Tracking Uncertainty During Uncertain Tracking

Anonymous

Abstract

Multiple object tracking is often studied in settings where objects are largely observable. However, tracking often occurs in settings with much greater uncertainty. Objects can frequently go in and out of view, requiring us to constantly update our estimates of where things might be, and assess whether or not something new has appeared. To accomplish this, people need to rely on top-down inferences to fill in the gaps of uncertainty. To study this phenomenon, we introduce a novel "firefly tracking" paradigm, in which people need to estimate the quantity and dynamics of an unknown objects under highly sparse observations. We model human behavior on this task and demonstrate how probabilistic inference in a generative model captures human uncertainty during challenging tracking tasks.

Keywords: Multiple Object Tracking, Probabilistic Inference, Mental Correspondence

A common maxim in cognitive science claims that "objects do not wink in and out of existence". In the visual world, however, objects disappear for uncertain amounts of time, occluded by other scene elements or passing out of view, and they may return after an unknown duration. Often, an object reappears with a significantly different appearance, yet perceptual systems must recognize and successfully re-identify the object in order to construct a coherent scene percept.

It is useful to cleave this problem into parts. At a minimum, lighting and visual noise will alter the raw pixels which describe an object, but more significant appearance changes due to pose and partial occlusion are ubiquitous in any typical video or natural environment. Tolerance to changes in *appearance* are often called "invariances", and they form the basis of standard approaches to understanding visual perception in humans and designing it in machines. The idea is that by learning all the ways an object might appear, its identity can be recognized despite confounding visual factors. While appearance invariance is an important ingredient in a successful perceptual system, this focus diminishes the problem of *structural* changes in a scene. In dynamic settings, objects disappear completely due to (full) occlusion, moving out of view, etc., and the problem is how to incorporate a new observations of an object – either to re-identify it or to decide that it is a new scene element. Especially in computational work, it is difficult to formulate and solve this problem, despite its ubiquity in everyday life. Invariance to appearance provides part of the solution to this problem, but it is not enough: assumptions about the nature of the physical and visual world and their relationship seem to be required.

To make this point more salient, imagine you're sitting at home and you hear the unmistakable buzz of a fly. You look up and observe it whizzing around the room for several seconds, before losing it in a shaded area. A few quiet but tense minutes later, you wander into the kitchen, only to once again encounter a fly buzzing around. At this moment, your brain engages in a critical evaluation of whether this is the same fly that you saw before, or whether you're in a more drastic fly predicament than you'd hoped.

In order to carry out this correspondence computation, you must evaluate several hypotheses. For example, if there were two or more flies in your house — given how chaotic a fly's dynamics can be — maybe you would expect to observe a fly more often, even if you only see one at a time. You can also speculate about the paths it must have taken (i.e. you think it probably didn't come back into this room, but it also couldn't have gone too far before coming back to the kitchen). This kind of reasoning is something we engage in all the time, and reflects a critical intersection of top-down and bottom-up perception. To make sense of what we see, we need to actively integrate what we know about the world around us. Crucially, in these scenarios, you're rarely 100% sure — but you can make an approximate, probabilistic estimate.

In order to study this phenomenon, we introduce a novel "Firefly Tracking" task. At its core, this task is similar to many Multiple Object Tracking (MOT) paradigms. Several indistinguishable circles move around a screen with some dynamics, and participants need to track the objects. However, unlike standard MOT tasks, participants don't initially know how many objects are in the scene. Fireflies can blink on and off with different blinking rates, and can move with different dynamics. We then measure a participant's inferences and uncertainty over time about properties such as the number of fireflies present in the scene, or their approximate positions. We demonstrate how approximate probabilistic inference can capture a range of perceptual behavior, and model the dynamics of correspondence decisions people make while engaging in this online tracking task.

Related Works

Multiple Object Tracking is a classic paradigm in psychology, and a popular task in the field of computer vision.

Psychologists have used MOT tasks for decades in order to study the limits of visual attention, working memory, and the underlying nature of object representations (Pylyshyn & Storm, 1988; Balaban, Smith, Tenenbaum, & Ullman, 2024; vanMarle & Scholl, 2003). In a standard MOT task in psychology, the participant is presented with several identical objects on a screen (typically generic shapes, like circles or squares). At the start of each trial, one of those items will be singled out as the target to track. All the objects will then start to move around, and after several seconds of motion the participant needs to indicate which item corresponds to the target object. These studies often focus on characterizing cognitive limitations through statistical or qualitative descriptions, such as quantifying the number of swaps (mixing up a target and distractor) or drops (where a participant loses the target entirely) (Drew, Horowitz, & Vogel, 2013).

