Vision Research 115, 71-82
doi:10.1016/j.visres.2015.07.006

Visual categorization of surface qualities of materials by capuchin monkeys and humans

Chihiro Hiramatsu[*,1], Kazuo Fujita

Department of Psychology, Graduate School of Letters, Kyoto University, Yoshidahonmachi, Sakyo-ku, Kyoto 606-8501, Japan.

[*] Corresponding author at: Department of Human Science, Faculty of Design, Kyushu University, 4-9-1 Shiobaru, Minamiku, Fukuoka 815-8540, Japan.

E-mail addresses: chihiro@design.kyushu-u.ac.jp (C. Hiramatsu), kfujita@bun.kyoto-u.ac.jp (K. Fujita).

[1] Present address: Department of Human Science, Faculty of Design, Kyushu University, 4-9-1 Shiobaru, Minamiku, Fukuoka 815-8540, Japan.

Abstract

Visually identifying and categorizing the material composition of objects before actually interacting with them is an important skill for operating smoothly and safely in the world. This ability is assumed to have been shaped by evolution; therefore, non-human animals should share similar categorization abilities. Little is known, however, about how non-human animals do this. We tested whether tufted capuchin monkeys (*Cebus apella*) were able to visually categorize images that represented nine different materials (metal, ceramic, glass, stone, bark, wood, leather, fabric, and fur), and we compared their performance with that of humans. Capuchins showed excellent categorization abilities for images of fur, which is a familiar material to captive monkeys. Humans showed a tendency to confuse material categories that resembled each other visually and/or semantically. Correlation analyses on reaction time showed that both species made correct choices rapidly in selecting glossy categories like metal and ceramic compared with matte categories like fabric and stone, which contain minute patterns. Overall, our results suggest that monkeys share similar perceptual tendencies with humans in visual categorization of material images to some extent and the potential to categorize materials frequently encountered in their daily lives by visual observation.

Keywords

Material perception; Texture; Comparative perception; Capuchin monkeys

1. Introduction

We routinely classify and identify the material composition of objects visually on the basis of their distinctive surface qualities, which are formed by the reflection of light on the materials. Material categories vary from natural (e.g., wood and stone) to artificial (e.g., metal and glass), and some category names refer to surfaces of animals (e.g., fur and leather). Visual recognition of materials facilitates proper and adaptive action with their objects. Progress has been made in understanding the perception of surface qualities of materials in humans (Anderson, 2011; Maloney & Brainard, 2010; Motoyoshi et al., 2007; Sharan, Rosenholtz, & Adelson, 2014). Sharan and colleagues demonstrated that material categorization is as rapid and accurate as object and scene categorization and one of a basic abilities of the visual system (Sharan, et al., 2014). Another study (Wolfe & Myers, 2010), using visual search based on surface qualities of materials, showed that materials cannot draw attention automatically. An investigation of the semantic aspects of materials showed that humans represent material classes similarly in the visual and semantic domains (Fleming, Wiebel, & Gegenfurtner, 2013).

From a biological point of view, our perception of materials should have been shaped largely through evolutionary processes. Arguably, visual perception of the surface qualities of materials is extremely useful for survival among diurnal animals that use vision as the primary sensory modality. For example, perceiving surface qualities such as glossiness and transparency should be helpful for identifying fresh fruits and water, especially when other cues such as color or odor are unreliable.

The ecological importance and evolutionary foundation of the perception of surface qualities have received recent support from several physiological studies. Neurophysiological studies have found neurons and brain areas responsive to surface qualities, such as glossiness and texture, in monkey brains (Freeman et al., 2013; Nishio, Goda, & Komatsu, 2012; Nishio et al., 2014; Okazawa, Goda, & Komatsu, 2012; Okazawa, Tajima, & Komatsu, 2015). An fMRI study reported that macaque brains represent real-world material categories (e.g., metal, wood, fur) in a way similar to humans (Goda et al., 2014). These studies suggest that primates may share a similar perception of surface qualities of materials. Although experience ought to modify material perception, more fundamental processes are likely to have considerable evolutionary origin. However, few studies have asked how non-human animals perceive materials, and therefore very little information is available to discuss evolutionary backgrounds of such perception.

In the present study, we aimed to investigate how non-human primates visually perceive and categorize materials humans encounter in daily life. We tested this ability in tufted capuchin monkeys (*Cebus apella*), a species of New World monkeys that separated from Old World monkeys about 40 million years ago (Kiesling et al., 2014). Although capuchin monkeys are phylogenetically more distant from humans than are Old World monkeys such as macaques, they show habitual tool-using behavior such as cracking nuts with stones (Ottoni & Izar, 2008) and use visual information effectively to conduct various tasks (Paukner, Huntsberry, & Suomi, 2009; Wright, 1999). They also show remarkable omnivorous tendency; feed on small-sized species of amphibians and reptiles, young birds and birds' eggs, as well as various kinds of fruit and insects (Izawa, 1975, 1978). Therefore, they may benefit from recognizing materials such as stones and textures of foods with cryptic coloration visually. They share many perceptual properties with humans (e.g., preference for regularity, perceptual completion) (Anderson et al., 2005; Fujita & Giersch, 2005), but a difference has also been detected (e.g., perceptual grouping) (Spinozzi, De Lillo, & Castelli, 2004). Because of moderate similarity and differences between two species, they are good candidates to compare visual material perception from an evolutionary perspective. In this study, we observed how similarly (or differently) monkeys and humans behave in visual matching task based on material properties and discussed what kind of factors, e.g., visual features, saliency and experience, influence their performance. The comparison between the two species would shed light on the evolutionary processes of material perception in primates.

2. Experiment 1

2.1. Methods

2.1.1. Animal subjects

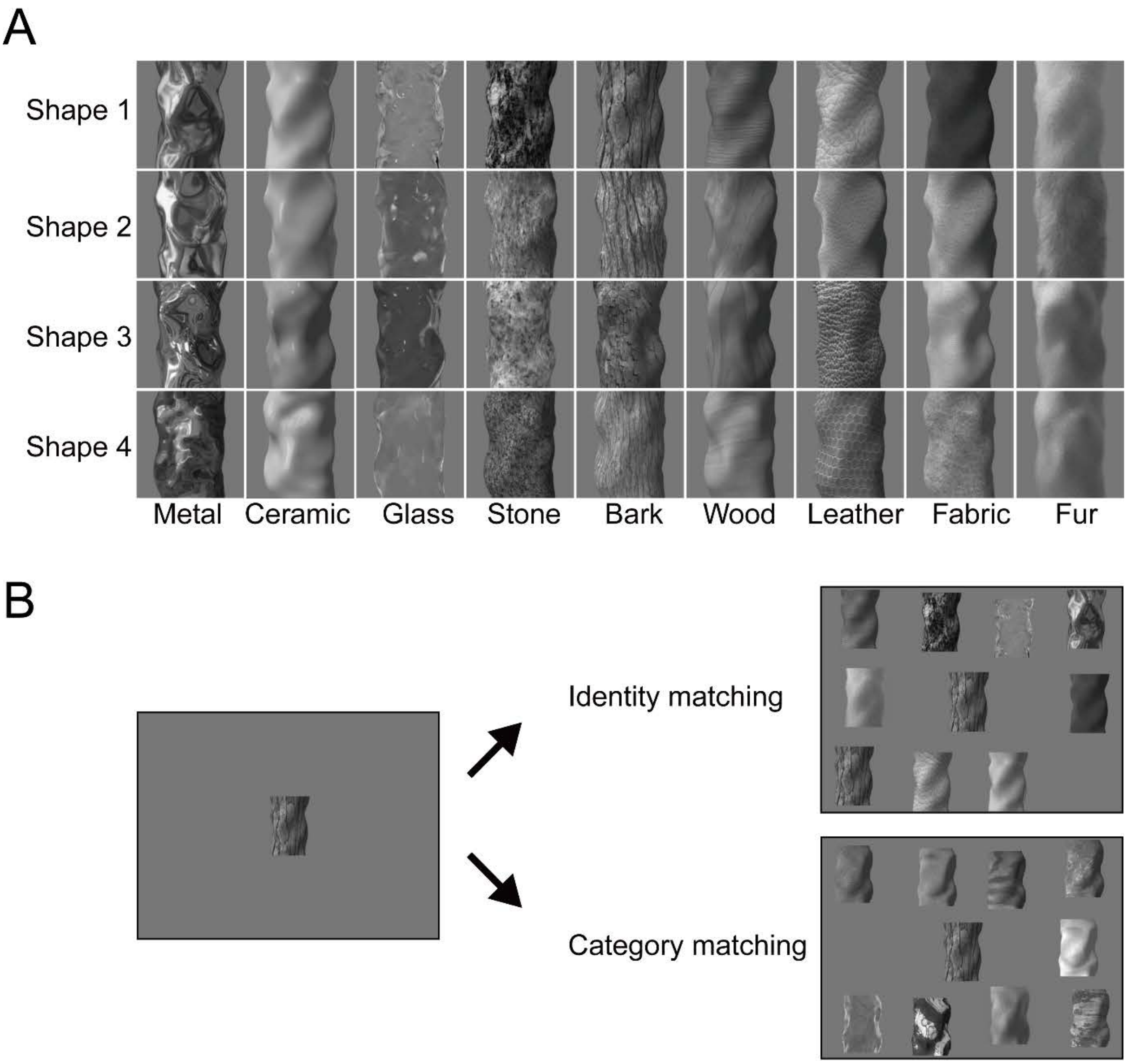
The animal care and experiment were conducted according to the principles of the ARRIVE (Animal

Research: Reporting of In Vivo Experiments) guidelines (Kilkenny et al., 2010). Seven tufted capuchin monkeys participated in the experiment. Among them, two 2-year-old females conducted experimental design 1 (see experimental design). They had been trained to match to sample using simple shapes (circle and cross) for 1 year but had never experienced experiments on visual perception before the training. Other five adult monkeys (8–18 years old, three females) conducted experimental design 2. They had experienced various types of visual and cognitive experiments (Fujita, 2009; Fujita & Giersch, 2005) with touch-sensitive monitors and were highly skilled at matching-to-sample tasks. The monkeys were not food deprived but received a portion of their daily diet during testing and the remainder in their home cage after testing each day. In the home cage, monkeys had free access to water. No animals were sacrificed in this study. The experiment was approved by the Animal Experiments Committee of the Graduate School of Letters, Kyoto University (permit number 11-04) in accordance with the European Directive 2010/63 on the Protection of Animals in Scientific Experimentation.

2.1.2. Stimuli and apparatus

We used material images created by the computer graphics software LightWave 3D (NewTek, San Antonio, TX, USA). The images were of nine material categories (metal, ceramic, glass, stone, bark, wood, leather, fabric, and fur; Fig. 1A). Each category had eight exemplars with different surfaces and slightly different meaningless shapes (shapes one to eight). In total, there were 72 gray-scale material images (Fig. S1). A color version of these images was used in the previous fMRI study with human subjects (Hiramatsu, Goda, & Komatsu, 2011). The psychological analysis in the previous study showed that exemplars of metal, ceramic and glass share glossy appearance and those of other categories share matte appearance (Hiramatsu, et al., 2011). Because capuchin monkeys are known to have highly polymorphic color vision (Jacobs, 2007), we used gray-scale images in our experiments to eliminate the effect of color-vision differences. All images in the current study were resized to 180 × 180 pixels (ca. 9.5 × 9.5 degrees at a 15-cm viewing distance). The images were presented on a touch-sensitive LCD monitor (TSD-CT157-MN; Mitsubishi, Japan) (1024 × 768 pixels). The image presentation, response detection, and food delivery were controlled by a custom program written with Visual Basic 2008 programing software (Microsoft, Redmond, WA, USA) installed on a built-to-order computer (CPU: Core 2 Duo 2.93 GHz; Intel, Santa Clara, CA, USA). The monitor was calibrated with the i1 Display Pro calibration tool (X-rite, Grand Rapids, MI, USA). The background of the material images was uniformly gray ($x = 0.311$ and $y = 0.330$, 30 cd/m^2). The monitor was placed at the front of a transparent operant box (45 × 45 × 45 cm) where the monkeys performed the tasks. The experiment was conducted in a dark room with low illumination by an incandescent bulb (7 lux at monitor location) attached to the operant box. White noise was presented during the experiment so that monkeys were not disturbed by noise from outside the operant box.

