Causal and compositional generative models in online perception

Ilker Yildirim*[1] (ilkery@mit.edu), **Michael Janner***[1] (janner@mit.edu), **Mario Belledonne**[1] (belledon@mit.edu), **Christian Wallraven**[2] (christian.wallraven@gmail.com), **Winrich Freiwald**[3] (wfreiwald@rockefeller.edu), **Joshua B. Tenenbaum**[1] (jbt@mit.edu)

[1] Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA; [2] Brain and Cognitive Engineering, Korea University, Seoul, Korea; [3] Laboratory of Neural Systems, Rockefeller University, New York, NY

Abstract

From a quick glance or the touch of an object, our brains map sensory signals to scenes composed of rich and detailed shapes and surfaces. Unlike the standard pattern recognition approaches to perception, we argue that this mapping draws on internal causal and compositional models of the outside physical world, and that such internal models underlie the generalization capacity of human perception. Here, we present a generative model of visual and multisensory perception in which the latent variables encode intrinsic properties of objects such as their shapes and surfaces in addition to their extrinsic properties such as pose and occlusion. These latent variables can be composed in novel ways and are inputs to sensory-specific causal models that output sense-specific signals. We present a novel recognition network that performs efficient inference in the generative model, computing at a speed similar to online perception. We show that our model, but not an alternative baseline model or a lesion of our model, can account for human performance in an occluded face matching task and in a cross-modal visual-to-haptic face matching task.

Keywords: generative models; recognition models; visual perception; multisensory perception; analysis-by-synthesis

Introduction

Human perception is unmatched in its generalization capacity. Unlike modern machine perception systems, we acquire novel object categories from one or very few examples, process previously unseen or unusual objects and scenes, and parse complicated scenes into their underlying individual components and perceive their three-dimensional (3D) structure. Humans can imagine novel objects and scenes by composing together previously acquired shape parts, texture parts, and objects (Lake, Salakhutdinov, & Tenenbaum, 2015). Perception is online and remarkably fast – we process a novel scene to know its elements, including shapes, textures, and occluding relationship within hundreds of milliseconds.

Consider the image of a horse in Fig 1. Despite the horse being mostly occluded, your percept of it will be rich: you will know the three-dimensional shape of the horse and its pose. You will also know that the zebra-like patterns are because the cloth has such texture and not because of an actual zebra in the image. Moreover, if you were tempted to lift the cloth to see more of the horse, you could imagine where and how you would grasp the cloth to do so.

The standard machine approaches to perception – vision in particular – are mainly pattern recognition systems that are trained for object categorization or identification from vast image datasets (Yamins et al., 2014; Krizhevsky, Sutskever, & Hinton, 2012). These approaches compute quickly, similar to the speed of online perception, but they don't attempt to explain the variation in the input, and are restricted in their generalization capacity (Fig 1). For example, Krizhevsky's Alexnet (Krizhevsky et al., 2012) that is trained to recognize both horses and zebras incorrectly sees a "zebra" in Fig 1. In fact, for such pattern recognition systems, generalization to novel tasks would require non-trivial amounts of data and that amount would have to scale with the number of novel tasks.

Figure 1: It may be the first time you are seeing a horse under a zebra outfit, but you will have no difficulty knowing, and knowing quickly, that the scene is composed of a horse draped under a cloth.

We postulate that to deeply understand scenes underlying sensory inputs and to generalize strongly, a computational model of perception needs to explain the variation in the inputs as opposed to ignoring it (Dayan, Hinton, Neal, & Zemel, 1995). Our guiding desiderata for a computational model of perception are compositionality, causality, and efficiency (Lake, Ullman, Tenenbaum, & Gershman, in press). Causality refers to the relationship between cause and effect. A perception system should implement forward models of individual sense modalities, such as the optics processes describing how elements of a scene (e.g., shapes, textures, lighting, camera, etc.) give rise to images (Yildirim & Jacobs, 2013). Compositionality refers to combining primitives and parts to obtain new parts and wholes. A perceptual system should allow composing causal elements such as shapes and textures and their parts (e.g., Biederman, 1987). And finally, given an input, a perceptual system should compute its causal and compositional explanation efficiently without extensively iterative algorithms, in line with our online perceptual experience.

