one under illumination condition 3. The observers were asked to report which of the object images looked odd by clicking one of the images. Each participant made one judgment for each condition. The other steps of the procedure were the same as those in the laboratory experiment.

Data analysis

For each oddity task, we computed the proportion that each participant got correct. The chance level of the correct proportion was 0.25 for illumination condition 1 and 0.33 for illumination conditions 2 and 3. We computed the sensitivity d' from each correct proportion by using a numerical simulation to estimate the sensitivity of the oddity task (Craven, 1992). We used the Palamedes data analysis library for the simulation (Kingdom & Prins, 2010; Kingdom & Prins, 2016; Prins & Kingdom, 2018). To avoid values of infinity, we converted the one probability according to the total trial number in the simulation; that is, we corrected the one value to $1 - (1/2N)$, where N is the total trial number (Macmillan & Kaplan, 1985). For the laboratory experiment, we computed the sensitivity (d') of each observer and averaged it across observers. For the crowdsourcing experiment, because each observer engaged in each task one time, we computed the proportion correct for each task from all of the observers' responses and used that value to compute d'.

Results

In this section, we describe the results of our benchmark data acquisition. First, we evaluate the environment dependency of our experiment—the performance difference between the online and

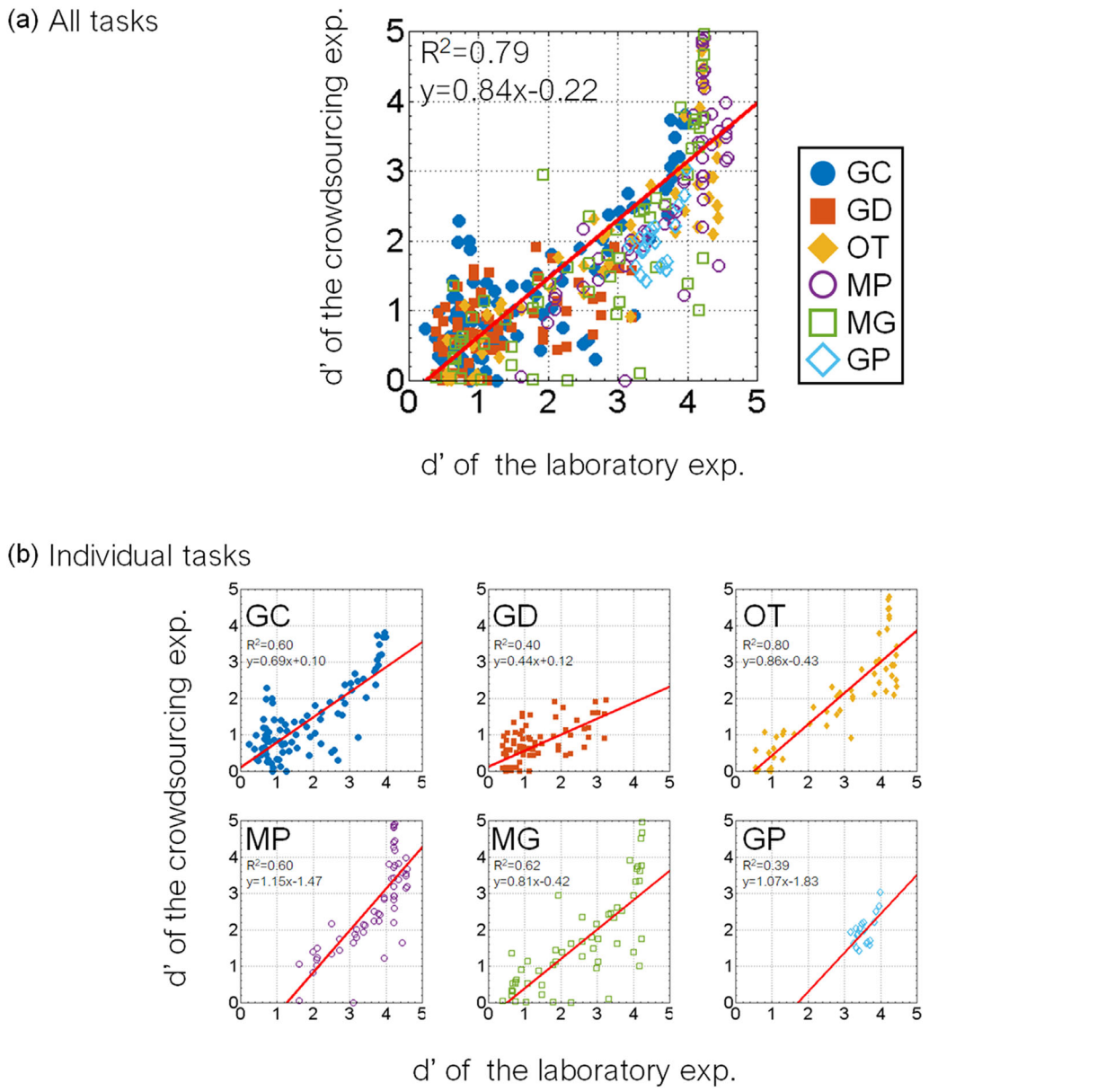

Figure 7. Results of laboratory and crowdsourcing experiments. The sensitivity d' in each task in the crowdsourcing experiment is plotted as a function of that in the laboratory experiment. (a) Results of all tasks. Each plot indicates a task with an object, an illumination, and a difficulty. The red line indicates the linear regression between the crowdsourcing and laboratory results. The coefficient of determination (R^2) of the regression and the equation are shown in the legend. The results show that the tasks were generally robust across experimental conditions. (b) Results of individual tasks. Different panels indicate tasks involving different materials. Each plot in a panel indicates a task with an object, illumination, and difficulty. The red line indicates the linear regression between the laboratory and crowdsourcing results. The coefficient of determination (R^2) of the regression and the equation are shown in the legend. The accuracy of task 2 (GD) in the crowdsourcing experiment was generally lower than that in the laboratory experiment. The correlation of task 6 (GP) between the laboratory and crowdsourcing experiments was the worst.

laboratory experiments. Then, we describe the illumination and geometry effect on each task. After discussing each task, we show how intermediate visual features contribute to task performance. In the end, we analyze the individual difference in each task.

Environment dependence

For cross-cultural, cross-species, brain-dysfunction, and developmental studies, stimulus presentation on a monitor cannot always be strictly controlled because of apparatus or ethical limitations. Therefore,

a performance validation of each task across different apparatuses is critical to decide which tasks the users of our database should select in their experimental environment. Figure 7a shows the results of the correlation analysis between the laboratory and crowdsourcing experiments. The coefficient of determination (R^2) of the linear regression between the sensitivity (d') in the laboratory experiment and that of the crowdsourcing experiment is 0.79, indicating a high linear correlation. However, the slope of the regression is less than 1. This shows that the sensitivity of the crowdsourcing experiment was worse than that of the laboratory experiment, with many repetitions in general. These findings suggest that the present tasks maintain relative performance across different experimental environments.

Figure 7b shows the results for each task of the laboratory and crowdsourcing experiments in more detail. The coefficients of determination (R^2) in tasks 1 to 6 were 0.60, 0.40, 0.86, 0.60, 0.62, and 0.39, respectively. The coefficient of task 6 (GP) was the worst, followed by task 2 (GD). As in the latter section, task 6 (GP) also showed large individual differences; thus, the correlation between the laboratory and crowdsourcing experiments was decreased. The slope of the linear regression on task 2 (GD) was 0.44, and the proportion correct in the crowdsourcing experiment for task 2 was generally lower than that in the laboratory for task 2. In the laboratory experiment, we used a 30-inch liquid-crystal display monitor, and the stimulus size of each image was presented at a size of 8.5° × 8.5°, which we expected to be larger than when participants on the web observed the image on a tablet or PC. Task 2 (GD) is related to the DOI of the specular reflection; thus, the spatial resolution might have affected the accuracy of the observers' responses, although the relative difficulty for task 2 (GD) even in the crowdsourcing experiment was similar to that in the laboratory experiment. These findings suggest that the absolute accuracy of task 2 (GD) depends largely upon the experimental environment.

Illumination and geometry

Figures 8 to 13 show the performance of each task in the laboratory experiment. Different panels depict results obtained for different objects. Different symbols in each panel depict different illumination conditions. The results of the crowdsourcing experiment are shown in Appendix A. For tasks 1 to 5 (Figures 8 to 12), we parametrically changed the material parameters (e.g., the contrast dimensions for task 1, GC). Results show that the discrimination accuracy increased as the target material parameters deviated from the non-target one. This trend is most readily observed for illumination 1 on each task condition. In contrast, the accuracy did not change much with the material parameters for some

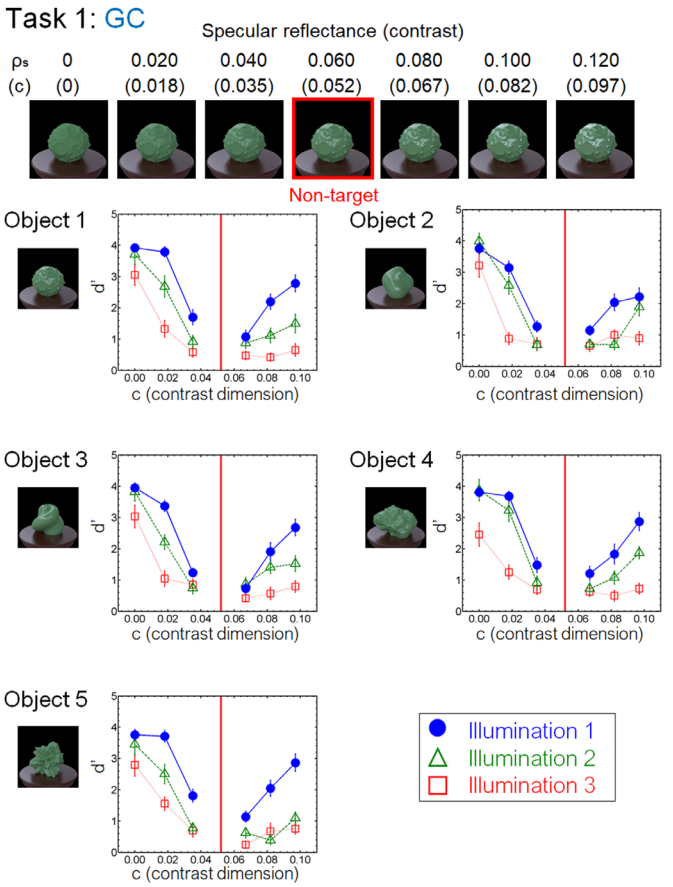

Figure 8. Results of task 1 (GC) in the laboratory experiment. Different panels show different objects. Different symbols in each panel depict different illumination conditions. The vertical red line in each panel indicates the parameter of the non-target stimulus. Error bars indicate ±1 *SEM* across observers.

conditions. This trend can be observed for illuminations 2 and 3 of task 1 (GC) and objects 4 and 5 of task 2 (GD). For task 6 (GP), the relation of target and non-target stimuli differed from that of the other tasks. In this task, the non-target stimulus was made for each material parameter (i.e., DOI). As shown in Figure 13, this material parameter did not affect the task difficulty.