Fig. 1.



C

Design 1

Phase	Trial type	Image set
1	Identity-training	Shapes 1-2
2	Identity-training	Shapes 3-4
3	Identity-baseline Identity-test Category-test	Shapes 5-8
4	Identity-training	Shapes 5-8
5	Identity-baseline Category-test	Shapes 1-8

Design 2

Phase	Trial type	Comparison image set
1	Identity-training	Typical exemplars
2	Identity-baseline Category-test	Typical exemplars

Stimuli and task design. (A) Examples of the stimulus images used in this study. Each stimulus set (shape one to shape four) consisted of one exemplar from nine categories. See Fig. S1 for the complete stimulus set. (B) Schematic drawings of the nine-choice matching-to-sample task. After touching a sample image, the task split into one of two cases: Identity matching is the case where a sample image is surrounded by comparisons from the stimulus set that contains the same image as the sample. Category matching is the case where a sample image is surrounded by comparisons from a different stimulus set that does not contain the same image as the sample. (C) Summary of the experimental designs 1 and 2. Note that image sets used in phases 1 and 2 of design 1 were counterbalanced between monkeys. The sample and comparison images were chosen from image sets indicated in each phase. In design 2, only nine typical exemplars from each category (see Figs. 7 and S1) were used as comparison images.

2.1.3. Experimental design

The experiments asked whether monkeys would generalize the identity-matching performance learned in training phases to similar images in test phases. There were two types of test trials: identity matching and category matching. Performance of identity matching would indicate the ability to directly apply the strategy learned in training phases. In contrast, performance of category matching would imply the ability to generalize the learned concept to slightly different images that share similar surface qualities, i.e. material category.

Briefly, trials started with the monkey’s pressing a lever down for 1 sec, which resulted in appearance of a sample image. After the monkey touched the sample image three times, nine comparison images, one exemplar from each category, appeared (Fig. 1B). Trials ended when the monkey touched one of the comparison images. More details are described in supplementary methods.

We used two experimental designs (designs 1 and 2). Fig. 1C summarizes each experimental design. The design 1 consisted of five phases. The first two phases were training, in which the monkeys had to choose the same image as the sample from nine comparison images (identity matching) that consisted of a stimulus set to which the sample image belong. Two stimulus sets were used for the first and second training phases, respectively (shapes one to four). The third phase consisted of baseline and test trials. The baseline trials were identity matching with learned images in the first and second phases. In the test trials, new stimulus images from four stimulus sets, shapes five to eight, were used. There were two types of test trials: identity matching and category matching. In identity matching, comparison images comprised of a stimulus set that contained one image identical to the sample. In category matching, comparison images were chosen from a different stimulus set that consisted of nine images belonging to different material categories but sharing the same shape. The correct answer in a category matching trial was to choose a comparison with the same category to the sample but with a different material texture together with a different shape (Fig. 1B). Since no identical image appeared in category matching trials, the monkeys were expected to choose the most similar image from comparisons. In the fourth phase, we trained the monkeys on identity matching using the same stimulus sets, shapes five to eight, used in the third phase. In the fifth phase, we re-tested category matching using all stimulus sets.

Experimental design 2 consisted of two phases. The first phase was training and we used nine typical exemplars selected by image analysis (see supplementary methods) from each category as a stimulus set. In the second phase, we introduced new images to test whether monkeys could demonstrate understanding of the concept of material categories. In the test trials, a new image appeared as the sample, and monkeys were required to select one image from the nine typical exemplars (comparisons) for which identity matching was trained in the first phase. The main difference of design 2 from design 1 was that the comparison images were always the same nine images those were typical exemplars of each category.

In training phases of both designs, monkeys were rewarded only if they selected the correct identical image and moved to the next phase after their performance exceeded 80% correct for two consecutive sessions. In test phases, about 25% of test trials were randomly inserted in baseline trials. In test trials, monkeys were always rewarded irrespective of their choice. This non-differential reinforcement procedure was employed to test genuinely how monkeys would generalize the learned concept to choose the most similar image to a sample from comparisons to the new stimuli by minimizing the opportunity for explicit direction on matching. In baseline trials, however, monkeys were rewarded only if they selected the correct identical image, to maintain the motivation of monkeys to choose correct stimuli. Each session consisted of 72 to 135 trials depending on phases. In total, each monkey conducted 360 and 189 test trials in designs 1 and 2, respectively. In test trials of design 1, each image appeared seven times as a sample and 63 times as one of comparison images. In contrast, in test trials of design 2, each image used as a sample appeared three times and never appeared as a comparison image. Instead, nine typical exemplars from each category always appeared as comparison images.

2.1.4. Analysis

All analyses described in this manuscript were conducted using MATLAB R2010b (MathWorks, Natick, MA, USA). For experimental design 1, we observed learning curve of each individual because two participant monkeys with little experience of visual tasks were suitable to see how they learn visual matching based on material properties. For experimental design 2, in which five monkeys participated, we examined whether correct choices and confusion errors between categories were statistically above chance (100/9 ≈ 11.1 %) or threshold (chance + (100-chance)/2 ≈ 55.6%) levels via a one-tailed t-test with a significance level of 0.05.

For both designs, we created confusion matrices based on monkeys’ choices in the trials. Each row of the matrix indicates the category presented as the sample and each column the category chosen from comparisons. The color of each cell reflects the percentage of choices for each combination between the sample and the chosen category averaged across all exemplars used as the sample. Diagonal blocks from upper left to lower right represent responses that human experimenters consider correct categorization.

2.2. Results

2.2.1. Result of experimental design 1

The number of sessions necessary to reach the performance criterion (> 80% correct for two consecutive sessions) was drastically decreased from the first phase to the second phase in both monkeys (Kojilo: 41 to 6; Zephie: 25 to 7). The learning curves in the first and second phases for each category are shown in Fig. 2A as the moving average across 50 trials. Performance improved faster for metal and fur than for the other materials in both monkeys, although Kojilo performed well on the ceramic quality from early in session 1 (Fig. 2A). By contrast, the performance for both monkeys on fabric and wood remained around the threshold level. These tendencies are reflected in the confusion matrices (Fig. 2B), which show trends of monkeys' choices when each material category was the sample during the last 10 sessions in the first phase. Both monkeys performed well for metal and fur but showed many confusion errors between wood and fabric.

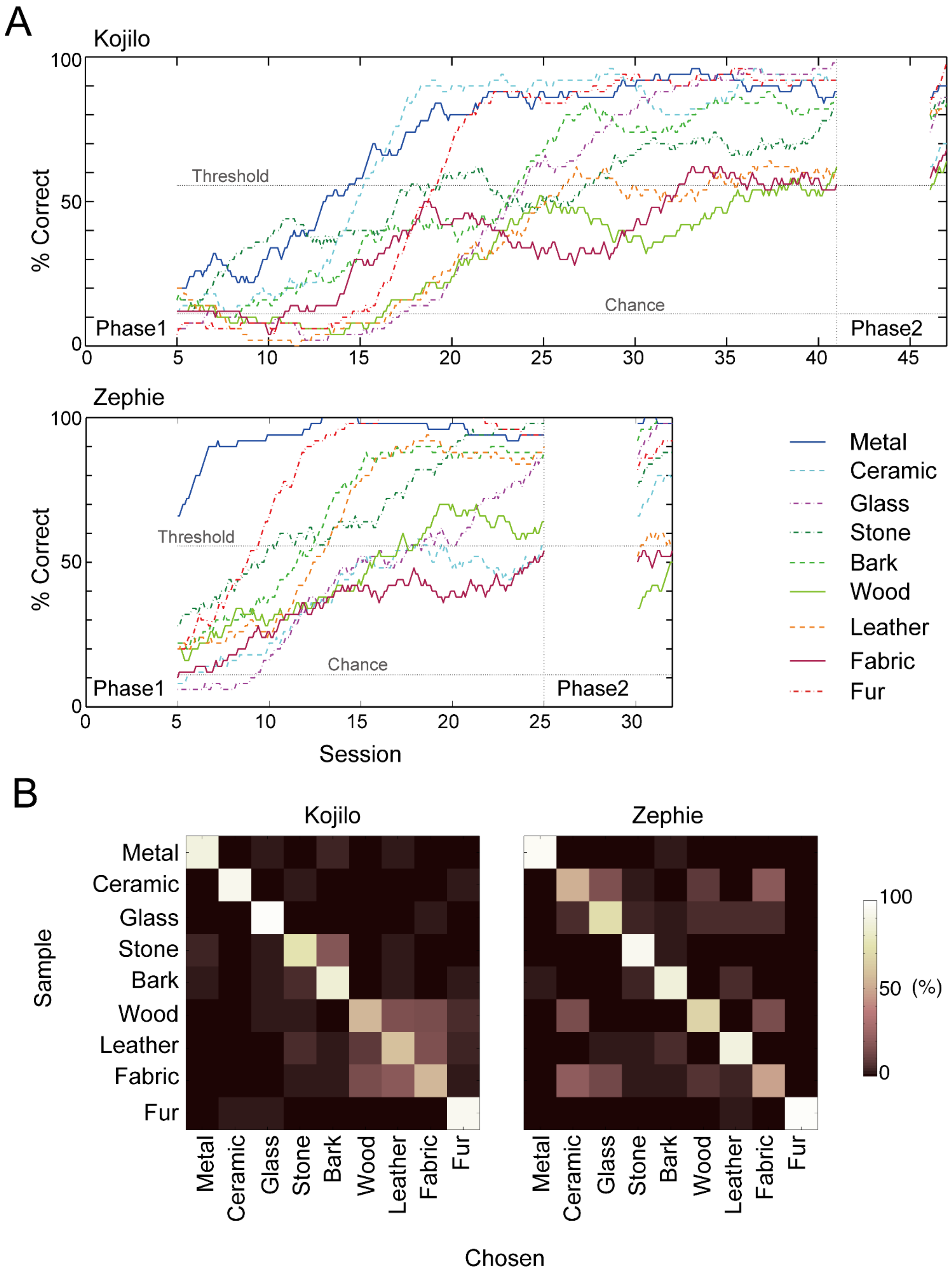

A
Kojilo
% Correct
100
50
0
Threshold
Chance
Phase1
Phase2
0
5
10
15
20
25
30
35
40
45
Zephie
% Correct
100
50
0
Threshold
Chance
Phase1
Phase2
0
5
10
15
20
25
30
Session
Metal
Ceramic
Glass
Stone
Bark
Wood
Leather
Fabric
Fur
B
Kojilo
Zephie
Sample
Metal
Ceramic
Glass
Stone
Bark
Wood
Leather
Fabric
Fur
Chosen
100
50 (%)
0

Fig. 2. Transition of performance by category and confusion matrices in the first phase of experimental design1. (A) Moving average (across 50 trials) of performance by each category from the first to second phase for Kojilo (top panel) and Zephie (bottom panel). Dotted horizontal lines indicate the chance level (11.1%) and threshold level (55.6%). (B) Confusion matrices of each monkey indicating how frequently monkeys matched sample images to each category in comparisons. Frequency of choice was averaged across the last 10 sessions and shown in percentages. The diagonal blocks from top left to bottom right represent correct performance.

