Title

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

Authors

Masataka Sawayama[1], Yoshinori Dobashi[2,3], Makoto Okabe[4], Kenchi Hosokawa[1], Takuya Koumura[1], Toni Saarela[5], Maria Olkkonen[6], & Shin'ya Nishida[1,7]

Affiliations

1. NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
2. Hokkaido University
3. Prometech CG Research
4. Shizuoka University
5. University of Helsinki
6. Durham University
7. Kyoto University

corresponding author: Masataka Sawayama (masa.sawayama@gmail.com)

Abstract

For elucidation of complex visual processing such as material perception of realistic objects, multiple research approaches should be combined. Shearing stimulus and response data is an effective strategy to make the results of different studies directly comparable, and to assist a wide range of researchers with different backgrounds to jump into the field. We therefore constructed a database of a variety of material images with standard human discrimination performance attached. We created various material images by using physically-based computer graphics techniques and conducted psychophysical experiments both in the laboratory and through crowdsourcing. The observer's task was material discrimination for six material dimensions (gloss contrast, gloss sharpness, translucent vs. opaque, gold vs. plastic yellow, glass vs. silver, and glossy vs. painted) with several task difficulties controlled by material parameters. We also varied illumination consistency and object geometry. We used a non-verbal procedure (oddity task) for our database to be widely used in diverse studies such as cross-cultural, cross-species, clinical and developmental studies. Results showed that discrimination performance was affected by illumination condition and object geometry, in agreement with previous gloss perception studies, although the pattern of effects was slightly different for some material dimensions. Specifically, we found that discrimination of translucency was easier for rugged shapes than for smooth shapes, while the opposite was the case for gloss perception. We also found that the ability to discriminate the spatial consistency of specular highlights on glossiness perception had a larger individual difference than other tasks. The performance obtained through crowdsourcing was highly correlated with that in the laboratory, which suggests that our database can be used even when the experimental conditions are not strictly controlled under non-laboratory environments. Several projects using our dataset are underway.

Introduction

We can visually recognize a variety of material properties of objects we daily encounter. Although the material properties, such as glossiness and wetness, substantially contribute to our recognition of objects, value-based decision making, and motor control, the computational and neural mechanisms underlying material perception had been unfairly overlooked until relatively recently — conventional vision science mainly used simple artificial stimuli to elucidate the underlying brain mechanisms. In the last two decades, however, along with the advancement in computer graphics and machine vision, material perception becomes one of major topics in vision science (Adelson, 2001; Fleming, 2017; Nishida, 2019).

Visual material perception can be considered as estimation of physical material-related properties from object image. For example, gloss/matte perception is visual computation of diffuse/specular refection of the surface. 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 et al. 2003), the other two main factors of image formation. Such estimation errors provided useful information as to what kind of image cues human uses to estimate gloss. In addition to gloss, a significant amount of psychophysical studies have been made on perception of other optical material properties (e.g., transparency, transparency and wetness) (Fleming et al., 2005; Motoyoshi, 2010; Xiao et al., 2014; Sawayama, Adelson, & Nishida, 2017), and mechanical properties (e.g., viscosity, elasticity) (Kawabe et al., 2015; Paulun et al., 2017; van Assen, Barla & Fleming, 2018). Neurophysiological and neuroimaging studies have revealed various neural mechanisms underlying material perception (Kentridge et al., 2012; Nishio et al., 2012, 2014; Miyakawa et al., 2017). Some recent studies also focus on developmental, environmental and clinical factors of material processing (Yang et al., 2015; Goda et al., 2016; Ohishi et al. 2018). For instance, Goda et al. (2016) showed in their monkey fMRI study that the visuo-haptic experience for material objects alters the visual cortical representation of objects. Besides, a large individual difference in visual color and material recognition for a photo (#TheDress) has attracted a broad range of interest, and provoked an intensive discussion (Brainard & Hurlbert, 2015; Gegenfurtner et al., 2015).

To deepen our understanding of material perception from a more global point of view, we need to promote such multidisciplinary studies that compare behavioral/physiological responses of humans or animals obtained under a variety of developmental, environmental, cultural and clinical conditions. There are two problems however. One lies in high degree of freedom in selecting experimental stimulus parameters and task procedures. Since material appearance depends not only on high-dimensional reflectance parameters, but also on geometry and illumination, both of which are also high dimensional, use of different stimuli (and different tasks) in different studies could impose serious limitations in direct comparing the data. The other problem is technical expertise necessary for rendering realistic images, which could be an obstacle for researchers unfamiliar with graphics to start a new material perception study. Aiming at removing the obstacles, the present study attempted to build a standard image set for material recognition, and

collect the behavioural responses of a large number of "standard" observers participating in general purpose discrimination tasks.

To make our database as generally useful as possible, we used the diversity of material attributes. We controlled material attributes in six different dimensions by using computer graphics techniques (Fig.1). Three of them are related to gloss (Fig. 1, Task 1: GC, Task 2: GD, and 6: GP), the most widely investigated material attribute in the past (Pellacini et al., 2000; Fleming et al., 2003; Motoyoshi et al., 2007; Olkkonen & Brainard, 2010; Doerschner et al., 2011; Kim et al., 2011; Marlow et al., 2011; 2012; Kentridge et al., 2012; Sun et al., 2015; Nishio et al., 2014; Adams et al., 2016; Miyakawa et al., 2017). We controlled two dimensions of gloss: contrast and sharpness, as in previous studies (Pellacini et al., 2000; Fleming et al., 2003; Nishio et al., 2014). For instance, Nishio et al. (2014) found neurons in the inferior temporal cortex (ITC) of the monkey that selectively and parametrically respond to the gloss changes in the two dimensions. Another stimulus manipulation for gloss perception (Fig. 1, Task 6: GP) was spatial consistency of specular highlights: by breaking the consistency, some highlights look like white paint (albedo change) (Beck & Prazdny, 1981; Kim et al., 2011; Marlow et al., 2011; Sawayama & Nishida, 2018). Besides gloss perception, translucency perception has been most widely investigated (Fleming & Bülthoff, 2005; Motoyoshi, 2010; Nagai et al., 2013; Gkioulekas et al., 2013; Xiao et al., 2014; Chadwick et al., 2018). We adopted the task of discriminating opaque from translucent objects by controlling the thickness of the translucent media (Fig. 1, Task 3: OT). Furthermore, we adopted the tasks of plastic-yellow/gold discrimination (Okazawa et al., 2011, Task 4: MP) and glass/silver discrimination (Kim & Marlow, 2016; Tamura et al., 2019, Task 5: MG).

To evaluate the human discrimination performance for each material dimension, we used an oddity task (Fig. 3). We chose this task, since it requires neither complex verbal task instruction to the observer, nor verbal response by the observer. Therefore, the same task can be applied to a wide range of observers including infants, animals and machine vision algorithm, and the task performance can be directly compared across them. Indeed, several research projects using our dataset are underway (see Discussion).

To control the task difficulty, we changed the parameter value on each material dimension. In addition, we introduced two stimulus manipulations that affected the task difficulty. We used three display conditions: a set of stimuli with different pose under the identical illumination environment (Fig. 2a, Display condition 1); a set of stimuli with identical pose under the different, but similar to one another, illumination environment (Fig. 2a, Display condition 2); and a set of stimuli with identical pose under the largely different illumination environment (Fig. 2a, Display condition 3)..We also used the five different geometries of objects for each material task (Fig. 2b).

We wish to collect the data of a large size of observers. By a laboratory experiment, one can achieve a good control over stimulus presentation environment, but cannot easily collect large-scale human data. In contrast, one can collect large-scale data through crowdsourcing, but loose a reliable stimulus control. To overcome this trade-off, we conducted identical psychophysical experiments both in the laboratory and through crowdsourcing. This enables us to evaluate

individual difference distribution, along with the effects of environmental factors on the task performance.

In sum, here we made a large set of image stimuli for evaluation of discrimination performance along six material dimensions (gloss contrast, gloss sharpness, translucency-opaque, plastic-gold, glass-silver and glossy-painted), and measured the performance of the oddity task for large number of adult human observers both in the lab experiments and through crowdsourcing, using three display conditions and five object geometries. Although the main purpose of this project was to make the standard stimulus-response dataset of material recognition for promotion of multidisciplinary studies, it should be added that this is the first systematic comparison of the effects of display condition and object geometry, as well as of the individual variation, across a variety of material dimensions. Our data includes several novel findings in this respect as we see below.

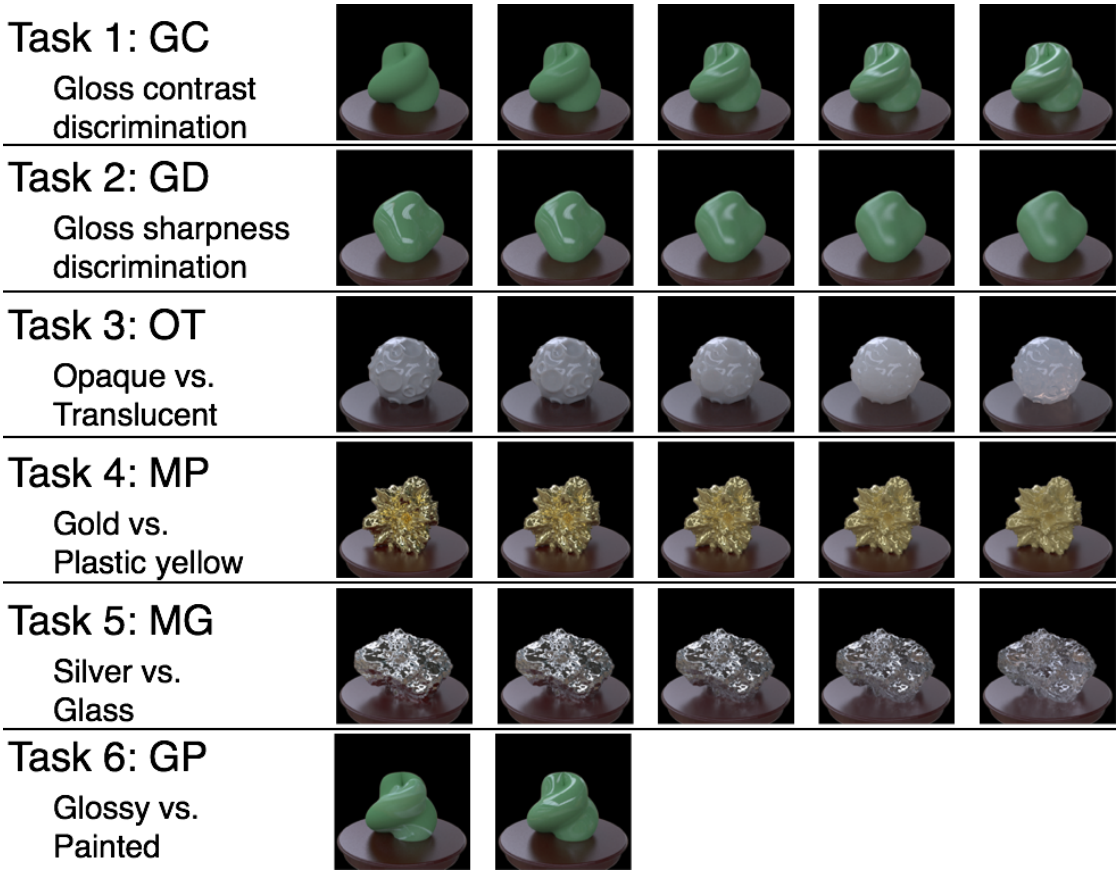

Figure 1. A schematic overview of the tasks recorded in the database. The six material tasks were conducted as in the legend.

