TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 8 (2016) 335–348

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12189

Attention Modulates Spatial Precision in Multiple-Object Tracking

Nisheeth Srivastava, Ed Vul

Department of Psychology, University of California, San Diego

Received 26 October 2015; accepted 27 October 2015

Abstract

We present a computational model of multiple-object tracking that makes trial-level predictions about the allocation of visual attention and the effect of this allocation on observers' ability to track multiple objects simultaneously. This model follows the intuition that increased attention to a location increases the spatial resolution of its internal representation. Using a combination of empirical and computational experiments, we demonstrate the existence of a tight coupling between cognitive and perceptual resources in this task: Low-level tracking of objects generates bottom-up predictions of error likelihood, and high-level attention allocation selectively reduces error probabilities in attended locations while increasing it at non-attended locations. Whereas earlier models of multiple-object tracking have predicted the big picture relationship between stimulus complexity and response accuracy, our approach makes accurate predictions of both the macro-scale effect of target number and velocity on tracking difficulty and micro-scale variations in difficulty across individual trials and targets arising from the idiosyncratic within-trial interactions of targets and distractors.

Keywords: Visual cognition; Multiple-object tracking; Attention dynamics; Metacognition; Bayesian models of cognition; Computational cognitive science

1. Introduction

Pylyshyn's multiple-object tracking (MOT) paradigm is one of the most prominent testbeds for studying visual cognition (Pylyshyn & Storm, 1988). In a typical MOT task (Fig. 1), subjects see a number of objects, typically circles, moving onscreen. Some subset of the objects are marked as targets before the trial begins, but during the trial, all

Correspondence should be sent to Nisheeth Srivastava, Department of Psychology, University of California, San Diego, 9500 Gilman Dr., La Jolla, CA 92093, USA. E-mail: nisheeths@gmail.com

objects turn to a uniform color and move haphazardly for several seconds. The task is to keep track of which objects were marked as targets at the start of the trial so that they can be identified at the end of the trial when the objects stop moving.

How is MOT even possible? It was not historically intuitive that MOT would be possible at all (Scholl, 2009). Classical theories of attention have tended to assume the existence a single attention "spotlight," suggesting that tasks involving selective attention at multiple locations simultaneously should be impossible. Yet, in practice, humans can track multiple objects quite successfully. How do they do it?

There are three theoretical accounts that shed light on different elements of the overall process. The first, Pylyshyn's FINST theory, was the primary motivator for the design of the MOT task itself. In this account, it was argued that people assign "pointers" to various objects and then track them, using a hitherto unknown pre-attentive mechanism. Pylyshyn's deep insight was to disconnect the task of indexing an object from attending to it. After all, he reasoned, how can you pay attention to *something* unless you have pre-attentively assigned it to be *some thing* (Pylyshyn, 1989)?