Here, we present an efficient analysis-by-synthesis model of visual and cross-modal perception. The latent variables in this model are 3D shape and texture, extrinsic scene variables such as pose and lighting, and occluders. The model can compose 3D shapes and occluders in novel ways. The model consists of causal sensory-specific forward models that can project these latent variables to visual and haptic signals. In order to perform inference in this highly expressive model, we introduce a structured recognition model. The recognition model is trained using samples from a generative model

to map inputs to their underlying latent variables as accurately as possible. Once trained, the model can infer the underlying components and causes efficiently: exclusively through feed-forward computations without any iterative algorithms.

Our computational model builds upon a number of recent studies: (1) analysis-by-synthesis approaches to vision (e.g., Kulkarni, Kohli, Tenenbaum, & Mansinghka, 2015; Egger et al., 2016) and (2) efficient analysis-by-synthesis in which neural networks markedly speed up inference in expressive generative models (e.g., Eslami et al., 2017; Yildirim, Kulkarni, Freiwald, & Tenenbaum, 2015; Moreno, Williams, Nash, & Kohli, 2016). Further, our model goes beyond these lines of works with efficient and occlusion-aware 3D scene reconstruction, applying it to a cross-modal recognition task, and by showing that such a model can account for the detailed gradations of human subjects' responses.

We study two perceptual tasks where perception as explanation should be crucial. First is the task of matching objects across occluded and unoccluded scenes. Second is the task of matching objects across their visual and haptic presentations (cross-modal recognition). In both cases, we show that our model closely predicts human behavior, to a much better extent than all the alternative models considered here.

In this study, we concentrate on faces as our object category and various everyday objects (e.g., window blinds, fences, curtains) as occluders. Using faces has several advantages: (1) we can incorporate existing high quality 3D generative shape and texture models (Blanz & Vetter, 1999), (2) faces present us with a rich space of possible shapes and textures and constitute a behaviorally significant class of stimuli, and (3) the study of faces drew extensive attention across the fields of psychology (e.g., Bruce & Young, 1986), neuroscience (e.g., Freiwald & Tsao, 2010), and computer vision (e.g., Parkhi, Vedaldi, & Zisserman, 2015), and therefore concentrating on faces under our occlusion and cross-modal tasks provides the opportunity to make various contact points with researchers from these fields.

Model

Generative Model

Our generative model of faces is parametrized by face shape S, texture T, pose P, and lighting conditions L. As in (Kulkarni et al., 2015), we use the 3D Morphable Face Model (MFM) as a prior distribution over S and T (Blanz & Vetter, 1999). The MFM provides independent linear models of face shape and texture derived from 200 laser-scanned faces. The model consists of means $\mu_{\{shape,tex\}}$ and covariance matrices $\Sigma_{\{shape,tex\}}$ for both the shape and texture. New faces can be drawn from the model by sampling from a multivariate Gaussian with the MFM parameters: $S \sim \mathcal{N}(\mu_{shape}, \Sigma_{shape})$ and $T \sim \mathcal{N}(\mu_{tex}, \Sigma_{tex})$. The prior distribution over face pose is uniform from -25° to 25° in azimuth.

Given face parameters from the MFM, an approximate rendering engine $g(\cdot)$ provides a projection of the face into image space, $I_{face} = g(\{S,T,P,L\})$. Occlusion of such faces is treated as an image-level operation, in which the pixels of the occluding object mask those of the rendered face: $I_{scene} = I_{occl} \oplus I_{face}$.

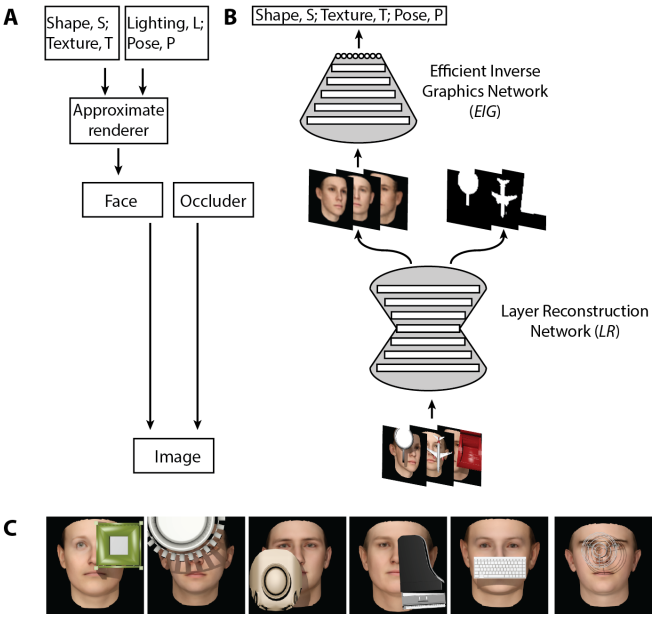

Figure 2: A) Overview of the generative model producing occluded face images. B) Recognition model structure, in which an input image is first decomposed into face and occluder layers and then the unoccluded face reconstruction is used to make latents estimations. C) Samples from the training set.