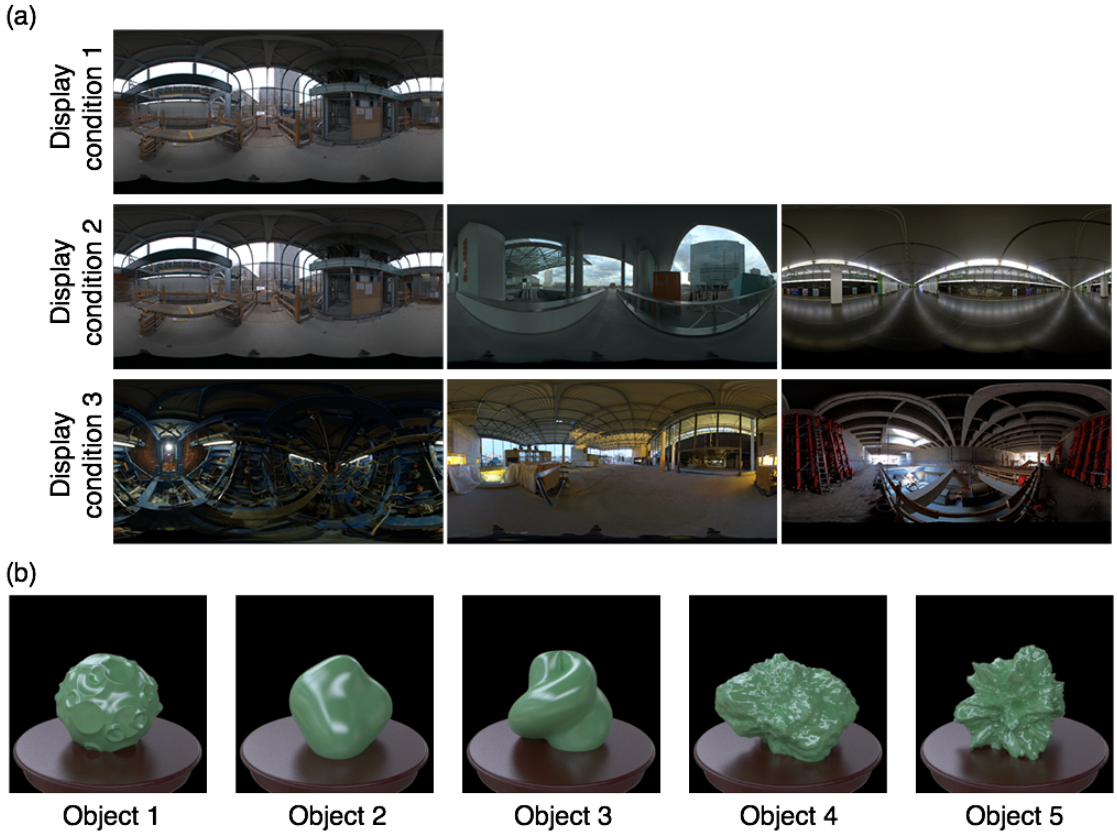

Figure 2. (a) The illumination environment used in rendering. We used three display conditions. Under the display condition 1, the objects presented on the monitor in a task were rendered by using one identical illumination environment. Under the display condition 2, the objects presented in a task were rendered by using three illumination environments, which are similar to each other in pixel histogram similarity. Under the display condition 3, the objects presented in a task were rendered by using three largely different illumination environments. (b) The objects used in rendering. We used five objects for each material task under each display condition.

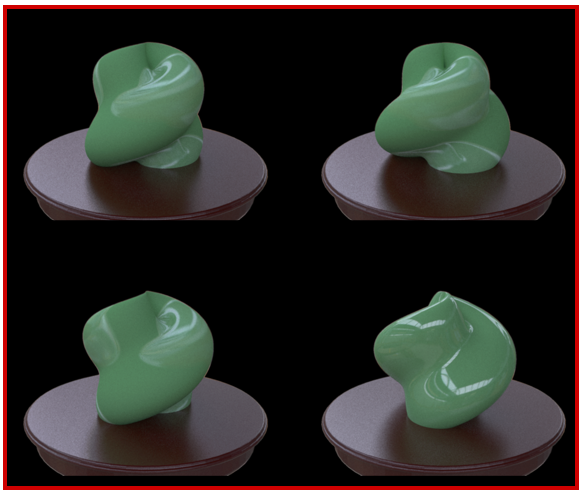

Figure 3. A display example presented in a task for collecting the standard observer data. The observers were asked to select which image is an oddity one in the four images. We did not instruct that the experiment is for material recognition. We conducted the experiments both in the laboratory and the crowdsourcing.

Methods

To collect the observer's performance for material recognition, we used six material tasks according to previous material studies: 1) Gloss discrimination (contrast dimension) (GC); 2) Gloss discrimination (sharpness dimension) (GD); 3) Opaque vs. Translucent (OT); 4) Gold vs. Plastic Yellow (MP); 5) Silver vs. Glass (MG); 6) Glossy vs. Painted (GP). For each task, we used 5 geometry models and 6 global illuminations. We conducted behavioral experiments with adopting an oddball procedure, which can be used even for human babies, animals, or brain-injured participants without complex verbal instructions. In the experiment, the observers were asked to select the oddity stimulus in the multiple object stimuli. They were not given any feedback about whether their responses were correct or not. We also controlled task difficulty by changing the illumination and material parameters. To test the generality of the database, we conducted an identical experiment both in the laboratory and the crowdsourcing.

Image generation for standard image database

To make object images with various materials we utilized the *Mitsuba* renderer, which is a physically-based rendering software (Jakob 2010). We controlled the material properties in the following six different material dimensions.

Material for the tasks of 1) Gloss discrimination (contrast dimension) (Task 1: GC) and 2) Gloss discrimination (sharpness dimension) (Task 2: GD)

To control the material property of gloss discrimination tasks, we used the bidirectional reflection distribution function (BRDF) model of Pellacini et al. (2000). The model is derived from the Ward BRDF model (Ward, 1992) so that the parameter space of the Ward model can be perceptually uniform one. The model has two parameters, i.e., the contrast (c) and the sharpness (d), and the difficulty of the two gloss discrimination tasks was controlled by separately modulating the two parameters. The parameter space of the BRDF model can be described from the Ward model as follows.

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

where ρ(θi,φi,θo,φo) is the surface reflection model, and θi, φi, and θo, φo are the incoming and outgoing directions, respectively. There are three parameters in the model; ρd, ρs, and α, where ρd is the diffuse reflectance of a surface, ρs is the energy of its specular component, and α is the spread of the specular lobe. Pellacini et al. (2000) defined the parameters *c* and *d* based on the Ward model's three parameters as follows.

$$d = 1 - \alpha$$

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

Although more physically-feasible BRDF models have been proposed for gloss simulation (Ashikmin et al., 2000; Walter et al., 2007), we adopted the present model because it has been widely used in the previous studies of psychophysics and neuroscience (Nishio et al., 2014).

For the task of gloss discrimination in contrast dimension, the contrast parameter *c* was changed in a range from 0.00 to 0.12 in 0.02 steps while keeping the value of *d* constant (*d*=0.94) (Fig. 4, *Task 1: GC*). As the contrast *c* was close to 0, the object appears as a matte one. The contrast *c* of the non-target stimulus on the task was 0.06.

For the experiment of gloss discrimination in sharpness dimension, the sharpness parameter *d* was changed in a range from 0.88 to 1.00 in 0.02 steps while keeping the value of *c* constant (*c*=0.06) (Fig. 4, *Task 2: GD*). As the sharpness *d* was close to 1.00, the highlights of the object appear as sharp ones. The sharpness *d* of the non-target stimulus on the task was 0.94.

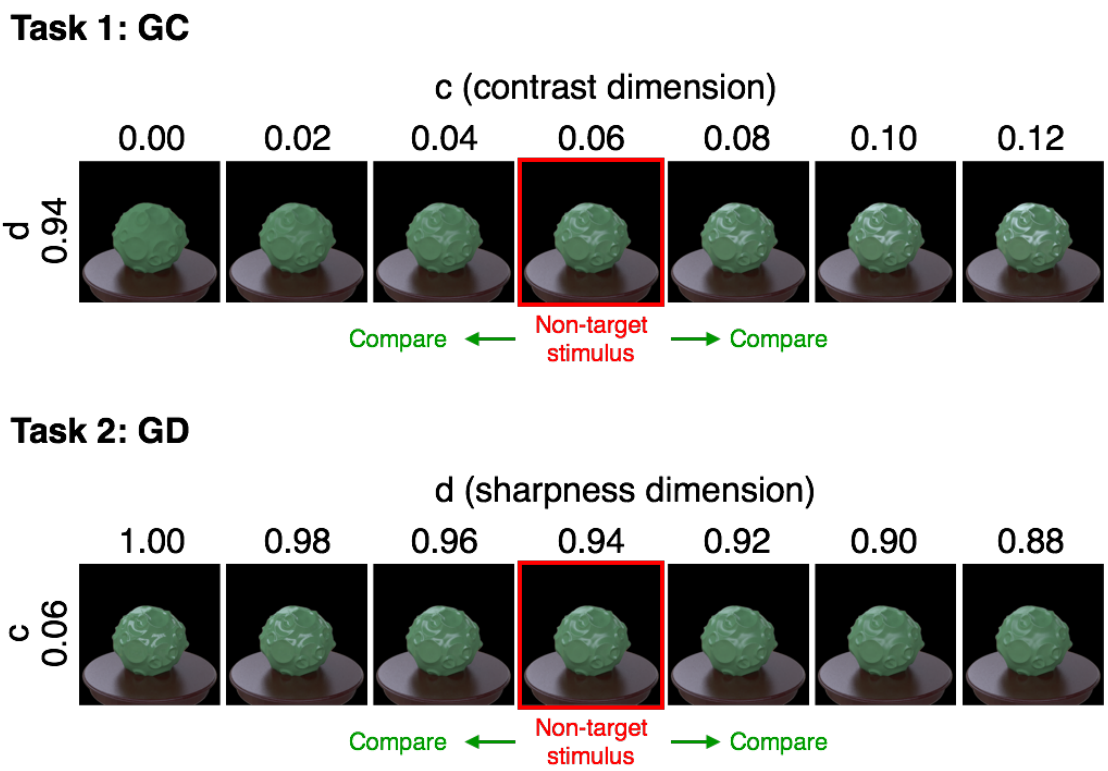

Figure 4. Material examples of the tasks 1 and 2 in the database. For the task 1, the contrast of the specular reflection changes from 0.00 to 0.12. The non-target stimuli that

were presented as the error objects in each task had the contrast of 0.06. For the task 2, the sharpness of the specular reflection changes from 1.00 to 0.88. The non-target stimuli had the sharpness of 0.94.