By comprehensively assessing material recognition performance across different stimulus conditions, we found novel properties that have been overlooked in the previous literature, one of which pertains to the geometrical dependence of material recognition. When object images changed in the gloss DOI discrimination task (task 2, GD) (Figure 9), the observers could detect the material difference better for smooth objects (objects 2 and 3) than for rugged objects (objects 4 and 5). In contrast, when the object images changed in the gloss contrast discrimination task (task 1, GC) (Figure 8), little geometrical dependence was found. We also found little geometrical dependence when observers detected highlight-shading consistency (task 6, GP) (Figure 13). Although geometrical dependencies

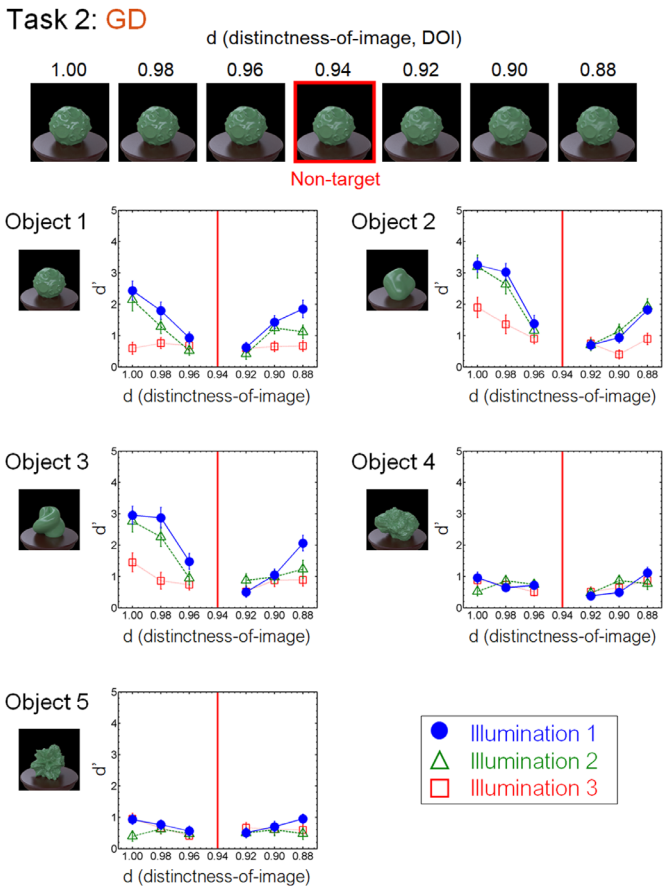

Figure 9. Results of task 2 (GD) in the laboratory experiment.

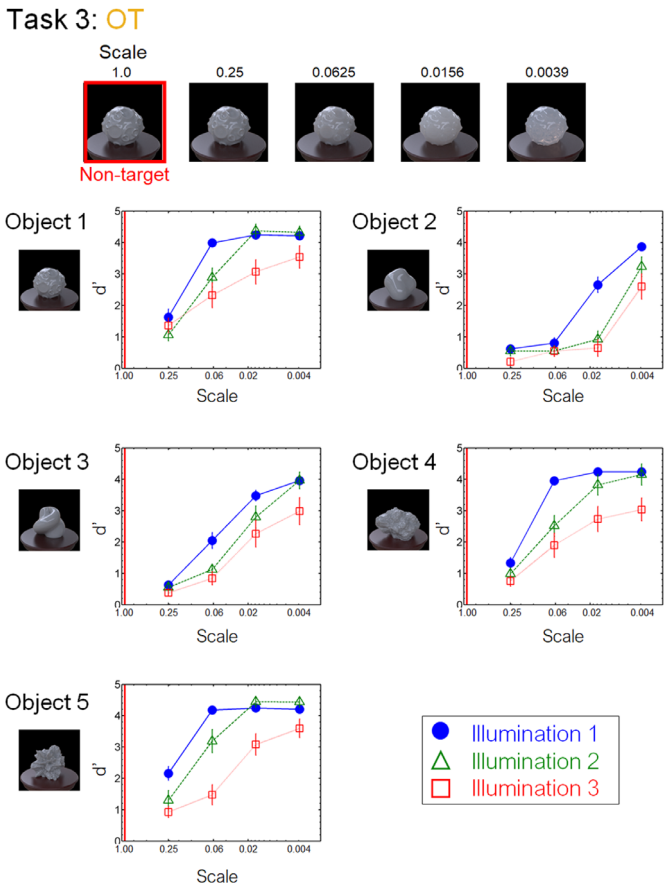

Figure 10. Results of task 3 (OT) in the laboratory experiment.

of glossiness perception have been reported before (Nishida & Shinya, 1998; Vangorp et al., 2007), they were mainly about the effects of shape on apparent gloss characteristics, not on gloss discrimination. Furthermore, our results also show a geometrical dependence of translucency perception (task 3, OT) (Figure 10). Similar to the dependence on the DOI dimension, the sensitivity changed between the smooth objects (objects 2 and 3) and rugged objects (objects 4 and 5), but in the opposite way. Specifically, the translucent difference was more easily detected for the rugged objects than for the smooth objects (Figure 10).

We also found an illumination dependence in material recognition. We used three illumination conditions, wherein the illumination environments used in a task were identical (illumination 1), similar to each other (illumination 2), or largely different from each other (illumination 3). The results showed that task accuracy decreased as the difference in light probes across the images increased from illumination 1 to illuminations 2 and 3 (Figures 8–13). This finding not only confirms the large effect of illumination on gloss perception reported before (Fleming et al., 2003; Motoyoshi & Matoba 2012; Zhang, de Ridder, Barla, & Pont, 2019) but also demonstrates similarly strong effects of illumination on other material discrimination tasks (OT, MP, and MG). The observers' data is stored in the repository with the image dataset (Appendix B).

Intermediate visual feature analysis

One may raise the concern that our observers might have made oddity judgments based on differences in low-level superficial image properties such as the mean color of the object. We did not explicitly ask the observers to select one object image in terms of the material appearance. This procedure could lead observers to take a simple strategy unrelated to material judgment. A related question is that, if not such simple properties, are there any intermediate image features in hierarchical visual processing that can explain the observers' responses? Recent studies have shown that the intermediate processing in the ventral visual stream of humans and monkeys encodes the higher-order image features as computed in texture synthesis algorithms or deep convolutional neural networks (Freeman et al., 2013; Okazawa, Tajima, Komtsu, 2015; 2017; Yamins & Dicarlo, 2016). It has been suggested that the processing in the visual ventral stream also

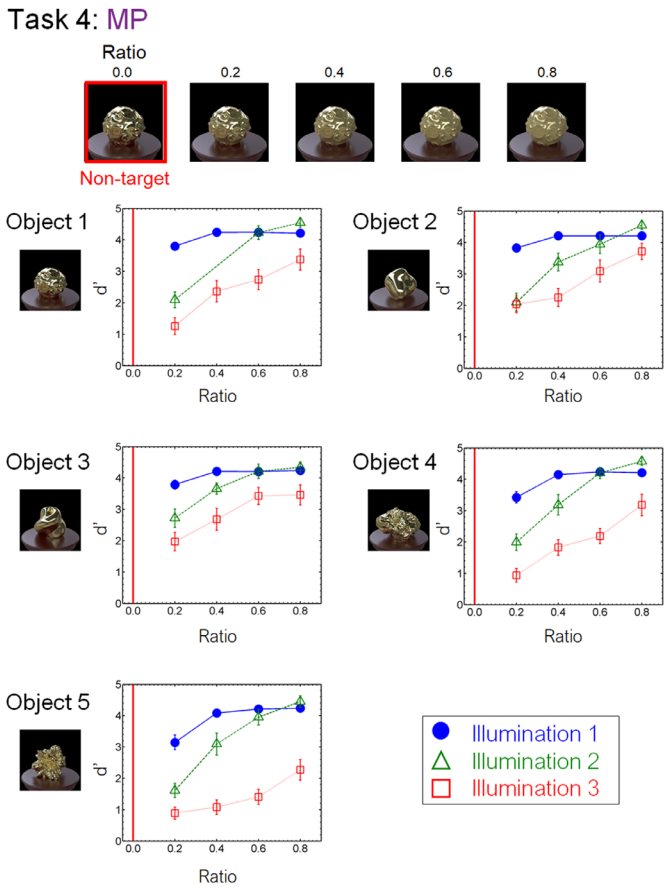

Figure 11. Results of task 4 (MP) in the laboratory experiment. Data for one of the observers for object 1 and illumination 2 are missing due to a mistake in the stimulus presentation.

