Research Statement
Ilker Yildirim

My research is driven by a basic puzzle of perception: How can our perceptual experiences be so rich in content, so robust to environmental variation, and so fast to construct from sensory inputs— all at the same time? When we see an object we do not only see it as a kind of thing—a book, a bookcase, a table, a teacup, a car, or a tree. From a quick glance, the touch of an object, or a brief sound snippet, our brains map raw sensory signals into rich mental representations of physical scenes, shapes, surfaces, and spatial layouts. I study visual and multisensory perception to understand both the brain’s representations of physical objects and the algorithms used to construct them. I study these representations via novel tasks such as recognizing visually presented objects by touch or under heavy occlusion, recognizing objects under unusual viewing conditions, estimating physical object properties (e.g., weight and friction) from how objects interact in a dynamic scene, and predicting the outcomes of those interactions. My work weaves together computational, behavioral, and neural approaches to build a unified understanding of higher-level perception and its links with cognition.

My computational work embraces an old approach to studying how sensory inputs give rise to rich mental representations—the approach of analysis-by-synthesis. In this approach perception is viewed as the agreement of the physical sensory inputs and their computational re-construction or re-creation from abstract hypotheses. Despite its long history and unique appeal, this approach lagged behind other approaches to perception. My work modernizes and efficiently re-instantiates analysis-by-synthesis by drawing on methods from multiple paradigms in psychology, neuroscience and artificial intelligence. These methods include probabilistic programs, deep neural networks, and forward models such as video game engines. Together these methods provide not only a unified account of perception in neural and cognitive terms, but also have applications in building more human-like machine and robotic perceptual systems.

The goal of my behavioral work is to study mental representations by tasks that require combining very limited sensory inputs with internal models of the world to make inferences about unobservable physical properties. I approach this goal with novel experimental paradigms that rely on advanced simulation software and 3D printing technology for visual, non-visual, and multisensory stimulation. For example, in one experiment I showed subjects life-like images of an unknown object (e.g., a chair) draped in a cloth where they had to make inferences about the hidden object; in another experiment, I had them feel an unfamiliar 3D-printed object and match it to images. These paradigms show that people can infer rich physical object properties from very limited sensory inputs, and provide the data to test specific models.

The goal of my computational neuroscience work is to establish links between the neural activity and physical object representations. Because internal models of the physical world can be built from such representations, my work is well positioned to uncover the “mental life of neurons” or to relate neural activity to cognition. I have developed multiple collaborations with primate electrophysiology researchers, which involve a combination of analyzing data they have already collected, jointly planning new experiments to test the models I build, and using the results of those experiments to inspire new model designs. I use these and alternative models to make inferences about the neural activity obtained during tasks that require representing objects such as vision-based planning to grasp objects.

In the following sections I describe several of these findings progressing from strictly visual object representations through multisensory representations up to the richest but most abstract forms of physical object representations. At the end of each section, I outline my future research plans.

1. Visual object representations

Vision is often well characterized as the solution to an inverse problem: inverting the process of how three-dimensional (3D) scenes give rise to images to reconstruct the scene that best explains a given observed image. Yet the neural circuits underlying this inverse mapping is poorly understood. Leading neural models of vision concentrate on the task of object identification or categorization, and do not attempt to explain how vision also constructs rich descriptions of scenes with detailed 3D shape and surface appearance. Alternative inverse graphics or analysis-by-synthesis approaches aim to recover 3D scene content, but are typically implemented using iterative computations that are hard to map onto neural circuits and much too slow for the speed of natural perception. It remains a mystery how vision constructs such rich scene descriptions of 3D shape and texture, so quickly—on the order of a hundred milliseconds.
To address this challenge, I developed a computational model that recovers 3D object shape and texture from a single image similar to conventional analysis-by-synthesis approaches, but does so quickly using only feedforward computations without the need for iterative algorithms. The model combines a probabilistic generative model based on a 3D graphics program for image synthesis with a recognition model based on a deep convolutional neural network \cite{5, 4}. As a test case, we apply this model in the domain of face recognition where, as a rarity, data from brain imaging, single-cell recordings, and behavior all come together to constrain models. However, the scheme of the model is more general than faces and can apply to any object class.

We found that this approach provides the best account of a high-level function in the brain (face recognition), spanning neural activity and behavior. We found that our model ("Latents") accounted well for the qualitative patterns across populations of neurons in the monkey IT cortex (Figure 1B and C) equally well if not better than an alternative state-of-the-art machine vision system ("VGG Face"; \cite{28}) that is trained using millions of images of faces to discriminate thousands of identities. We also found that in challenging face recognition tasks (Figure 1D), only our model performed on par with human subjects. Two alternative discriminative models (VGG Face network and the "Identity" model that is similar to VGG Face model but is trained with images rendered by the generative model) performed worse than human performance in most case. Moreover, we found that our model –on a trial-by-trial bases– predicted human performance consistently well across all experiments, again unlike the alternative models. Our results provide a new way to think about perception as explanation \cite{17}, and show how visual processing can naturally support a broad range of cognitive operations such as causal inference, imagination, communication and action planning in addition to object recognition.