Material for the task of 3) Opaque vs. Translucent (Task 3: OT)

To make translucent materials we used the function of homogeneous participating medium implemented in the Mitsuba renderer. In the function, a flexible homogeneous participating medium is embedded into each object model. The intensity of the light that travels in the medium is decreased by scattering and absorption and increased by nearby scattering. The parameters of absorption and scattering coefficients of the medium describe how the light is decreased. In our database, we used the parameters of the "Wholemilk" measured by Jensen et al. (2001). The parameter of the phase function describes the directional scattering properties of the medium. In our database, we used the isotropic phase function. To control the task difficulty of translucent discrimination, we modulated the scale parameter. The parameter describes the density of the medium. The smaller the scale parameter is, the more translucent the medium is. The scale parameter was changed in a range of 0.0039, 0.0156, 0.0625, 0.25, and 1.00 (Fig. 5, Task 3: OT). The scale parameter of the non-target stimulus on the task was 1.00. In addition, the surface of the object was modeled as a smooth dielectric material to produce strong specular highlights, used in the previous studies (Gkioulekas, I. et al, 2013; Xiao et al., 2014).

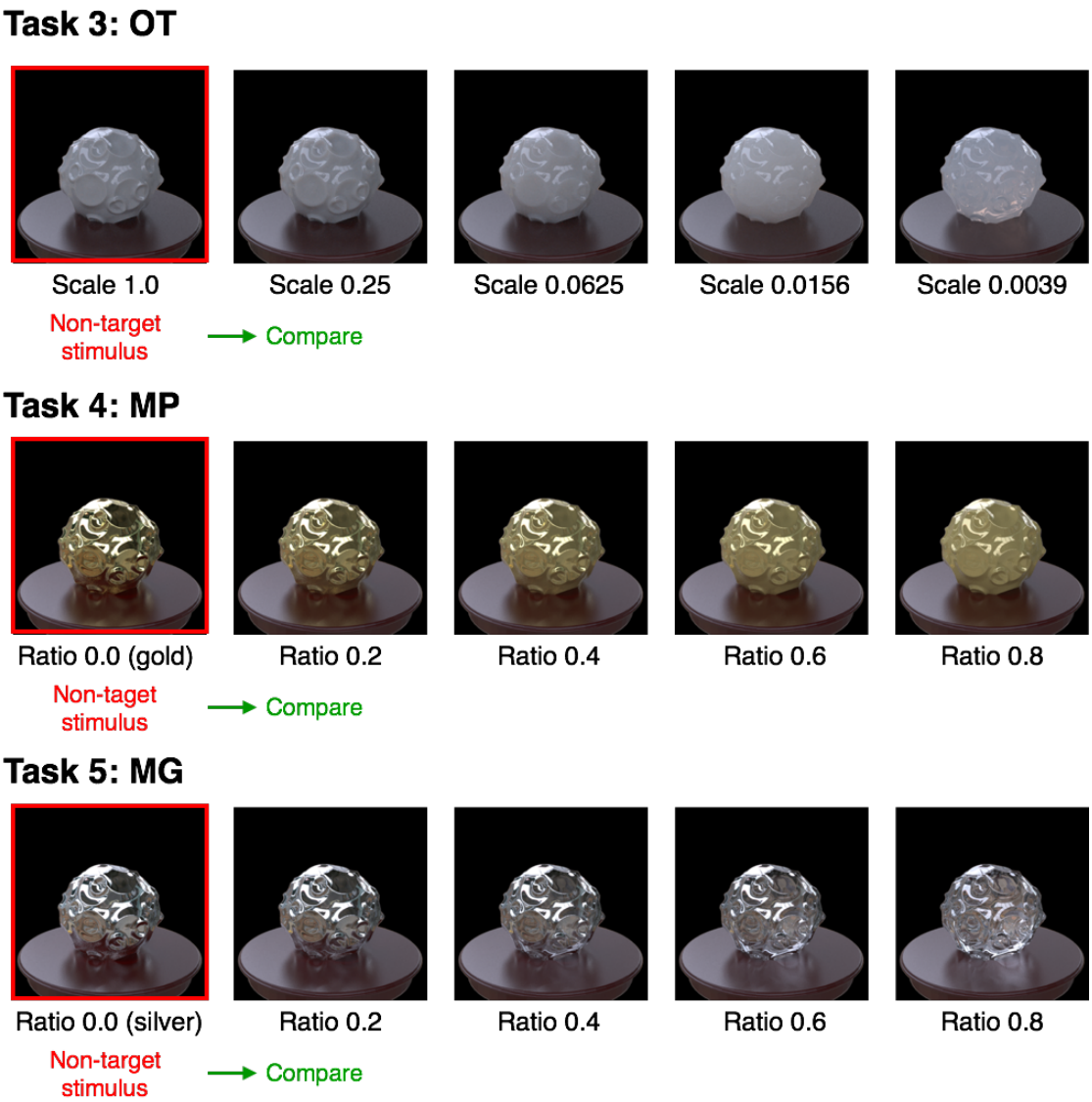

Figure 5. Material examples of the tasks 3, 4, and 5 in our database. For the task 3, the scale of the volume media that consisted from milk changed from 1.0 to 0.0039. For the task 4 and 5, the blending ratio of the two materials changed from 0.0 to 0.8. The non-target stimuli in the tasks were shown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

Material for the task of 4) Gold (Metal) vs. Plastic Yellow (Task 4: MP)

To control the materials between gold and plastic yellow, we utilized the linear combination of the gold and plastic BRDFs, which is 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). For the task of yellow-gold discrimination, the weight was changed in a range from 0.00 to 0.80 in 0.20 steps (Fig. 5, Task 4: MP). The parameter of the non-target stimulus on the task was 0 where the material was pure gold one.

Material for the task of 5) Silver (Metal) vs. Glass (Task 5: MG)

Similar to the material task 4), we utilized the linear combination of the dielectric glass and silver materials, which is also implemented in the Mitsuba renderer. The weight of the combination was changed in a range from 0.00 to 0.80. The parameter of the non-target stimulus on the task was 0 where the material was pure silver one (Fig. 5, Task 5: MG).

Material for the task of 6) Glossy vs. Painted (Task 6: GP)

The skewed intensity distribution due to specular highlights of an object image can be one of the diagnostic cues for gloss perception (Motoyoshi et al., 2007). However, when the specular highlights are inconsistent in position and/or in orientation with the diffuse shading component, they look more like white blobs produced by surface reflectance changes even if the intensity distribution is kept constant. Based on the observation, a lot of previous studies have investigated the spatial consistency of specular highlights (Beck & Prazdny; 1981; Anderson & Kim, 2009; Kim et al., 2011; Marlow et al., 2011; Sawayama & Nishida, 2018). Our database also includes these object images, i.e., glossy objects vs. matte objects with white paints. First, as in the task of gloss discrimination, we rendered the glossy objects based on the Pellacini et al. (2000)'s model. The parameter *c* was fixed at 0.08, and the parameter *d* ranged from 0.88 to 1.00 in 0.04 steps (Fig. 6, lower).

To make object images with inconsistent highlights (white paints), we extracted the highlight-only image of a glossy object by setting the diffuse reflectance to 0 under an display condition. By using the highlight-only image as a texture with a diffuse material, we mapped the texture to the object geometry and rendered it as a matte object under the same display condition (Fig. 6, lower). After we render the highlight inconsistent image, the color histogram of the image was set to that of the consistent glossy object image by using a standard histogram matching method (Sawayama & Nishida, 2018).

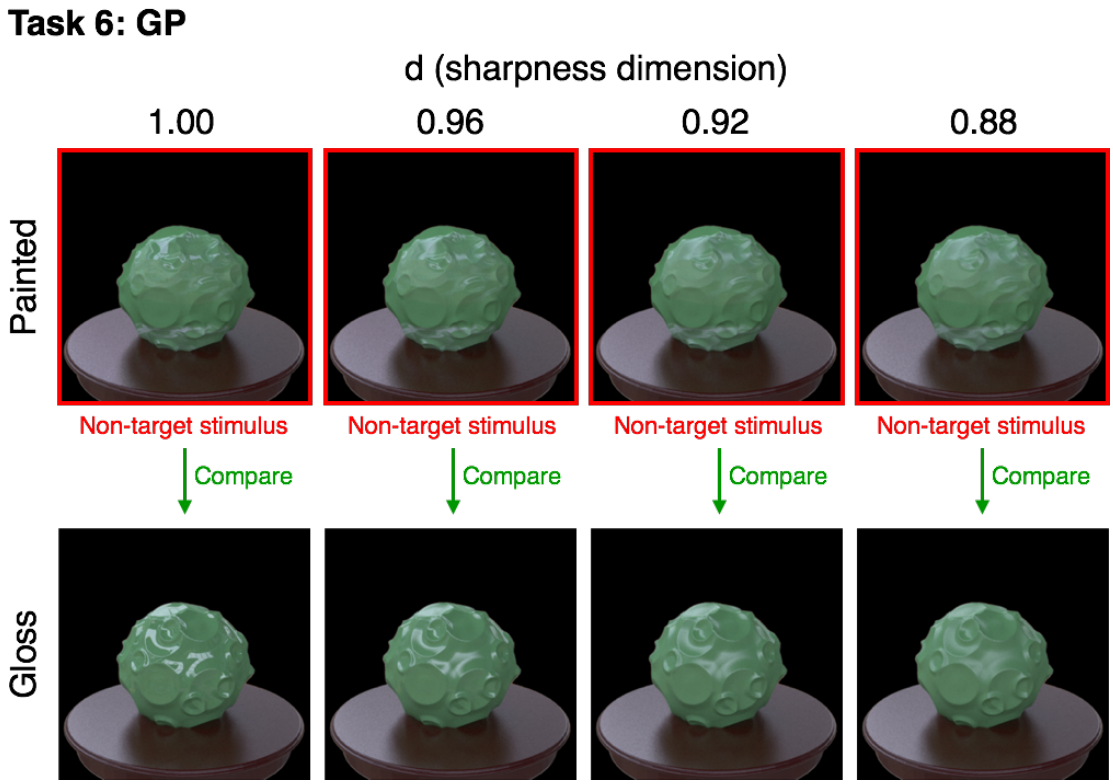

Figure 6. Material examples of the task 6 in the database. For the task 6, the sharpness of the specular reflection changed from 1.00 to 0.88. The non-target stimuli in each task were the painted objects with each specular sharpness value.

Geometry

For each material, we rendered the object images by using five different abstract geometries (Fig. 2b). These geometries were made from a sphere by modulating each surface normal direction by different noises (see also ShapeToolbox: https://github.com/saarela/ShapeToolbox) (Saarela & Olkkonen, 2016, Saarela, 2018). Specifically, the Object_1 is made from the modulations of a 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). The Object_2 is a bumped sphere modulated by a low-pass band-pass noise. The Object_3 is a bumped sphere modulated by a sine-wave noise. The Object_4 and Object_5 are bumped spheres modulated by Perlin noises with different parameters.

