

Modeling Physical Perception in Virtual Interactions

Elyse D. Z. Chase, *Member, IEEE* and Marcia K. O'Malley, *Fellow, IEEE*

Abstract—Humans can interact effectively with complicated environments, seamlessly taking actions to learn about the objects around them and build individual cognitive world models. If robots of the future are to easily collaborate with humans on tasks in a range of dynamic environments, those robots must be able to learn from human interaction and understand personalized mental models in near real-time. These interactions are inherently multisensory, leading to layers of complexity. As a step towards understanding multisensory human mental models from interactions, we gathered pilot data from interactions and probed density judgments in virtual reality with pseudohaptic illusions. We then implemented a particle filtering workflow to estimate each individual's mental model. Future work could expand this to consider more sensory information in different tasks.

I. INTRODUCTION

People's perceptions and decisions arise from the dynamic interplay of sensory channels, including vision, audition, touch, and proprioception. People continually integrate streams of uncertain, partially redundant signals, applying prior knowledge about object properties, physics, and their own actions to form a mental model of the world. Capturing that complexity is critical if we hope to build robots that can easily collaborate with humans in the physical world, anticipate their intentions, and adapt to the nuanced ways people explore and learn about their environment.

Virtual Reality (VR) is an effective way to probe these multisensory learning processes. Unlike purely physical setups, VR allows us to systematically manipulate both the *true* physical properties of objects (e.g., mass, inertia) and the sensory cues that participants receive (e.g., vision, haptics), holding other factors constant. It also provides a platform to record and track all user interactions, such as grasping, lifting, squeezing, and dropping, temporally reflecting how people take action to learn about their environment.

This work introduces a Bayesian estimation framework for modeling human multisensory cognition of object density in VR. We ran a pilot study ($n = 4$) where people interacted with three different objects. In addition to considering people's interactions with each object, we also periodically elicit participants' judgments and confidence ratings of a hidden property of the object: density. In the model, a population of particles represents each user's evolving belief about object properties, each encoding a possible ordering of objects by

This work was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at Rice University administered by Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence (ODNI).

Chase & O'Malley are with the Department of Mechanical Engineering at Rice University, Houston, TX, USA, 77006. ec100@rice.edu

their physical attribute of interest. As participants interact with the objects, we update the particle weights via a tailored transition and observation model, incorporating observed and self-reported feedback.

After fitting user-specific parameters in an initial training round, we test the model's predictive accuracy in a second round of interactions. We demonstrate how our approach captures individual differences in multisensory learning by comparing the model's inferred orderings against participants' actual responses. Our work is a simple but important step towards adaptive robotic systems that can understand and respond to human mental models in real-time.

II. RELATED WORK

A. Haptic & Virtual Interactions

We use our hands and sense of touch to perceive and learn. Researchers have identified that we perform exploratory procedures with our hands that are directly tied to the information we want to gather [1]. Later work compared touch with and without vision, and found that haptic information tells us more about material properties than shape.

In virtual interactions, haptic feedback is usually absent without additional hardware. However, pseudohaptic illusions allow designers of virtual environments to provide the sense of touch with only visual manipulations. One such illusion is the Control to Display (C/D) ratio, in which the mapping between the user's and display's motions is modified with a ratio α where 1 is precisely the user's input, $\alpha < 1$ slows the user's input, and $\alpha > 1$ speeds the user's input. This illusion works because our proprioceptive sense (the body's sense of position and movement in space) is not as accurate without visual feedback. In VR, multiple studies have found that the C/D ratio influences people's perception of mass [2], [3], with values $\alpha < 1$ increasing mass and $\alpha > 1$ decreasing mass perception. Here, we use C/D ratios to give signals about mass.

B. Density Perception & Modeling

Density is the ratio of mass to volume. While we can assess shape with vision alone, humans are not skilled at estimating volume. Some say we have *elongation bias*, wherein our perception depends on the object's longest linear dimension [4], [5]. A study comparing a tetrahedron, cube, and sphere found significant biases, seemingly due to people incorrectly using the surface area as a proxy for volume [6]. Alternatively, mass, which is difficult to assess visually, can be evaluated haptically through statically holding objects [1].

Our experience interacting with objects creates forces and resistances, which combine with our expectations from



Fig. 1. A participant is seated in an empty study space wearing a Quest 3 HMD. Virtual items and question panels are shown semi-transparently. Participants see three virtual objects and a question panel asking them to rate the relative density of the objects and their confidence in each ranking.

planning before taking that action, to produce our overall perception [7]. Some researchers have developed models of how people interpret relative mass from two objects hitting each other. Cohen & Ross used probability distributions and Markov Chain Monte Carlo to model participant responses for such simple mass ratios [8]. More work is needed to fully understand how humans process these complex signals and how to model their perceived responses.