Recognition Model

Following (Yildirim et al., 2015), we use a recognition model to invert the generative model of face generation. However, in the current work, our scenes take on extra structure, in the form of an occluding object, that makes direct mapping from images to face latent variables difficult. Indeed, we find that the occluders consistently interfere with shape and texture predictions from pretrained neural models, and that such models are not easily adapted to be invariant to a general class of occluders. In order to deal with this issue, we explicitly represent occlusion in a structured manner in our model.

The first stage of our recognition model, shown in Fig 2B, consists of two Layer Reconstruction (*LR*) networks. These networks separate an input image into their underlying independent layers for the face and the occluder mask: $I_{face} = \text{LR}_{face}(I_{scene})$ and $I_{occl} = \text{LR}_{occl}(I_{scene})$. As opposed to predicting I_{occl}, which is essentially an assignment problem where pixels are denoted as "occluder" or "other", predicting I_{face} requires reconstructing portions of the face not visible in the input image. LR_{face} reconstructs the face image at its original pose. Both LR networks are implemented as convolutional autoencoders, with 5 layers for the occluder model and 11 layers for the face model, and trained via stochastic gradient descent.

We use the face layer output from LR_{face} as input to the

second stage of our recognition model, deemed the Efficient Inverse Graphics (*EIG*) network. EIG consists of linear mappings from a generically-trained neural model to the latent variables of our generative model. We found the second fully connected layer of the AlexNet architecture (Krizhevsky et al., 2012) trained on the ImageNet database (Deng et al., 2009) to yield the best performance. By taking as input unoccluded reconstructions of the input image as opposed to the raw occluded scene, we avoid issues with the occluder affecting the predictions of EIG.

Data Generation

A dataset of 40,000 occluded face images was generated using the Blender rendering engine. Accompanying each occluded image was the rendered unoccluded face and a single-channel mask of the occluder locations for training the LR networks.

Faces were produced from random samples from the MFM and occluders were selected from the ShapeNet core dataset (Chang et al., 2015). For each image, an occluder category was selected uniformly at random from the 55 available and then a model was selected uniformly from the chosen category. Our test occluders, for both our models and behavioral stimuli, were not a part of ShapeNet core and were not seen during training.

It is worth noting that Blender's rendering is more comprehensive than the pixel-wise masking process of occlusion in our generative model. Not existing in such simple masking, for example, is ray-tracing, meaning that the Blender-rendered images contain shadows on the face from the occluder that our generative model's outputs lack. This masking approximation allows for occlusion to be encoded in the generative model in a straightforward manner while still retaining the most important attributes of the data generation process.

Experiments

The goal of our behavioral experiments is to test human subjects' performance in matching occluded to unoccluded face images and to test the generative model's ability to predict humans' detailed behavioral patterns. In all experiments, subjects were asked to judge whether a test face is the same as a study face (Fig 5).

Participants

A total number of 178 participants were recruited from Amazon's crowdsourcing service Mechanical Turk. The experiment took about 12 minutes to complete. Participants were paid $1.5. The study was approved by the Massachusetts Institute of Technology IRB. All participants were 18 years or older and provided their informed consent.

Stimuli

Stimuli were 200x200 color images of photo-realistic faces. The view of the faces could be occluded by five types of objects: jail bars, window blinds, a curtain, a fence, and a half-transparent door. Face models were generated using the MFM and occluding object models and their textures were downloaded from individual online model banks. They were then rendered orthographically using ray-tracing in Blender, an open-source computer animation software. Prior to rendering, face models were randomly rotated along the z-axis with angles spanning the range of 25° to 25°.