In the third phase, we examined the monkeys' ability to transfer the concept of matching based on visual material properties into new stimulus sets. The confusion matrices indicate that both monkeys made many confusion errors with the new stimulus sets in both identity matching (Fig. 3A, middle) and category matching (Fig. 3A, right), but high performance was maintained for learned stimulus sets (baseline trials, Fig. 3A, left). It is worth noting, however, that performance for metal in both identity and category matching was above the threshold level (75% and 66.7%) for Kojilo and above the chance level (50% and 50%) for Zephie (Fig. 3A). It is also notable that both monkeys performed at 41.7% for fur in category matching. Zephie's matching accuracy was above the chance level for all categories in identity matching. Her performance was also higher than chance for all categories except glass and wood in category matching. Kojilo performed above the chance level only for metal, glass and wood in identity matching and for metal, ceramic, bark, leather and fur in category matching despite her high performance for metal. She also committed many confusion errors with metal and other categories.

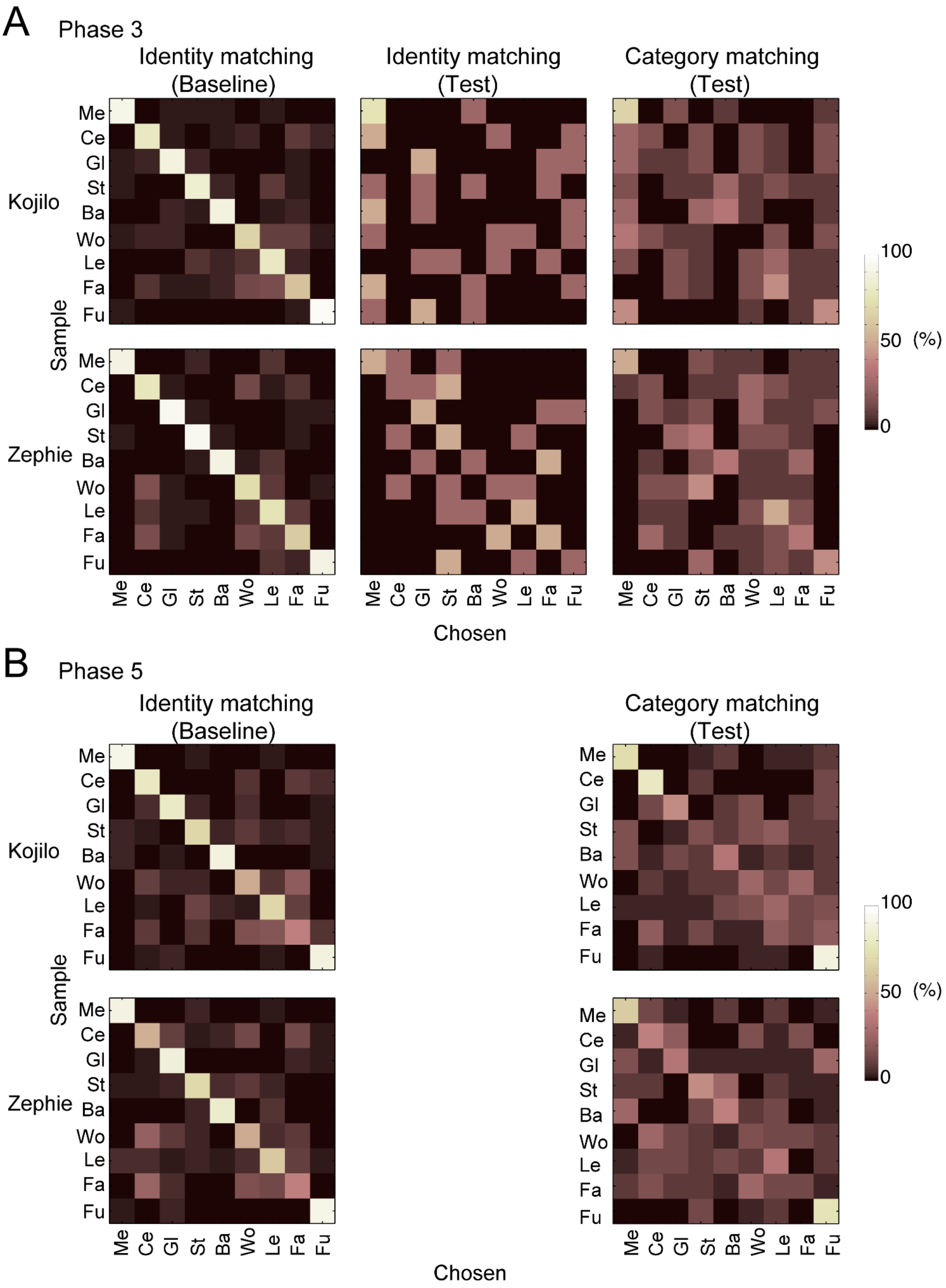

A
Phase 3
Identity matching (Baseline)
Identity matching (Test)
Category matching (Test)
Kojilo
Zephie
Sample
Me
Ce
Gl
St
Ba
Wo
Le
Fa
Fu
Chosen
100
50 (%)
0
B
Phase 5
Identity matching (Baseline)
Category matching (Test)

Fig. 3. Confusion matrices of each participant in the third and fifth phase of experimental design 1. Confusion matrices: (A, left) for identity-matching baseline trials in the third phase. (A, middle) those for identity-matching test trials in the third phase. (A, right) for category-matching test trials in the third phase. (B, left) for identity-matching baseline trials in the fifth phase. (B, right) for category-matching test trials in the fifth phase. Note that there were no identity-matching test trials in the fifth phase. Abbreviations; Me: metal, Ce: ceramic, Gl: glass, St: stone, Ba: bark, Wo: wood, Le: leather, Fa: fabric, Fu: fur.

In the fifth phase, category matching was tested again for learned stimulus sets. The left panel of Fig. 3B shows the confusion matrix for baseline trials (learned identity matching) and the right panel shows the matrix for category-matching trials in the fifth phase. In category matching, both monkeys showed performance above the threshold level for metal (Kojilo: 70.8%, Zephie: 62.5%) and fur (Kojilo: 87.5%, Zephie: 75%). Kojilo also showed high performance for ceramic (79.2%). The performance in all other categories exceeded the chance level (Fig. 3B, right). However, both monkeys committed considerable numbers of confusion errors (more than the chance level) between fabric and wood (Kojilo: 16.7%, Zephie: 16.7%) and between fabric and leather (Kojilo: 18.1%, Zephie 13.9%), even in the baseline condition (Fig. 3B, left).

2.2.2. Result of experimental design 2

The monkeys required 17–35 sessions (mean = 29, SD = 9) to reach the criterion in the first phase. Fig. 4 shows confusion matrices across the nine sessions in the second phase, averaged across the five monkeys. Confusion matrices for each individual are provided as Fig. S2. Table 1 shows the statistics of the second phase. The mean performance of identity matching in baseline trials was significantly above chance for all categories and above the threshold level for metal, ceramic, glass, bark and fur, but not for stone, wood, leather or fabric (Table 1, Fig. 4, left). There were confusion errors significantly above the chance level between stone and bark when stone was the sample (t = 2.26, p = 0.0435, mean = 26.7%) and between leather and wood when leather was the sample (t = 2.40, p = 0.0373, mean = 21.7%). In category matching, performance was significantly above the chance level for metal, ceramic, stone, bark and fur, but no category showed significantly higher performance than the threshold level (Table 1, Fig. 4, right). There was no category confused with any other categories at a level significantly above chance. Apparent confusion errors between metal and other categories in category matching were in fact because of one monkey's (Zen's) tendency to select metal irrespective of sample category (see Fig. S2).

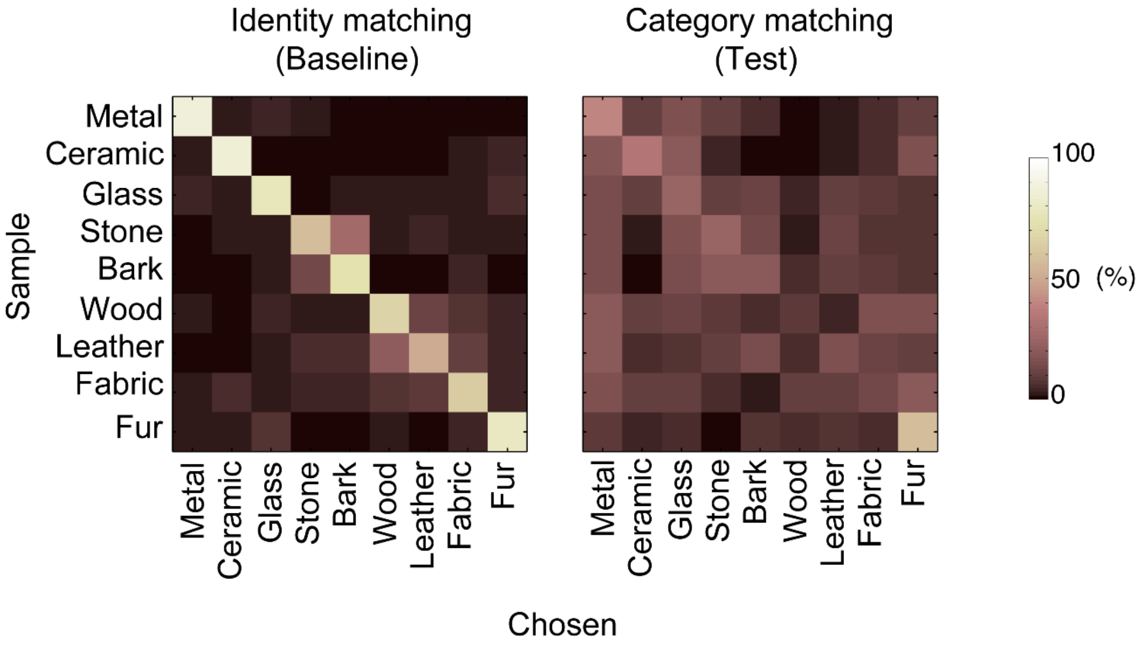


Fig. 4. Averaged confusion matrices in the second phase of experimental design 2. Left: Confusion matrix for identity-matching baseline trials. Right: Confusion matrix for category-matching test trials in the second phase. Both matrices are presented as the mean value of the five monkeys' performance. There were no identity-matching test trials in design 2.

Table 1. Statistics on performance in identity- and category-matching trials in the second phase of experimental design 2 in experiment 1.