Illumination and pose

We used six HDR light probe images as illuminations for rendering. These images were obtained from Bernhard Vogl's light probe database (http://dativ.at/lightprobes/). To prepare several task difficulties, we used three display conditions (display conditions 1, 2, and 3, Fig. 2a). Under the display condition 1, the observers selected one oddity from four images in a task. We rendered the images by using the identical light probe (i.e., 'Overcast Day/Building Site (Metro Vienna)'). We prepared five poses for each task of the display condition 1 by rotating each object with 36-degree steps, and four of them were randomly selected in each task.

Under the display condition 2, the observers selected one oddity from three images at a task. We created the images by using different, but similar in pixel histogram similarity, light probes (i.e., 'Overcast Day/Building Site (Metro Vienna)', 'Overcast day at Techgate Donaucity', and 'Metro Station (Vienna Metro)'). The task procedure of the display condition 3 was the same as that of the display condition 2. For the display condition 3, we created the three images by using largely different light probes in pixel histogram similarity ('Inside Tunnel Machine', 'Tungsten Light in the Evening (Metro Building Site Vienna)', and 'Building Site Interior (Metro Vienna)'). We computed the histogram similarity for each illumination pair and used it as the distance for the multidimensional scaling analysis (MDS). We could extract three largely different light probes in the MDS space and used them for the display condition 3. We also selected three similar light probes in the space and used them for the display condition 2.

Rendering

To render the images we used the integrator of the photon mapping method for the task 1, 2, 4, 5, and 6, and used the integrator of the simple volumetric path tracer implemented in the Mitsuba renderer for the task 3 (OT). The calculation was conducted using single-float precision. Each rendered image was converted to sRGB format with the gamma of 2.2 and saved as an 8-bit .png image.

Behavioral experiments for standard observer database

By using the generated stimuli, the behavioral experiments were conducted both in the laboratory and in crowdsourcing.

Laboratory experiment

The participants who were participated in the experiment of the experimental room were 20 paid volunteers. All the observers were Japanese. Before staring the experiment, we confirmed that all the observers had normal color vision by using the Famsworth–Munsell 100 Hue Test, and that

all had normal or corrected-to-normal vision by using 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 EIZO color monitor (ColorEdge CG303W) controlled with an NVIDIA video card (Quadro 600). In the experiments, a 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 x 512 px was presented at the size of 8.5 x 8.5 degrees.

In each trial, four (Display 1) or three (Display 2 & 3) object images chosen from each task were presented on the monitor (Fig. 3). Measurements of different display conditions were conducted in different blocks. Under the display condition 1, the four different object images were presented in different orientations each other. Under the display conditions 2 and 3, the three different object images have different illuminations each other. The order of the illumination 1, 2, and 3 was counterbalanced across observers. The observers were asked to select which of the object images looked oddity by pushing one of the keys. The stimuli were presented until the participant made a response. The task instructions were simply to compare oddity. The observers were not given any feedback about whether their response was correct or not. All participants made 10 judgments for each task of the display condition 1. Seventeen participants made 10 judgments for each task of the display condition 2, and three participants did 7 judgments due to the experiment time limitation. Seventeen participants made 10 judgments for each task of the display condition 3, and three participants did 7 judgments due to the experiment time limitation.

Crowdsourcing experiment

On the web experiment, 416, 411, and 405 paid volunteers were participated in the tasks of the display conditions 1, 2, and 3, respectively. We recruit observers by using a Japanese commercial crowdsourcing service. All the observers participated in the display condition 3 were also participated in the display conditions 1 and 2. All participated in the display condition 2 were also done in the display condition 1. The experiment was approved by the Ethical Committees at NTT Communication Science Laboratories.

Each observer used own PC's or tablet's web browser to participate in the experiment. In the experiments, we asked observers to see the monitor from a distance of about 60 cm. We present each object image by the size of 512 x 512, but we cannot know about their precise visual angles.

The procedure was similar to that of the experiment in the experimental room. In each trial, four or three object images chosen depending on the task were presented on the monitor as in Fig. 3. On the web experiment, the measurement of the display condition 1 was first conducted, followed by the display condition 2, then the display condition 3. As in the experiment on the

experimental room, the observers were asked to select which of the object images looked oddity by clicking one of the images. The stimuli were presented until the participant made a response. In the task, no response feedback was given. Each participant made one judgment for each condition. The other methods were the same as those in the experiment in the experimental room.

Data records

The database is available at https://www.dropbox.com/s/6bh1ncm8mv3i7dx/material_swym.zip?dl=0 [Currently, the database is placed on a Dropbox folder, but we will share this from our project page later]. Figure 7 shows the data structure in the database. The standard data is first divided into three folders according to the display conditions. In each display condition folder, the folders of the material tasks (Task 1 to 6) are stored. Each material task folder includes the experimental task folders. Each experimental task folder corresponds to one task in the behavioural experiments. The name of each folder indicates the display condition, the object, the material task, and the task level. For instance, the name "Il1_obj1_Task1_06_12" indicates the display condition 1 (i.e., Il1), the object 1 (i.e., obj1), the task 1 (Task1), the contrast of 0.06 for the non-target stimulus, and the contrast of 0.12 for the comparison stimulus.

In each task folder, the two folders named "1" and "0" are stored. The images in the folder "0" indicate the non-target stimuli, while the images in the folder "1" are the target stimuli. Under the display condition 1, three images were randomly selected from the folder "0", and one correct image was selected from the folder "1". The five images with different poses are stored in each "1" or "0" folder for the display condition 1, while the three images with different illuminations are stored for the display conditions 2 and 3. The images in the database are .png format with a size of 512 x 512 px. In addition, standard observers data were placed on the top layer in the database as a .csv file. The file includes the observer data about the probability of the correct response and the sensitivity d' for each task on the crowdsourcing and the laboratory experiments.

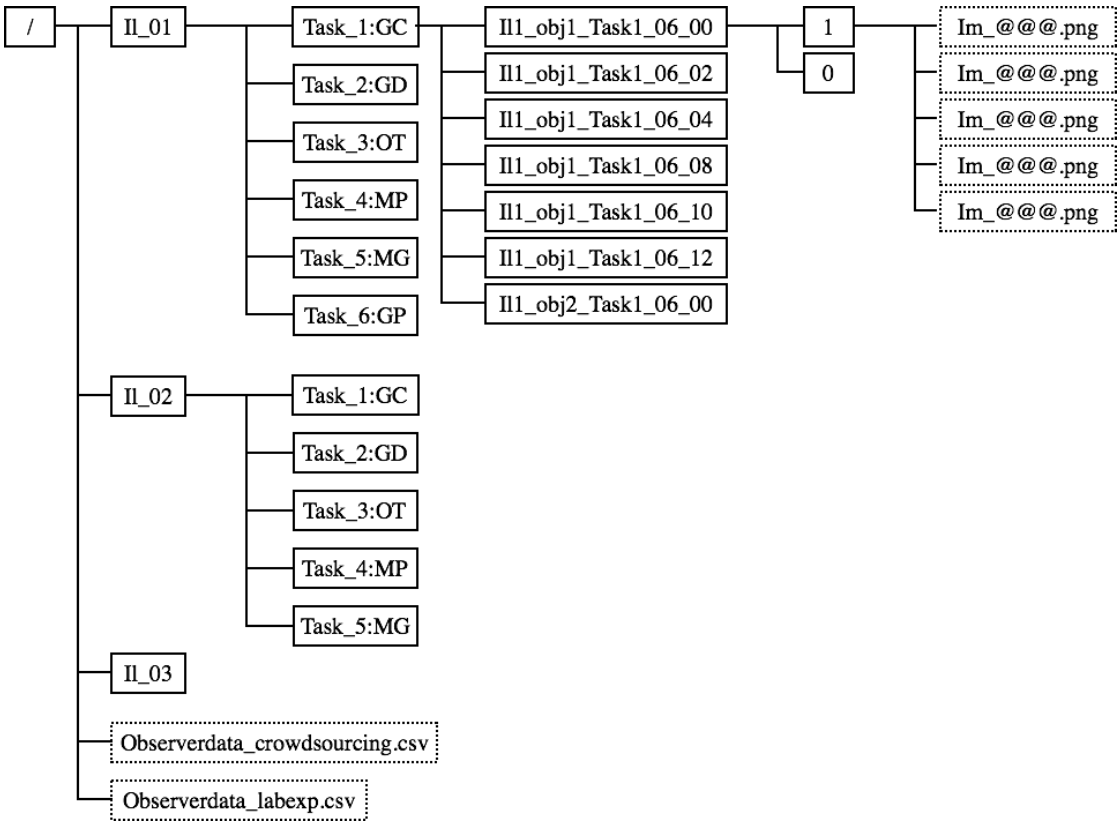

Figure 7. The data structure in the database. Each solid square indicates a folder, while the dashed one indicates a file.

Data analysis

For each oddity task, we computed the proportion correct on each participant. The chance level of the proportion correct was 0.25 for the display condition 1, and 0.33 for the display conditions 2 and 3. We computed the sensitivity d' from each proportion correct by using numerical simulation to estimate the sensitivity for an oddity task (Craven, 1992). We used a data analysis library "Palamedes" for the simulation (Kindom & Prins, 2010; 2016; Prins & Kingdom, 2018). To avoid the infinity, the proportion correct of 1 was replaced to 0.999 in the simulation. For the laboratory experiment, we computed the sensitivity d' on each observer and averaged it across observers. For the crowdsourcing experiment, since each observer engaged in each task one time, we computed the proportion correct of each task from all observers' responses and used it to compute the sensitivity d'.

Results

Environment dependence

For cross-cultural, cross-species, brain-dysfunction, and development studies, the stimulus presentation on a monitor cannot be always controlled strictly because of apparatuses or ethical limitations. Therefore, the performance validation of each task across different apparatuses is critical for the users to decide which tasks the users should select in their experimental environment. Figure 8a shows the results of the correlation analysis between the laboratory and crowdsourcing. The coefficient of determination (R^2) of the linear regression between the sensitivity d' on the laboratory and the crowdsourcing is 0.83, indicating a high linear correlation. However, the slope of the regression is lower than 1. This shows that the sensitivity of the crowdsourcing experiment is worse than that of the laboratory experiment with many repetitions in general. These findings suggest that the present tasks keep relative performance across different experimental environments in general.

In more detail, Figure 8b shows the results of each task between the laboratory and the crowdsourcing. The coefficients of determinant (R^2) on the task 1 to 6 are 0.66, 0.38, 0.81, 0.67, 0.67, and 0.33, respectively. The coefficient of task 6 (GP) was the worst, followed by task 2 (GD). As in the latter section, task 6 (GP) also shows a large individual difference and thus the correlation was decreased. The slope of the linear regression on task 2 (GD) was 0.34, and the proportion corrects in the crowdsourcing for tasks 2 were generally lower than those in the laboratory for tasks 2. In the experiment in the laboratory, we used an LCD monitor of 30 inches, and the stimulus size of each image was presented at the size of 8.5 x 8.5 degrees, which is expected to be larger than when participants on the web observe the image through ordinary monitors of iPad or PCs. Task 2 (GD) is related to the sharpness of the specular reflection, and thus the spatial resolution might affect the accuracy of the observers' responses, although the relative difficulty for task 2 (GD) even in the crowdsourcing was similar to that in the laboratory. The finding suggests that the absolute accuracy of task 2 (GD) largely depends upon the experimental environment.