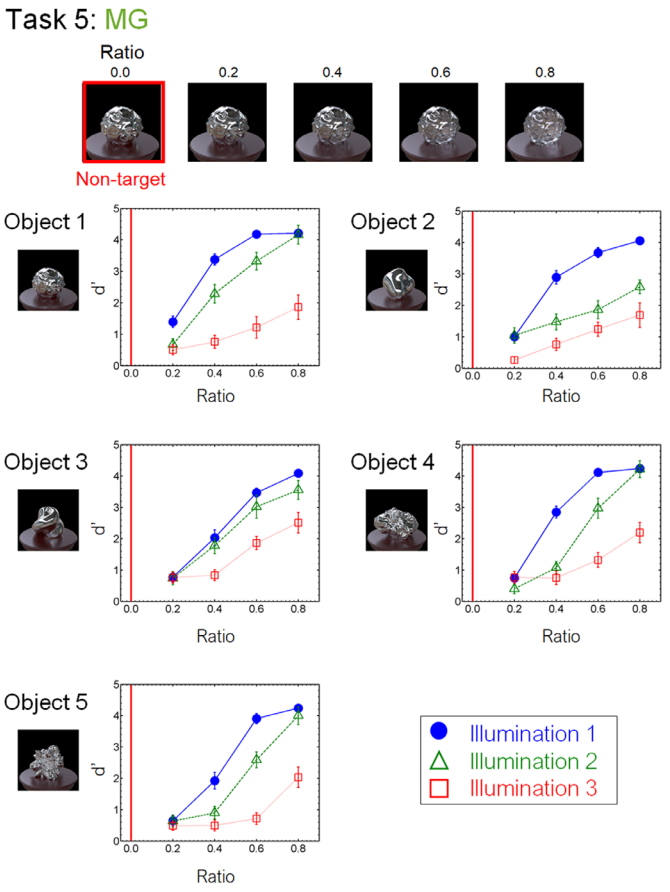

Figure 12. Results of task 5 (MG) in the laboratory experiment.

mediates material recognition for static objects (Nishio et al., 2012; Nishio et al., 2014; Miyakawa et al., 2017). We asked how such intermediate features possibly processed in material computation could explain the observers' responses.

More specifically, we analyzed how various image feature differences on each task could explain the observers' task performance. Each task (i.e., a material dimension with an object under an illumination condition) included a set of material objects with different combinations of poses (illumination condition 1) or illuminations (illumination conditions 2 and 3). These combinations were used as repetition for the behavioral experiment. In the analysis, we chose all combinations for each task and calculated the mean feature distance. We calculated this distance metric using various image features (e.g., pixel statistics, texture statistics) as described below in detail. If the distance metric of each image feature is correlated with human performance, then the feature can be diagnostic for human judgments.

We linearly regressed the discrimination sensitivity d' for each task using the distance metric calculated from various image features. Specifically, we used the texture parameters originally proposed in the literature of texture synthesis by Portilla & Simoncelli (2000). They suggested that natural textures can be synthesized by the probabilistic summary statistics derived from the pixel histogram and the subband distribution, including higher-order statistics such as the correlations across the subband filter outputs. More recently, many studies have shown that the intermediate visual processing in the ventral stream, such as V2 or V4, encodes these texture parameters (Freeman et al., 2011; Okazawa et al., 2015). Following the previous studies (Okazawa et al., 2015), we reduced the original texture parameters by removing redundant features because a large number of parameters can make the fitting unreliable. Specifically, we conducted the same reduction as Okazawa et al. (2015), except that (1) we included the mean, *SD*, and kurtosis of the marginal statistics, as well as the skewness, and (2) we calculated these statistics not only for grayscale images (CIE L* image) but also for color images (CIE a* and CIE b* images). We defined the white XYZ value averaging the diffuse white sphere rendered under each illumination condition and used it to calculate the CIE L*a*b* of each image. We extracted the center 128 × 128 pixels of each image and calculated the texture parameters

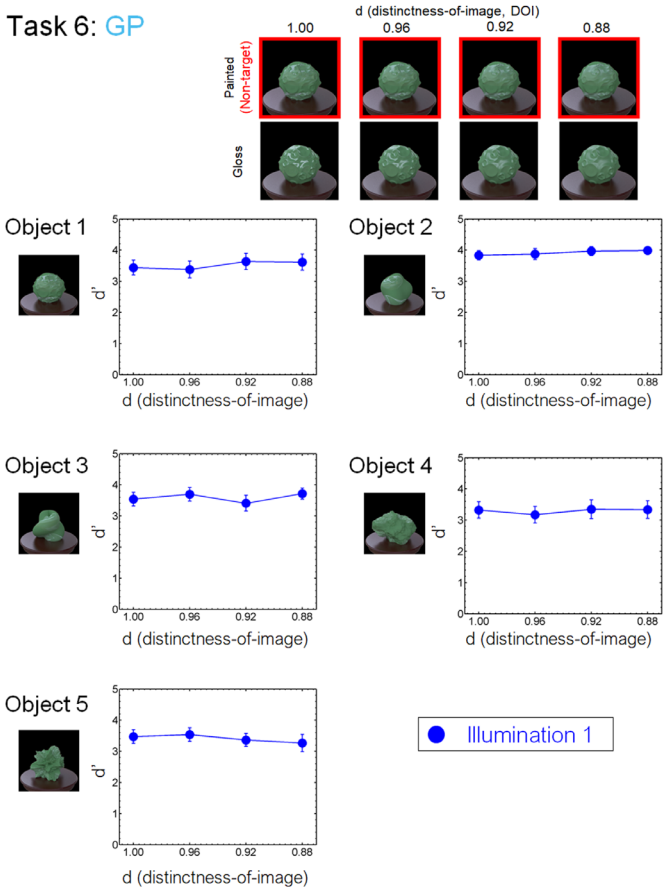

Figure 13. Results of task 6 (GP) in the laboratory experiment.

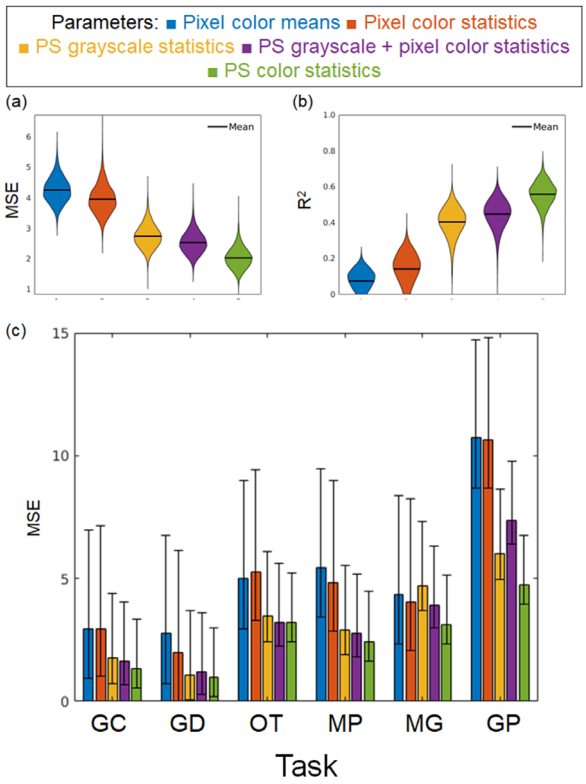

Figure 14. Results of the linear regressions using different parameters. We regressed the human discrimination performance on pixel color means (three parameters, blue), pixel color statistics (12 parameters, red), PS grayscale texture statistics (regularized 18 parameters, yellow), PS grayscale statistics and pixel color statistics (regularized 18 parameters, purple), or PS color statistics (regularized 18 parameters, purple). (a) Results of the *MSE* for each regression. (b) Results of the *MSE* for each regression. These results are shown using a violin plot. (c) Results of the *MSE* for each task. The error bars indicate the bootstrap 95% confidence intervals.

using the texture synthesis algorithm by Portilla & Simoncelli (2000) with four scales and four orientations. We reduced these original texture parameters of each CIE L*, a*, or b* image to 32 parameters following Okazawa et al. (2013). More details are described in the supplementary tables S1 and S2 of Okazawa et al. (2013). In total, we used 96 parameters for the regression analysis.

We conducted five regressions with different types of parameters to explore the contribution of different statistics. Specifically, we used (1) pixel color means, (2) pixel color statistics, (3) Portilla–Simoncelli (PS) grayscale texture statistics, (4) PS grayscale and pixel color statistics, and (5) PS color statistics. The pixel color means and the pixel color statistics were the marginal statistics in the PS texture statistics. The pixel color means indicated the averaged pixel values of each L*a*b* channel. The pixel color statistics indicated the mean, *SD*, skewness, and kurtosis of each color channel. The numbers of these parameters were 3 and 12, respectively. For the two conditions, we used a linear regression without regularization to fit the discrimination sensitivity (blue and red in Figure 14). For the three PS texture statistics conditions (yellow, purple, and green in Figure 14), we used the compressed PS statistics as described above. Because the number of parameters for these conditions was large (32, 48, and 96, respectively), we used L1-penalized linear least-squares regression (i.e., lasso) to avoid overfitting. We controlled the hyperparameters so that the number of independent variables was 18, where the regression of the PS grayscale statistics condition showed the minimum mean-squared error (*MSE*).

We divided all tasks into training and test datasets with a ratio of four to one, respectively, and conducted the above five regressions. The task ratio was kept constant across the training and test datasets. For the training dataset on the lasso regressions, we regressed the discrimination sensitivity using 5-fold-cross validation. Figure 14 shows the *MSE* and the determinant coefficient for the test datasets. We resampled the training and test datasets 10,000 times and have depicted the distribution using a violin plot. The predictions based on the color mean statistics did not match the observers' discrimination