The prominence of MOT in visual cognition has inspired a number of computational models. For example, one line of prior work describes a resource rational, probabilistic model to account for the patterns of errors people exhibit as you increase the number of objects they need to track (Vul, Alvarez, Tenenbaum, & Black, 2009). Yet, despite the many variations of this task people have explored, there are still a number of basic questions and disagreements that remain about the mechanisms and limits of our ability to track multiple objects simultaneously (Holcombe, 2023).

One salient difference between our paradigm and classic MOT tasks is that objects are not consistently visible. While a number of studies have specifically looked at how humans track invisible or occluded objects, these experiments typically use short periods of occlusion or invisibility, and subjects have initial knowledge of the number of objects on screen (Horowitz, Birnkrant, Fencsik, Tran, & Wolfe, 2006; Franconeri, Pylyshyn, & Scholl, 2012; Teichmann, Moerel, Rich, & Baker, 2022). In our work, participants are asked to make a joint inference over the number of objects and their approximate locations. Our emphasis on that joint inference, as well as tracking perceptual uncertainty over these states present a novel set of human experiments.

Multiple Object Tracking has also occupied a central role in the field of computer vision. In contrast with the highly controlled stimuli in psychology, MOT tasks in computer vision emphasize tracking every instance of a target object in real world stimuli. One of the most common approaches to MOT tasks in deep learning is a basic framework known as "tracking-by-detection" (Luo et al., 2021). Tracking by detection works as follows - in every frame, an image processing model extracts all the visible target objects. To associate detections across frames, a common approach to associate over time is to use a MAP estimate based on some dynamics model (often a Kalman filter). Because of the heavy bottom-up approach of these models, they tend to struggle in the absence of dense detections, and frequently make use of heuristics to determine when to drop a track that can't find suitable detections for several frames (Rajasegaran, Pavlakos, Kanazawa, & Malik, 2022).

There is a sub-genre in the field of Multiple-Object-Tracking, known as "Multiple Target Tracking" (MTT) which is more directly relevant to this work. The multiple target tracking problem is to explicitly estimate the number of targets in a region of interest and the state of each target, and many approaches over the years have made use of various tools from Bayesian modeling to handle uncertainty over said states (Bell, Corwin, Stone, & Streit, 2013). This project can be viewed an exploration of how humans engage in multiple target tracking.

Stimuli

Our stimuli were carefully designed by hand to elicit dynamic judgments over time - for example, scenes where two fireflies start off in a similar location, and then diverge, or start in different locations and then converge. All our stimuli, and their associated "ground truth" states can be viewed at this url.

Within each trial there was a fixed number of fireflies, and participants were informed that fireflies could not fly out of frame, or die. The trials were each 100 frames rendered at 20 frames per second, for a 5-second video.

We created 40 stimuli in total, consisting of four groups of 10 stimuli each. Each group contained a different number of fireflies: 1, 2, 3, or 4 fireflies per stimulus. Each participant saw all 40 stimuli, in a shuffled order. We also included 3 attention trials with high blinking rates and slow moving fireflies, and excluded participants who failed more than one attention trial.

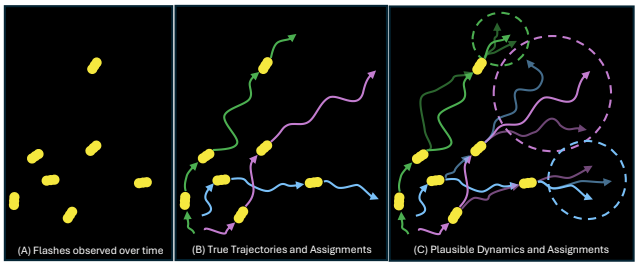

Figure 2: An illustrative example of our stimuli. (A) Sample flashes a participant might see over time. In between flashes, objects disappear. (B) Ground truth trajectories of fireflies and assignments of flashes to those fireflies in that trial. (C) Tracking requires joint inference over a distribution of possible assignments and trajectories at each time step.

Experiments

Experiment 1: Dynamic Counting Experiment

Our first experiment is a counting experiment, where the goal is to track the dynamics of how people estimated the number of objects in a scene. These judgments are deeply related to correspondence, since a new observation doesn't necessarily represent a new object. Participants need to account for the dynamics of currently tracked, invisible fireflies in order to decide whether a new flash corresponds to a new firefly or not.

In our counting experiment, participants were encouraged to report an initial guess as soon as they saw a blink, and to update their guesses as soon as they inferred new fireflies were present. After each trial, they were prompted to enter a final guess. If their final guess was incorrect, they received 0 points. Otherwise they were scored based on the average number of frames in which they had the correct guess. This enabled us to bias participants to respond as early, and as accurately as possible.