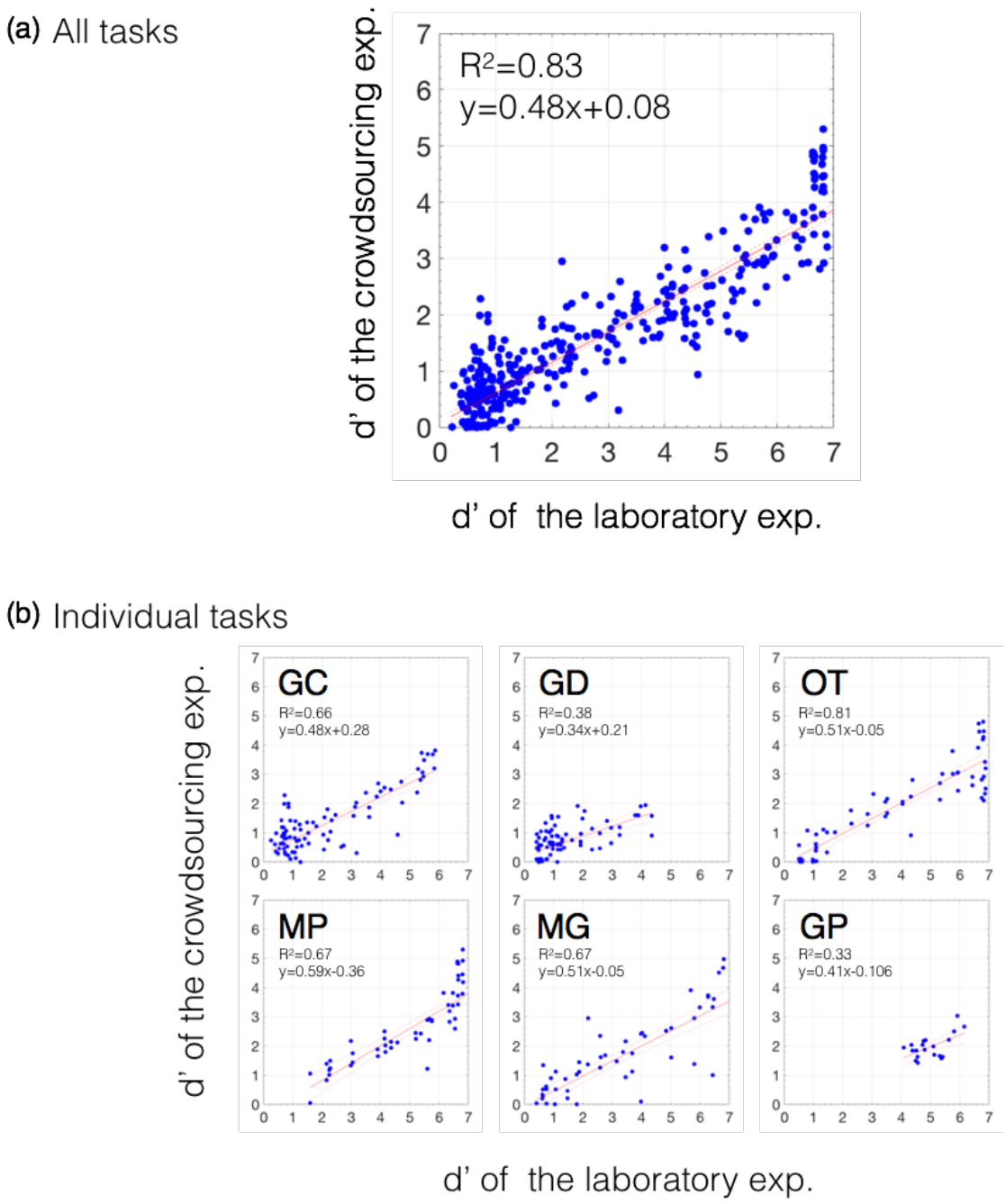

Figure 8. Results of the experiments in the laboratory and the crowdsourcing. The sensitivity d' on each task in the crowdsourcing is plotted as a function of that in the laboratory. (a) The results of all tasks. Each plot indicates a task with an object, an illumination, and a difficulty. The red line indicates the linear regression between the results on the crowdsourcing and the laboratory. The coefficient of determination (R^2) of the regression and the equation are shown in the legend. The results show that the present tasks are robust across different experimental environments in general. (b) The results of individual tasks. Different panels indicate different material tasks. Each plot in a panel indicates a task with an object, an illumination, and a difficulty in the material task. The red line indicates the linear regression between the results between the laboratory

and the crowdsourcing. 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 was generally lower than that in the laboratory. The correlation of task 6 (GP) between the laboratory and the crowdsourcing was the worst.

Performance for each task

Figures 9 to 14 show the human performance of each task in the laboratory experiment. Different panels depict different objects. Different symbols in each panel depict different display conditions. The results of the crowdsourcing experiment were shown in Appendix A. For each task in the identical display and object conditions, i.e., for the identical symbols of each panel in Fig. 9 to 14, human discrimination accuracy was monotonically improved as the material parameter of the target stimulus deviated from that of the non-target stimulus. The finding suggests that our material manipulation appropriately controlled the task difficulty.

Since we comprehensively collect the human data of material recognition across different stimulus conditions, our database also shows novel findings of material recognition, which is overlooked in the previous literature. Specifically, we found geometry dependence on material recognition. The geometry dependence on glossiness perception has been reported in several previous studies (Nishida & Shinya, 1998; Vangorp, Laurijssen, & Dutré, 2007). Our results also show the geometry dependence on glossiness perception. When the object images change with glossiness sharpness dimension (task 2:GD, Fig. 10), observers could detect the material difference better for the plane geometries (Object 2 & 3) than for the rugged objects (Object 4 & 5). In contrast, when the object images change with the glossiness contrast dimension (task 1: GC, Fig. 9), little geometry dependence was obtained. It is noteworthy that our results also show the geometry dependence on translucency perception (task 3: OT, Fig. 11). Similar to the dependence on glossiness sharpness, the sensitivity changed between the smooth objects (Object 2 & 3) and the rugged objects (Object 4 & 5). Specifically, the translucent difference was easily detected for the rugged objects than for the smooth objects (Fig. 11).

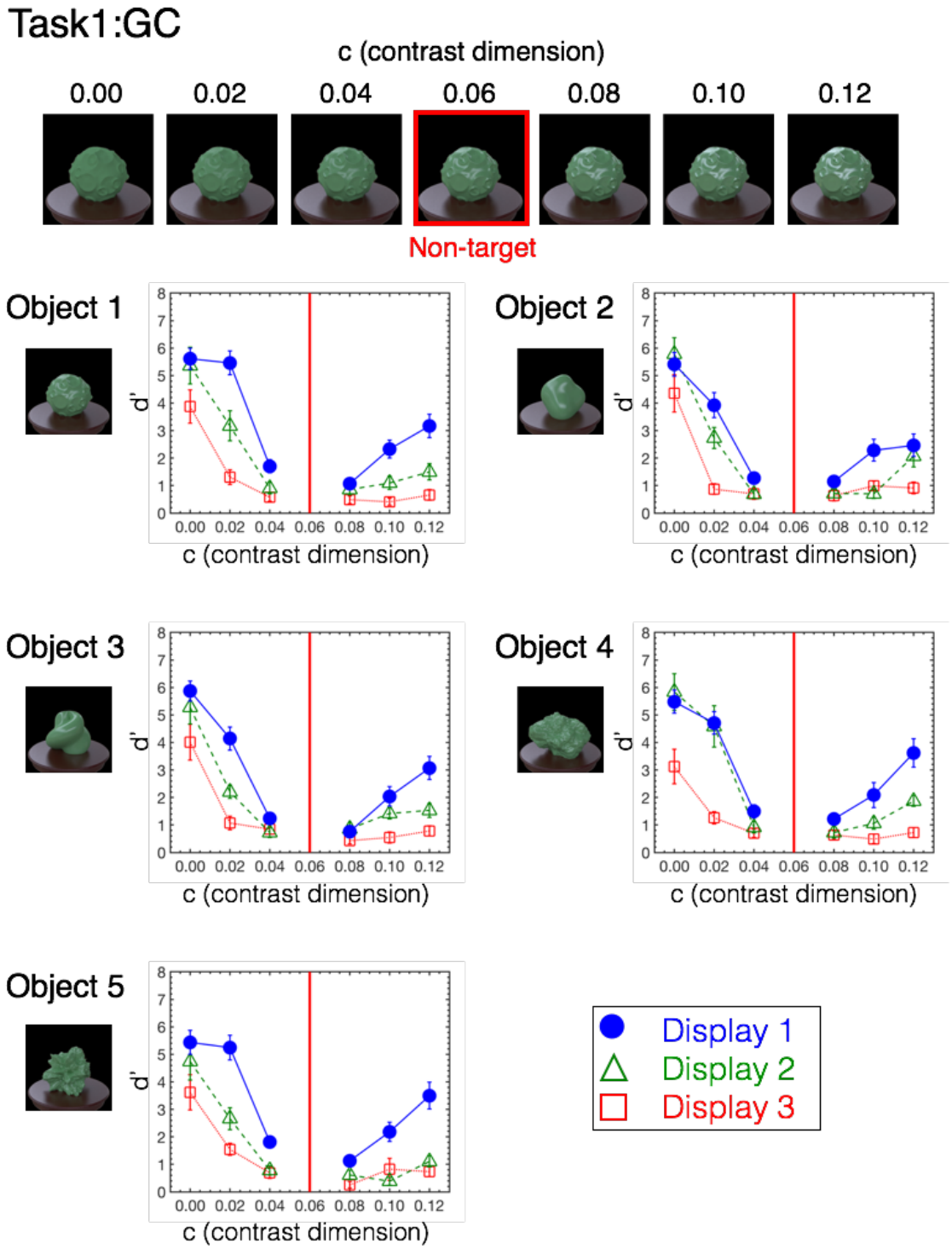

Figure 9. Results of task 1 (GC) in the laboratory experiment. Different panels show different objects. Different symbols in each panel depict different display conditions. The vertical red line in each panel indicates the parameter of the non-target stimulus.