				Comparison to chance level (11.1%)		Comparison to threshold level (55.6%)	
	Category	Mean (%)	SD (%)	t-value	*p*-value	t-value	*p*-value
Identity matching	Metal	86.9	13.9	12.16	**0.0001**	5.03	**0.0037**
	Ceramic	85.6	8.1	20.46	**<0.0001**	8.25	**0.0006**
	Glass	78.1	20.2	7.41	**0.0009**	2.49	**0.0338**
	Stone	56.7	18.3	5.56	**0.0026**	0.14	0.4493
	Bark	74.7	14.4	9.89	**0.0003**	2.98	**0.0204**
	Wood	66.1	16.1	7.63	**0.0008**	1.46	0.1086
	Leather	50.8	8.8	10.07	**0.0003**	-1.2	0.8514
	Fabric	63.3	18.3	6.39	**0.0015**	0.95	0.1977
	Fur	79.2	13.6	11.15	**0.0002**	3.87	**0.009**
Category matching	Metal	40	30.2	2.14	**0.0497**	-1.15	0.843
	Ceramic	33.3	11.7	4.26	**0.0065**	-4.26	0.9935
	Glass	23.8	18.1	1.57	0.0962	-3.91	0.9913
	Stone	23.8	12.6	2.25	**0.0436**	-5.63	0.9976
	Bark	20	6.2	3.2	**0.0164**	-12.81	0.9999
	Wood	8.6	6.2	-0.91	0.7939	-16.92	1
	Leather	17.1	12	1.13	0.1611	-7.19	0.999
	Fabric	13.3	10.3	0.48	0.3277	-9.15	0.9996
	Fur	57.1	21.8	4.72	**0.0046**	0.16	0.4393

T-values and *p*-values reflect the results of t-tests on differences between actual performance and chance-level or threshold-level performance, respectively. *P*-values in bold font indicate significant results at the alpha level of 0.05. N = 5.

2.3. Discussion

Interestingly, in the first test phase (the third phase) of experimental design 1, monkeys showed good categorization of metal and, to a lesser degree, fur, but showed poor performance in the other categories. The improvement of performance and the ability to shift to categorization after intensive identity matching in the fourth phase varied across categories. Monkeys showed good improvement for ceramic, glass, and stone, although their performance for ceramic was split (Kojilo did well, while Zephie made confusion errors with fabric). The monkeys showed consistently superior categorization for metal and fur than for the other categories, and their confusion between wood and fabric did not improve.

In the third phase of design 1, there were two types of test trials, identity matching and category matching. The latter was expected to be a much stricter test for exploring whether observers had a concept of the tested material category, whereas the former task can be achieved by simply matching the images without any conceptual knowledge of material category. However, we did not find prominent differences in performance between the two types of trials. This result indicates that applying the concept of matching based on material properties to new images was difficult even in identity matching for most categories. However, it is likely that the monkeys tried to apply the learned strategy as they kept good performance in baseline trials.

Metal and fur are materials consistently present in the captive environment of our monkeys; the cage is made of metal, and of course, each animal has its own fur. In contrast, opportunities to view the other material categories are much less frequent; therefore, it is reasonable to think that the monkeys' high accuracy for metal and fur and not for other categories is owing to experience. However, it is also possible that metal and fur have distinctive visual features that make these categories salient and differentiate them from the other categories.

The main difference in test trial procedures between designs 1 and 2 was that the comparison images in design 2 were always the same learned images in the first phase, whereas novel images were used in design 1. The procedures in design 2 were expected to enhance the monkeys' strategy of matching based on category-specific features of the typical exemplars. However, performance was not improved relative to category matching in the third phase of design 1, although fur obtained better performance in design 2 (mean = 57.1%) than in design 1 (mean = 41.7%). The high categorization ability for metal compared with other materials was not observed in design 2. This might be attributable to the procedures of design 2 that used the typical exemplars as comparison or to the individual differences in monkeys. In sum, monkeys showed similar tendencies in the two designs: consistently high accuracy for fur and many confusion errors for the other categories.

3. Experiment 2

The material categories used in experiments 1 represent materials that humans encounter in daily life, but with the exception of fur and metal, these materials are not ubiquitous in the environment of captive monkeys. Therefore, humans should perform better on most material categories used in this study, and similarity and differences between humans and monkeys might become clearer if their categorization abilities were compared by the same design. Therefore, we conducted a comparable matching-to-sample task with human participants.

3.1. Methods

3.1.1. Participants

Six male and six female human participants with an average age of 23.8 years (SD = 2.8 years) participated voluntarily. The experiment was conducted according to the guidelines for human research in Kyoto University and the principles of the Declaration of Helsinki, and each participant signed an informed consent before participating. All participants were naive to the purpose of the experiments and had normal or corrected-to-normal visual acuity.

3.1.2. Stimuli and procedure

The main procedure of the matching-to-sample task was the same as that in experiments 1, with a few modifications. First, responses were mouse clicks. Second, the sample image disappeared when it was clicked. Third, all responses were simply extinguished; i.e., no reward or timeout followed. Participants performed one session that consisted of 72 identity and 504 category matching trials (see supplementary methods for details).

3.1.3. Analysis

We examined whether correct choice and confusion errors between categories were statistically above chance (11.1%) or threshold (55.6%) levels by a one-tailed t-test with a significance level of 0.05.

To compare the temporal aspect of humans' and monkeys' tendencies in the visual categorization of materials, we analyzed reaction time (RT) represented as the latency of the participant's choice of comparison stimuli after the onset of comparison images. We calculated mean RTs for each category and for each individual in the test phases (third and fifth phases of design 1, second phase of design 2 in experiment 1 and the single phase in experiment 2). We used RT in correct trials and analyzed identity-matching and category-matching trials separately. First, we examined whether there were differences in RT between the two trial types in each species. Then we conducted a one-way ANOVA to examine whether there was a main effect of category in each trial type in each species. If there was a main effect, we further conducted multiple comparisons to look into detailed differences of categories in RT. Since the apparatus, body size, time schedule of a trial and physical movement for choice were different between monkeys and humans, we did not directly compare the RT between two species. Instead, we conducted a correlation analysis to see whether there were similar propensities in the two species. Using Pearson correlation analysis, we correlated the mean RT for each category across participants in monkeys (n = 7) with that in humans (n = 12).

3.2. Results

3.2.1. Material categorization by humans

Fig. 5 shows the mean confusion matrices averaged across the 12 human participants. The statistics of experiment 2 are shown in Table 2. The performance of humans in identity matching (Fig. 5, left) was significantly above the threshold level for all categories (p<0.0001 for all categories). In fact, there were no error trials in identity matching of ceramic. In category matching (Fig. 5, right), the performance was also significantly above the threshold level for all categories. There was no category that was confused with other categories at significantly greater than the chance level in either identity or category matching. However, in category matching, there were confusion errors marginally significantly higher than the chance level between stone and bark when stone was the sample (t = 1.58, p = 0.071, mean = 15.2%) and between glass and metal when glass was the sample (t = 1.37, p = 0.099, mean = 15.3%). In addition, in category matching, there were considerable but not statistically significant confusion errors between fabric and wood when fabric was the sample (mean = 12.6%), between fabric and leather when fabric was the sample (mean = 13.2%) and between bark and stone when bark was the sample (mean = 11.2%). In identity matching, although not statistically significant, considerable confusion emerged between bark and wood when the sample was wood (mean = 11.5%).

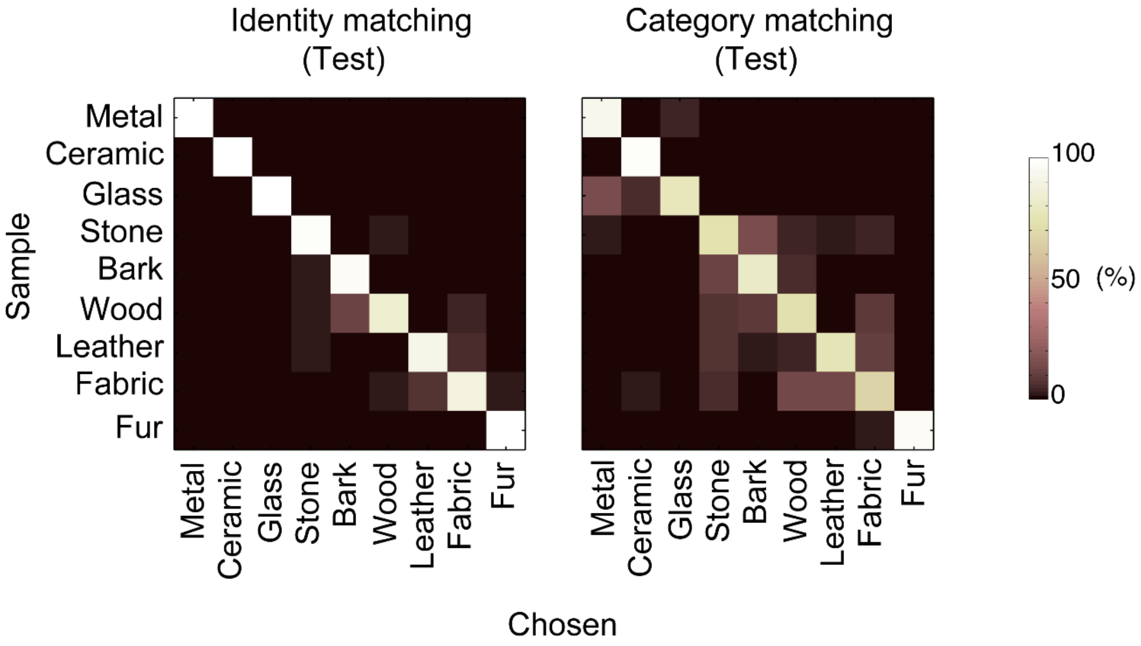


Fig. 5. Averaged confusion matrices in experiment 3. Left: Confusion matrix for identity-matching test trials. Right: Confusion matrix for category-matching test trials. Both matrices are presented as the mean value for the 12 human participants.

Table 2. Statistics on performance in identity- and category-matching trials in experiment 2.

				Comparison to chance level (11.1%)		Comparison to threshold level (55.6%)	
	Category	Mean (%)	SD (%)	t-value	*p*-value	t-value	*p*-value
Identity matching	Metal	99	3.6	84.33	**<0.0001**	41.67	**<0.0001**
	Ceramic	100	0				
	Glass	99	3.6	84.33	**<0.0001**	41.67	**<0.0001**
	Stone	96.9	5.7	52.55	**<0.0001**	25.32	**<0.0001**
	Bark	95.8	8.1	36.05	**<0.0001**	17.14	**<0.0001**
	Wood	83.3	9.7	25.71	**<0.0001**	9.89	**<0.0001**
	Leather	90.6	13.2	20.88	**<0.0001**	9.21	**<0.0001**
	Fabric	87.5	18.5	14.33	**<0.0001**	5.99	**<0.0001**
	Fur	99	3.6	84.33	**<0.0001**	41.67	**<0.0001**
Category matching	Metal	93.3	6.5	43.73	**<0.0001**	20.08	**<0.0001**
	Ceramic	97.8	3.7	80.27	**<0.0001**	39.1	**<0.0001**
	Glass	78	12.7	18.27	**<0.0001**	6.13	**<0.0001**
	Stone	73.5	13.5	16.03	**<0.0001**	4.61	**0.0004**
	Bark	81.1	7.9	30.86	**<0.0001**	11.26	**<0.0001**
	Wood	72	15.6	13.48	**<0.0001**	3.65	**0.0019**
	Leather	76.3	16	14.16	**<0.0001**	4.51	**0.0004**
	Fabric	65.9	18.8	10.08	**<0.0001**	1.91	**0.0416**
	Fur	95.4	5	58.84	**<0.0001**	27.81	**<0.0001**

T-values and *p*-values reflect results of t-tests on differences between actual performance and chance-level or threshold-level performance, respectively. *P*-values in bold font show significant results at the alpha level of 0.05. There were no errors in identity-matching trials of ceramic. N = 12.