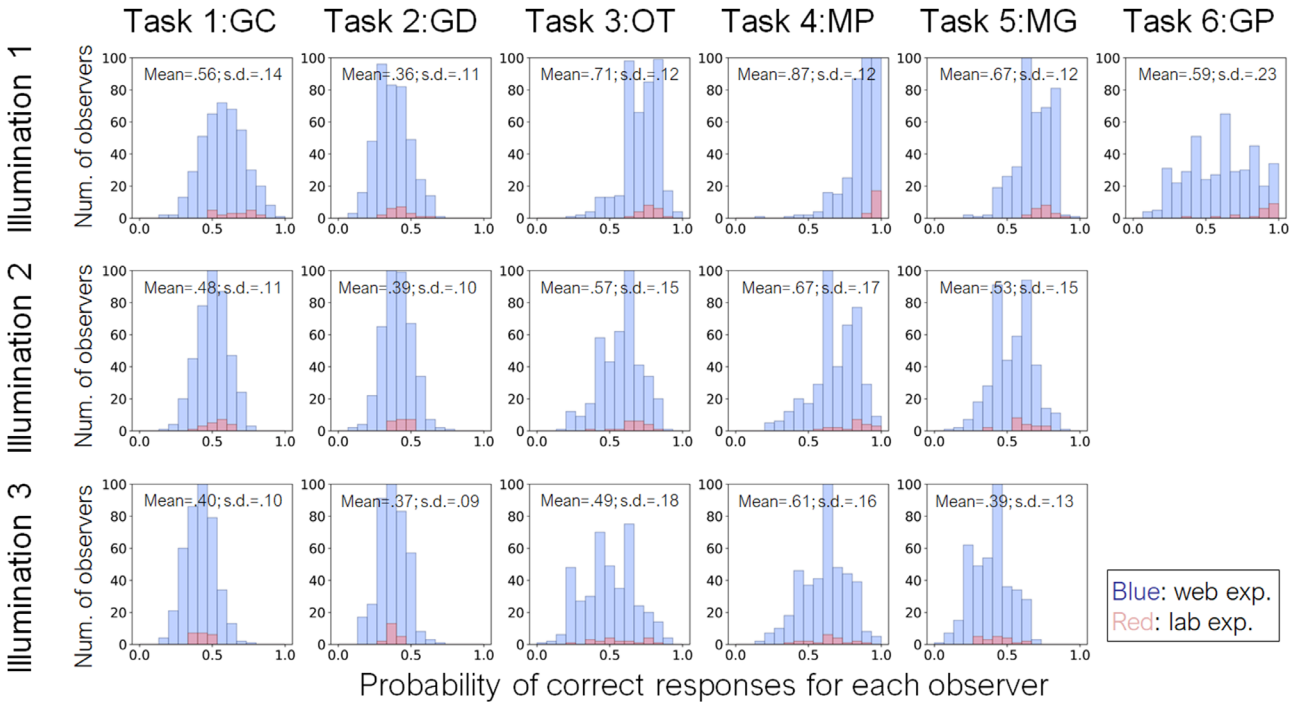

Figure 15. Histogram of response accuracy for each observer in the crowdsourcing (blue) and lab (red) experiments. Different panels indicate different material tasks and illumination conditions. For each condition, the probability of a correct response was calculated by averaging the responses of each observer across objects and task difficulties. The histograms of crowdsourcing and lab experiments are overlaid in each panel. The mean and standard deviation of each distribution are shown in each panel.

sensitivity at all (Figs. 14a, 14b). These results suggest that the observers did not simply rely on the mean differences to perform the oddity tasks. The MSE and the determinant coefficient for the marginal statistics condition were more improved when we added the higher-order statistics (marginal statistics condition, PS grayscale statistics condition, and PS color statistics condition). Because the regularization parameter was controlled under the PS color and grayscale statistics conditions, these results cannot be ascribed to the number of independent variables. It is noteworthy that, even when all of the PS color statistics were used, the prediction was not sufficient to explain the observers' discrimination performance. This finding suggests that human material judgments also rely on higher-order features the PS statistics do not cover. One possible future direction is to use the intermediate activation of the deep neural networks. To support this direction, we include in our database the activation data of VGG-19, a feedforward convolutional neural network, for our image dataset and the analysis about how the dataset is represented in each layer (see Appendix C). In short, our dataset images were clustered in higher layers of the pretrained network according to object differences, and the material differences were represented in each object cluster.

Individual differences

Next, we evaluated the individual differences of each task in the Japanese adult population. Figure 15 shows the histogram of the response accuracy for each observer in the crowdsourcing and laboratory experiments. For the crowdsourcing experiment, the number of observers for illumination conditions 1, 2, and 3 was 416, 411, and 405, respectively. For the laboratory experiment, the number of observers was 20. For each condition, the probability of a correct response was calculated by averaging the responses of each observer across objects and task difficulties. The standard deviations of tasks 1 to 6 under illumination condition 1 were 0.14, 0.11, 0.12, 0.12, 0.12, and 0.23, indicating a particularly large individual difference for task 6 (GP). The standard deviation under illumination conditions 2 and 3 ranged from 0.09 to 0.18. It should also be noted that most of the conditions showed unimodal distributions, whereas task 6 (GP) showed a nearly uniform distribution. This finding suggests that individual differences in the discrimination ability of the spatial consistency of specular highlights are larger than those for other material properties, including glossiness contrast and DOI (GC and GD).

Discussion

The present study aimed to construct a database of material images annotated with the results of human discrimination tasks. We created material images that varied in six different material dimensions on the basis of the previous material-recognition studies. Our dataset includes various objects and illuminations so that users can comprehensively investigate the effects of these physical causes on material recognition. The results of psychophysical experiments showed that most of the task difficulty could be appropriately controlled by manipulating the material parameters. Furthermore, analysis of visual feature showed that the parameters of higher-order color texture statistics (Figure 14, PS color statistics) can partially, but not completely, explain task performance. One crucial point of our dataset is that we used a nonverbal procedure to collect the observers' data. Because this procedure is widely used in babies, brain-injured participants, and animals, the current behavioral data can be a benchmark for more diverse research fields.

Because we comprehensively investigated the material recognition using a structured dataset, our dataset itself revealed novel findings about material recognition. For instance, the present results indicate that the performance of the tasks in the crowdsourcing experiment was strongly correlated with that in the laboratory experiment. This suggests that the dataset has enough tolerance to conduct new experiments involving a variety of observers and experimental conditions. Another is that geometry dependency on material recognition emerges similarly in different material attributes such as gloss DOI or translucency (Figure 10). Specifically, the translucency discrimination sensitivity was high when the object had rugged surfaces (e.g., objects 1, 4, and 5). Some studies have shown that physically prominent features of translucent objects appear around sharp corners on the surface (Fleming & Bülthoff, 2005; Gkioulekas et al., 2013). One possibility is that the diagnostic features for translucent perception lie in the edge/corner of a translucent object and our rugged objects included much information to judge translucency. More recently, Xiao, Zhao, Gkioulekas, Bi, and Bala (2020) investigated the effect of geometry on translucency perception. In their experiments, they changed the smoothness of the object edges. In agreement with our findings, the edge modulation was critical to the translucency perception. Specifically, the object with the smooth edge was perceived as more translucent than the sharp one.

Another finding is that the ability to discriminate the spatial consistency of specular highlights in glossiness perception has large individual differences, although other glossiness discrimination tasks do not show such large differences. Some studies suggest that image statistics are diagnostic for glossiness perception (Adelson, 2001; Motoyoshi et al., 2007). However, when specular highlights of an object image are inconsistent in terms of their position and/or orientation with respect to the diffuse shading component, they look more like white blobs produced by surface reflectance changes (Beck & Prazdny, 1981; Kim et al., 2011; Marlow et al., 2011). This is why the highlight-inconsistency effect is considered to be a counterexample to the image statistics explanation. The large individual differences suggest that the discrimination of the spatial consistency of specular highlights may be mediated by a different, and possibly more complicated, mechanism than that responsible for glossiness contrast/DOI discrimination. In agreement with this notion, Sawayama and Nishida (2018) showed that highlight inconsistency is discriminated by image gradient features different from those used in the human material computation. This suggests that the glossiness computation is mediated by multiple stages, such that one step is to discriminate different materials on a surface for extracting a region-of-interest (ROI) and another is to compute the degree of glossiness in the ROI as shown in Motoyoshi et al. (2007).

One may have a concern that the intermediate objects in tasks 4 and 5 are physically not feasible because they are a mixture of two physically distinct materials. However, our stimuli do not look so unrealistic. The dielectric/metal materials are distinct material categories when considering an object with a uniform single material, but many daily objects surrounding us are a mixture of various materials, and we often see a plastic object coated by a metallic material. We can regard our intermediate materials as an approximation of such coated materials. In addition, continuously connecting distinct categories is common in various research fields, such as speech recognition (e.g., Grey & Gordon, 1978) or face recognition (e.g., Turk, Heatherton, Kelley, Funnell, Gazzaniga, & Macrae, 2002), especially to elucidate what stimulus image features are involved in the processing. Considering the literature, we think our intermediate approach is reasonable.

Although our database includes diverse material dimensions, they are still not enough to cover the full range of natural materials. One example is cloth (Bi & Xiao, 2016; Xiao, Bi, Jia, Wei, & Adelson, 2016; Bi, Jin, Nienborg, & Xiao, 2018; Bi, Jin, Nienborg, & Xiao, 2019). Cloth materials are ubiquitous in everyday environments. A reason we did not include this class of materials is that it has been shown that the cloth perception strongly relies on dynamic information (Bi et al., 2018; Bi et al., 2019). Because of the limited experimental time, our database currently focuses on static images. This is why other materials related to dynamic information (reviewed by Nishida, Kawabe, Sawayama, & Fukiage, 2018) related to the perception of liquidness (Kawabe, Maruya, & Nishida, 2015), viscosity (Kawabe, Maruya, Fleming, & Nishida, 2015, van Assen & Fleming, 2016), and stiffness (Paulun

et al., 2017; Schmid & Doerschner, 2018), among others, were not used in the current investigation. In addition, the perception of wetness (Sawayama et al., 2017a) and the fineness of surface microstructures (Sawayama, Nishida, & Shinya, 2017b) were not investigated because of the difficulty of continuously controlling physical material parameters by using identical geometries of other tasks. Because we only used five geometries, material perceptions derived from object mechanical properties were also not investigated (Schmidt, Paulun, van Assen, & Fleming, 2017). A crucial point is that we can share our source code to reproduce images. We hope to remove obstacles to constructing a new dataset and contribute to future work on material recognition. Sharing the datasets with the source code should allow researchers to easily conduct new studies within this literature. For example, we measured the discrimination sensitivities in our experiments from one side of the materials in tasks 3, 4, and 5 (i.e., opaque, gold, and silver). The sensitivities from the other side (i.e., transparent, plastic, and glass) could be slightly different from the current results. Researchers can easily render new images of different material parameters in the same scene condition and conduct a new investigation.

Our datasets also highlight the difficulty of choosing appropriate parameters that cover the full range of the material sensitivity. We chose the stimulus parameters based on the preliminary experiments. We tried to choose the parameters so that we can measure the sensitivity of each task in the full range, from the level of chance to maximum accuracy. However, we found large individual differences in some tasks (e.g., task 6), and they resulted in the partial measurement of the narrow sensitivity range. This unpredictability is one of the difficulties of producing the large size of the dataset. The current findings should contribute to future attempts to create material image datasets.