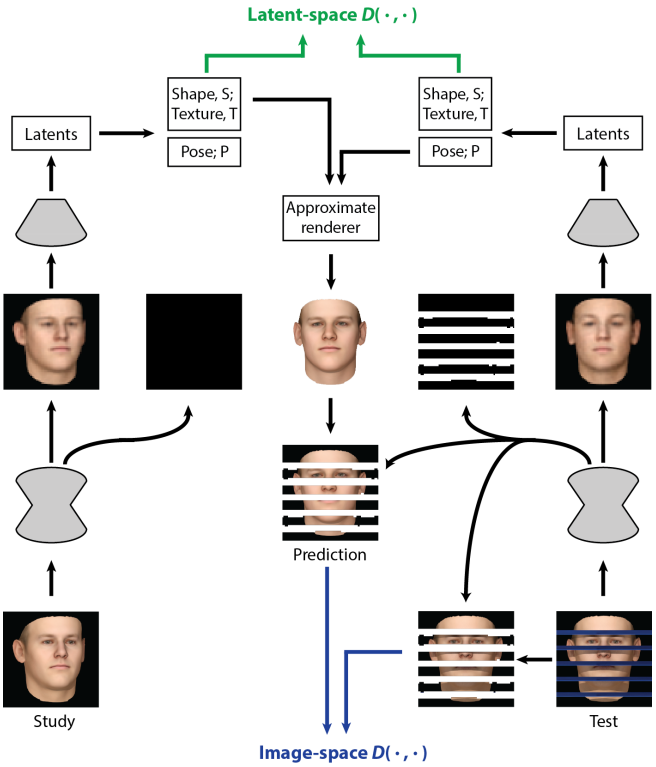

Figure 3: Evaluation pipelines contrasting the full model, resulting in an image-space comparison, to the recognition model alone, which results in a comparison in the latents space of the MFM.

The face images could be occluded at three levels: low (15% of all the face pixels), medium (35% of all the face pixels), and high (55% of all the face pixels). The occlusion levels were obtained by rotating and modifying the base occluder models: by increasing the diameter of the jail bar cylinders, by rotating the window blinds along the x-axis (i.e., closing the window blinds), by translating the curtain along the z-axis (i.e., closing the curtain), by increasing the diameter of the fence mesh, and by rotating the door along the z-axis (i.e., closing the door).

Procedure

Participants were assigned to one of the two groups: either they matched occluded study items to unoccluded test items (Oc $\rightarrow$ Un) or they matched unoccluded study items to occluded test items (Un $\rightarrow$ Oc; Fig 5A). In both groups, the study items were always randomly rotated along the z-axis, but the test item was always frontal.

The study item was presented for 250 msecs, followed by

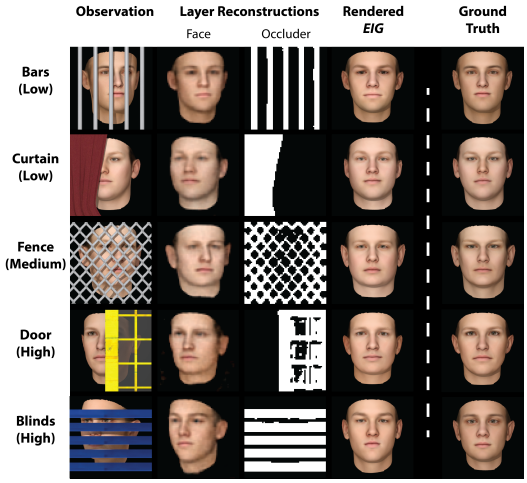

Figure 4: Sample model outputs on the behavioral set occluders. "Layer Reconstructions" are the interior representations of the recognition model. "Rendered *EIG*" shows the result of transforming the latents estimations from the *EIG* network into a textured face mesh and rendering it in frontal orientation.

a mask image (a scrambled face image) for 500 msecs. Then the test item appeared and stayed on the screen until the user entered their response: either an "f" for "same" or a "j" for "different".

There were a total of 96 trials, divided equally between "same" and "different" pairs. The study identities (shape and texture latents in the MFM) were unique to each trial, and trials were randomly shuffled for each participant. Each participant saw each occluder type in 18 to 20 trials, and they saw each level of occlusion equally often (32 trials each).