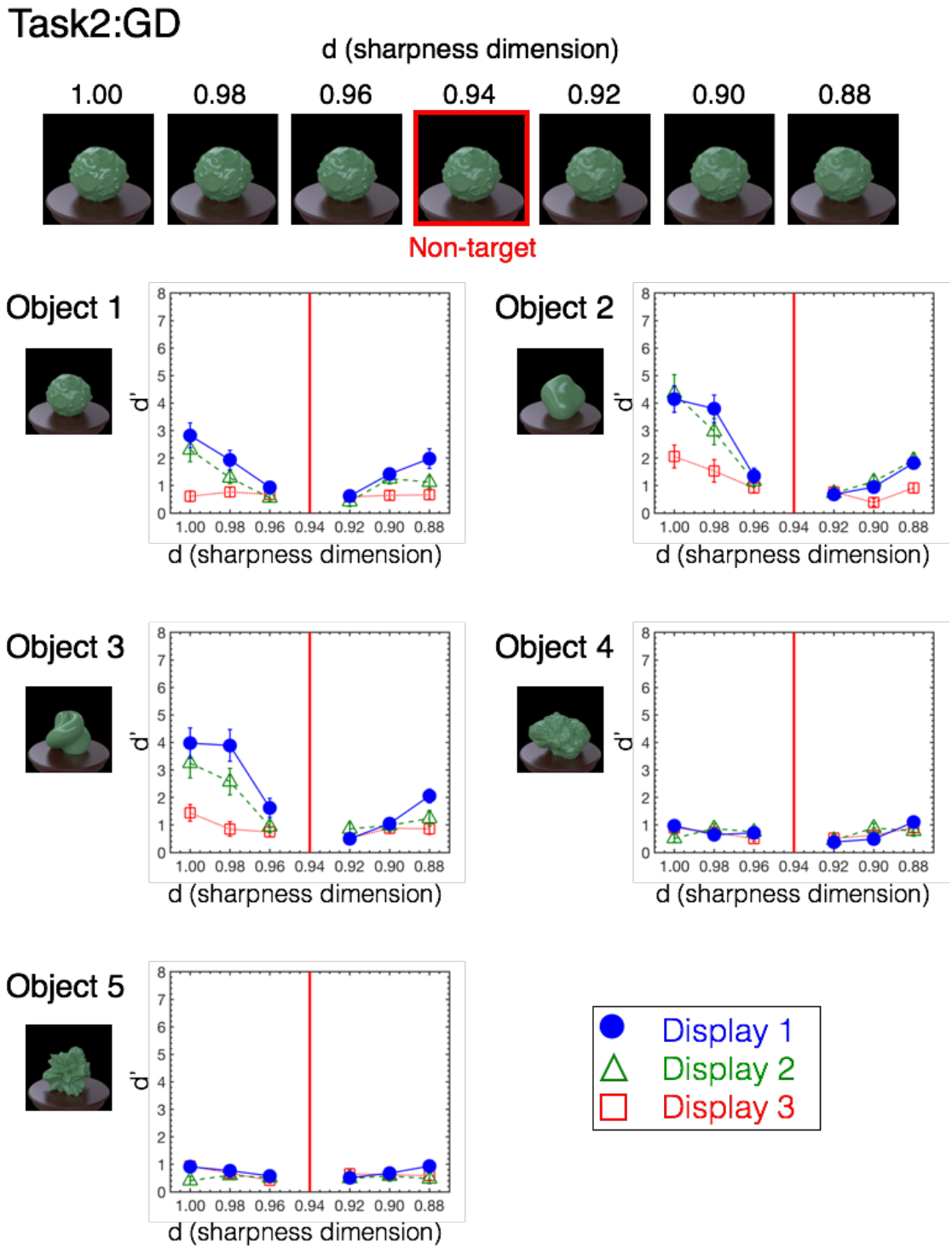

Figure 10. Results of task 2 (GD) in the laboratory experiment.

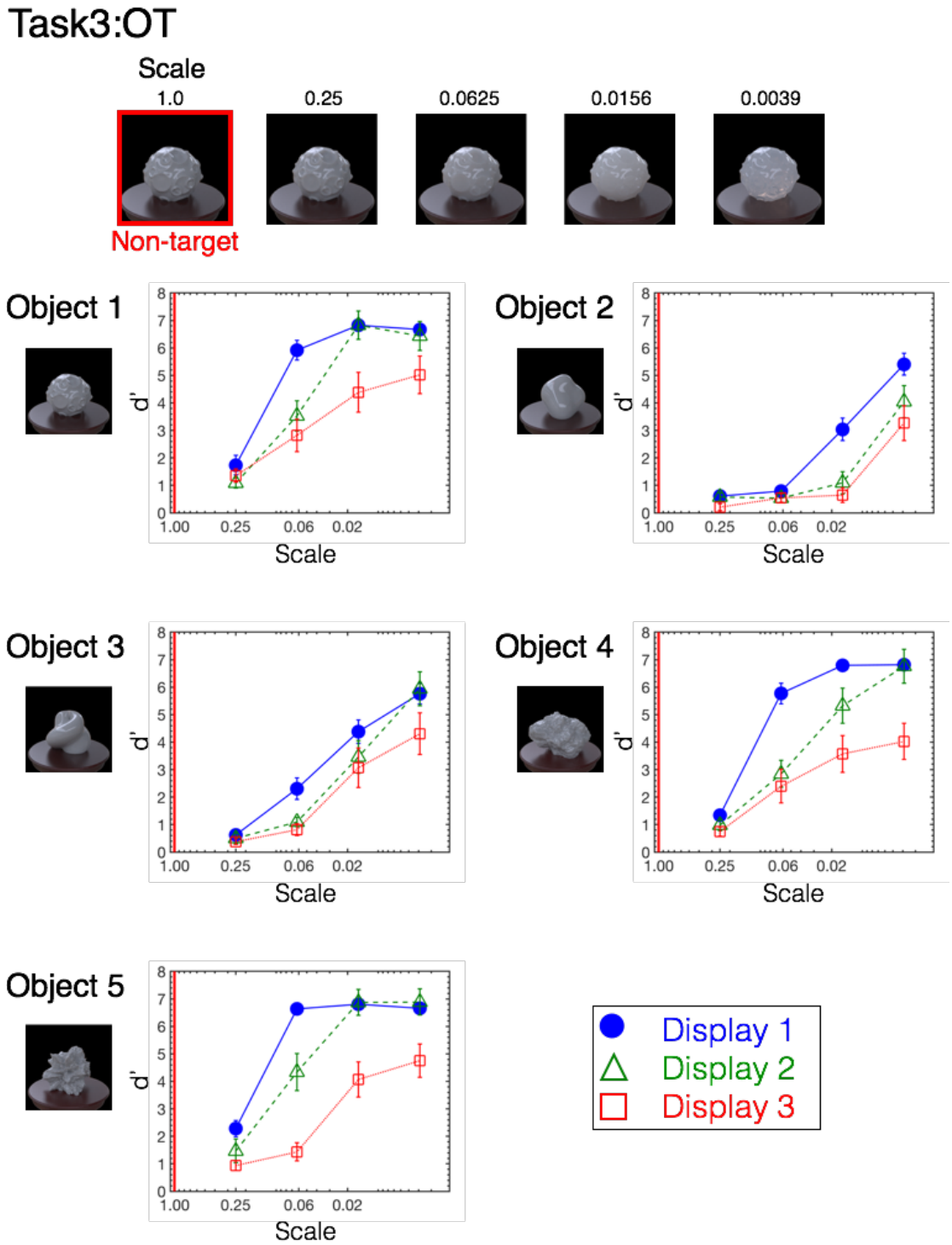

Figure 11. Results of task 3 (OT) in the laboratory experiment.

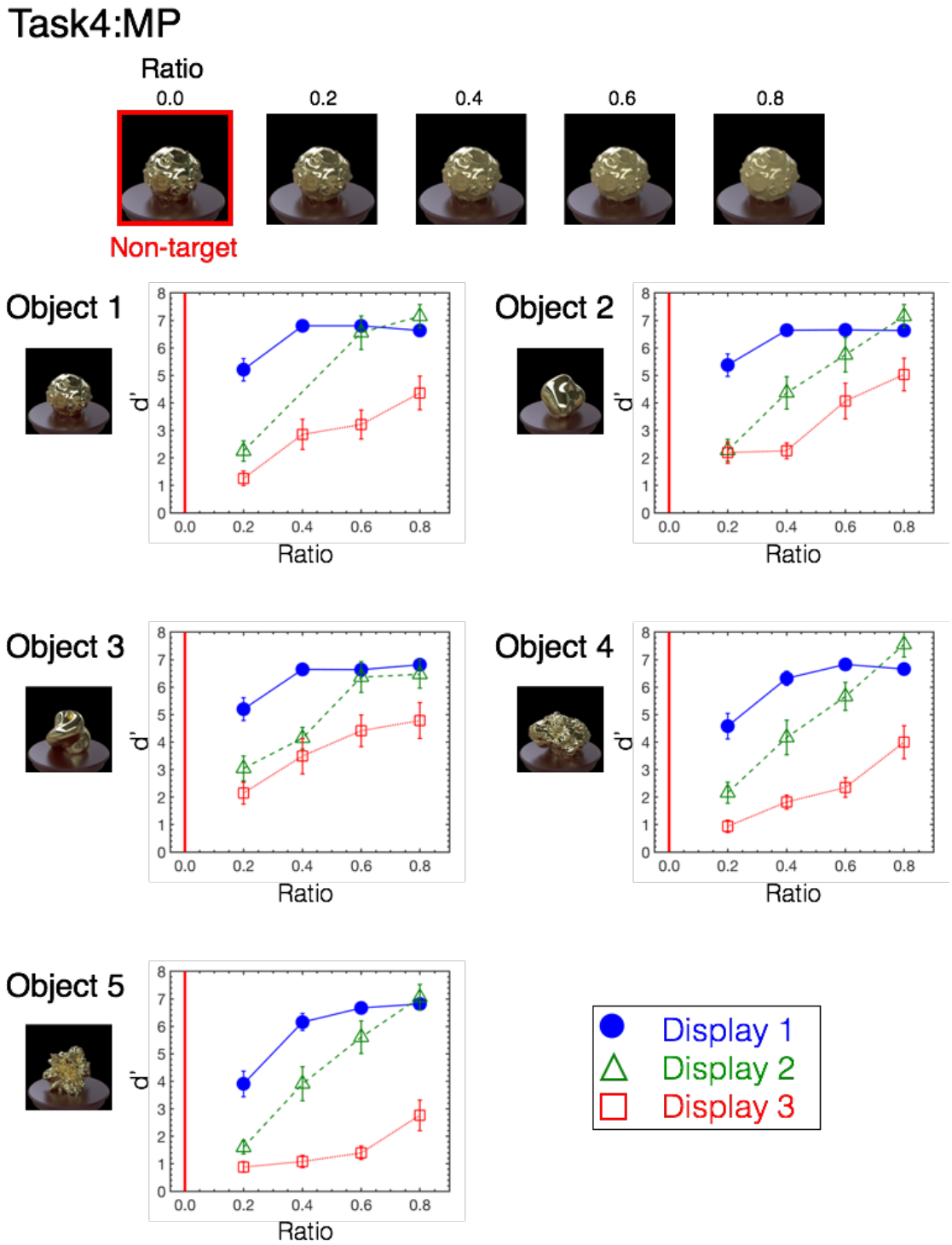

Figure 12. Results of task 4 (MP) in the laboratory experiment. One of the observer data on Object 1 and Display 2 is missing due to a mistake of stimulus presentation.