3.2.2. Comparison of reaction time between monkeys and humans

We looked into participants' RT patterns for correct responses in each category condition. Fig. 6A demonstrates the mean RT across participants (red horizontal lines) and distribution of individual data (blue dots) for each category in identity-matching trials (left) and for category-matching trials (right) in test phases. Upper and lower panels illustrate RTs for monkeys and for humans, respectively. Both species tended to require more time in category-matching trials compared with identity-matching trials. In humans, RTs in all categories except for metal were significantly longer in category-matching trials than identity-matching trials ($p < 0.05$ for all categories except metal, one-tailed paired t-test, df = 11). Within each trial type (identity matching and category matching), there was a significant main effect of categories on RT for identity trials (one-way ANOVA, $p < 0.0001$, F = 5.79, df = 8) and for category trials ($p < 0.0001$, F = 8.35, df = 8) in humans. Post-hoc multiple comparison analysis showed that there were significant differences in RT between one of glossy materials (metal, ceramic and glass) and other categories (wood, fabric, leather and fur) for identity trials. In category trials, there were also significant differences between one of glossy materials (metal and ceramic) and other categories (stone, wood, leather and fabric) (Fig. 6A, lower panels). In monkeys, RTs in all categories except metal, glass and wood were significantly longer in category-matching trials than identity-matching trials ($p < 0.05$ for all categories except metal, glass and wood; one-tailed paired t-test, df = 6). There was no significant main effect of category in monkeys' RTs.

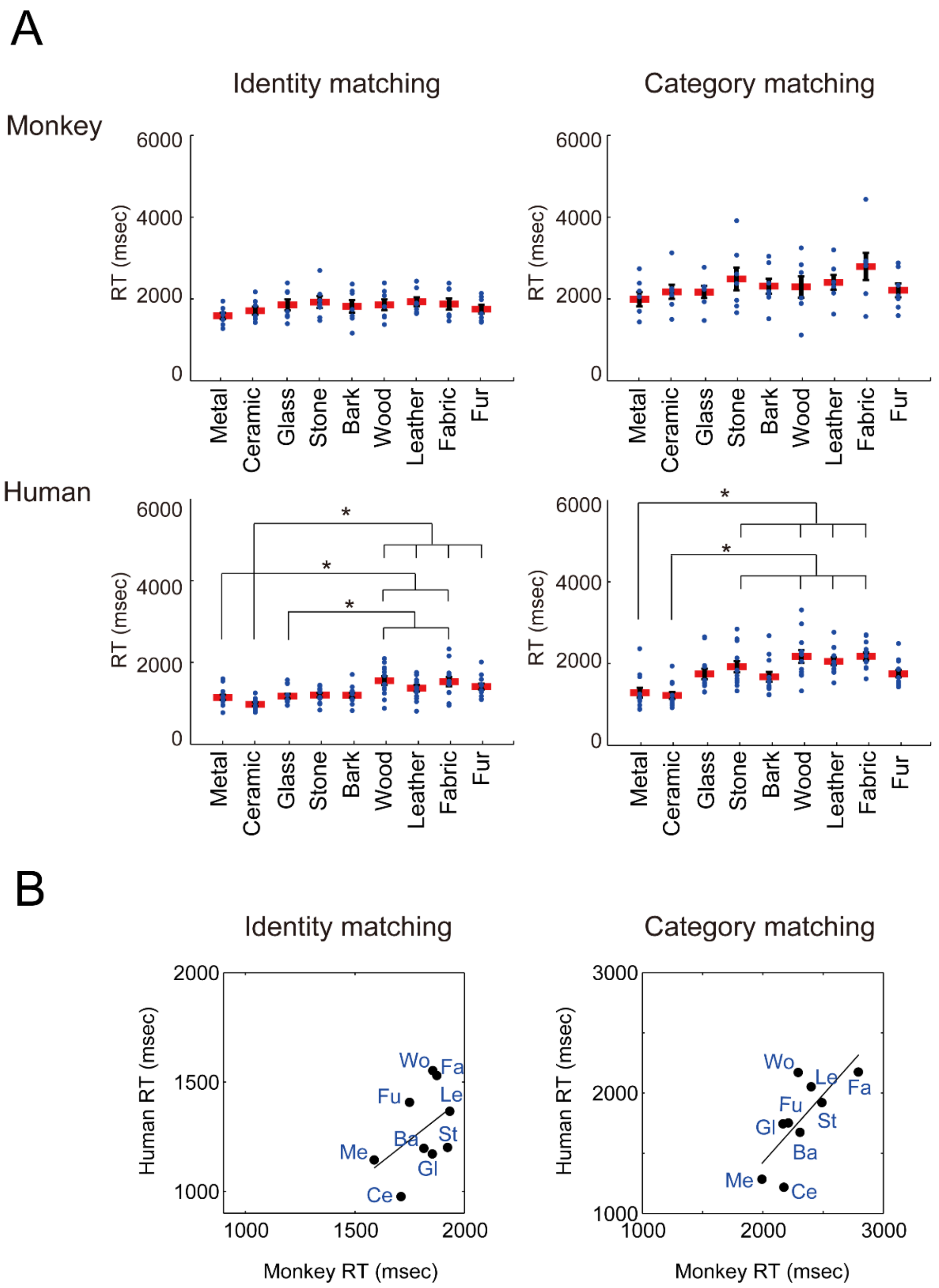


Fig. 6. Reaction-time analysis. (A) Mean reaction time (red horizontal line) ± SEM (black vertical line with ticks) for each category across participants and distribution of mean reaction time (RT) of each participant (blue dot). Top panels are RTs for monkeys and bottom panels are for humans for identity matching (left) and for category matching (right) during test phases. Asterisks indicate significant differences in mean RTs at the alpha level of 0.05

in multiple comparison analyses. (B) Correlation between monkey RT and human RT for identity matching (left) and for category matching (right). Solid lines indicate regression lines.

There was a significant positive correlation of RT between monkeys and humans in category-matching trials (*r* = 0.74, *p* = 0.023, Fig. 6B, right). Although not significant, the correlation was also positive in identity-matching trials (*r* = 0.45, *p* = 0.22, Fig. 6B, left). There was a consistent pattern that RTs for fabric, stone and leather were relatively longer, while those for ceramic and metal were shorter in both species in both category- and identity-matching trials (Fig. 6B). We note that similar results were obtained when incorrect trials were included in the RT analyses, although correlation coefficients were smaller.

3.3. Discussion

The performance of humans was generally high in both identity- and category-matching trials. This is to be expected because material images used in this experiment were originally created on the basis of human categorization (Hiramatsu, et al., 2011). However, humans made considerable confusion errors between glass and metal; the error rate was marginally significantly above chance in category matching (Fig. 5). Similar errors were not prominent in monkeys. Glass and metal may share similar semantics such as gloss, hardness and artificiality (Hiramatsu, et al., 2011). Therefore, this result may suggest that the semantics of categories had some influence for human performance and this aspect would be an interesting topic in future investigations.

In RT analyses for correct trials, both monkeys and humans generally showed longer RTs for category-matching trials compared with identity-matching trials (Fig. 6A). This suggests that both species took more time to make their decisions in difficult trials. Interestingly, in both species, there was no difference in RTs between identity matching and category matching in the metal condition, and metal had the fastest RTs in monkeys and the second-fastest RTs in humans. This indicates that metal images were easy to categorize for both species. However, there was no significant main effect of category in the RTs of monkeys. This is probably due to the small sample size. Intensive identity-matching training before test phases might also attenuate differences of difficulty across categories in identity trials. However, we found an interesting significant correlation between monkeys and humans in category-matching trials, suggesting similar patterns of difficulty in categorizing materials visually for the two species. For humans, metal and ceramic were easy, whereas fabric, wood and leather were difficult to categorize. For monkeys, metal was the easiest and fabric was the most difficult category although difficulty of other categories implied by RTs was not clear.

4. General discussion

4.1. Visual features related to performance

Monkeys showed good categorization ability for metal and fur but not for the other categories in design 1 of experiment 1. This result may suggest that frequent experience with certain material categories enhances the ability to classify them categorically. However, it is also possible that metal and fur have distinctive visual features that make these categories salient and differentiate them from the other categories. If this is true, these visual features must be consistent across samples, although it is unknown whether such dimensions are low-level features (e.g., luminance contrast, orientation, spatial frequency) or higher-level, material-specific features. We examined exemplars of each category with low-level visual features using 20 parameters (generated from orientation, spatial frequency, and pixel statistics of luminance histogram of images) obtained by texture analysis (see supplementary methods). Fig. 7 shows the distribution of each exemplar by classical multi-dimensional scaling (cMDS) in a two-dimensional space, using the 20 parameters of low-level visual features. This analysis shows fur exemplars to be similar to each other within the category and this may help the monkeys to match to sample accurately. However, this does not explain why other categories (bark, wood, stone, leather and fabric) similarly clustered in the cMDS space were difficult to classify by monkeys. Probably, those difficult categories would require more complex texture analysis and/or frequent opportunities to see the materials. One point of caution for interpreting our results is that the outline of furry objects was fuzzy and thus different from the clear edges of the other categories. We allowed this exception in our stimuli because creating clear outlines for furry objects greatly impaired the perception of fur in a preliminary experiment with humans. Therefore, we cannot deny the possibility that monkeys paid attention to the outline of fur when the sample was fur. Future studies should control this factor, and using an aperture to view furry objects might be one solution.

Although metal exemplars are not clustered in the cMDS space, their extreme dissimilarity in low-level visual features might be highly distinguishable from other categories. In the cMDS space, several exemplars that belong to the glossy materials (metal, glass and ceramic) located peripherally compared to others. If saliency in low-level visual features had been the key to conduct the task, monkeys should have performed better for these exemplars compared to the other exemplars belonging to the glossy categories but locating centrally in the space. Therefore, we compared the performance for glossy peripheral exemplars (indicated by yellow circles in Fig. 7) with that for other glossy exemplars (glossy central) in the test category-matching trials. In both designs 1 and 2, there was a tendency that monkeys performed better for peripheral exemplars (mean performance: glossy central: 32.5±8.9 %, glossy peripheral: 44.9±21.6 %, n=7). Six monkeys out of seven performed better for peripheral exemplars although there was no statistically significant difference between peripheral and central exemplars (*p* = 0.14, two-tailed Wilcoxon signed rank test). More investigations are necessary to clarify the contribution of experience and visual features for material categorization.

An fMRI study by Goda et al. (2014) that used the same stimuli as in this study and investigated neural representation of material categories in macaque monkeys demonstrated that low-level image features of metal, ceramic, and fur were relatively strong contributors to the neural representation of materials in V4 and the posterior part of the IT, where discrimination of natural textures can be processed (Arcizet, Jouffrais, & Girard, 2008; Koteles et al., 2008). Interestingly, capuchin monkeys in the current study showed better category-matching performance for metal, ceramic, and fur than for other categories (see Figs. 3B, right and 4, right). This coincidence might indicate that mid-level neural processing based on low-level image features is related to the feasibility of categorizing specific materials. Because there were no common characteristics among those categories in the cMDS distribution based on low-level image statistics (Fig. 7), it remains unknown how image features of those categories were related to neural representation and performance in monkeys.

It is notable that both species tended to show better and faster categorization for glossy categories like metal and ceramic compared with matte categories like fabric, leather and stone. This result was consistent with the human psychological experiment that showed strong connection of glossy appearance to rapid classification of materials (Nagai et al., 2014). The perceptual salience and categorization of glossy objects might be hard-wired in the primates' visual system. A study using a preferential looking paradigm in human infants showed that 7- to 8-month-olds looked longer at glossy objects than at matte objects (Yang et al., 2011), which supports this idea. From an evolutionary perspective, monkeys may have come to perceive glossiness as an important cue to survive in their natural environment. For example, fruits and leaves, which are the main food resources of many primate species, have various glossiness levels depending on their maturity and species. Therefore perceiving glossiness appears helpful to foragers for recognizing edible fruits or leaves. Future studies should explore to what extent material properties like glossiness are hard-wired by evolutionary processes and experience can modulate perception of them.