Our dataset focuses on expanding the previous findings regarding material recognition into more diverse research fields. From the view of a global standard dataset, our dataset has several limitations as described above. However, it did contribute to this expansion purpose. Specifically, several research groups of behavioral science, computer science, and neuroscience have ongoing projects utilizing our dataset, and some findings have already been reported at conferences and journals. Kawasaki et al. (2019) used our dataset to explore the role of the monkey inferior temporal cortex on material perception by using electrocorticography recordings. Tsuda, Fujimichi, Yokoyama, and Saiki (2020) investigated the role of working memory on material processing using our dataset. Koumura, Sawayama, and Nishida (2018) explored how mid-level features in deep convolutional neural networks can explain human behavioral data.

Conclusion

We constructed image and observer database for material recognition experiments. We collected observation data about material discrimination in tasks that had a nonverbal procedure for six material dimensions and several task difficulties. The results of psychophysical experiments in laboratory and crowdsourcing environments showed that the performance of the tasks in the crowdsourcing experiment was strongly correlated with the performance of the tasks in the laboratory experiment. In addition, by using the above comprehensive data, we obtained novel findings on the perception of translucence and glossiness. Not only can the database be used as benchmark data for neuroscience and psychophysics studies on the material recognition capability of healthy adult humans, but it can also be used in cross-cultural, cross-species, brain-dysfunction, and developmental studies of humans and animals.

Keywords: material perception, image and observer database, visual psychophysics, computer graphics, crowdsourcing

Acknowledgments

Supported by Grants-in-Aid for Scientific Research on Innovative Areas from the Japan Society of Promotion of Science (JSPS KAKENHI JP15H05915, JP15H05924, JP20H05954, and JP20H05957 to SN and YD).

Commercial relationships: none.
Corresponding author: Masataka Sawayama.
Email: masa.sawayama@gmail.com.
Address: NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, Kanagawa, Japan.

References

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

Adelson, E. H. (2001). On seeing stuff: The perception of materials by humans and machines. *Proceedings of SPIE 4299: Human Vision And Electronic Imaging VI* (pp. 1–12). Bellingham, WA: SPIE.

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/jov.9.11.10.

Ashikmin, M., Premože, S., & Shirley, P. (2000). A microfacet-based BRDF generator. *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 65–74). New York: Association for Computing Machinery.

Beck, J., & Prazdny, S. (1981). Highlights and the perception of glossiness. *Attention, Perception, & Psychophysics, 30*(4), 407–410.

Bi, W., & Xiao, B. (2016). Perceptual constancy of mechanical properties of fabrics under variation of external force. *Proceedings of the ACM Symposium on Applied Perception* (pp. 19–23). New York: Association for Computing Machinery.

Bi, W., Jin, P., Nienborg, H., & Xiao, B. (2018). Estimating mechanical properties of cloth from videos using dense motion trajectories: Human psychophysics and machine learning. *Journal of Vision, 18*(5):12, 1–20, https://doi.org/10.1167/18.5.12.

Bi, W., Jin, P., Nienborg, H., & Xiao, B. (2019). Manipulating patterns of dynamic deformation elicits the impression of cloth with varying stiffness. *Journal of Vision, 19*(5):18, 1–18, https://doi.org/10.1167/19.5.18.

Brainard, D. H., & Hurlbert, A. C. (2015). Colour vision: understanding #TheDress. *Current Biology, 25*(13), R551–R554.

Chadwick, A. C., Cox, G., Smithson, H. E., & Kentridge, R. W. (2018). Beyond scattering and absorption: perceptual unmixing of translucent liquids. *Journal of Vision*. *18*(11):18, 1–15, https://doi.org/10.1167/18.11.18.

Craven, B. J. (1992). A table of d′ for M-alternative odd-man-out forced-choice procedures. *Perception & Psychophysics, 51*(4), 379–385.

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

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

Fleming, R. W., & Bülthoff, H. H. (2005). Low-level image cues in the perception of translucent materials. *ACM Transactions on Applied Perception, 2*, 346–382.

Fleming, R. W., Dror, R. O., & Adelson, E. H. (2003). Real-world illumination and the perception of surface reflectance properties. *Journal of Vision, 21*(13):3, 1–23, https://doi.org/10.1167/jov.21.13.3.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*(9), 1195–1201.

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature Neuroscience, 16*(7), 974–981.

Gegenfurtner, K. R., Bloj, M., & Toscani, M. (2015). The many colours of 'the dress.' *Current Biology, 25*(13), R543–R544.

Gigilashvili, D., Thomas, J.-B., Hardeberg, J. Y., & Pedersen, M. (2021). Translucency perception: A review. *Journal of Vision, 21*(8):4, 1–41, https://doi.org/10.1167/jov.21.8.4.

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

Goda, N., Yokoi, I., Tachibana, A., Minamimoto, T., & Komatsu, H. (2016). Crossmodal association of visual and haptic material properties of objects in the monkey ventral visual cortex. *Current Biology, 26*(7), 928–934.

Grey, J. M., & Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America, 63*(5), 1493–1500.

Hunter, R. S. (1937). *Methods of determining gloss*, Research Paper RP958. Washington, DC: National Bureau of Standards, U.S. Department of Commerce.

Jakob, W. (2010). Mitsuba: Physically Based Renderer. Retrieved from https://www.mitsuba-renderer.org/download.html

Jensen, H. W., Marschner, S. R., Levoy, M., & Hanrahan, P. (2001). A practical model for subsurface light transport. *SIGGRAPH '01: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 511–518). New York: Association for Computing Machinery.

Kawabe, T., Maruya, K., & Nishida, S. (2015). Perceptual transparency from image deformation. *Proceedings of the National Academy of Sciences, USA, 112*(33), E4620–E4627.

Kawabe, T., Maruya, K., Fleming, R. W., & Nishida, S. (2015). Seeing liquids from visual motion. *Vision Research, 109*, 125–138.

Kawasaki, K., Miki, H., Anzai, K., Sawayama, M., Matsuo, T., & Suzuki, T., …Okatani, T. (2019). Spatial and time-frequency representations of glossy material properties in the monkey inferior temporal cortex. *Neuroscience 2019*. Washington, DC: Society for Neuroscience.

Kentridge, R. W., Thomson, R., & Heywood, C. A. (2012) Glossiness perception can be mediated independently of cortical processing of colour or texture. *Cortex, 48*(9), 1244–1246.

Kim, J., & Marlow, P. J. (2016). Turning the world upside down to understand perceived transparency. *i-Perception, 7*(5), 2041669516671566.

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.

Kingdom, F. A. A., & Prins, N. (2010) *Psychophysics: A practical introduction*. London: Academic Press.

Kingdom, F. A. A., & Prins, N. (2016). *Psychophysics: A practical introduction* (2nd ed.). London: Academic Press.

Koumura, T., Sawayama, M., & Nishida, S. (2018). Explaining behavioral data of visual material discrimination with a neural network for natural image recognition. *28th Annual Conference of Japanese Neural Network Society (JNNS2018)*, Fukuoka, Japan: Japanese Neural Network Society.

Liao, C., Sawayama, M., & Xiao, B. (2022). Crystal or jelly? Effect of color on the perception of translucent materials with photographs of real-world objects. *Journal of Vision, 22*(2):6, 1–23, https://doi.org/10.1167/jov.22.2.6.

Macmillan, N. A., & Kaplan, H. L. (1985). Detection theory analysis of group data: estimating sensitivity from average hit and false-alarm rates. *Psychological Bulletin, 98*(1), 185.

Marlow, P., Kim, J., & Anderson, B. L. (2011). The role of brightness and orientation congruence in the perception of surface gloss. *Journal of Vision, 11*(9):16, 1–12, https://doi.org/10.1167/11.9.16.

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

Miyakawa, N., Banno, T., Abe, H., Tani, T., Suzuki, W., & Ichinohe, N. (2017). Representation of glossy material surface in ventral superior temporal sulcal area of common marmosets. *Frontiers in Neural Circuits, 11*, 17.

Motoyoshi, I. (2010). Highlight-shading relationship as a cue for the perception of translucent and transparent materials. *Journal of Vision, 10*(9):6, 1–11, https://doi.org/10.1167/10.9.6.

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., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*(7141), 206–209.

Nagai, T., Ono, Y., Tani, Y., Koida, K., Kitazaki, M., & Nakauchi, S. (2013). Image regions contributing to perceptual translucency: A psychophysical reverse-correlation study. *i-Perception, 4*(6), 407–428.

Nishida, S. Y. (2019). Image statistics for material perception. *Current Opinion in Behavioral Sciences, 30*, 94–99.

Nishida, S. Y., Kawabe, T., Sawayama, M., & Fukiage, T. (2018). Motion perception: From detection to interpretation. *Annual Review of Vision Science, 4*, 501–523.

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 and Image Science, and Vision, 15*(12), 2951–2965.

Nishio, A., Goda, N., & Komatsu, H. (2012). Neural selectivity and representation of gloss in the monkey inferior temporal cortex. *Journal of Neuroscience, 32*(31), 10780–10793.

Nishio, A., Shimokawa, T., Goda, N., & Komatsu, H. (2014). Perceptual gloss parameters are encoded by population responses in the monkey inferior temporal cortex. *Journal of Neuroscience, 34*(33), 11143–11151.

Oishi, Y., Imamura, T., Shimomura, T., & Suzuki, K. (2018). Visual texture agnosia in dementia with Lewy bodies and Alzheimer's disease. *Cortex, 103*, 277–290.

Okazawa, G., Koida, K., & Komatsu, H. (2011). Categorical properties of the color term "GOLD". *Journal of Vision, 11*(8):4, 1–19, https://doi.org/10.1167/11.8.4.

Okazawa, G., Tajima, S., & Komatsu, H. (2017). Gradual development of visual texture-selective properties between macaque areas V2 and V4. *Cerebral Cortex, 27*(10), 4867–4880.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proceedings of the National Academy of Sciences, 112*(4), E351–E360.

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.