Experiment 2: Location Estimation

In addition to uncertainty over number of objects, we're also interested in how people update uncertainty over latent states

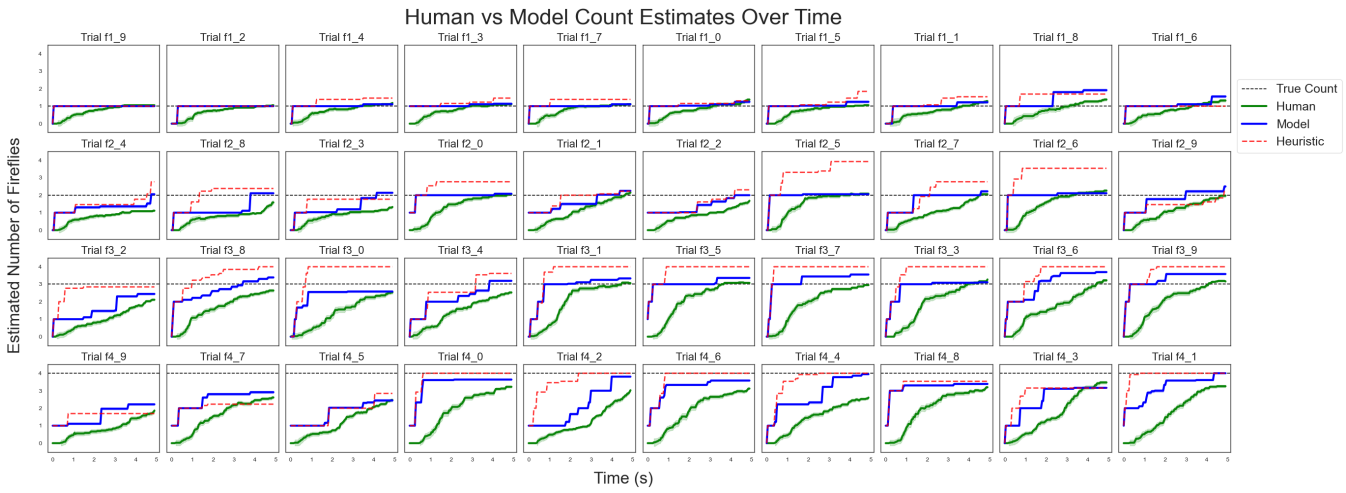

Figure 1: Our model (blue) also nicely accounts for the dynamics of count inferences, indicating that our model is making correspondence judgments similar to humans (in green). Because we don't model physical reaction time, we observe a slight delay between human and model time-courses, though the overall trends are similar. We also plot our heuristic baseline (in red), which often overshoots both human and model estimates.

like position. In this experiment, participants watch the 5 second video, at which point the canvas turns black, and all firefly motion freezes. Participants then draw circles to "catch" each of the fireflies in the scene. A net can only catch one firefly, but they can be dynamically resized to cover a wider area.

This allows us to measure both their final count inference and the spread of uncertainty over location. Participants were scored after each trial based on the number of fireflies they caught (ie; whether all fireflies were inside individual nets) and the average size of their nets. The exact scoring function was implemented as follows:

$$totalScore = \frac{\sum_{nets} \mathrm{I}[fireflyCaught] - \frac{NetArea}{CanvasArea}}{\max(\#fireflies, \#nets)} \quad (1)$$

where $\mathrm{I}[fireflyCaught]$ is the indicator function, which is 1 if a firefly is inside the net and 0 otherwise. The $\frac{NetArea}{CanvasArea}$ term penalizes the size of the net, such that a net covering the full screen would result in 0 points (even if it caught a firefly). The denominator takes the max of the number of ground truth fireflies and nets drawn, which penalizes overcounting (where there are more nets than fireflies) and undercounting (where there are fewer nets than fireflies). Before scoring, to handle overlap, nets are assigned to fireflies using the Hungarian Algorithm, a popular optimization method used to find the best assignment relative to some cost (Kuhn, 1955).

Smaller nets scored more points than larger nets, but larger nets scored more points than empty nets. This creates an intuitive incentive to draw the correct number of nets and - critically - to scale them based on uncertainty over locations of fireflies.

Each experiment contained several practice trials, a comprehension check, and attention checks during the experiment. Experiments were run on Prolific, and subjects were paid a minimum of $15/hr, with a possible bonus of up to $2 based on performance. Participants who failed attention checks were filtered from the final datasets.