III. METHODS

We conducted a pilot study in VR where participants interacted with three objects at a time and ranked them by perceived density. Additionally, we built a model to determine each participant’s belief flexibility and sensory noise, which we then used to predict responses. Below, we introduce the necessary software, explain the experimental parameters, and describe the cognitive model.

A. Study Design

To control and record the factors involved in this work, we ran the study in VR with a Meta Quest Oculus 3 head-mounted display and paired controllers (Figure 1). We built the interaction with the Unity Game Engine (Version 2022.3.15f1) and the Meta Interaction SDK.

The study focuses on interaction methods and how that might allow future systems to intuit the human’s mental model of a series of objects through their interactions alone. We wanted a task that required both vision and touch (or proprioception). Volume alone can be judged visually (with varying accuracy), and mass requires interpreting weight from our sense of touch. Thus, density is a more complex material property that combines information across the senses.

As highlighted by prior work, people are not ideal at estimating volumes of different shapes, and various illusions (e.g., size-weight, color-size) can affect those estimations further [9]. To reduce those factors, we selected three objects with equal volume presented with the same untextured, matte material and color, to avoid any cues about thickness or material properties. Participants interact with a cube, sphere, and cylinder in the study. We selected these three shapes as

they are difficult to judge the volume of one compared to the other, so we forced participants to make a qualified mass judgement.

People interacted with two sets of the three objects in two rounds, with the volume randomly assigned to one of two values ($v = 43, 91 \text{ cm}^3$; corresponding to a cube with side lengths of 3.5 and 4.5 cm). We also randomized the objects’ relative locations between rounds.

However, shape alone should not inform participants about density. So, we manipulate the mass of each object ($m = 150, 300 \text{ g}$) – again randomized between the two rounds. Additionally, we introduce one pseudo-haptic illusion – alteration of the C/D ratio during volitional movements of the user holding on to different objects. We chose two different C/D ratios ($cd = 0.115, 0.125$). For each C/D ratio, objects have three possible values: no alteration (1) and two options near perceptual detection ($1 + cd$ & $1 - cd$) in either direction from no modulation.

We gave participants a fixed, six-second exploration window [1] which began upon the first interaction in each trial. After the exploration window, we provided a randomized reflective delay (1.5 - 2.5 s) before the question panel would appear. This short delay should allow us to capture an “online” belief without memory decay or strategic thinking.

B. Cognitive Modeling

Our goal is to capture each participant’s evolving belief about the object densities. We assume a simple and intuitive model with two user-specific parameters: ϕ (belief flexibility) and σ_m (sensory noise). Since the primary feedback from each user is a relative ranking of the objects and their confidence in each position, we represent the belief as a particle filter over the discrete space of all object ranking permutations [10]. We use the parameters from the first round of trials to predict out-of-sample behavior in the second round. Finally, we compare these predictions with the user’s self-reported ranking.

1) *User-Specific Parameters*: The first parameter, ϕ , represents how likely the user is to change their belief:

High ϕ : many swaps, user is flexible.

Low ϕ : few swaps, user is conservative in updating.

The second parameter, σ_m , quantifies the assumed noise in motion feedback:

Low σ_m : feedback treated as precise.

High σ_m : feedback treated as noisy.

2) *Particle Filtering for Bayesian Belief Estimation*: Particle filtering is a sequential Monte Carlo method for approximating the posterior distribution over latent states in a Bayesian state-space model. At each time step t , the filter maintains a set of N particles $f_{\pi_t^{(i)}}, w_t^{(i)} g_{i=1}^N$ that together approximate the belief

$$p(\pi_t | \text{data}_{1:t}) \propto \prod_{i=1}^N w_t^{(i)} \delta(\pi_t - \pi_t^{(i)}),$$

where π_t is the hidden state and $w_t^{(i)}$ are normalized weights. Each particle $\pi_t^{(i)}$ encodes one hypothesis about the true

state, and the population of particles defines a nonparametric representation of the full posterior.

Filtering proceeds in two alternating steps:

Prediction (Propagation) Each particle is *propagated* forward through the transition model $\pi_t^{(i)} \sim p(\pi_t^{(i)} | \pi_{t-1}^{(i)})$, thereby forming a prior over the new state.

Correction (Weighting & Resampling) Upon observing new data o_t , each particle receives a weight proportional to the likelihood $p(o_t | \pi_t^{(i)})$. The weights are then normalized, and particles are *resampled* with replacement in proportion to their weights, yielding a new equally weighted set that focuses computational effort on high-probability regions.

The hidden state is a permutation (ordering) of the three objects by density. Each particle is one candidate ordering, and the weight update incorporates both:

- 1) a *sensory likelihood* comparing the observed motion cue o_t to the expectation under that ordering, and
- 2) a *ranking likelihood* measuring agreement between the particle’s ordering and the user’s reported ranking r_t .

After resampling, the cloud of particles represents the user’s posterior belief, from which we read out the most probable ordering and its implied confidence.