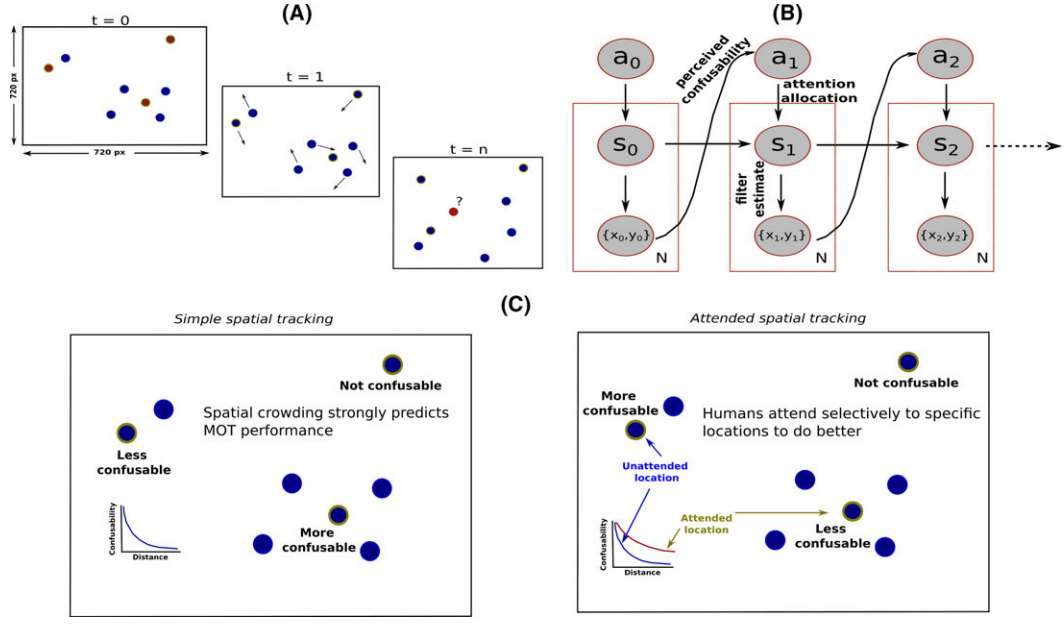

Fig. 1. (A) Schematic representation of a typical multiple-object tracking (MOT) task. (B) Graphical description of a hierarchical model for tracking *N* objects simultaneously. The low-level state estimate is computed using a bank of Kalman filters which predict particle locations with an accuracy that is influenced by their spatial resolution. (C) The scale of spatial resolution for a filter at any time step is determined by the attention allocated to it by the top-level model of attention dynamics. This model obtains information about the confusability of tracking targets from the filter predictions and rationally allocates attention to minimize overall confusability constrained by its attention budget.

However, if FINST (FINgers of INSTantiation) pointers are pre-attentive, it is reasonable to consider that a finite number of them preexist and are assigned when needed to environmental cues. Such accounts predict a sharp capacity constraint on the number of targets that observers can track—perfect tracking up to the number of pointers available and collapse thereafter. It is seen, however, that at sufficiently slow speeds, a relatively large number of targets can be accurately tracked (Alvarez & Franconeri, 2007), whereas at high speeds, even one or two objects cannot be tracked (Holcombe & Chen, 2012). The capacity constraint, then, is gradual, not steep, which implicates the allocation of some finite cognitive resource as the source of the MOT capacity constraint.

Cognitive resource models, as proposed in Alvarez and Franconeri (2007), explain the gradual nature of performance degradation with increasing target load. But this is not the only manipulation that requires explanation. Objects that move faster are harder to track, but if they are moved far apart in the visual field, they can be tracked even at high speeds, suggesting that spatial crowding also limits tracking (Franconeri, Lin, Pylyshyn, Fisher, & Enns, 2008). Additionally, objects that move for longer durations are harder to track, even at constant speeds (Oksama & Hyona, 2004). To account for such results, Franconeri and colleagues have proposed that the primary source of MOT difficulties is the distance that objects have to travel, with longer trajectories naturally providing more chances for identification errors (Franconeri, Jonathan, & Scimeca, 2010).

It is easy to find counterexamples casting doubts on strong versions of each of these theories. A claim that MOT is entirely pre-attentive and restricted to using a finite set of three to four object pointers (Pylyshyn, 1989) would fail to explain failures of tracking as few as one or two objects at high speeds (Holcombe & Chen, 2012) and the ability to track up to eight objects at sufficiently low speeds (Alvarez & Franconeri, 2007). The claim that spatial interference accounts for all MOT difficulties (Franconeri et al., 2010) fails to account for MOT failures in experimental setups designed to permit no spatial interference (Holcombe & Chen, 2012). Claims that cognitive resource limitations alone are responsible for MOT failures are difficult to set up in the first place, since it is still not clear what the resource being utilized in MOT is (Vul, Frank, Alvarez, & Tenenbaum, 2009). But any such claims would need to deal with the array of perceptual manipulations that lead to systematic changes in MOT performance we listed above. The effects of cognitive resource allocation cannot conceivably account for the impact of, say, changed object spacing, on MOT performance. Such effects necessarily have a perceptual basis.

Thus, while each of these theories brings useful insights into the MOT phenomenon, none can yet claim to explain it in its totality, and it is likely that true understanding will incorporate ideas from each one of them. So, for instance, it is possible to envisage a scheme wherein multiple attention foci are generated by object salience cues as required by FINST theory, are resolved to varying degrees of spatial precision through competitive allocation of endogenous attention as resource models predict, and suffer from errors of spatial interference when distractors approach the targets within an attention focus at a distance too close for the focus-specific spatial resolution to differentiate them.

In this article, we operationalize this particular theoretical synthesis computationally and test its predictions. The proposed model uses recursive Bayesian estimation of position coordinates to model the consequences of perceptual uncertainty and controls the effective length scales on which these estimators work as a function of the amount of *attention* allocated to them by a high-level controller. Our model follows the phenomenological intuition that humans are able to make finer-grained judgments of relative position when they attend more to a particular location and that such targeted covert attention is a scarce resource—resolution gain in the attended patch is bought at the expense of coarser-grained resolution elsewhere.

