Probing semantic and visual representations in material perception through psychophysics and unsupervised learning

Chenxi Liao (cl6070a@american.edu)
Department of Neuroscience, American University
4400 Massachusetts Ave NW, Washington, DC, 20016, USA

Masataka Sawayama (masataka_sawayama@ipc.i.u-tokyo.ac.jp)
Graduate School of Information Science and Technology, The University of Tokyo,
Tokyo, Japan

Bei Xiao (bxiao@american.edu)
Department of Computer Science, American University
4400 Massachusetts Ave NW, Washington, DC, 20016, USA

Abstract

We investigate the relationship between visual judgment and language expression in material perception to understand how visual features relate to semantic or categorical representations. We use deep generative networks to construct an expandable image space to systematically sample familiar and unfamiliar materials. We compare the perceptual representations of materials from two tasks, visual material similarity judgments, and verbal descriptions, and discover a moderate correlation between vision and language within individuals. However, we also find a gap between these two modalities, signifying that while verbal descriptions capture material qualities on the coarse level, they may not fully convey visual nuances. Furthermore, we examine the image representation of materials derived from various data-rich neural network models and demonstrate that the distilled image features from these models have the potential to capture the human representation of materials.

Keywords: Human perception; Transfer learning; Generative model; Unsupervised learning; Large Language Models

Introduction

Recognizing materials and estimating their properties (e.g., softness, edibleness) from visual input is essential for humans to plan interactions with the environment (Fleming, 2017). Along with vision, language allows us to communicate relevant information about the materials. Furthermore, probing the connection between visual judgment and semantic description may unveil communicable features in materials. Although we can visually discriminate various materials, we might find it challenging to effectively describe their appearances with words. To what extent do words encapsulate the richness of visual material perception? We answer this by measuring material perception with visual similarity judgments and semantic descriptions using AI-generated images. When contrasting the representations derived from behavioral results with those from pre-trained data-rich vision models, we observed alignment and misalignment between human material perception and the task-agnostic deep features extracted by these models.

Methods

Space of Morphable Material Appearance We developed an unsupervised image synthesis framework to create an extensive range of familiar and unfamiliar materials in a controllable manner. Our framework is based on StyleGAN2-ADA (Karras et al., 2020), which captures the statistical regularity of the images with its multi-scale generative network (G) and layer-wise latent space (W). We transferred the pre-trained StyleGAN2-ADA model from a large dataset D_{soap} (8085 photos of 60 soaps) to the smaller datasets D_{rock} (3180 photos 24 rocks/crystals) and D_{toy} (1900 photos of 15 squishy toys), via separately applying end-to-end fine-tuning (Figure 1A). As a result, the Soap Model (W_{soap}, G_{soap}) turned into Rock

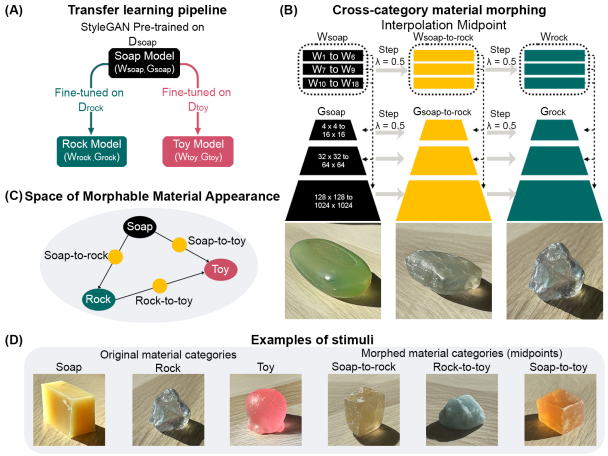

Figure 1: Synthesis pipeline for morphable material appearances. (A) Transfer learning from photographs of one material (soap) facilitates synthesizing other materials (rocks/crystals) and (toys). (B) Illustration of cross-category material morphing. (C) Illustration of the Space of Morphable Material Appearance. (D) Examples of synthetic stimuli used in the psychophysical experiments.

(W_{rock}, G_{rock}) and Toy Models (W_{toy}, G_{toy}), and can synthesize images of corresponding materials. We also created ambiguous materials by linearly interpolating between the latent codes of two materials (e.g., w_{soap} and w_{rock}) while also interpolating corresponding material generators' parameters (e.g., G_{soap} and G_{rock}) (Figure 1B). This enables us to build an expandable Space of Morphable Material Appearance, which includes three original materials (i.e., soap, toy, rock) and three morphed materials at morphing midpoints (i.e., soap-to-rock, rock-to-toy, and soap-to-toy) (Figure 1C).

Psychophysical Experiments We sampled 72 stimuli from six image categories of material (12 from each) from the Space of Morphable Material Appearance (Figure 1D). We conducted two behavior tasks within individuals: (1) Multiple Arrangement (MA): participants (N=16) arranged materials based on the similarity judgment of material properties (Kriegeskorte & Mur, 2012); (2) Verbal Description (VD): the same participants described the same images with free-form text input from five aspects: material name, color, optical properties, mechanical properties, and surface texture (Figure 2A).