3) *Estimate–Predict Workflow*: Our modeling proceeds in two distinct phases for each participant: (1) *parameter estimation* using Round 1 data, and (2) *open-loop prediction* on Round 2 data. Holding other parameters fixed, we fit the two free parameters, the transition flexibility ϕ and the sensory-noise standard deviation σ_m .

Round 1: Parameter Estimation

We jointly estimate ϕ and σ_m by maximizing the total log-likelihood under our particle-filter model. On each trial t we know:

- the stimulus parameters f, m_t, v_t, cd_t, g , yielding an expected motion cue $\mu_t = \text{sort}(cd_t/m_t)$ (descending),
- the observed motion feedback o_t ,
- the user’s reported ranking r_t and confidence c_t .

We maintain N particles $f\pi^{(i)}g$ over permutations, initialized uniformly, and propagate each via a transition kernel:

$$\pi_t^{(i)} \sim \text{T}(\pi_{t-1}^{(i)}, \phi) \quad \text{with} \quad \text{T}(\pi^0, \phi) \propto e^{-d_K(\cdot; \cdot)}.$$

Each particle is then re-weighted by the product of three likelihood terms:

$$w_t^i \propto \underbrace{N \frac{o_t; \mu_t^{(i)}, \sigma_m^2}{\{Z\}}}_{\text{sensory}} \underbrace{\exp \left(-\frac{d_K(r_t; \pi_t^{(i)})}{\tau_r} \right)}_{\text{ranking}} \underbrace{N \frac{c_{t,r}; \hat{c}_{t,r}, \sigma_c^2}{\{Z\}}}_{\text{confidence}}.$$

where τ_r and σ_c are hyperparameters used to calibrate the model to user feedback. $\hat{c}_{t,r}$ is the percentage of particles that agree on the modal object at rank r . Summing and logging these weights across all T trials yields a scalar objective $L(\phi, \sigma_m)$, which we minimize via L-BFGS, initialized with

method-of-moments initial values. This fit uses *all* user feedback (sensory, ranking, and confidence) to identify both how precise the user’s motion perception is (σ_m) and how readily they revise their ranking beliefs (ϕ).

Round 2: Open-Loop Prediction With $\hat{\phi}, \hat{\sigma}_m$ fixed, we make the model predict Round 2 behavior *open-loop*, i.e., *without using any of the user’s new rankings or confidences*. We initialize the same particle set and, on each new trial t , perform only:

- 1) **Expectation**: compute μ_t from the new (m_t, v_t, cd_t) .
- 2) **Proposal**: draw $\pi_t^{(i)} \sim \text{T}(\pi_{t-1}^{(i)}, \hat{\phi})$.
- 3) **Sensory update**: re-weight each particle by $N(o_t; \mu_t^{(i)}, \hat{\sigma}_m^2)$ only.
- 4) **Normalize & resample**, then extract $\hat{r}_t = \text{mode} f\pi_t^{(i)}g$ as the predicted ranking.

Because the filter gets no feedback in Round 2, the model must *generalize* its Round 1 parameters to novel stimuli.

IV. PILOT STUDY

Four participants ($\mu_{age} = 24$, 2 female) completed the pilot study in alignment with our protocol, which was approved by the Rice University Institutional Review Board (IRB-FY2019-49). Over 15 minutes, participants interacted with three objects in VR in two rounds.

We randomized each participant’s object location, mass, volume, and C/D ratio between rounds. Participants were instructed to pick up objects with their right hand/controller and judge density. They could interact with objects in any order and in any desired way.

Participants freely explored the objects in each 6-s trial. They completed 10 distinct interactions and response trials for each round. After each trial, there was a randomized reflective delay before the questions appeared. The question panel (Figure 1) had drop-down menus to rank the shapes from most to least dense. Then, participants used sliders to determine their confidence in each ranking.

During all trials, we recorded rankings, confidence, response times, and which objects they interacted with, for how long, and how much they manipulated them.

V. RESULTS & DISCUSSION

In this pilot study, we have two dependent variables (Figure 2): relative density ranking and confidence for each ranking. There are also several independent variables, including mass, volume, and C/D ratio. We also recorded the real and *proxy* (virtual – with C/D adjustments) hand transformations and rotations during the study.

We implemented our model in Python 3.8 on a Google Colab CPU runtime, using SciPy’s L-BFGS-B optimizer to fit the two free parameters, ϕ and σ_m , to Round 1 data. We held the particle-filter size at $N = 500$ and performed a small grid search over $\tau_r \in \{0.1, 1, 10\}g$ and $\sigma_c \in \{5, 10, 50\}g$. Reported results use the hyperparameter settings that maximized Round 2 top-1 accuracy; per-participant runtimes (20 trials) averaged under 30-s.

We evaluate predictive performance in Round 2 using two metrics. **Top-1 accuracy** is the fraction of trials where the