Model

We compare human responses for each task to the output of an approximate Bayesian filter in a hidden Markov model (HMM) (Murphy, 2012). Our HMM posits simple probabilistic motion and blinking dynamics for the fireflies. As the number of fireflies is a priori unknown, the state variables of the HMM also keep track of the total number of fireflies. Our model assumes that the stimulus images are created via a simple pixelization process. The task of the filter is to approximately infer the Bayesian posterior on the state variables of the HMM at any time step given all observed images up to (and including) that step. We use sequential Monte Carlo (SMC) (Chopin, Papaspiliopoulos, et al., 2020) methods to implement the filter. The details of the generative model and inferences are as follows. In our notation, $[n]$ denotes the set $\{1,\ldots,n\}$ and for a tuple t the expression $\pi_i(t)$ denotes its i-th element.

HMM details. Our HMM is conditional on the number of flashes observed at every time step. That is, we assume we are given a sequence of non-negative integers $(k_t)_{t=1}^{T}$ where T is the total number of observed frames and k_t is the number of flashes at frame t. The state vector of the HMM at time t is (n_t, a_t, X_t). The random variable $n_t \in \mathbb{N}$ denotes the number of hypothesized fireflies at time step t. Given n_t and k_t, the random variable a_t denotes the assignment from the observed flashes to the hypothesized fireflies [1]. Lastly, given n_t, the random matrix $X_t \in \mathbb{R}^{4\times n_t}$ contains the position and velocity

[1] We assume an arbitrary but fixed ordering of the pixels of each observed frame, so that we can speak of the first flash, second flash, etc.

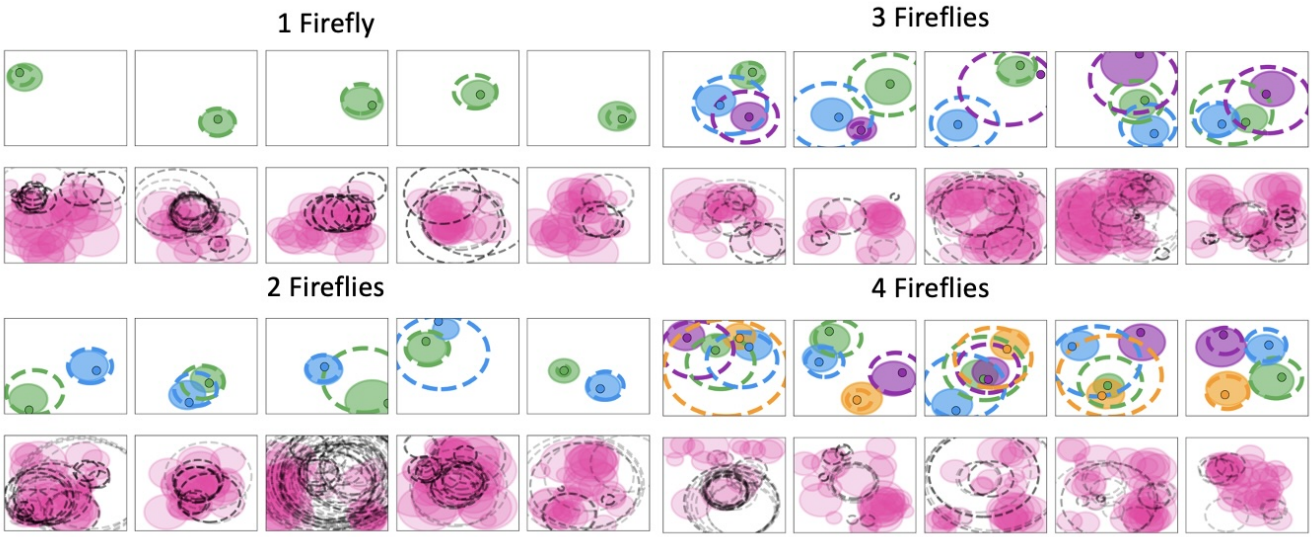

Figure 3: Average human nets and model nets for successfully caught fireflies (top row), as well as distribution of missed nets (bottom row, in pink). Model predictions are plotted with dotted lines, and human predictions are filled in. Circles are colored based on the firefly assignment. A net is a miss if it either fails to catch a firefly or catches one which has already been assigned to a different net. On many trials, our model almost perfectly matches human predictions for caught fireflies, and frequently captures interesting structure in the distribution of misses.

Algorithm 1 The assignment prior of the HMM

procedure ASSIGNMENTPRIOR(k, n, β, M)
  $c \leftarrow (1, \ldots, \min(n+1, M))$ ▷ possible next assignments
  for $i = 1, \ldots, k$ **do**
    $w \leftarrow \left(\beta^i (1-\beta)^{i-1}\right)_{i=1}^{|c|}$
    $j \sim$ CATEGORICAL(w) ▷ unnormalized weights
    $a(i) \leftarrow c_j$
    $c \leftarrow$ REMOVEINDEX(c, j)
    if $n < a(i) < M$ **then**
      $c \leftarrow$ APPEND($c, a(i)+1$)
    end if
  end for
  return a
end procedure

of the hypothesized fireflies. We denote the j-th column of X_t by X_t^j, and interpret the event $X_t^j = [x\ y\ v_x\ v_y]^\top$ to mean that the j-th firefly is at position (x, y) and has velocity (v_x, v_y) at time t.