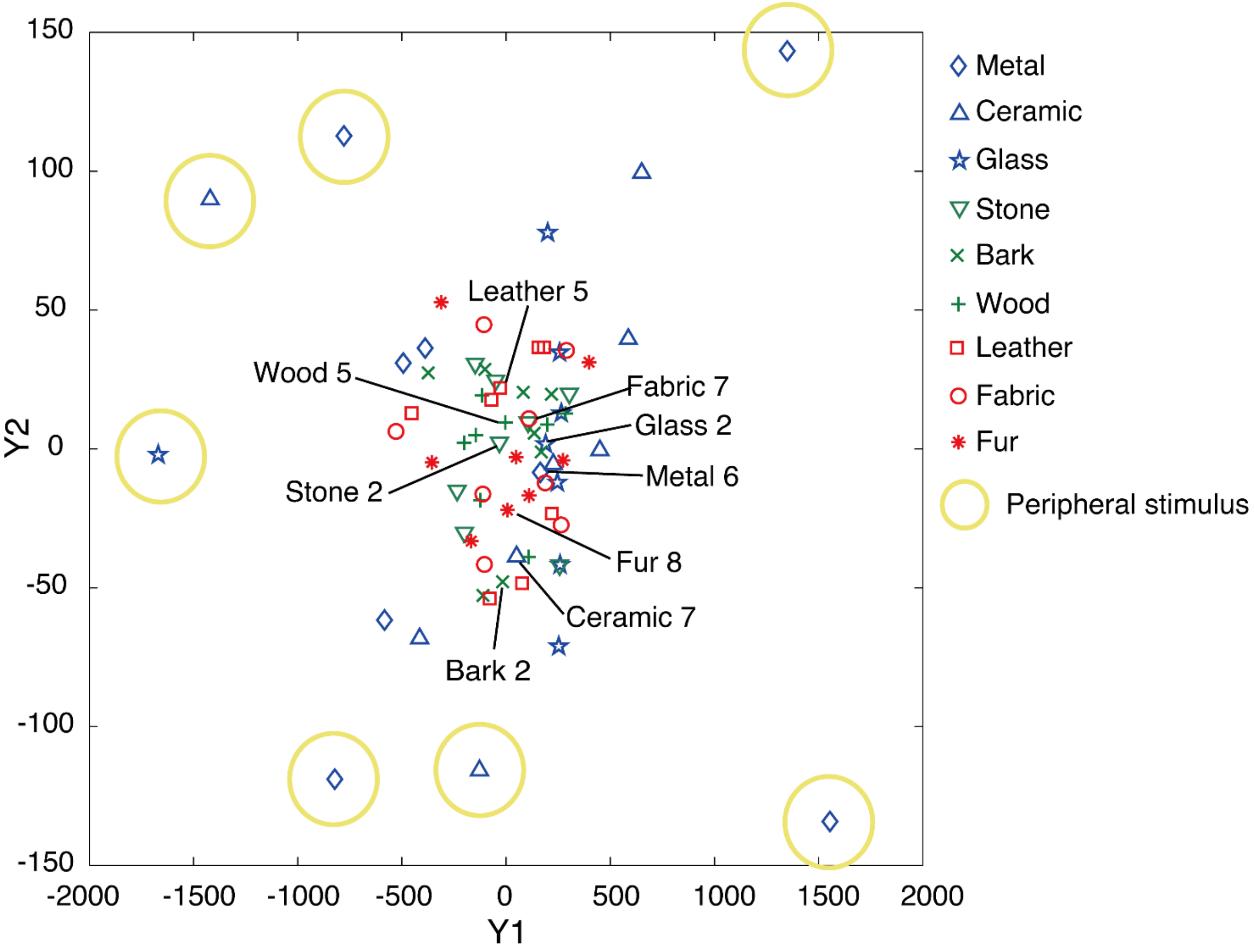


Fig. 7. Distribution of all images in a two-dimensional cMDS space based on the 20 low-level image statistics by texture analysis. The typical exemplars used in experimental design 2 are labeled. Exemplars indicated by yellow circles are peripheral exemplars.

4.2. Effect of color vision

Capuchin monkeys, like most New World monkeys, are known to have polymorphic color vision (Jacobs, 2007). Generally, there are dichromatic and trichromatic individuals in one social group due to variation of the X-linked red-green opsin gene. Females who are heterozygous on this gene become trichromats, while homozygous females and hemizygous males ought to be dichromats (Mollon, 1989; Hiramatsu et al., 2005). Because of this unique characteristic of New World monkeys, we used gray-scale images in our experiments to eliminate the effect of color-vision differences. However, one might think that there may still be correlated behavioral polymorphism with regard to the extraction of surface properties, because dichromats are less capable of extracting information from color and may rely more on the features present in the images used here. Therefore, we conducted color-vision typing for monkeys by DNA analyses described elsewhere (Hiramatsu, et al., 2005; Hiramatsu et al., 2008). In fact, there were three trichromatic females (Kojilo, Kiki and Zen) and four dichromats (Zephie, Heiji, Zinnia and Zilla) among our monkey participants. Fig. S3 shows confusion matrices averaged across participants with the same color-vision type in test phases (third phase of experimental design 1 and second phase of design 2). There was a trend that trichromatic females tended to choose metal irrespective of sample categories (Fig. S3, 2nd row, right). However, this trend was not consistently significant for all categories. Therefore, it is difficult to conclude that the trend is attributable to differences in color vision rather than to some other individual variation. Performance for correct-choice (diagonal blocks of confusion matrices) and RT analyses showed no significant differences between dichromats and trichromats.

In experiment 2, one male participant was diagnosed as having anomalous trichromatic color-vision type by the Ishihara color test, but we observed no apparent performance differences between this participant and other trichromatic participants (see Fig. S3). Because the sample size was small, it is not clear whether there are differences in perception of surface qualities related to color vision. This will be an interesting topic for future investigation.

5. Conclusions

As far as we know, this is the first study that has tested visual categorization ability for surface qualities of materials in non-human animals. Overall, our results showed that capuchin monkeys and humans share some perceptual tendencies in the categorization of surface qualities of materials by visual inspection. Glossy material like metal and ceramic seemed easy to categorize for both species compared with matte categories like wood, fabric and leather. Detailed texture patterns accompanied by those matte categories may require careful observation for categorization. Although inter-species differences emerged in other aspects, our results suggest the possibility that human perception of surface qualities has been shaped through primate evolution. To deepen our understanding of the ways that the perception of materials has evolved in animals, more comparative studies are needed that focus on three areas: (1) visual features and neural mechanisms that animals share for perceiving surface qualities, (2) the influence of experience on material perception and (3) the role of interaction with other sensory modalities, such as the tactile sense.

Acknowledgments

This work was supported by a Grant-in-Aid for JSPS Fellows (no. 10J04395) to C. Hiramatsu and JSPS Grants-in-Aide for Scientific Research Nos. 20220004 and 25240020 to K. Fujita. We thank the anonymous reviewers for their valuable comments and suggestions to improve the manuscript. We are grateful to H. Kuroshima, T. Matsuno and A. Takimoto for initial training of the matching-to-sample task for the monkeys; to T. Matsuno for providing a program for the task; to N. Goda for providing programs for the analysis; and to H. Komatsu for helpful comments and discussions.

References

Anderson, B. L. (2011). Visual perception of materials and surfaces. *Curr Biol, 21*(24), R978-R983. doi: 10.1016/j.cub.2011.11.022

Anderson, J. R., Kuwahata, H., Kuroshima, H., Leighty, K. A., & Fujita, K. (2005). Are monkeys aesthetists? Rensch (1957) revisited. *J Exp Psychol Anim Behav Process, 31*(1), 71-78.

Arcizet, F., Jouffrais, C., & Girard, P. (2008). Natural textures classification in area V4 of the macaque monkey. *Exp Brain Res, 189*(1), 109-120. doi: 10.1007/s00221-008-1406-9

Fleming, R. W., Wiebel, C., & Gegenfurtner, K. (2013). Perceptual qualities and material classes. *J Vis, 13*(8): 9, 1-20. doi: 10.1167/13.8.9

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. *Nat Neurosci, 16*(7), 974-981. doi: 10.1038/nn.3402

Fujita, K. (2009). Metamemory in tufted capuchin monkeys (*Cebus apella*). *Anim cogn, 12*(4), 575-585. doi: 10.1007/s10071-009-0217-0

Fujita, K., & Giersch, A. (2005). What perceptual rules do capuchin monkeys (*Cebus apella*) follow in completing partly occluded figures? *J Exp Psychol Anim Behav Process, 31*(4), 387-398.

Goda, N., Tachibana, A., Okazawa, G., & Komatsu, H. (2014). Representation of the material properties of objects in the visual cortex of nonhuman primates. *J Neurosci, 34*(7), 2660-2673. doi: 10.1523/jneurosci.2593-13.2014

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *Neuroimage, 57*(2), 482-494. doi: 10.1016/j.neuroimage.2011.04.056

Hiramatsu, C., Melin, A. D., Aureli, F., Schaffner, C. M., Vorobyev, M., Matsumoto, Y., & Kawamura, S. (2008). Importance of Achromatic Contrast in Short-Range Fruit Foraging of Primates. *Plos One, 3*(10), e3356. doi: 10.1371/journal.pone.0003356

Hiramatsu, C., Tsutsui, T., Matsumoto, Y., Aureli, F., Fedigan, L. M., & Kawamura, S. (2005). Color vision polymorphism in wild capuchins (*Cebus capucinus*) and spider monkeys (*Ateles geoffroyi*) in Costa Rica. *Am J Primatol, 67*(4), 447-461. doi: 10.1002/ajp.20199

Izawa, K. (1975). Foods and feeding behavior of monkeys in the upper Amazon basin. Primates, 16(3), 295-316. doi: 10.1007/BF02381557

Izawa, K. (1978). Frog-eating Behavior of Wild Black-capped Capuchin (*Cebus apella*). Primates, 19(4), 633-642. doi: 10.1007/BF02373631

Jacobs, G. H. (2007). New world monkeys and color. *Int J Primatol, 28*(4), 729-759. doi: 10.1007/s10764-007-9168-y

Kiesling, J. N. M., Yi, S. V., Xu, K., Sperone, G. F., & Wildman, D. E. (2014). The tempo and mode of New World monkey evolution and biogeography in the context of phylogenomic analysis. *Mol Phylogenet Evol*, doi: 10.1016/j.ympev.2014.03.027

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLoS Biol, 8*(6), e1000412. doi: 10.1371/journal.pbio.1000412

Koteles, K., De Maziere, P. A., Van Hulle, M., Orban, G. A., & Vogels, R. (2008). Coding of images of materials by macaque inferior temporal cortical neurons. *Eur J Neurosci, 27*(2), 466-482. doi: 10.1111/j.1460-9568.2007.06008.x

Maloney, L. T., & Brainard, D. H. (2010). Color and material perception: achievements and challenges. *J Vis, 10*(9): 19. doi: 10.1167/10.9.19

Mollon, J. D. (1989). Tho she kneeld in that place where they grew... The uses and origin of primate color-vision. *J Exp Biol, 146*, 21-38.