Paulun, V. C., Schmidt, F., van Assen, J. J. R., & Fleming, R. W. (2017). Shape, motion, and optical cues to stiffness of elastic objects. *Journal of Vision, 17*(1):20, 1–22, doi:10.1167/17.1.20.

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000). Toward a psychophysically-based light reflection model for image synthesis. *SIGGRAPH '00: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 55–64). New York: Association for Computing Machinery.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex

wavelet coefficients. *International Journal of Computer Vision, 40*(1), 49–70.

Prins, N., & Kingdom, F. A. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes Toolbox. *Frontiers in Psychology, 9*, 1250.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.

Saarela, T. (2018). ShapeToolbox: Creating 3D models for vision research. *Journal of Vision, 18*(10), 229, https://doi.org/10.1167/18.10.229.

Saarela, T., & Olkkonen, M. (2016). ShapeToolbox. Retrieved from https://github.com/saarela/ShapeToolbox.

Sawayama, M., Adelson, E. H., & Nishida, S. (2017a). Visual wetness perception based on image color statistics. *Journal of Vision, 17*(5):7, 1–24, https://doi.org/10.1167/17.5.7.

Sawayama, M., & Nishida, S. Y. (2018). Material and shape perception based on two types of intensity gradient information. *PLoS Computational Biology, 14*(4), e1006061.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*.

Sawayama, M., Nishida, S., & Shinya, M. (2017b). Human perception of subresolution fineness of dense textures based on image intensity statistics. *Journal of Vision, 17*(4):8, 1–18, https://doi.org/10.1167/17.4.8.

Schmid, A. C., & Doerschner, K. (2018). Shatter and splatter: The contribution of mechanical and optical properties to the perception of soft and hard breaking materials. *Journal of Vision, 18*(1):14, 1–32, https://doi.org/10.1167/18.1.14.

Schmid, A. C., Barla, P., & Doerschner, K. (2021). Material category of visual objects computed from specular image structure. *bioRxiv*, 2019–12.

Schmidt, F., Paulun, V. C., van Assen, J. J. R., & Fleming, R. W. (2017). Inferring the stiffness of unfamiliar objects from optical, shape, and motion cues. *Journal of Vision, 17*(3):18, 1–17, https://doi.org/10.1167/17.3.18.

Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour, 5*(10), 1402–1417.

Sun, H.-C., Ban, H., Di Luca, M., & Welchman, A. E. (2015). fMRI evidence for areas that process surface gloss in the human visual cortex. *Vision Research, 109*, 149–157.

Tamura, H., Prokott, K. E., & Fleming, R. W. (2019). Distinguishing mirror from glass: A 'big data' approach to material perception. *arXiv*, arXiv:1903.01671v1.

Tsuda, H., Fujimichi, M., Yokoyama, M., & Saiki, J. (2020). Material constancy in perception and working memory. *Journal of Vision, 20*(10):10, 1–16, https://doi.org/10.1167/jov.20.10.10.

Turk, D. J., Heatherton, T. F., Kelley, W. M., Funnell, M. G., Gazzaniga, M. S., & Macrae, C. N. (2002). Mike or me? Self-recognition in a split-brain patient. *Nature Neuroscience, 5*(9), 841–842.

van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current Biology, 28*(3), 452–458.

van Assen, J. J. R., & Fleming, R. W. (2016). Influence of optical material properties on the perception of liquids. *Journal of Vision, 16*(15):12, 1–20, doi:10.1167/16.15.12.

van Assen, J. J. R., Nishida, S. Y., & Fleming, R. W. (2020). Visual perception of liquids: Insights from deep neural networks. *PLoS Computational Biology, 16*(8), e1008018.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11).

Vangorp, P., Laurijssen, J., & Dutré, P. (2007). The influence of shape on the perception of material reflectance. *ACM Transactions on Graphics, 26*(3), 267–276.

Walter, B., Marschner, S. R., Li, H., & Torrance, K. E. (2007). Microfacet models for refraction through rough surfaces. *Proceedings of the 18th Eurographics Conference on Rendering Techniques* (pp. 195–206). New York: Association for Computing Machinery.

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

Xiao, B., Walter, B., Gkioulekas, I., Zickler, T., Adelson, E., & Bala, K. (2014). Looking against the light: how perception of translucency depends on lighting direction. *Journal of Vision, 14*(3):17, 1–22, https://doi.org/10.1167/14.10.1316.

Xiao, B., Bi, W., Jia, X., Wei, H., & Adelson, E. H. (2016). Can you see what you feel? Color and folding properties affect visual–tactile material discrimination of fabrics. *Journal of Vision, 16*(3):34, 1–15, https://doi.org/10.1167/16.3.34.

Xiao, B., Zhao, S., Gkioulekas, I., Bi, W., & Bala, K. (2020). Effect of geometric sharpness on translucent material perception. *Journal of Vision, 20*(7):10, 1–17, https://doi.org/10.1167/jov.20.7.10.

Yang, J., Kanazawa, S., Yamaguchi, M. K., & Motoyoshi, I. (2015). Pre-constancy vision in infants. *Current Biology, 25*(24), 3209–3212.

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*(3), 356–365.

Zhang, F., de Ridder, H., Barla, P., & Pont, S. (2019). A systematic approach to testing and predicting light-material interactions. *Journal of Vision, 19*(4):11, 1–22, https://doi.org/10.1167/19.4.11.

Appendix A. Crowdsourcing experiment

The results of the crowdsourcing experiment are shown in Figures A1 to A6. The same experiments were also conducted in the laboratory environment, and their results are shown in Figures 8 to 13 in the text.

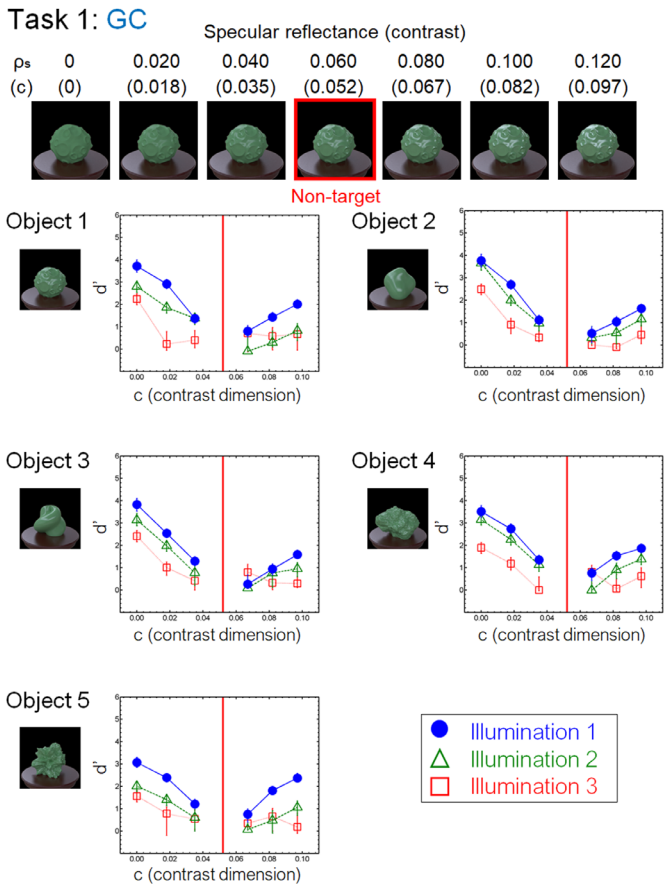

Figure A1. Results of task 1 (GC) in the crowdsourcing experiment. Different panels show different objects. Different symbols in each panel depict different illumination conditions. The vertical red line in each panel indicates the parameter of the non-target stimulus. Error bars indicate the 95% bootstrap confidence intervals.

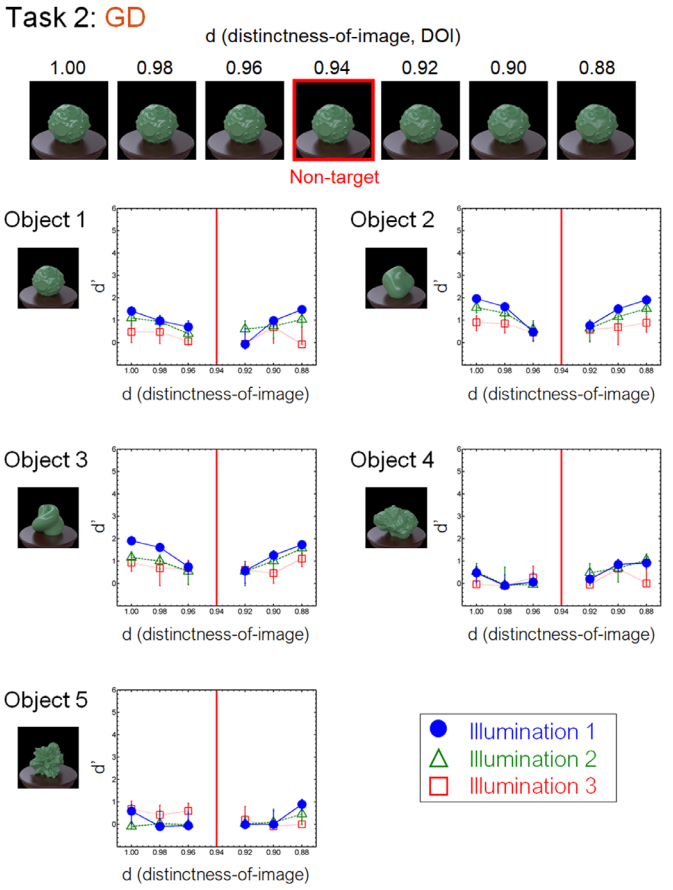

Figure A2. Results of task 2 (GD) in the crowdsourcing experiment.