At the first frame, we deterministically set $n_1 = k_1$, $a_t(i) = i$, and independently sample $X_1^i \sim \mathcal{N}(0, \Sigma_0)$ where Σ_0 is a fixed prior hyperparameter. At every subsequent step t, upon receiving k_t the state variables are updated as follows:

$$
\begin{aligned}
&a_t \sim \text{ASSIGNMENTPRIOR}(k_t, n_{t-1}, \beta, M) \\
&n_t := \max(n_{t-1}, a_t(1), \ldots, a_t(k_t)) \\
&X_t^j \sim \mathcal{N}(AX_{t-1}^j, \Sigma_a) && 1 \le j \le n_{t-1} \\
&X_t^j \sim \mathcal{N}(0, \Sigma_0) && n_{t-1} < j \le n_t.
\end{aligned}
$$

The matrix A encodes the Newtonian dynamics of each firefly with random acceleration and has the block form

$$
A := \begin{bmatrix} I_{2\times 2} & \Delta t \cdot I_{2\times 2} \\ O_{2\times 2} & I_{2\times 2} \end{bmatrix},
$$

where Δt denotes the time between two frames. The distribution ASSIGNMENTPRIOR samples a new assignment, while breaking the inherent symmetry of associating flashes to fireflies (Bell et al., 2013). The hyperparameter $M \in \mathbb{N}$ specifies an upper bound on the number of fireflies. In all experiments we take $M = 4$. The hyperparameter $\beta \in (0, 1)$ biases the prior to favor explaining new flashes with existing fireflies, which serves to implicitly model blinking frequency. Algorithm 1 describes the details of ASSIGNMENTPRIOR.

At every time step, the emissions of the HMM are produced by a simple pixelization procedure: we represent emissions as an $H \times W$ binary matrix depicting flashing fireflies in the rectangular region $(y_{\min}, y_{\max}) \times (x_{\min}, x_{\max})$. We denote the emission at time t by Y_t and enforce that pixel (i, j) takes value 1 if and only if for some $k \in [n_k]$ we have $X_t^k = [x\ y\ v_x\ v_y]^\top$ such that

$$
\left\lfloor \frac{(x - x_{\min})W}{x_{\max} - x_{\min}} \right\rfloor = j \qquad \left\lfloor \frac{(y - y_{\min})H}{y_{\max} - y_{\min}} \right\rfloor = i.
$$

Inference details. Our SMC inference algorithm is a pseudo-marginal particle filter (Lew, Cusumano-Towner, & Mansinghka, 2022; Naesseth, Lindsten, & Schon, 2015). At time t, each particle is of the form $(\widehat{n}_t, \widehat{a}_t)$ where $\widehat{n}_t$ is the particle's hypothesis for the number of fireflies n_t, and $\widehat{a}_t$ is the particle's hypothesis for the assignment a_t of fireflies to flashes. In addition to the particles, our SMC algorithm carries for each particle the parameters

$$
\theta_t := \left(\widehat{\mu}_t^j, \widehat{\Sigma}_t^j\right)_{j=1}^{n_t}
$$

of Gaussian approximations to the conditional distribution of X_t^j given $n_t = \widehat{n}_t$ and $a_t = \widehat{a}_t$. After receiving the observation Y_t, the proposal kernel of the particle filter uses θ_{t-1} to propose the values of n_t and a_t. In turn, the value of θ_t is deterministically updated from θ_{t-1} via expectation propagation (EP) (Minka, 2013), given Y_t and the proposed values of

$\widehat{n}_t$ and $\widehat{a}_t$. As these updates are deterministic, we get a simple rule for updating the importance weights of the particles. These importance weights are then used for multinomial resampling of the particles at each step.

At the first step, after observing Y_1 we initialize each particle with

$$\begin{aligned}\widehat{n}_1 &= k_1 \\ \widehat{a}_1(i) &= i \qquad\qquad 1 \le i \le \widehat{n}_1 \\ \theta_{1,\_} &= \text{EPUPDATE}\left(Y_1, (0,\Sigma_0)_{j=1}^{\widehat{n}_1}, \widehat{n}_1, \widehat{a}_1\right)\end{aligned}$$

The function EPUPDATE updates the Gaussian parameters from one step to the next. The Gaussian parameters of fireflies that are not associated with flashes remain unchanged. For those fireflies that are associated with a flash, first their position variables are updated by moment matching their previous Gaussian restricted to the pixel area of their associated flash. This moment matching procedure is done deterministically using quasi-Monte Carlo integration (Lemieux, 2009), and produces an estimate of the marginal likelihood of the supplied image as an intermediate result. This marginal likelihood estimate is the second return value of EPUPDATE. Next, each of these updated position Gaussians is joined with the exact conditional distribution of the velocity of the relevant firefly to get a full updated Gaussian.