Motoyoshi, I., Nishida, S., Sharan, L., & Adelson, E. H. (2007). Image statistics and the perception of surface qualities. *Nature, 447*(7141), 206-209. doi: 10.1038/nature05724

Nagai, T., Matsushima, T., Koida, K., Tani, Y., Kitazaki, M., & Nakauchi, S. (2014). Temporal properties of material categorization and material rating: visual vs non-visual material features. *Vision Res*, doi: 10.1016/j.visres.2014.12.011

Nishio, A., Goda, N., & Komatsu, H. (2012). Neural selectivity and representation of gloss in the monkey inferior temporal cortex. *J Neurosci, 32*(31), 10780-10793. Doi: 10.1523/jneurosci.1095-12.2012

Nishio, A., Shimokawa, T., Goda, N., & Komatsu, H. (2014). Perceptual gloss parameters are encoded by population responses in the monkey inferior temporal cortex. *J Neurosci, 34*(33), 11143-11151. doi: 10.1523/jneurosci.1451-14.2014

Okazawa, G., Goda, N., & Komatsu, H. (2012). Selective responses to specular surfaces in the macaque visual cortex revealed by fMRI. *Neuroimage*, 63(3), 1321-1333. doi: 10.1016/j.neuroimage.2012.07.052.

Okazawa, G., Tajima, S., & Komatsu, H. (2015). Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc Natl Acad Sci U S A, 112*(4), E351-360. doi: 10.1073/pnas.1415146112

Ottoni, E. B., & Izar, P. (2008). Capuchin monkey tool use: overview and implications. Evolutionary Anthropology: Issues, News, and Reviews, 17(4), 171-178.

Paukner, A., Huntsberry, M. E., & Suomi, S. J. (2009). Tufted capuchin monkeys (*Cebus apella*) spontaneously use visual but not acoustic information to find hidden food items. *J Comp Psychol, 123*(1), 26. doi: 10.1037/a0013128.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *J Vis, 14*(9): 12. doi: 10.1167/14.9.12

Spinozzi, G., De Lillo, C., & Castelli, S. (2004). Detection of grouped and ungrouped parts in visual patterns by tufted capuchin monkeys (*Cebus apella*) and humans (*Homo sapiens*). *J Comp Psychol*, *118*(3), 297-308.

Wolfe, J. M., & Myers, L. (2010). Fur in the midst of the waters: visual search for material type is inefficient. *J Vis, 10*(9): 8. doi: 10.1167/10.9.8

Wright, A. A. (1999). Visual list memory in capuchin monkeys (*Cebus apella*). *J Comp Psychol*, *113*(1), 74.

Yang, J., Otsuka, Y., Kanazawa, S., Yamaguchi, M. K., & Motoyoshi, I. (2011). Perception of surface glossiness by infants aged 5 to 8 months. *Perception, 40*(12), 1491-1502.

Supplementary material

Supplementary Methods

1. General procedures of a matching-to-sample task in experiment 1

Each trial started with the illumination of one of two levers below the monitor. The monkey's pressing the lever down for 1 sec resulted in the appearance of a sample image at the center of the monitor. The monkeys had been trained to keep pressing the lever down throughout a trial by one hand. After the monkey touched the sample image three times by the other hand, nine comparison images, one exemplar from each category, appeared in locations randomly chosen from 10 possible cells surrounding the sample image (Fig. 1B). The position of the comparisons within each cell was slightly changed from trial to trial to avoid monkeys' persistently touching the same locations on the screen. The trial ended and all stimuli disappeared when the monkey touched one of the comparison images. There was no time limit on a trial and the monkeys were allowed to choose a material image from comparisons as long as they kept pressing the lever down. Releasing the lever before the end of a trial aborted the trial and the same trial repeated after a 3-sec inter-trial interval.

2. Procedures of experimental design 1 in experiment 1

The experimental design 1 consisted of five phases. The first was a training phase, in which the monkeys had to choose the same image as the sample from nine comparison images (i.e., identity matching) taken from two stimulus sets (shapes one and two for Kojilo and shapes three and four for Zephie). The order of the material category of the sample and the stimulus set (shape) were randomized within a session. In this phase, a piece of sweet potato (a reward) accompanied by a 0.5-sec electronic doorbell sound (Horohoro buzzer; Panasonic, Japan) was presented if the monkeys touched the correct comparison, whereas a 5-sec timeout accompanied by a 0.5-sec buzzer sound different from the doorbell sound was given if the monkeys touched an incorrect comparison. The light on the operant box was turned off during timeout periods. Inter-trial intervals of 3 sec followed each reward and timeout. Each session consisted of 90 trials, and each image was presented as the sample five times within a session. Each monkey worked for one session per day, 6 days per week. Monkeys moved to the next phase after their matching accuracy exceeded 80% for two consecutive sessions.

In the second phase, new two stimulus sets (shapes three and four for Kojilo and shapes one and two for Zephie) were used as stimulus images. The procedure was exactly the same as that in the first phase.

The third phase consisted of baseline and test trials. Each session consisted of 54 baseline trials and 18 randomly inserted test trials. The baseline trials were identity matching with learned images in the first and second phases. In the test trials, new stimulus images from four stimulus sets, shapes five to eight, were used and monkeys were rewarded irrespective of their choice. There were two types of test trials: identity matching and category matching. In identity matching, comparison images comprised a stimulus set that contained one image identical to the sample, i.e., a stimulus set with the same shape as the sample image. In category matching, comparison images were from a different stimulus set from the sample, i.e., a stimulus set with a shape that was different from the sample (Fig. 1B). Each new image was presented once as the sample in identity-matching test trials and three times (with a different combination of stimulus sets between the sample and comparison images) in category-matching test trials during eight sessions.

In the fourth phase, we trained the monkeys on identity matching using the same stimulus sets, shapes five to eight, as were used in the third phase. Sessions consisted of 90 trials. All procedures were the same as those in the first phase; only the images differed. Monkeys moved to the fifth phase after their performance exceeded 80% correct for two consecutive sessions.

In the fifth phase, we re-tested category matching using stimulus sets of shapes one to eight. Each session consisted of 27 category-matching test trials and 108 identity-matching baseline trials. All category-matching combinations with different stimulus sets for each category were presented once during the eight sessions. Monkeys were again rewarded on all test trials irrespective of their choice.

3. Procedures of experimental design 2 in experiment 1

Procedure 2 consisted of two phases. The first phase was a training phase and we used nine typical exemplars from each category as a stimulus set in the training phase. To assess typicality, we analyzed each image using 20 parameters of low-level visual features (see image analysis section below). We defined the typical exemplar as the one that was nearest to the middle of the eight exemplars of a category in the two-dimensional space constructed by classical multi-dimensional scaling (cMDS) (Fig. 7). Typical exemplars for each category were: metal shape 6, ceramic shape 7, glass shape 2, stone shape 2, bark shape 2, wood shape 5, leather shape 5, fabric shape 7, and fur shape 8 (Figs 7 and S1). In the first phase, the monkeys were trained to choose the same image (one of exemplar images) from the nine typical exemplars from each category as comparison images. Each session consisted of 90 trials (10 identity matching per category). Monkeys moved to the next phase after their performance exceeded 80% correct for two consecutive sessions.

In the second phase, we introduced new images to test whether monkeys could demonstrate understanding of the concept of material categories. The new stimuli were 63 images out of 72 that had not been used as typical exemplars (7 exemplars × 9 categories). In test trials, a new image was presented as the sample, and monkeys were required to select one image from nine typical exemplars (comparisons) for which identity matching was trained in the first phase. Each session consisted of 21 test trials with new images (category matching) and 72 baseline trials trained in the first phase (identity matching). Each monkey participated in nine sessions. Each new stimulus was tested once within three consecutive sessions and in total three times within nine sessions. The sequence of trials was randomized in each session and for each monkey. Monkeys were always rewarded as in experimental design 1 in test trials irrespective of their choice; i.e., they were not trained for new images. In baseline trials, monkeys were rewarded only when they selected the correct identical image, to maintain the motivation of monkeys to choose correct stimuli.

4. Detailed methods of experiment 2

Enlarged versions of the same gray-scale images from experiments 1 were used. The size was 300 × 300 pixels on the 27-inch (1920 × 1080 pixels) hardware calibration LCD monitor (Color Edge CG275W; EIZO, Hakusan, Japan) to maintain a visual angle (ca. 9 × 9 degrees when viewed from 30 cm) similar to that for monkeys, who performed the task from approximately a 15-cm viewing distance. Calibration of the monitor was conducted by Color Navigator (version 6.0.0) calibration software (EIZO) with the built-in color calibration sensor of the monitor. The heads of participants were not restricted by a chin rest, so this experimental condition matched that used with the monkeys. A delay of 0.5 sec was inserted to prevent a ceiling effect of human accuracy. Ceiling effects were expected if sample and comparison images were presented simultaneously, because the material images had been created in the previous study (Hiramatsu, et al., 2011) so that they would be classified consistently into each category by humans.

Each stimulus set paired with a sample image consisted of nine categories with the same shape (e.g., stimulus set "shape 1"). Among 576 trials of one session, 72 trials (9 categories × 8 exemplars × 1 stimulus set) were identity matching and the remaining 504 trials (9 categories × 8 exemplars × 7 stimulus sets) were category matching. In identity-matching trials, all comparison images had the same shape and the correct choice was the image that was identical to the sample image. In contrast, in category-matching trials, comparison images were composed using the nine categories of one of the stimulus sets that was different from the stimulus set of the sample image. Therefore, the correct choice in a category-matching trial was intended to be the image that belonged to the same category but had a different shape and pattern from the sample image. Before the test session, participants learned the procedures in a 10-trial practice session. During the practice session, the instruction was given verbally to look at the sample and to find and click a similar or identical image from the comparison set. The experiment was carried out in a dark room. Participants were allowed to rest after every 192 trials and to restart the experiment at their own pace.

5. Image Analysis

To examine how the nine exemplars from each category resemble each other at the image level, we analyzed images using the Portilla–Simoncelli model of texture analysis algorithm (Portilla & Simoncelli, 2000). In the analysis, the center part of each material image (96 × 96 pixels for images used with monkeys) within the object contour was transformed using the "steerable pyramid" (Simoncelli & Freeman, 1995). This method decomposes images into 14 subbands (12 oriented, plus high-pass and low-pass residuals) and calculates six pixel statistics (mean, variance, skew, kurtosis, minimum, and maximum values of the image pixels) of the luminance histogram. Then we applied classical multi-dimensional scaling (cMDS) and calculated the pairwise Euclidian distance between exemplars, using the z-scored 20 parameters of low-level visual features obtained by the texture analyses and plotted these in a two-dimensional space (Fig. 7).

References

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *Neuroimage*, *57*(2), 482-494. doi: 10.1016/j.neuroimage.2011.04.056

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on Joint statistics of complex wavelet coefficients. *Int'l J Comp Vis*, *40*(1), 49-71.

Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: A flexible architecture for multi-scale derivative computation. *IEEE International Conference on Image Processing*. 3444-3447.