Results

Comparing visual judgment with verbal description Within each participant, we compared Representational Dissimilarity Matrices (RDMs) between the visual judgment (i.e., Vision RDM) and verbal description (i.e., Text RDM) results, by applying the Representational Similarity Analysis (RSA) (Figure 2B). For each participant, a Vision RDM is created based on the Euclidean distances of pairwise comparisons in MA, and a Text RDM by encoding the images' text descriptions with a pre-trained large language model and computing

Figure 2: Behavioral and modeling results. (A) Illustration of behavioral experiments: visual material similarity judgment via Multiple Arrangement, and Verbal Description (B) Individual and group average perceptual RDMs (Top: Vision RDMs, Bottom: Text RDMs). The Spearman's correlation (r_s) between the corresponding Vision and Text RDMs are annotated in the box below. (C) Representational similarities between individual participants' visual judgments and selected pre-trained data-rich models. Left: RDM of 72 stimuli based on latent feature extracted from a pre-trained vision encoder: OpenCLIP-ViT-L/14. Right: Spearman's Correlation between each participant's Vision RDM or the model's image-feature RDM and the group average Vision RDM. (D) Optimal transportation plans (Γ). The purple diagonal indicates the perfect alignment on the image-to-image level.

their cosine distances (text embedding from CLIP (Radford et al., 2021) is shown as the main result). We found that all of the participants' verbal responses exhibited a significant correlation with their own MA behavior but with substantial individual variances (min Spearman's correlation $r_s = 0.10$, max $r_s = 0.52$, all $p < 0.001$, FDR-corrected). We observed a stronger correlation when comparing the group average Vision and Text RDMs ($r_s = 0.74, p < 0.001$).

We scrutinized the more nuanced-level alignment between the two perceptual spaces using the unsupervised alignment method, Gromov-Wasserstein Optimal Transport (GWOT) (Kawakita, Zeleznikow-Johnston, Tsuchiya, & Oizumi, 2023). To quantify the structural similarity between group average Vision and Text RDMs, GWOT yields the optimal transportation plan matrix, Γ. Each element in Γ indicates the probability of a sample in one similarity structure corresponding to another in the other similarity structure. As shown in Figure 2D (Left), the optimal Γ significantly deviates from being a diagonal matrix. Misalignments tend to occur within images synthesized from the same material generator (e.g., within soaps) or between the morphed materials (e.g., soap-to-toy) and the materials they morphed from (e.g., soap and toy). Participants could use similar words to describe samples within each material category that exhibit shared visual attributes (e.g., translucency). Our analysis suggests that visual judgment and language are relatively consistent at the coarse categorical level but the nuanced visual difference between material samples cannot be precisely described with words.

Comparing human visual judgment with models How do we find the features that are "missing" from the verbal descriptions? Recent weakly-supervised and self-supervised models show the plausibility of narrowing the behavioral difference between human and machine vision (Geirhos et al., 2021). Here, we assessed whether the representations learned by these pre-trained models align with human visual material judgments. If their image-feature representations approximate human perception, these computational models may provide insights into searching vision-specific features in material reasoning tasks. We confronted various pre-trained models (e.g., visual-semantic models OpenCLIP (Ilharco et al., 2021)) with our stimuli and obtained material similarity representations based on the models' deep image features (Figure 2C Left). Using both RSA and GWOT, we found that the image-feature representations from the tested models moderately correlated (Figure 2C Right) with human visual judgment at the level of coarse categories, yet precise mapping at the image level is still lacking (Figure 2D Right). Fitting a linear regression model, we found that joining the participant's own Text RDM with the image-feature RDM (e.g., OpenCLIP-ViT-L/14) significantly improves ($p < 0.001$ for all participants) the prediction of the participant's own Vision RDM. This implies that the image features distilled from such a model may contain relevant visual information about materials.

Conclusion

Using a novel image synthesis framework and two behavioral tasks, we found that human semantic representation of materials is crucial in their visual similarity judgments. However, there is a gap in using words to capture the nuanced visual differences between diverse material samples. Data-rich computational models that capture human visual material judgment may provide cues explaining visual nuances.

References

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

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *34*, 23885–23899.

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., ... Schmidt, L. (2021, July). *Openclip.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5143773 (If you use this software, please cite it as below.) doi: 10.5281/zenodo.5143773

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., & Aila, T. (2020). Training generative adversarial networks with limited data. *Advances in Neural Information Processing Systems*, *33*, 12104–12114.

Kawakita, G., Zeleznikow-Johnston, A., Tsuchiya, N., & Oizumi, M. (2023). Comparing color similarity structures between humans and llms via unsupervised alignment. *arXiv preprint arXiv:2308.04381*.

Kriegeskorte, N., & Mur, M. (2012). Inverse mds: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in psychology*, *3*, 245.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).