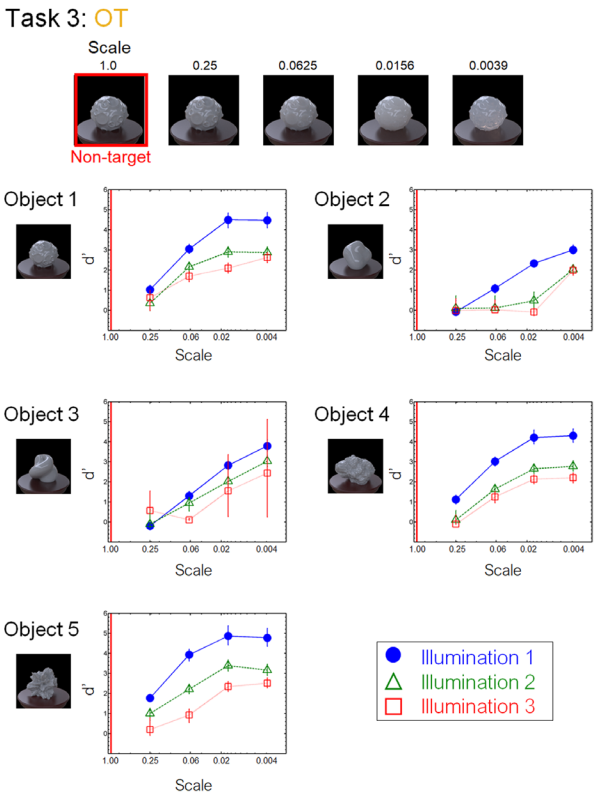

Figure A3. Results of task 3 (OT) in the crowdsourcing experiment.

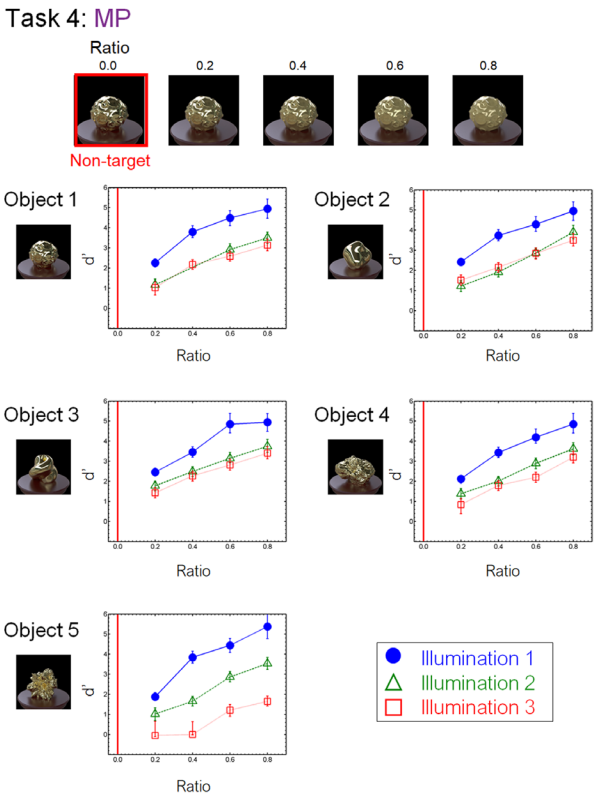

Figure A4. Results of task 4 (MP) in the crowdsourcing experiment.

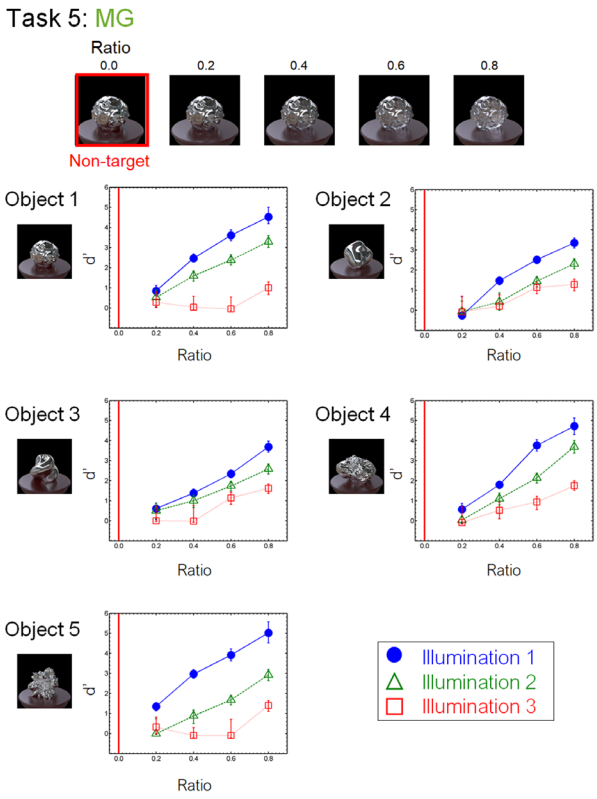

Figure A5. Results of task 5 (MG) in the crowdsourcing experiment.

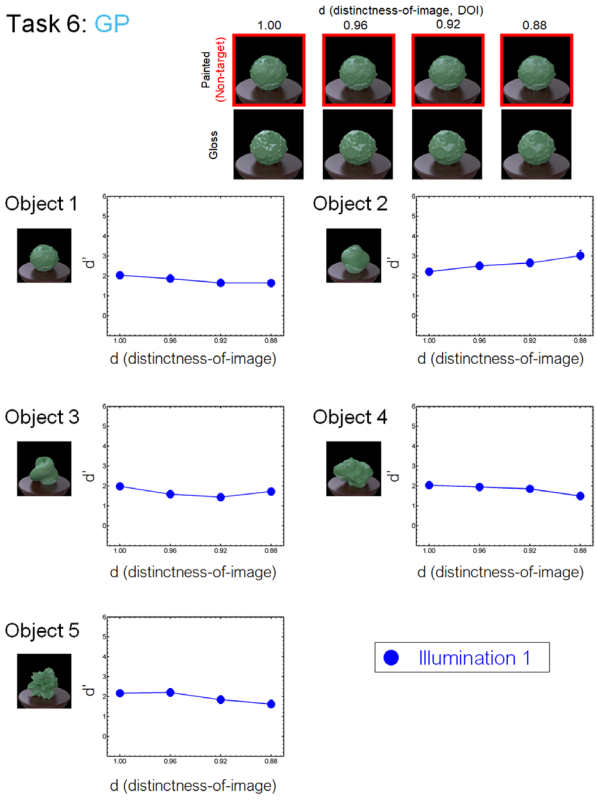

Figure A6. Results of task 6 (GP) in the crowdsourcing experiment.

Appendix B. Data records

The database is available at https://github.com/mswym/material_dataset. Figure B1 shows the data structure. The standard data are divided into three folders according to the illumination conditions. Each illumination condition folder contains folders of the material tasks (tasks 1 to 6). Each material task folder includes experimental task folders. Each experimental task folder corresponds to one task in the behavioral experiments. The name of each folder indicates the illumination condition, object, material task, and task level. For example, the name "Il1_obj1_Task1_06_12" indicates illumination condition 1 (Il1), object 1 (obj1), task 1 (Task1), contrast of 0.06 for the non-target stimulus, and contrast of 0.12 for the comparison stimulus.

Each task folder contains the two folders named "1" and "0". The images in folder "0" indicate the non-target stimuli, and the images in folder "1" are the target stimuli. Under illumination condition 1, three images are randomly selected from folder "0", and one correct image is selected from folder "1." Five images with different poses are stored in each "1" or "0" folder for illumination condition 1, whereas three images with different illuminations are stored for illumination conditions 2 and 3. The images in the database are in PNG format and have a size of 512 × 512 pixels. In addition, standard observer data are placed on the top layer in the database in a CSV file. The file contains observer data including the probability of the correct response and the sensitivity d' for each task in the crowdsourcing and laboratory experiments.

Appendix C. Convolutional neural networks

We analyzed how our datasets are represented in convolutional neural networks (CNNs). We extracted the visual features from each intermediate layer of a CNN. We used VGGNet16 (Simonyan and Zisserman, 2015), pretrained for the object recognition task using ImageNet 2012 (Russakovsky et al., 2015), and computed the activation of 30 convolution layers and three fully connected layers of the model. To reduce the number of dimensions, we spatially averaged the activation of each channel. Thus, we obtained the multidimensional activation vector for each layer with the dimension number of the channels.

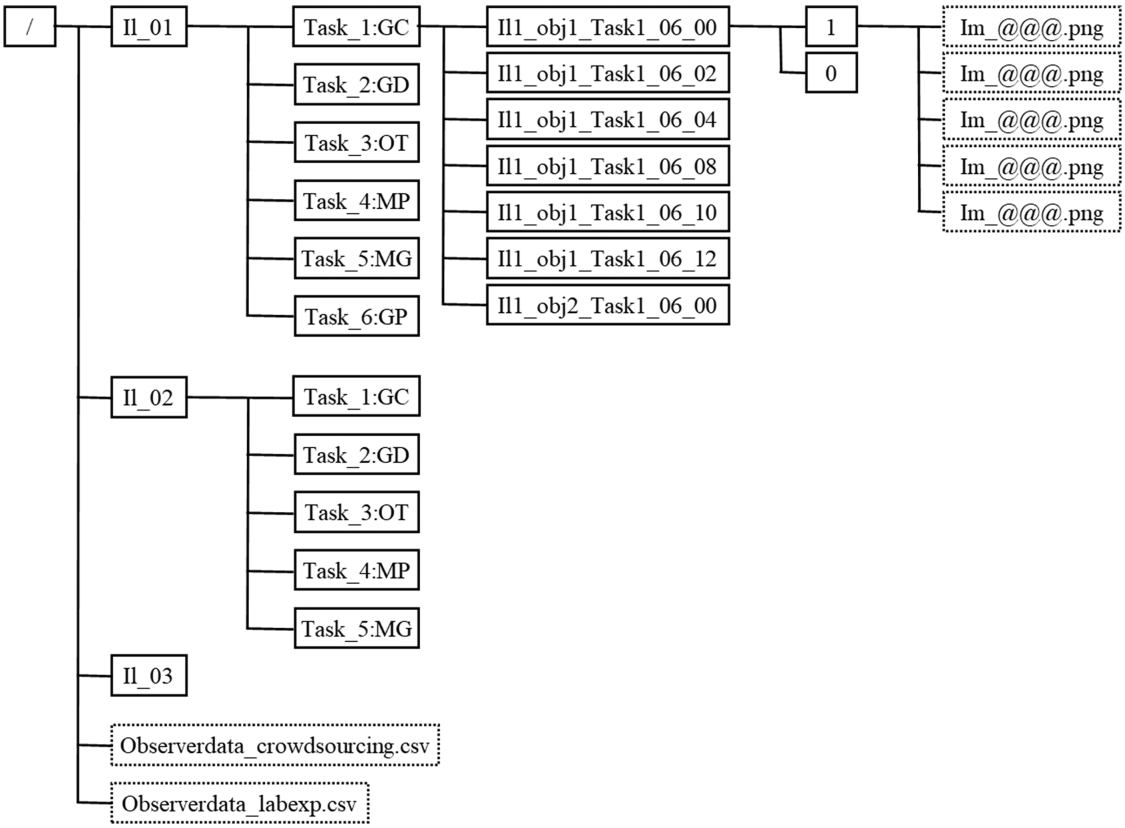

Figure B1. Data structure in the database. Solid rectangles indicate a folder; dashed ones indicate a file.