Model evaluation pipelines and the alternative models

We evaluated the generative model and two alternative models (described below) on the identical stimuli that the participants saw. We obtained a "same" or "different" response from the generative model in the following way (Fig 3). First, using the recognition model on the study item, I, we obtain a point estimate of its shape and texture latent variables, S^*, T^*. We also use the recognition model on the test item, O, to obtain a point estimate of its occlusion mask and pose, M^*, P^*. We then reconstruct S^*, T^*, and P^* using the approximate renderer. We layer M^* on top of the rendered face, giving the model's reconstruction image, R^*. The resultant score is the similarity of reconstruction, R^*, to the test item, O. Finally, we independently run the same pipeline by reversing the study and test items – i.e., by reconstructing the study item from the test item – and obtain the final score as the average of the two directions. We visualize the the interior layer representations, as well as the rendered EIG predictions, in Fig 4.

We considered two alternative models: The first alternative model is obtained by the lesion of the generative components of our model and uses only the recognition model (evaluation pipeline resulting in latent-space comparisons shown in Fig 3). We also consider a second baseline model using the VGG Face network, a state-of-the-art machine face recognition system. The VGG Face network is a particularly deep Convnet, with more than twice as many convolutional layers as our EIG network, trained with millions of labeled images of faces to identity thousands of individuals. This network is evaluated by using the similarity of activations at its second fully connected layer between the study and test items. (We found that there was no other layer in the network that performed better than its second fully connected layer.)

Results

Performance of the subjects in the two groups was strongly above chance. Participants in the Oc→Un group performed with 77%, 73%, and 67% accuracy under low, medium, and high occlusion levels. Participants in the Un→Oc group performed with 78%, 76%, and 70% accuracy under low, medium, and high occlusion levels. We did not find a statistically significant difference between the two groups' performance ($p = 0.08$, two-tailed t-test).

We tested how well each of the models captured the average subject responses, Pr(Same). We compared the models and humans on a trial-by-trial basis for all six conditions (two directions and three levels of occlusions with a total number of 96 trials in each). In order to quantify how well each model accounted for the behavior, we performed a bootsrapping analysis where we sampled individual subjects with replacement 1000 times. For each bootstrap sample, Pr(Same) was obtained for each trial as the proportion of the "same" responses in that sample. We found that our model captured subjects responses consistently better than the alternative models across all six conditions ($p < 0.05$ using direct hypothesis testing with the bootstrap samples for all comparisons between the generative model and each of the alternative models; Fig 5B).

We also analyzed the inter-subject consistency in our behavioral data using bootstrapping analysis. We generated 1000 random equal splits of the subjects into two exclusive sets for each condition. We then correlated the Pr(Same) vector of one set with the Pr(Same) vector of the other set. We found relatively high-levels of inter-subject consistency despite the difficulty of our task (Fig 5B; all values in the range of $r = 0.78$ to $r = 0.85$). We also found that our model's account of the data was close to this effective noise-ceiling level in half of the conditions: Oc→Un and low occlusion, Un→Oc and low occlusion, Un→Oc and medium occlusion. The model seemed to perform worse for, but yet still better than the alternatives, for Oc→Un and high occlusion.

Cross-modal recognition

Multisensory representations and perception systems based on causal generative models are closely related through the multisensory hypothesis, which states that people extract the intrinsic and modality-independent properties of objects and