In my future research within the next few years, I plan to further explore rich 3D object representations in the mind and brain by developing computational models and novel experimental setups as described below.
Figure 2: (A) A compositional space of 3D object shapes. Basic parts can be combined in infinitely different ways to obtain new objects. (B) Schematics of the visual forward model and haptic forward model. The visual forward model projects a 3D model to images using various viewing angles. A haptic forward model outputs grasps of a 3D model (e.g., joint angles of the hand at the time of a stable grasp of the model). (C) Humans’ cross-modal object categorization performance. Left panel shows their performance at the end of the final training block. Right panel shows their cross-model test performance. (D) Our proposed computational pipeline for cross-modal face recognition. (E) Example simulated grasps of everyday objects. Our system can generate plausible grasps using a large 3D model bank including more than 50000 shapes. This simulated dataset can spur development of recognition models for the haptic sense.

Future Research 1: Familiar and unfamiliar object recognition. The empirical literature on object recognition but in particular on face processing documents a striking distinction between the processing of familiar and unfamiliar faces [18]. However, the computational foundation of this difference is significantly under-explored. In collaboration with Ken Nakayama (Harvard) and Josh Tenenbaum (MIT), we are developing a unified account of the familiar and unfamiliar object recognition, and in particular of face recognition, both computationally and psychophysically [9, 11]. Our computational framework brings together the efficient analysis-by-synthesis idea (Figure 1A) with non-parametric Bayesian methods in order to implement a flexible and growing memory of familiar items. In parallel to this computational efforts, we are designing behavioral experiments that will control a subject’s familiarity with a given object identity, and will measure their recognition abilities afterwards as a function of their level of familiarity.

2. Multisensory object representations

Despite the different sensory modalities through which we perceive objects, we experience a coherent and unified physical world. This observation indicates an important perceptual invariance—namely, modality-invariance. I developed a computational framework of object recognition that integrates sensory inputs across modalities into unified representations and produces modality-invariance as a by product [10, 3, 2, 6, 8]. This framework has three components. First, it constructs a compositional space of objects or events such as 3D shape or spatial sequential patterns (e.g., localized visual flashes or brief sounds moving on a circle). Figure 2A shows image of an example 3D object and how it is composed from subparts and primitive parts. Second, it projects such modality-independent objects or events to sensory domains using sensory-specific forward models. For example, it projects 3D shapes to images using a
graphics engine; it projects 3D shapes to touch using a grasping simulator (Figure 2B). Third, in this framework, perception is implemented as Bayesian inference whereby input sensory signals are mapped onto the unobserved modality-independent representations. In my work, I implemented this framework for various computational tasks such as 3D reconstruction from single images, cross-modal (i.e., visual-haptic) recognition of object shape categories [2, 8], and modality-independent reasoning about spatial sequential patterns [6].

We behaviorally tested and quantitatively confirmed several of the predictions that these models make: (i), if a person learns to categorize objects based on inputs from one sensory modality (e.g., vision), the person can categorize these same or similar objects when objects are perceived through another modality (e.g., touch) (Figure 2C) [2]; (ii) if a person learns to categorize spatial sequential patterns based on inputs from one sensory modality (e.g., audition), the person can categorize these same or similar patterns when they are observed through another modality (e.g., vision) [6]; (iii) a person’s similarity judgments of pairs of unfamiliar objects match within (e.g., the person sees both objects) and across different sensory modalities (i.e., the person sees one object and feels the other object) [8].

This work so far explored novel and fairly simple categories of objects and patterns. Within the next few years, I plan to extend this framework to settings including more complex and everyday shapes.

**Future Research 2: Cross-modal recognition of complex object shapes.** I plan to extend my multisensory object recognition framework to more complex shapes (Figure 2E), and in particular to faces. In collaboration with Christian Wallraven (Korea University), we are designing a cross-modal recognition experiment where human participants will see an image of a face before feeling a 3D printed face sculpture and judging whether the two faces belong to the same individual or to two different people. In parallel, I am extending my efficient analysis-by-synthesis model described in §1 such that the model will project its 3D shape inferences into haptic modality (Figure 2D). This project will be the first to quantitatively model human performance in a cross-modal complex object recognition task. The combination of the computational and behavioral methods in this project will be a stronger test of whether a unified representation of shape shared between vision and touch suffices to explain human behavior than our and others’ previous works.

**Future Research 3: Computational study of the object representations in the anterior intraparietal cortex (AIP).** What object-level transformations does the brain perform for visually guiding grasping? Within the next few years, I plan to infer a detailed picture of the object representations in AIP, which is the brain region that is implicated in vision-based grasp planning. Three reasons make AIP of particular interest: (i) AIP visually processes 3D shape [33]; (ii) AIP participates in visual and tactile integration [34]; and (iii) AIP participates in visually-guided grasping [31]. I established collaboration with Winrich Freiwald (The Rockefeller University) and members of his lab in order to guide the development and evaluation of the models using their existing neural data.