Supplementary Figures

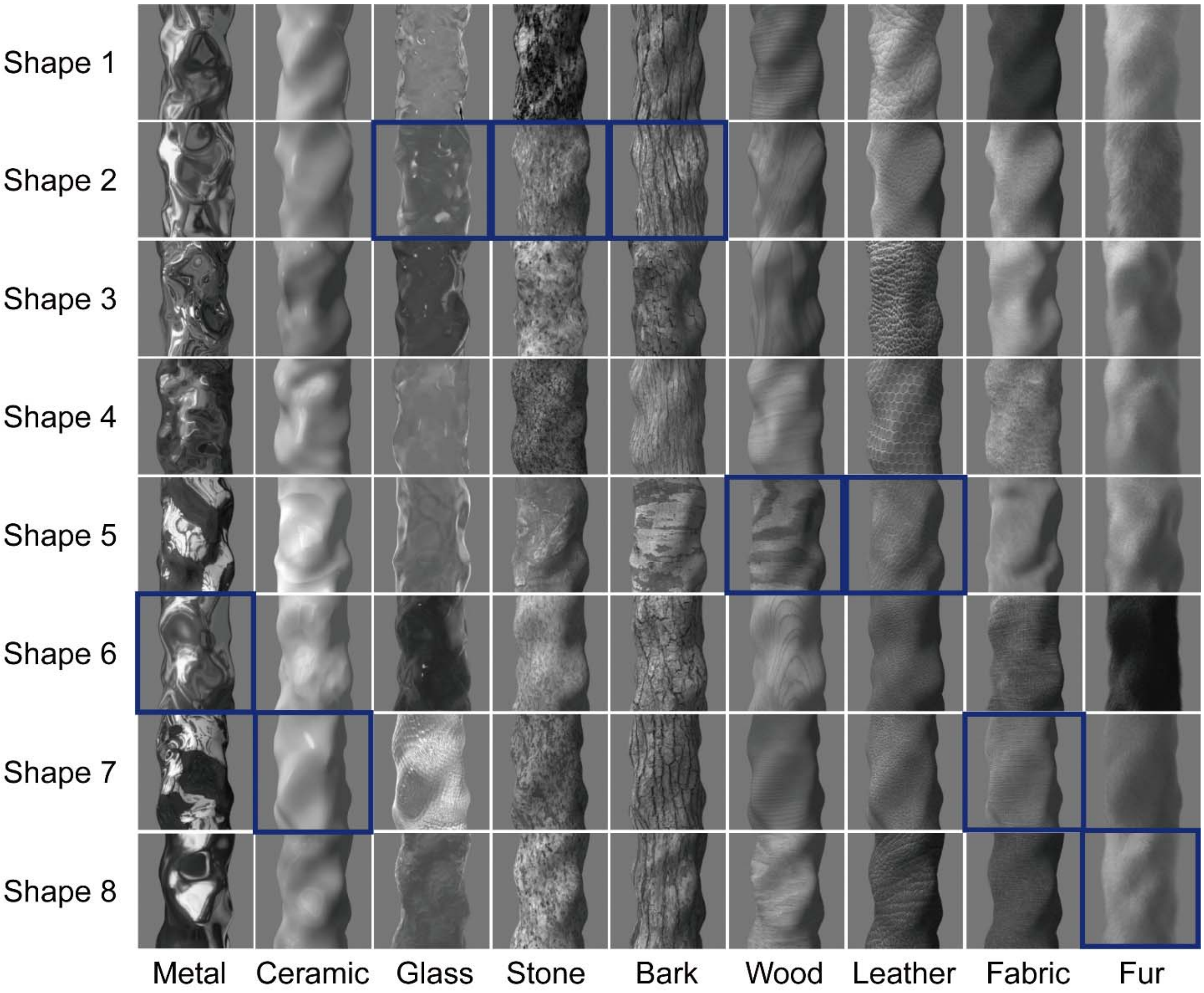


Fig. S1. Complete stimulus set used in this study.

The images outlined in blue were used as typical exemplars in experimental design 2 of experiment 1.

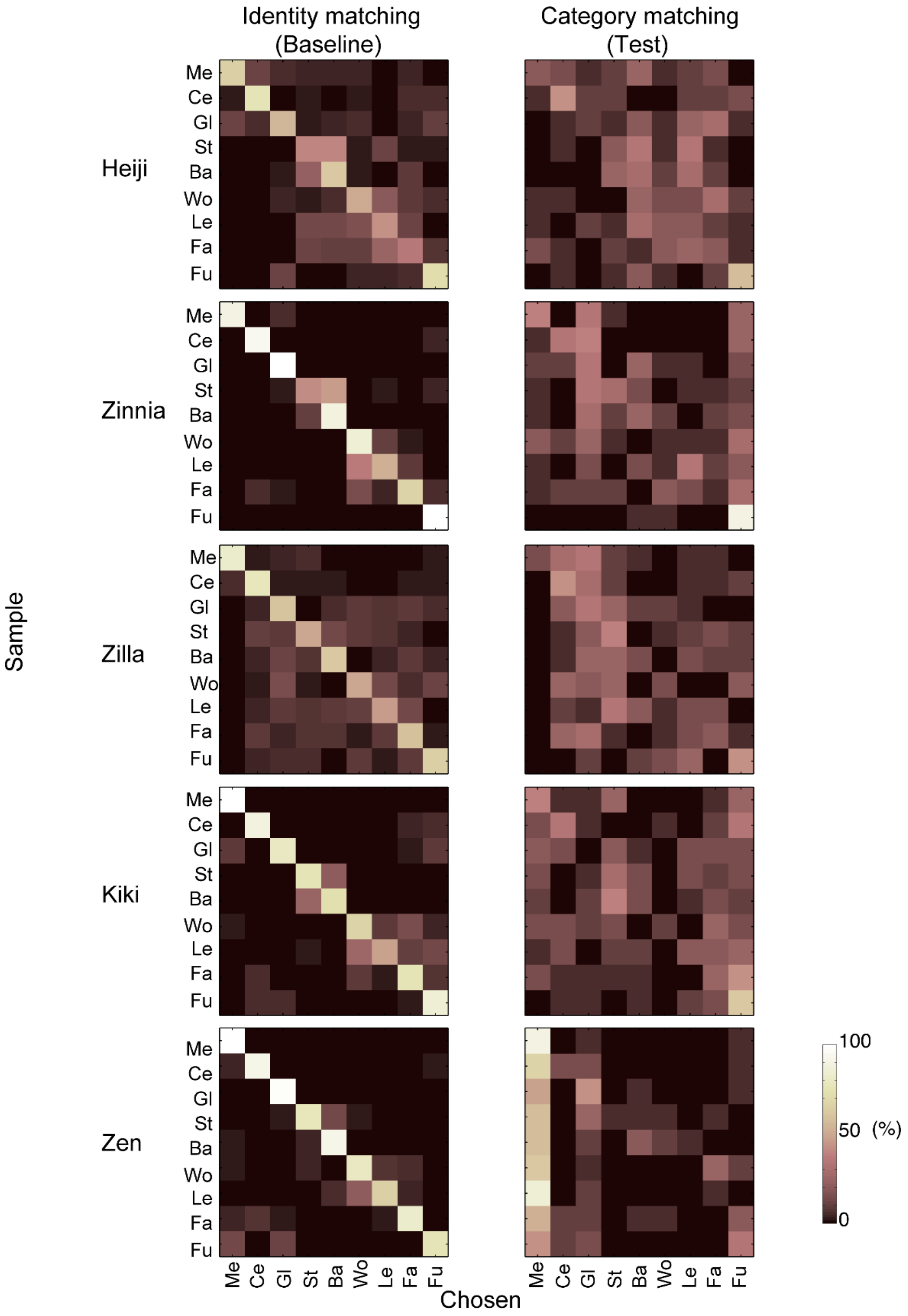


Fig. S2. Confusion matrices of each individual in the second phase of experimental design 2.

Left: confusion matrices for identity-matching baseline trials in the second phase.

Right: confusion matrices for category-matching test trials in the second phase.

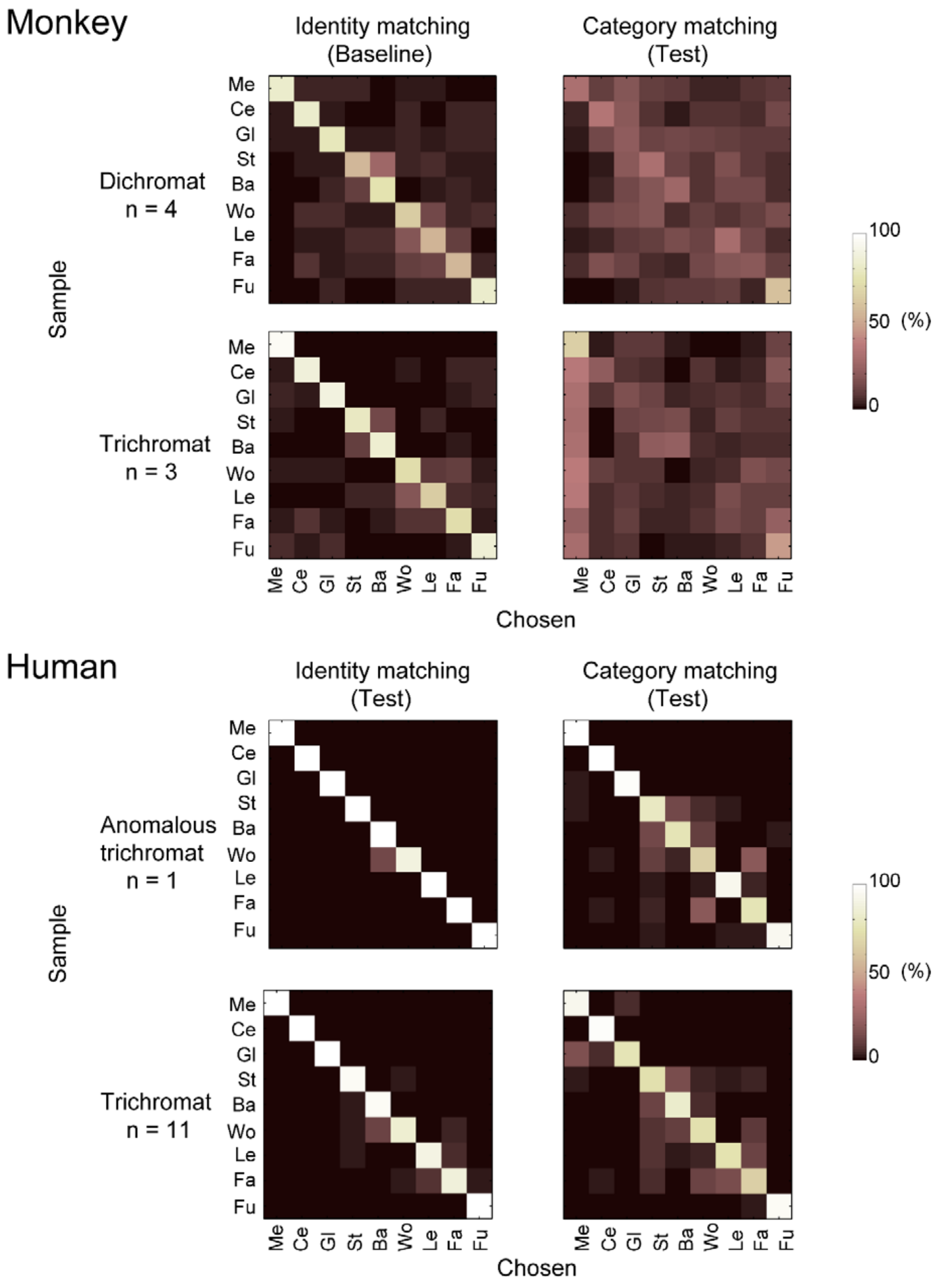


Fig. S3. Effect of color-vision type in confusion matrices.

From top to bottom, confusion matrices for dichromatic monkeys, for trichromatic monkeys, for an anomalous trichromatic human and for trichromatic humans in identity- (left) and category- (right) matching trials.