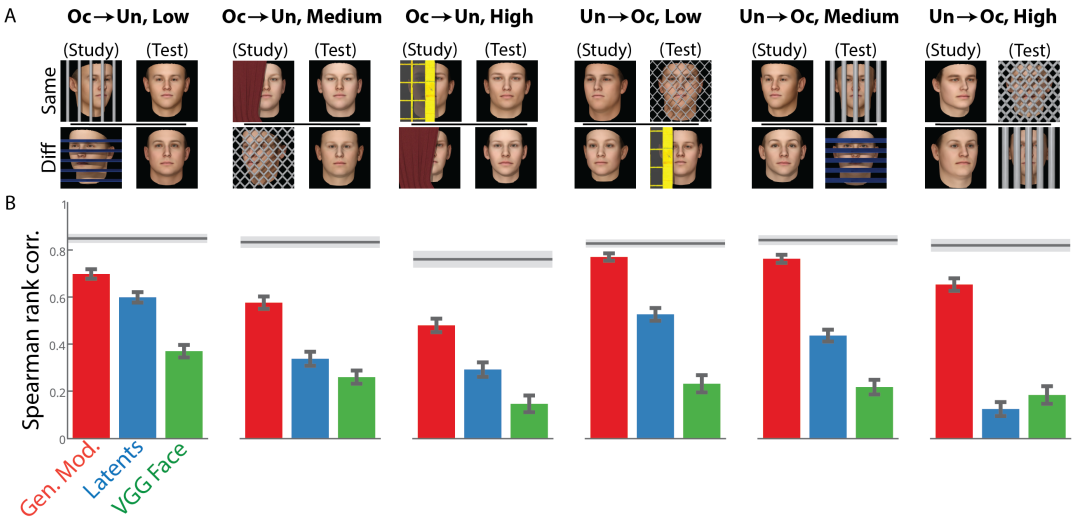

Figure 5: (A) We studied humans' occluded-unoccluded face recognition performance using "same"/"different" judgment tasks (B) Quantified Spearman rank correlations. Error bars indicate standard deviation of the correlations of the bootstrap samples. The horizontal lines indicate the average inter-subject consistency and its standard deviation calculated using bootstrap samples.

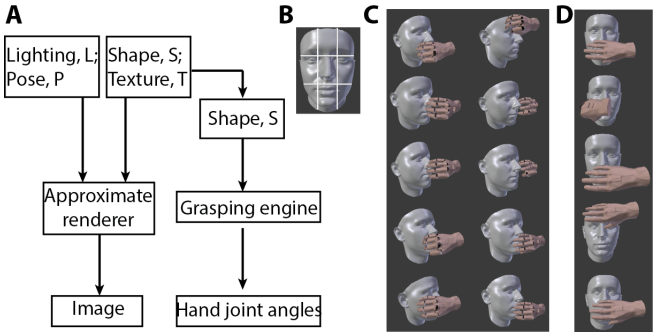

Figure 6: (A) A generative model of multisensory visual-haptic perception. The haptic generative model is a grasp synthesis engine that takes as input a 3D shape and outputs hand joint configurations. (B) In the pre-planning stage of grasp generation, we target each of these 9 cells on the face mesh. (C) Example grasps that passed filtering. (D) Example rejected grasps.

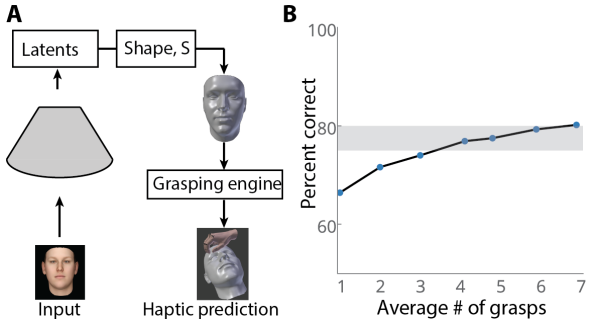

Figure 7: (A) The model evaluation pipeline for visual-to-haptic transfer. (B) Performance increases using more grasps per comparison. Shaded region indicates the approximated behavioral performance from (Dopjans et al., 2009).

events, and represent these properties in multisensory representations (Yildirim, 2014).

We obtained a model of multisensory perception by composing the face generative model (MFM) with a a causal forward model of human-hand grasp generation (Fig 6A). The latter model divides a 3D face mesh into 9 spatial cells (Fig 6B), heuristically evaluates a grasp position and rotation, and interfaces with Graspit!, a grasping engine, to execute an Autograsp (e.g. close the fingers of a hand at the given transformation). Each grasp consists of 16 joint-angles and a 3D mesh of the grasping hand (e.g., Fig 6C). We rejected any grasps that had less than 10 of the joint angles with 0.01 radians away from an open or closed hand; fully open grasps occurred due to the heuristic algorithm generating hand transformations that initially collided with the face (e.g., Fig 6D).

In order to test our model's performance, we simulated a visual-to-haptic recognition experiment. The model was tasked with judging whether an unoccluded face image (study item) and a 3D face mask (test item) were identical or different. The model inferred a point estimate of the 3D shape given the study item using its EIG network. The model then generated grasps on the estimated 3D mesh using its forward model (Fig 7A). The model also generated grasps on the test 3D face mask. That results in 9 grasps for each of the study and test items, but we keep the pairs of grasps that are accepted for both items. Object dis-similarity was computed as the average Euclidean distance between the volumetric properties of each of the accepted grasp pairs. Volumetric properties consisted of the hand grasp bounding box and its center.