The initial importance weight of each particle is computed as follows based on our model:

$$\begin{aligned}X_1^{ij} &\sim \mathcal{N}(\widehat{\mu}_1^j, \widehat{\Sigma}_1^j) \qquad 1 \le i \le N_w \\ W_1 &:= \frac{1}{N_w}\sum_{i=1}^{N_w} \frac{\ell(X_1^i; Y_1, \widehat{n}_1, \widehat{a}_1)\prod_{j=1}^{\widehat{n}_1}\mathcal{N}(X_1^{ij}; 0, \Sigma_0)}{\prod_{j=1}^{\widehat{n}_1}\mathcal{N}(X_1^{ij}; \widehat{\mu}_1^j, \widehat{\Sigma}_1^j)},\end{aligned}$$

where

$$\ell(X_t; Y_t, n_t, a_t) := \begin{cases} 1 & X_t^j \in \text{pixel area of } a_t^{-1}(j)\text{-th flash} \\ & \qquad \text{for all } j \in [n_t] \\ 0 & \text{o.w.}\end{cases}$$

is the likelihood function, and the hyperparameter $N_w \in \mathbb{N}$ controls the variance of the importance weight. At every subsequent step the particle filter proceeds as follows:

$$\begin{aligned}\widehat{n}_t, \widehat{a}_t &\sim \text{PROPOSAL}(Y_t, \theta_{t-1}, \widehat{n}_{t-1}, \widehat{a}_{t-1}) \\ \theta_{t,\_} &= \text{EPUPDATE}(Y_t, \theta_{t-1}, \widehat{n}_t, \widehat{a}_t) \\ X_t^{ij} &\sim \mathcal{N}(\widehat{\mu}_t^j, \widehat{\Sigma}_t^j) \qquad 1 \le i \le N_w \\ p_0 &:= \text{ASSIGNMENTPRIOR}(\widehat{a}_t; k_t, \widehat{n}_{t-1}, \beta, M) \\ q_0 &:= \text{PROPOSAL}(\widehat{n}_t, \widehat{a}_t; Y_t, \theta_{t-1}, \widehat{n}_{t-1}, \widehat{a}_{t-1}) \\ W_t &:= \frac{W_{t-1}p_0}{q_0 N_w}\sum_{i=1}^{N_w} \frac{\ell(X_t^i; Y_t, \widehat{n}_t, \widehat{a}_t)\prod_{j=1}^{\widehat{n}_t}\mathcal{N}(X_t^{ij}; AX_{t-1}^{ij}, \Sigma_a)}{\prod_{j=1}^{\widehat{n}_t}\mathcal{N}(X_t^{ij}; \widehat{\mu}_t^j, \widehat{\Sigma}_t^j)},\end{aligned}$$

where the density of the proposal distribution is given by

$$\begin{aligned}&\text{PROPOSAL}(\widehat{n}, \widehat{a}; Y_t, \theta_{t-1}, \widehat{n}_{t-1}, \widehat{a}_{t-1}) \propto \\ &\quad \pi_2(\text{EPUPDATE}(Y_t, \theta_{t-1}, \widehat{n}, \widehat{a})) \\ &\quad\quad \times \text{ASSIGNMENTPRIOR}(\widehat{a}; k_{t-1}, \widehat{n}_{t-1}, \beta, M).\end{aligned}$$

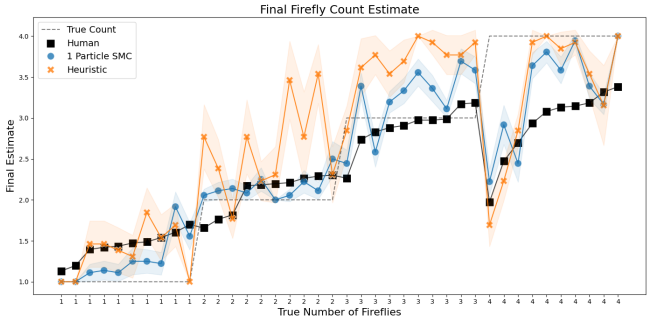

Figure 4: Across both catching and counting experiments, participants exhibited remarkably similar inferences (and similar errors) with respect to the number of fireflies in the scene. Our model (in blue) captures many of the same judgments as humans, compared to the baseline (orange).