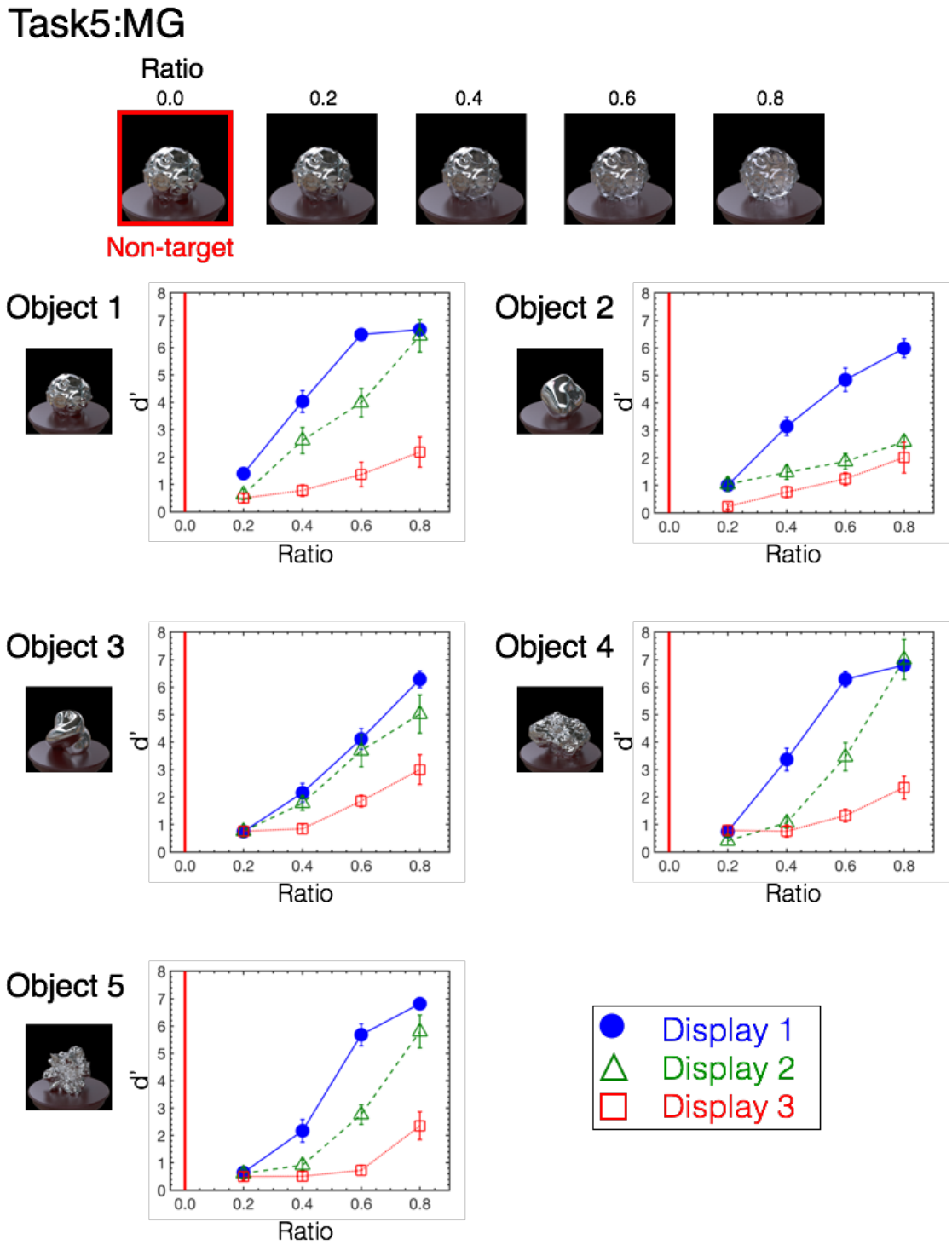

Figure 13. Results of task 5 (MG) in the laboratory experiment.

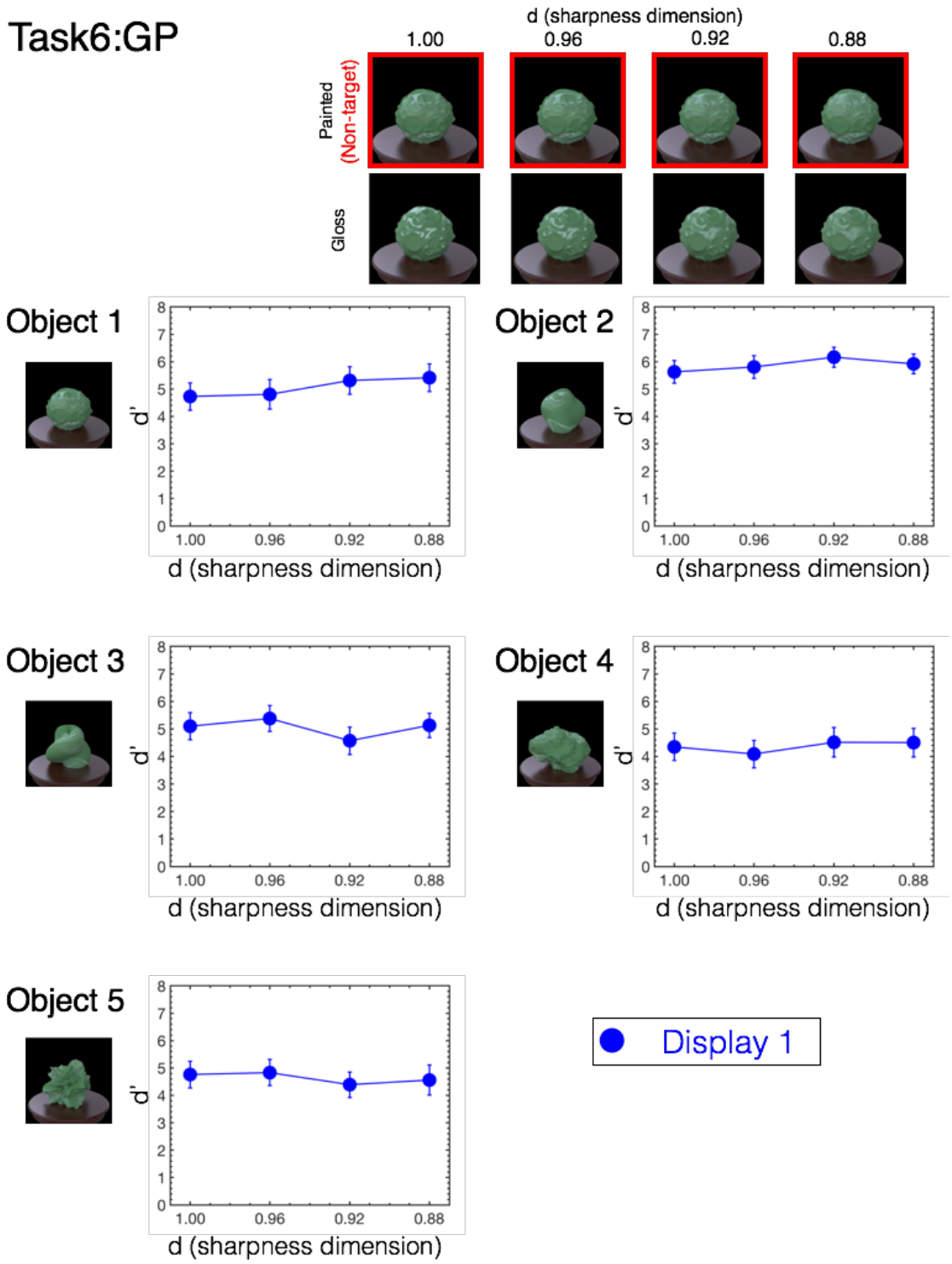

Figure 14. Results of task 6 (GP) in the laboratory experiment.

We also found illumination dependence on material recognition as in the previous studies (Fleming et al., 2003; Motoyoshi & Matoba 2012; Zhang et al., 2019). We used the three display conditions where the illumination environment used in a task was identical (Display 1), similar to each other (Display 2), or largely different each other (Display 3). The present results showed that

task accuracy decreases as the light probes across the images were largely different from each other (Figs. 9-14).

Material judgment over mean color judgment

One may concern that the observers just answered based on the simple color difference to decide an oddity because our procedure did not explicitly ask observers to select one object image based on the material appearance. To analyze this, Figure 15 shows the scatter plot between the sensitivity d' and the mean color difference of each task. The color difference was defined in the L*a*b* space by ΔE^*ab as follows.

$$\Delta E^*_{ab} = \left[\left(L^*_2 - L^*_1\right)^2 + \left(a^*_2 - a^*_1\right)^2 + \left(b^*_2 - b^*_1\right)^2 \right]^{1/2},$$

where L_1^*, a_1^*, and b_1^* are the CIE $L^*a^*b^*$ values averaged across the spatial mean values of all possible non-target stimuli for each task and L_2^*, a_2^*, and b_2^* are the CIE L*a*b* values averaged across the spatial mean values of all possible target stimuli for each task. Each plot in the Figure 15 indicates each task. As expected, when the color difference was large, the d' was generally high. In contrast, even when the color difference was small, the d' for some tasks was relatively high. These results suggest that the observers did not rely on simple color difference to select an oddity for each task and are consistent with that their judgments were based on the material difference for each task.

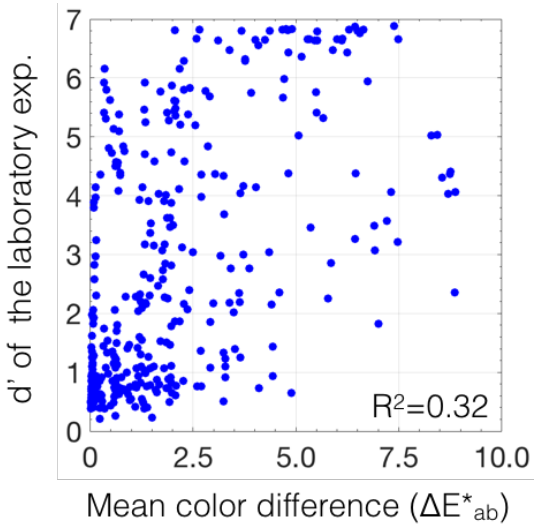

Figure 15. Comparison between the mean color difference of each task and the observers' sensitivity. The mean difference of the target and non-target stimuli was calculated in L*a*b* color space.

Individual difference

Next, we evaluated the individual difference of each task in a Japanese adult population. Figure 16 shows the histogram of the response accuracy for each observer in the crowdsourcing experiment. The number of observers of the display conditions 1, 2, and 3 was 416, 411, and 405, respectively. For each condition, the probability of the correct responses was calculated by averaging the responses in each observer across objects, task difficulties. The standard deviations of the task 1 to 6 on the illumination 1 are .14, .11, .12, .12, and .23, respectively, indicating a large individual difference on task 6 (GP). The standard deviation on the display conditions 2 and 3 ranged from .09 to .18.

The results show that most conditions have a unimodal distribution, while task 6 (GP) has a nearly uniform distribution. The finding suggests that the ability to discriminate the spatial consistency of specular highlights has a large individual difference, although the individual difference cannot be observed in other glossiness judgments such as task 1 (GC) and 2 (GD).