On a dataset of 96 pairs (half "same" , half "different"), the model responded "same" to pairs that had a distance less than the total average and "different" otherwise. We found that the model performed 80% of the trials correctly using all accepted grasps (corresponding to an average number of

6.89 grasps per comparison). We inspected the model's performance using different numbers of grasps per comparison. The model's performance was strongly above chance even when using one grasp per comparison, which resulted in an accuracy of 66%. However, increasing the number of grasps did not lead to much higher accuracy indicating the difficulty of the task (Fig 7B).

We found that the performance of our model approximately corresponded to the only behavioral report on cross-modal face recognition (Dopjans et al., 2009). Based on this study, we estimated the performance of the human subjects in the visual-to-haptic condition would be between $75-80\%$ (Fig 7B)[1]. Although they did not record the number of grasps participants performed per face, they reported that the haptic stimulus presentation time was 7 secs, which would indicate 7 to 10 grasps per stimulus under the assumption that a grasp on a face mask takes about 1 sec. If so, this suggests a reasonable level of correspondence between the maximum average number of grasps that our model used, 6.89, and the average number of grasps performed by the participants.

Discussion

We presented a generative model of visual and multisensory perception in which inference leads to *(1)* a causal explanation of sensory inputs (e.g., shape, texture, and layers underlying an image) and *(2)* these resulting scene elements can be composed in novel ways to then imagine new objects. Another way in which our model affords composition is through the integration of multiple generative models that share the same causal space (Grosse, Salakhutdinov, & William, 2012). Modality-invariance, an important perceptual invariance, falls out as a consequence of this kind of composition. We show that, unlike other models, this model accounts for human behavior across a wide range of scenarios – quantitatively in a visual same/different judgment task and qualitatively in a multisensory task.

Our results show that using causal generative models in the forward mode during recognition (a form of imagination) is crucial to account for behavioral performance. Future research is needed to understand how and where such generative models might be implemented in the brain.

Our work presents a neurally-plausible solution to the traditionally difficult problem of inference in rich and structured generative models by using a recognition model consisting of neuron-like units. This recognition model is trained with synthesized samples from the generative model in the style of Helmholtz machines (Dayan et al., 1995). Future work will explore variants of the recognition model where the training can be performed with less supervision from the generative model, such as a weakly supervised scheme in which the model may only know that certain objects don't change.

Future work will also explore a vision-to-touch transfer experiment based on the stimuli and "same"/"different" paradigm presented here.

[1]Although this study used face masks generated using MFM, their experiments differed from our simulated experiment both in terms of the identities and the total number of faces used.

Acknowledgments

This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. A high performance clustering environment for computations (Openmind) was provided by the McGovern Institute for Brain Research.

References

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on computer graphics and interactive techniques* (pp. 187–194).

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305–327.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... others (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Dayan, P., Hinton, G. E., Neal, R. M., & Zemel, R. S. (1995). The helmholtz machine. *Neural computation*, *7*(5), 889–904.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *Cvpr09*.

Dopjans, L., Wallraven, C., & Bulthoff, H. H. (2009). Cross-modal transfer in visual and haptic face recognition. *IEEE Transactions on Haptics*, *2*(4), 236–240.

Egger, B., Schneider, A., Blumer, C., Morel-Forster, A., Schönborn, S., & Vetter, T. (2016). Occlusion-aware 3d morphable face models. In *British machine vision conference (bmvc)*.

Eslami, S., Heess, N., Weber, T., Tassa, Y., Kavukcuoglu, K., & Hinton, G. E. (2017). Attend, infer, repeat: Fast scene understanding with generative models.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.

Grosse, R. B., Salakhutdinov, R., & William, T. (2012). Freeman, and joshua b. tenenbaum." exploiting compositionality to explore a large space of model structures.". In *28th conference on uncertainly in artificial intelligence* (pp. 15–17).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 4390–4399).

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (in press). Building machines that learn and think like people. *Behavioral and Brain Sciences*.

Moreno, P., Williams, C. K., Nash, C., & Kohli, P. (2016). Overcoming occlusion with inverse graphics. In *Computer vision–eccv 2016 workshops* (pp. 170–185).

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. In *British machine vision conference*.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624.

Yildirim, I. (2014). *From perception to conception: Learning multisensory representations*. Unpublished doctoral dissertation, University of Rochester.

Yildirim, I., & Jacobs, R. A. (2013). Transfer of object category knowledge across visual and haptic modalities: Experimental and computational studies. *Cognition*, *126*(2), 135–148.

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling neuronal representations. In *Annual conference of the cognitive science society*.