Results

We ran 48 distinct model runs with varying settings for assignment priors and dynamics priors, to simulate the diversity of responses we see in human data. We additionally ran all our experiments with 1, 2, 5, and 10 particle SMC, and found that all performed similarly. We therefore take inspiration from previous literature and present all our results with 1-particle SMC (Sanborn, Griffiths, & Navarro, 2010).

Baselines and Ablations

We compare our model against heuristic baseline models, inspired by previous work suggesting humans make use of simple heuristics to make correspondence judgments when tracking through occlusion (Franconeri et al., 2012). Our baseline counting model is parameterized by a window size w and computes the maximum number of blinks occurring in any w-length subsequence of frames. Since optimal window size can vary based on the stimulus due to varying blink rates, we average the final count over a range of window sizes. Our heuristic catching model takes the heuristic estimate n as input, finds the last n blinks, and scales a net around that location with some constant multiple for each additional frame.

We also perform two ablations: one that removes velocity estimation from our model (Velocity Ablation), and another that allows the position distribution to go out of bounds of the screen (Boundary Ablation).

Estimating Firefly Counts

Because both counting and catching experiments require humans to make a final count judgment, we combine and analyze the predicted counts over both experiments in Figure 4. We find that humans made remarkably similar judgments to each other, and both humans and the model are fairly accurate at estimating firefly counts. The heuristic baseline tends to overestimate by even more, and doesn't align as well with the human data.

For a more fine-grained look at the Counting experiment data, Figure 1 presents the firefly count estimates over time

for both people and the model. The model often follows a similar trend to humans, though humans tend to be a bit slower in updating their estimates, which can be at least partially attributed to constraints on reaction times. The heuristic model is again a poorer fit to humans, frequently estimating considerably larger counts than people.

To quantify these comparisons, we calculate the mean squared error (MSE) between average human estimates of firefly counts over time to that of our model, and present the results in Table 1. The MSE between humans and our SMC model is 1.12, while the heuristic is a far weaker fit to humans with an MSE of 2.75. The velocity ablation hinders performance of the model considerably, while the boundary ablation has little effect. The weaknesses of both the heuristic baseline and the velocity ablation suggest the importance of tracking not just recently observed positions but also dynamics when making accurate object correspondence judgments in the face of sparse and ambiguous information.

Catching Fireflies

To compare our model to humans in the catching experiment, we need a way to convert model beliefs - in the form of mixtures of Gaussians - into nets. We can take the mean of each Gaussian to be the center of each of the model's "nets", but we still need a principled method to determine how many standard deviations out to draw the net. To resolve this, we assume that people trade-off uncertainty and accuracy in part by estimating the expected utility under the scoring policy. We use a form of Monte Carlo posterior filtering to convert model beliefs into nets. For each net, we evaluate several possible sizes, and calculate the average score over a number of samples of possible firefly positions from the model's belief. We then normalize the distribution of scores for different net sizes, and sample a policy from that distribution. If a model has a small covariance for the position of some firefly, it can maximize the score by drawing a slightly larger net. Conversely, if a model has broad uncertainty about a firefly's position, there are diminishing returns for increasing the net sizes, and it can try to maximize its score by drawing a smaller (though riskier) net.

In Figure 3, average human and model nets that successfully capture fireflies are plotted for a representative set of stimuli. The model often draws very similar nets to those of humans. To compare the nets, we calculate the intersection-over-union (IOU) of successful human and model average nets for each firefly. This metric captures both the location accuracy and the similarity in scale between human and model nets. IOU is a popular metric in segmentation literature, and averaging this over trials (as well as over fireflies within a trial) gives a mean IOU between 0 and 1.

The resulting mean IOUs are presented in Table 1, where the model outperforms the heuristic with an IOU of 0.30 compared to 0.14. Unlike in the Counting experiment, in the Catching experiment the velocity ablation has little effect, while the boundary ablation lowers the performance considerably.

	Counting: MSE ↓	Catching: IOU ↑
1-Particle SMC	$\mathbf{1.12 \pm 0.27}$	$\mathbf{0.31 \pm 0.032}$
Velocity Ablation	1.98 ± 0.28	0.30 ± 0.032
Boundary Ablation	1.13 ± 0.26	0.20 ± 0.038
Heuristic	2.75 ± 0.33	$0.14 \pm .0310$

Table 1: In both counting and catching experiments, our model provides a better fit to human data than the baseline heuristic. This is measured by the Mean Squared Error against human dynamic count estimates (lower is better) and Mean IOU against nets drawn by humans (higher is better). 95% confidence intervals are included.