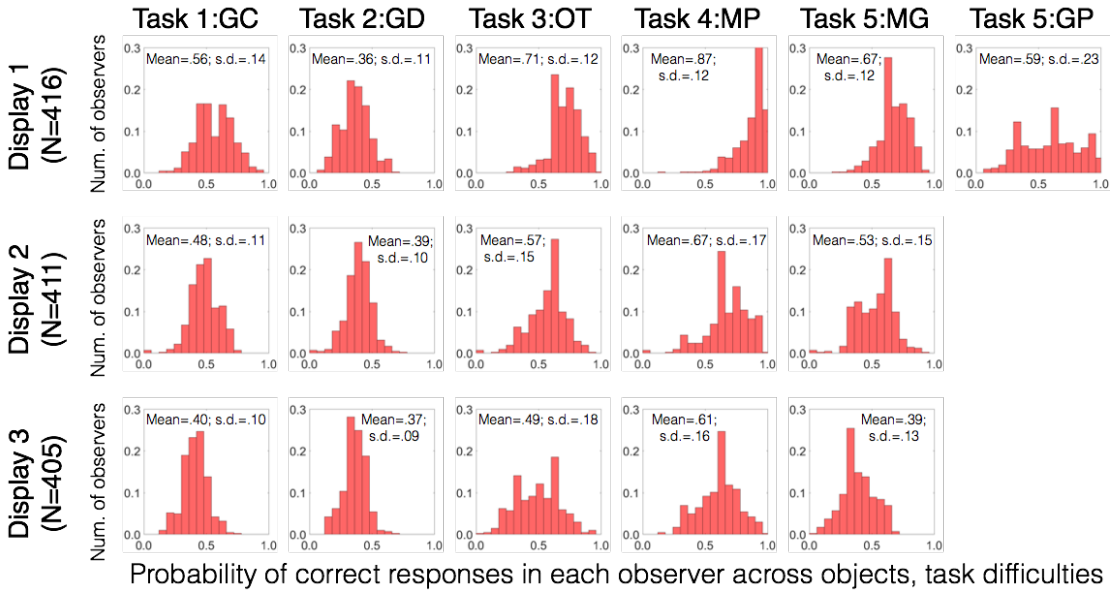

Figure 16. Histogram of the response accuracy for each observer in the crowdsourcing experiment. Different panels indicate different material tasks and display conditions shown as in the legend. For each condition, the probability of the correct responses was calculated by averaging the responses in each observer across objects, task difficulties. The mean and standard deviation of each distribution are shown in each panel.

Discussion

The present study aimed to construct a database of material images annotated by human discrimination tasks. We created material images in six different material dimensions, which are based on the previous works of material recognition. Our database includes various objects and illuminations so that the users can comprehensively investigate the effects of these physical causes on material recognition. Results of psychophysical experiments showed that the material manipulation appropriately controlled the task difficulty and the observers' performance cannot be explained by the simple color differences. One crucial point of our database is that we adopted a non-verbal procedure to collect the observers' data and evaluated the dependence of experimental conditions such as the laboratory and the crowdsourcing. Recently, a lot of studies have focused on the developmental and environmental factors on material recognition. Our database is not only for researches using healthy adult humans but also for researches using various humans and animals including babies or brain-injured participants. The present results showed that the performance in the crowdsourcing was highly correlated with that in the laboratory. This suggests that the database has a tolerance to conduct a new experiment for these observers in different experimental conditions. We also evaluated the possible effects of different experimental conditions on material recognition. For instance, we show that the absolute performance of specular sharpness discrimination (task 2) can change with the experiment environment.

Since we comprehensively investigated the material recognition using a structured dataset, our database itself showed some novel findings of material recognition. One is the geometry dependence on translucency perception (Fig. 11). Specifically, the sensitivity to translucent discrimination was high when the object has rugged surfaces (e.g., Object 1, 4, & 5). The researchers of translucency perception might empirically know the dependence because most of the researches have been conducted using natural complex objects (Motoyoshi, 2010; Xiao et al., 2012). Some studies show that the physical prominent features of translucent objects appear around sharp corners on the surface (Fleming et al., 2005; Gkioulekas et al., 2013). One possibility is that the rugged bumps for the present object produced a lot of local edges on the object image, whose features are similar to those shown in the previous studies, and the visual system strongly relies on the features for the translucent perception.

Another finding is that the ability to discriminate the spatial consistency of specular highlights on glossiness perception has a large individual difference, although other glossiness discrimination tasks do not have. Some studies suggest that the image statistics on an object image are diagnostic for glossiness perception (Adelson, 2001; Motoyoshi et al., 2008). However, when specular highlights of an object image are inconsistent in position and/or orientation with 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 was considered as one of the counterexamples for the image statistics explanation. For the discussion, the present findings suggest a possibility that the discrimination of the spatial consistency of specular highlights is mediated by a different mechanism from the glossiness contrast/sharpness discrimination. Similar to this notion, Sawayama and Nishida (2018)

showed that the highlight inconsistency can be discriminated by different image gradient features from those for material computation. This suggests that the human glossiness computation is mediated by multiple processings, i.e., one is to discriminate different materials on a surface for extracting a region-of-interest (ROI) of glossiness computation, and another is to compute the degree of glossiness in the ROI, as shown in Motoyoshi et al. (2007).

Although our database includes diverse material dimensions, they are still not enough to cover whole natural materials. One example is cloth (Xiao et a., 2016; Bi & Xiao, 2016; Bi et al., 2018; 2019). Perceiving the cloth material is ubiquitous in our daily life. A reason we did not include the material is that it has been shown that the cloth perception strongly relies on the dynamic information (Bi et al., 2018; 2019). Due to the limitation of the experimental time, our database currently has focused on static images. This is why other materials related to dynamic information (reviewed by Nishida et al., 2018) such as the perception of liquidness (Kawabe et al., 2015), viscosity (Kawabe et al., 2015, van Assen & Fleming, 2018), stiffness (Paulun et al., 2017), were also not used in the current investigation. Besides, the wetness perception (Sawayama, Adelson, & Nishida, 2017) and the fineness perception of surface microstructures (Sawayama, Nishida, & Shinya, 2017) are also not used due to the difficulty of continuously controlling physical material parameters by using the identical geometries to other tasks. Since we only used five geometries, material perception derived from object mechanical properties is also not investigated (Schmidt et al., 2017). However, a crucial point is that we share our database as an open-source. This should decrease obstacles to construct a new database and contribute to the future works about material recognition.

Currently, several research groups including behavioral sciences, computer sciences, and neurosciences have conducted the on-going projects by utilizing our database and some findings are reported in conferences. For instance, Kawasaki et al. (2019) have explored the role of the monkey ITC for material perception by using the electrocorticography (ECoG) recordings. Koumura et al. (2018) have investigated how the mid-level features in deep convolutional neural networks can explain human behavioral data.

Conclusion

We have constructed image and observer database for material recognition. We have collected the observers' data about material discrimination by using a non-verbal procedure for six material dimensions with several task difficulties. Results of psychophysical experiments in the laboratory and crowdsourcing environments showed that the performance in the crowdsourcing was highly correlated with that in the laboratory. Besides, from comprehensive data acquisition, we showed novel findings of translucent perception and glossiness perception. The database can be used as benchmark data for neuroscience and psychophysics about material recognition not only of healthy adult humans but also of various humans and animals such as cross-cultural, cross-species, brain-dysfunction and development studies.

Acknowledgements

This work was supported by Grant-in-Aid for Scientific Research on Innovative Areas from Japan Society of Promotion of Science to SN and YD (JSPS KAKENHI Grant Number JP15H05915 and JP15H05924, respectively).

Commercial relationships: none.

Competing interests

The authors declare no competing financial interests.

References

Adelson, E. H. in *Photonics West 2001-Electronic Imaging.* 1-12 (International Society for Optics and Photonics).

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

Ashikmin, M., Premože, S. & Shirley, P. in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques.* 65-74 (ACM Press/Addison-Wesley Publishing Co.).

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

Bi, W, and 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, NY: ACM.

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,

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

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

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

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. *et al.* Visual motion and the perception of surface material. *Current Biology* **21**, 2010-2016, doi:10.1016/j.cub.2011.10.036 (2011).

Fleming, R. W. (2017). Material perception. *Annual review of vision science*, 3, 365-388.

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

Fleming, R. W., Dror, R. O. & Adelson, E. H. Real-world illumination and the perception of surface reflectance properties. *Journal of vision* **3**, 3-3 (2003).

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

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**, 1-19, doi:10.1145/2516971.2516972.

Gkioulekas, I., Walter, B., Adelson, E.H., Bala, K., Zickler, T. (2015) On the appearance of translucent edges, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2015, pp. 5528–5536.

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**, 928-934.

Jakob, W. Mitsuba physically based renderer. mitsuba-renderer. Org. (2010).

Jensen, H. W., Marschner, S. R., Levoy, M. & Hanrahan, P. (2001) A practical model for subsurface light transport. *In Proceedings of the 28th annual conference on Computer graphics and interactive techniques,* 511-518.

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

Kawabe, T., Maruya, K., Fleming, R. W. & Nishida, S. (2015) Seeing liquids from visual motion. *Vision Reserch* **109**, 125-138, doi:10.1016/j.visres.2014.07.003.

Kentridge, R. W., Thomson, R. & Heywood, C. A. (2012) Glossiness perception can be mediated independently of cortical processing of colour or texture. *Cortex* **48**, 1244-1246, doi:10.1016/j.cortex.2012.01.011.

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**, doi:10.1167/11.9.4.

Kingdom, F. A .A. & Prins, N. (2010) Psychophysics: A Practical Introduction. Academic Press: an imprint of Elsevier, London.

Kingdom, F. A .A. & Prins, N. (2016) Psychophysics: A Practical Introduction, Second Edition. Academic Press: an imprint of Elsevier, London.

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**, doi: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**, 1909-1913, doi:10.1016/j.cub.2012.08.009.

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, 1–15.

Motoyoshi, I. (2010) Highlight-shading relationship as a cue for the perception of translucent and transparent materials. *Journal of Vision* **10**, 6, doi: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. pmid:17443193

Nagai, T. *et al.* Image regions contributing to perceptual translucency: A psychophysical reverse-correlation study. *i-Perception* **4**, 407-428 (2013).

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. & Shinya, M. Use of image-based information in judgments of surface-reflectance properties. *Journal of the Optical Society of America A* **15**, 2951-2965 (1998).

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. Perceptual gloss parameters are encoded by population responses in the monkey inferior temporal cortex. *Journal of Neuroscience* **34**, 11143-11151, doi:10.1523/JNEUROSCI.1451-14.2014 (2014).

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

Olkkonen, M. & Brainard, D. H. Perceived glossiness and lightness under real-world illumination. *Journal of Vision* **10**, 5-5, doi:10.1167/10.9.5 (2010).

Pellacini, F., Ferwerda, J. A., & Greenberg, D. P. (2000, July). Toward a psychophysically-based light reflection model for image synthesis. I*n Proceedings of the 27th annual conference on Computer graphics and interactive techniques* (pp. 55-64). ACM Press/Addison-Wesley Publishing Co.

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. doi: 10.3389/fpsyg.2018.01250

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

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

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

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

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, doi:10.1167/17.3.18.

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

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

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

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

Walter, B., Marschner, S. R., Li, H. & Torrance, K. E. (2007) Microfacet models for refraction through rough surfaces. *In Proceedings of the 18th Eurographics conference on Rendering Techniques,* 195-206.

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

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, doi:10.1167/14.3.17.

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, doi:10.1167/16.3.34.

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

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.