Mapping Core Similarity Among Visual Objects Across Image Modalities

Judith E. Fan[1*], Daniel Yamins[2†], James DiCarlo[2‡], & Nicholas B. Turk-Browne[1§]
[1]Department of Psychology, Princeton University, NJ, 08540
[2]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, 02139
*jefan@princeton.edu, †yamins@mit.edu, ‡dicarlo@mit.edu, §ntb@princeton.edu

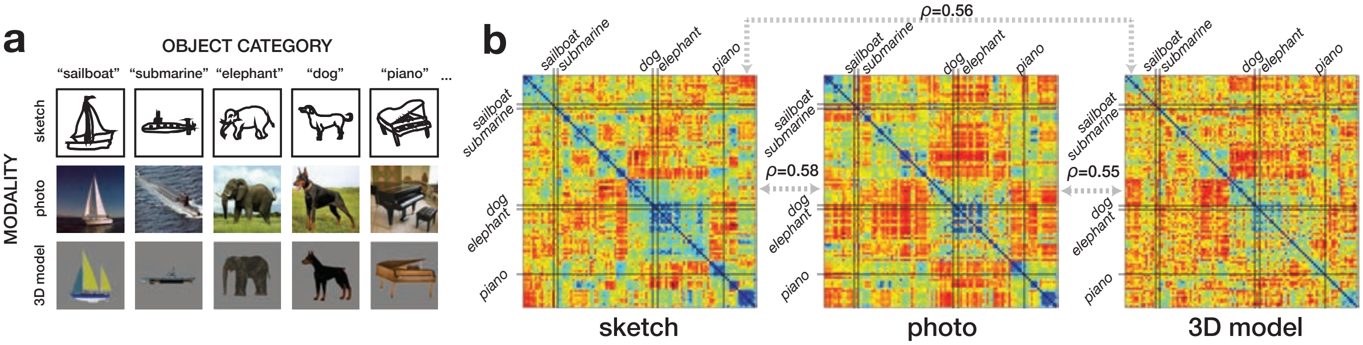

Figure 1. *(a) Sketch, photo, and 3D-model representations of common objects. (b) Representational Dissimilarity Matrices (RDMs) reflect the layout of object categories in high-dimensional feature space, separately for each image domain.* ρ *is the Spearman rank correlation between RDMs.*

1 Introduction

Humans have devised a wide range of technologies for creating visual representations of real-world objects. Some are ancient (e.g., line drawings using a stylus), while others are very modern (e.g., ptography and 3D computer graphics rendering). Despite large differences in the images produced by these differing modalities (e.g., sparse contours in sketches vs. continuous hue variation in photographs), all are effective at evoking the original real-world object.

What core visual properties are preserved across these diverse modalities such that reliable recognition is possible? Understanding the specific "pixel invariants" that are common to a photograph, line drawing, and 3D-rendered image of the same object is a challenge that lies at the heart of visual abstraction. In this work, we present a computational approach to extracting and quantifying such similarities.

2 Approach

We first assembled a multi-domain image set containing sketches, photographs, and synthetic rendered images (Fig. 1a). Using an existing sketch database [Eitz et. al. 2012], we obtained ~12,000 sketches of objects belonging to 147 common semantic categories. Using the Imagenet database [Deng et. al. 2009], we acquired ~200K natural photographs from corresponding categories. Finally, using 3D mesh models of objects in these same categories, we rendered ~200K synthetic images at high levels of object position, size, and pose variation.

We then applied a recently developed deep convolutional neural network architecture to extract features on these images [Yamins et al. 2013]. This network has been shown to achieve human-level performance on a challenging object recognition task, as well as provide an effective approximation of the neural population responses in high-level visual cortex. As such, it was an attractive candidate for capturing visual abstraction.

We next computed Representational Dissimilarity Matrices (RDMs) [Kriegeskorte 2008] for these extracted features. Each matrix entry in an RDM is the correlation distance (1-correlation) between the average feature vectors from the model for a pair of categories. Smaller values (cooler colors) reflect relatively proximal pairs of categories, whereas larger values reflect more distant category pairs. The three 147x147 RDMs (Fig. 1b) provide compact visualizations of the layout of all the categories in the high-dimensional feature space, separately for each of the three image domains.

All three RDMs individually show clear block-diagonality, indicating meaningful higher-order structure due to semantic clustering of object categories. The RDMs also show striking cross-domain similarities, both visually and as quantified by Spearman rank correlation comparisons (see figure caption). This indicates an underlying commonality in the feature representations for the three image modalities. Because of the large amount of variation in the photograph and 3d-model datasets, this similarity cannot be derived from low-level image statistics alone.

3 Implications

We have shown here how a deep neural network that exhibits sufficient capacity for visual abstraction can produce highly congruent category "maps" based upon images taken from three very different image domains. In the future, we plan to extend this work to build object recognition algorithms that automatically generalize across multiple image modalities. We also plan to apply machine learning techniques to capture the detailed mapping of features between the modalities, with the ultimate goal of producing image-level transformations that convert between them (e.g., automated "sketch-ification" of photos).

References

DENG, J., DONG, W., SOCHER, R., LI, L.-J., LI, K., & FEI-FEI, L. 2009. Imagenet: A large-scale hierarchical image database. *IEEE Computer Vision and Pattern Recognition (CVPR)*, 248–255.

EITZ, M., HAYS, J., & ALEXA, M. 2012. How do humans sketch objects? *ACM Transactions on Graphics (TOG)*, *31*(4), 44.

KRIEGESKORTE, N. 2008. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(4): 1-28.

YAMINS, D. L., HONG, H., & CADIEU, C., & DICARLO, J. 2013. Hierarchical Modular Optimization of Convolutional Networks Achieves Representations Similar to Macaque IT and Human Ventral Stream. *Advances in Neural Processing Systems.*

SIGGRAPH 2014, August 10 – 14, 2014, Vancouver, British Columbia, Canada.
2014 Copyright held by the Owner/Author.
ACM 978-1-4503-2958-3/14/08