Discussion

Our probabilistic model captures many of the inferences people make when engaging in uncertain tracking. In the counting experiment, our evidence supports the hypothesis that people do in fact use velocity to make probabilistic association judgments. When we ablate our model's ability to track velocity, performance drops drastically (though it still outperforms the heuristic). This suggests that, in order to decide if an observation is a new object or not, people consider how likely it is an existing object could have reached that point.

Similarly, in the catching experiment, our model outperforms the baseline model, as measured by the mean IOU. The model also performs better than the boundary ablated model, which tends to scale nets substantially larger due to its incorrect model of the scene dynamics. Interestingly, the velocity ablated model performs comparably in the catching experiment, and the boundary ablated model performs comparably in the counting experiment. This highlights the importance of our multiple experimental paradigms in teasing out the flexibility of our probabilistic model.

While our model does a reasonably good job capturing the patterns of uncertainty in our tracking tasks, there are still a number of limitations. For example, our model receives noiseless observations (that is, the true position of blinks). Humans, on the other hand, seem to have interesting structure in their patterns of misses, much of which might be a function of precise gaze location, and spatial uncertainty.

Conclusion

Our ability to compute correspondences between objects over discontinuous chunks of time and space is fundamental to our perception of the world as a stable, coherent entity. Moreover, we engage in this process nearly all the time, often unconsciously, yet the underlying computational mechanisms that allow us to do this efficiently and accurately remain a mystery.

Exploring how people track an uncertain number of noisy objects allows us to tap into the rich cognitive mechanisms underlying this ability. This work presents a novel setting for exploring the intersection of object tracking, mental correspondence and generative vision, and lays the groundwork

for further investigation in this direction.

References

Balaban, H., Smith, K., Tenenbaum, J., & Ullman, T. (2024). Electrophysiology reveals that intuitive physics guides visual tracking and working memory. *Open Mind*.

Bell, K., Corwin, T., Stone, L., & Streit, R. (2013). *Bayesian multiple target tracking, second edition*.

Chopin, N., Papaspiliopoulos, O., et al. (2020). *An introduction to sequential monte carlo* (Vol. 4). Springer.

Drew, T., Horowitz, T. S., & Vogel, E. K. (2013). Swapping or dropping? electrophysiological measures of difficulty during multiple object tracking. *Cognition*, *126*(2), 213-223. Retrieved from https://www.sciencedirect.com/science/article/pii/S0010027712002284 doi: https://doi.org/10.1016/j.cognition.2012.10.003

Franconeri, S. L., Pylyshyn, Z. W., & Scholl, B. J. (2012). A simple proximity heuristic allows tracking of multiple objects through occlusion. , *74*(4), 691–702.

Holcombe, A. (2023). *Attending to moving objects*. Cambridge University Press.

Horowitz, T. S., Birnkrant, R. S., Fencsik, D. E., Tran, L., & Wolfe, J. M. (2006). How do we track invisible objects? , *13*(3), 516–523. Retrieved from https://doi.org/10.3758/BF03193879 doi: 10.3758/BF03193879

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1-2), 83-97. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109 doi: https://doi.org/10.1002/nav.3800020109

Lemieux, C. (2009). *Monte carlo and quasi-monte carlo sampling* (Vol. 20). Springer.

Lew, A. K., Cusumano-Towner, M., & Mansinghka, V. K. (2022). Recursive monte carlo and variational inference with auxiliary variables. In *Uncertainty in artificial intelligence* (pp. 1096–1106).

Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T.-K. (2021). Multiple object tracking: A literature review. *Artificial Intelligence*, *293*, 103448.

Minka, T. P. (2013). Expectation propagation for approximate bayesian inference. *arXiv preprint arXiv:1301.2294*.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Naesseth, C., Lindsten, F., & Schon, T. (2015). Nested sequential monte carlo methods. In *International conference on machine learning* (pp. 1292–1301).

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. , *3*(3), 179–197. (Place: Netherlands Publisher: VSP) doi: 10.1163/156856888X00122

Rajasegaran, J., Pavlakos, G., Kanazawa, A., & Malik, J. (2022). Tracking people by predicting 3D appearance, location & pose. In *Cvpr*.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. doi: 10.1037/a0020511

Teichmann, L., Moerel, D., Rich, A. N., & Baker, C. I. (2022). The nature of neural object representations during dynamic occlusion. *Cortex*, *153*, 66-86. doi: https://doi.org/10.1016/j.cortex.2022.04.009

vanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects vs. substances. *Psychological Science*, *14*(5), 498–504.

Vul, E., Alvarez, G., Tenenbaum, J., & Black, M. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22). Curran Associates, Inc.