I will build computational models inspired by the efficient analysis-by-synthesis framework. Each model will embody a hypothesis about population-level representations in AIP. Two discrete dimensions will specify all models. First, the ultimate task of a model can be one of the following four: (i) object identification (or categorization), (ii) 3D shape recovery, (iii) grasp type identification, (iv) pre-shaping the hand for grasping. Second, the inputs to a model could be unsensory or multisensory: (i) vision-only inputs (i.e., images), (ii) visual-haptic inputs (i.e., both the images and simulated grasps of objects).

### 3. Physical object representations

Humans demonstrate remarkable abilities to predict physical events in dynamic scenes, and to infer the physical properties of objects from static images or dynamic videos [20, 24, 26]. As [19] noted, we intuitively know how objects rest on and support each other, what would happen if they collide or fall, and how much force would be required to move them. These observations raise a central question: how do visual inputs lead to rich physical scene understanding? I developed a computational model of physical scene understanding that takes as input real-world images or videos and infers physical properties of objects such as their mass, density and friction. Based on these inferences, the model makes predictions about the
Figure 3: (A) Frames from a representative subset of videos that our model and human subjects saw. (B) Integrated physical scene understanding model takes as input real world videos or images and interprets them using a convolutional neural network and a game engine that approximates expected interactions between physical objects. For a future project, we hypothesize that the convolutional neural network which serves as a vision pipeline represents the computations taking place in the visual cortex, and the simulation engine which serves as an approximation to intuitive physics knowledge represents computations taking place in the parts of the parietal cortex and the pre-motor cortex. (C) Simulation-based reconstructions of three videos by our model. Note that the model has to infer the physical properties of the objects in the scene. (D) A behavioral task that requires visual and physical understanding of the image. (E) Behavioral and computational results from a match-to-sample task where subjects had to match images of unoccluded objects to fully occluded scenes where an object is draped in cloth.

The future of the unfolding physical events in the scene [1, 13]. Frames from example real-world scenes that our model can process are shown in Figure 3A.

Figure 3B illustrates the schematic of our model. One of the two core elements of our model is a complex forward model, namely a 3D video game engine [25], operating on an object-based representation of physical properties, including mass, density, spatial position, 3D shape, and friction, and the interactions of these objects estimated using game engine physics. The other core element of our model is a deep neural network that maps visual inputs (e.g., images) to their physical object representations such as their shape and mass. This neural network is pre-trained on a large image corpus. We finetune it using MCMC-based inference results from the generative model. Note that this training procedure is a form of self-supervised or semi-supervised learning where the data required for training is obtained using the model itself, thus avoiding any costly data collection step (e.g., images of objects labeled by their weights). After training, the neural network and the generative model are combined into an integrated physical scene understanding model: the neural network serves as a recognition model – it is a feedforward vision pipeline that maps images or videos to parameters in the generative model such as mass and friction. The generative model
predicts feature outcomes for given inputs, generalizes across different scenarios, and can be used for causal and counterfactual inferences. Figure 3C shows frames from three input videos and frames from the corresponding simulation-based inference results. In behavioral experiments, we showed that both the model and humans can accurately predict the outcomes of object collisions and the relative mass of the objects.

This work opens up exciting future directions interfacing cognitive science, neuroscience and artificial intelligence.

**Future Research 4: Physical object representations and visual object recognition.** Does intuitive physical knowledge guide visual object recognition? Despite its centrality for everyday perception and cognition, the interaction between intuitive physical reasoning and visual object recognition is an under-explored research area. Consider a scenario where an object is draped in a cloth and so completely hidden. Can we recognize what that hidden object might be? In certain settings, this judgment is intuitive to us: for example, in Figure 3D, it is easy to tell which of the two objects might be a chair. In collaboration with Josh Tenenbaum (MIT), I am studying such intuitive judgments behaviorally and computationally [12]. Preliminary behavioral and computational results indicate that human participants successfully perform this difficult task, and that a generative model of perception can account for this behavior (Figure 3E).

**Future Research 5: Computational study of neural circuits responsible for intuitive physical inference.** Although the neural substrate supporting physical inferences such as estimating the velocity of an object or the weight of an object is studied, how the brain represents interactions of multiple physical objects is severely under-explored. Inspired by the work of Fischer et al. [19] and in collaboration with Winrich Freiwald (The Rockefeller University) and Josh Tenenbaum (MIT), we are planning behavioral experiments with monkeys that will leverage their physical intuitions. For example in one planned experiment, monkeys will “catch” a ball amongst multiple balls that are colliding with each other and with walls present in the scene. Single-cell recordings will be performed during these experiments from several pre-determined brain regions including parietal and pre-motor areas. Within the next few years, this joint work gives me the unique opportunity to computationally characterize the neural activity responsible for intuitive physical inferences in the brain.

**References**