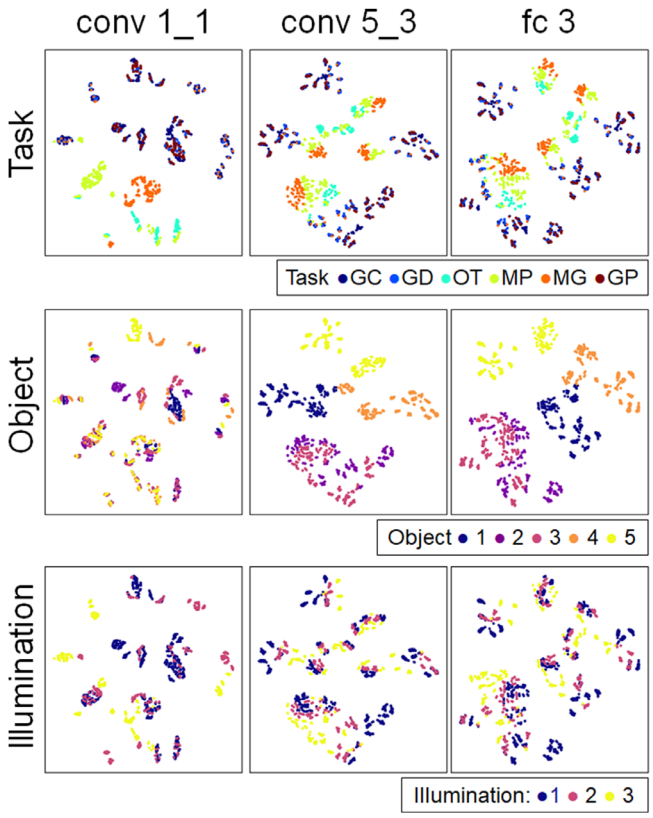

Figure C1. Embedding spaces of intermediate features of a deep neural network trained for object recognition. The top, center, and bottom rows show the same embedding spaces with different color symbols as indicated in the legend. The left, middle, and right columns are the results of the first convolution layer (conv1_1), the final convolution layer (conv5_3), and the third fully connected layer (fc 3), respectively.

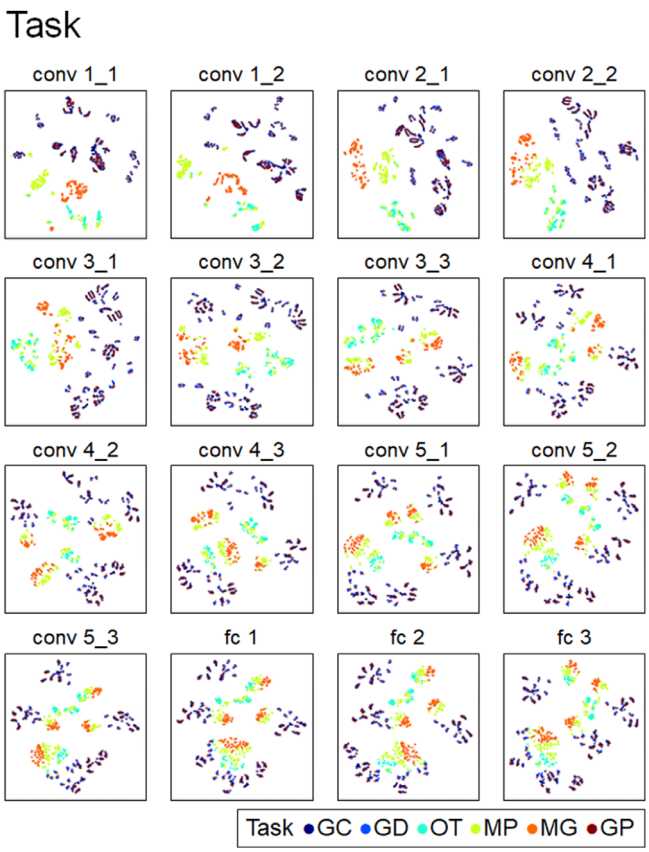

Figure C2. Embedding spaces of intermediate features of a deep neural network trained for object recognition. Results of all of the 16 layers are shown with coloring for different tasks.

Figures C1 to C4 show the t-SNE embedding of each layer (Maaten & Hinton, 2008). Figure C1 shows the results of the first convolution layer (conv 1_1), the last convolution layer (conv 5_3), and the third fully connected layer. Each plot indicates each material image. Different panels in each column indicate different labelings based on task, object, and illumination, as shown in the legends. Figure C2 shows the embeddings of all of the layers, which are colored by different tasks. Figures C3 and C4 show the same embeddings as Figure C2, except colored according to different objects and illuminations, respectively.

The embedding of the first convolution layer (conv 1_1) showed the clusters according to task differences, especially the MG, MP, and OT clusters. In contrast, this embedding did not show any object-based clusters. Earlier layers are generally sensitive to lower image features. Different tasks have different colors in our datasets, except that tasks GC, GD, and GP share similar green colors. In addition, some clusters of illumination condition 3 emerged in the first layer embedding. The pixel color distribution of illumination condition 3 is also largely different from the others. These results suggest that the first layer codes such lower image features.

The embeddings of the last convolution layer and the third fully connected layer show the clusters according to object differences. Different tasks and illuminations are separately distributed within each object cluster. Although the embedding is clustered according to object differences, it does not show the separation between objects 2 and 3. This finding is consistent with human discrimination performance. The results of behavioral experiments indicate that the task accuracies of objects 2 and 3 were similar to each other and different from other object conditions, especially for tasks GD and OT.

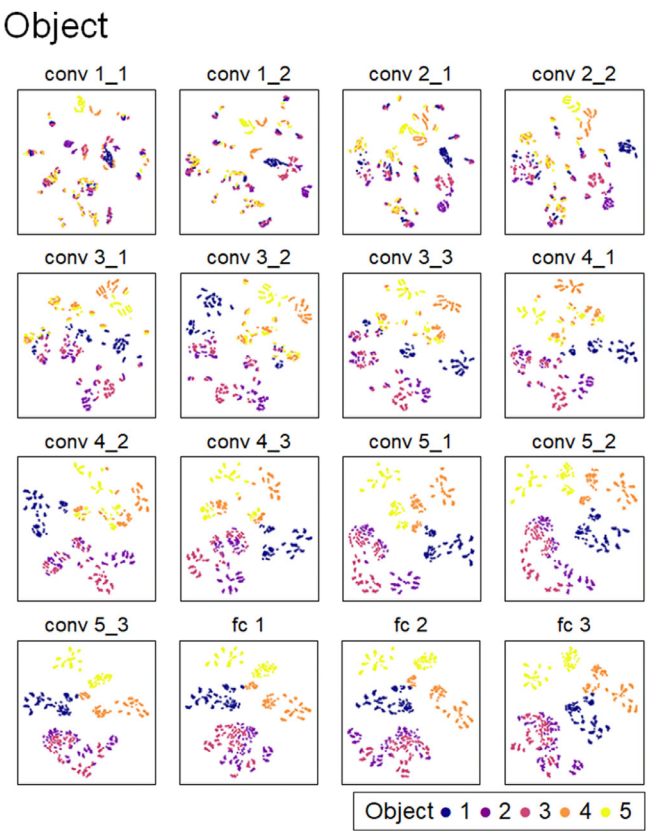

Figure C3. Embedding spaces of intermediate features of a deep neural network trained for object recognition. Results of all of the 16 layers are shown with coloring for different objects.

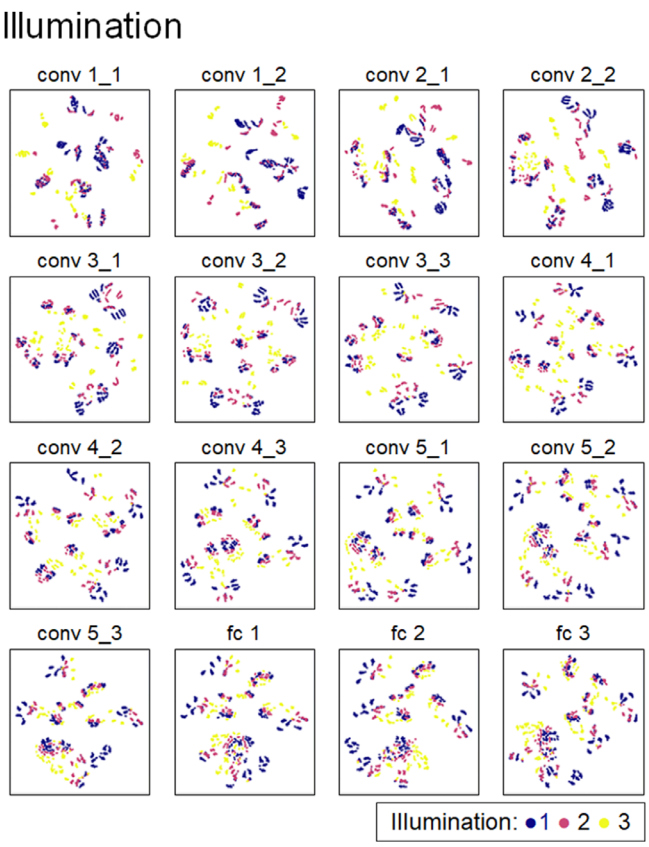

Figure C4. Embedding spaces of intermediate features of a deep neural network trained for object recognition. Results of all of the 16 layers are shown with coloring for different objects.