We demonstrate that adding a hierarchical controller that assigns spatial resolution to each of the low-level trackers out of a common pool of attention resource permits us to model MOT phenomena that reflect flexible cognitive resources, for example, the number of objects that can be tracked, and the profile of most common errors made by subjects. Successfully designing such a computational observer allows us to successfully model and predict human behavior at the millisecond level of resolution within MOT trials, unlike previous spatial interference studies (Franconeri et al., 2010), where model predictions apply to entire seconds-long trials at the finest level of analysis. Further, we show that people track different targets with variable spatial precision over time, following our models' predictions of strategic and dynamic allocation of cognitive resources, and that our model distinguishes between "dropping" and "swapping" errors (Drew, Horowitz, & Vogel, 2013), which we elicit behaviorally using a novel experimental manipulation.

2. Overview of flexible-resolution spatial tracking

We work within the framework of rational analysis, wherein models are strongly characterized by their computational goals.

The computational goal of the low-level controllers is to estimate individual object positions with statistical optimality given the noise/uncertainty of localizing objects in individual frames, following Vul et al. (2009). This assumption is entirely in line with existing ideas in Bayesian studies of visual perception and simply suggests that the low-level controllers behave as ideal Bayesian observers (Knill & Richards, 1996). In our account, each low-level controller tracks only one object in a MOT display; we discuss ways of relaxing this constraint in the Discussion.

We supplement this low-level controller with a finite resource, the allocation of which modulates the behavior of the ideal observer by changing local spatial uncertainty. This high-level controller operates on the assumption that humans can actively control the spatial resolution/uncertainty of individual percepts, but that localized spatial resolution magnification is bought at the expense of coarser resolution elsewhere. Intuitively, this can be visualized as the attention spotlight magnifying areas inside it, such that finer spatial discriminations become possible than would be possible without attention being focused on that region, as illustrated in Fig. 1c.

As the key contribution of this article, we formalized this intuition into a hierarchical model of inference, where low-level percept-tracking controllers learn the dynamics of individual objects and emit bottom-up signals identifying the likelihood of their tracking labels being lost, and a high-level metacognitive module uses these signals to rationally allocate attention to these controllers from a limited global pool, with the constraint that greater attention allocation permits finer spatial resolution. The top-down attention allocation, in turn, determines the uncertainty associated with low-level position measurements. Fig. 1b schematizes the computations involved in this model in graphical notation.

The computational goal of the high-level controller is to greedily reduce correspondence uncertainty, constrained by the total amount of attention resource available. While this is certainly not the only possible goal for metacognitive attention dynamics, constrained greedy optimization is rational in the context of dynamic resource allocation when the underlying demand distribution is non-stationary.

Our overall tracking model is based on coupling low-level controllers that iteratively solve the correspondence problem of observed objects across the visual space, with the high-level controller that allocates a finite resource that selectively improves localization precision at critical locations during the MOT trial.

2.1. Bayesian object tracking

We model individual object tracking as an ideal Bayesian observer learning a linear dynamical system. Given a state equation,

$$x_{t+1} = Hx_t + N(0, Q),$$

and a measurement equation,

$$z_t = Cx_t + N(0, R),$$

where Q is process noise and R is measurement noise, we implemented a Kalman filter that learns $\{H,C,Q,R\}$ at every time step using expectation-maximization–based parameter estimation (Ghahramani & Hinton, 1996). This filter serves as our perceptual ideal Bayesian observer for a single moving object. It takes the two-dimensional coordinates as the state observation $\{x,y\}$, predicts the future value of the latent state variable s, and thus generates predictions about future coordinates $\{x,y\}$.

A model completely faithful to the computational requirements of the MOT task would explicitly solve the correspondence problem: which observation should be associated with which filter, as in Vul et al. (2009). However, to account for human behavior, a simplification is possible: Rather than solving the correspondence problem at every time step, we can simply predict the ambiguity of correspondence at each time step and swap labels accordingly. This approach permits us to treat particle-filter bindings as known, instead of unknown, by default at every iteration, which greatly reduces the computational complexity of the model.

2.2. Rational attention allocation

The top-level attention model assumes that subjects possess a fixed amount of total attention, which can be represented as the scalar integer A. Following indexing-based ideas of object tracking (Pylyshyn, 1989), the model assigns indices p to all objects on the screen; and the amount of attention assigned to each object location at time t is a function $a_t(p)$, where $\sum_p a_t(p) = A$.

In every iteration, the model first determines the list of targets for which it will preferentially allocate attention by propagating the list of particles marked as targets (henceforth, the "target list") forward across time. While earlier indexing-based models of MOT have tried to retain the individual identities of each of the target particles, empirical results show that humans find it much easier to track target/non-target compared with tracking numbered target identities across the same trial duration (Pylyshyn, 2004). In light of this observation, we used binary target/distractor class identification for all particles.

At every time step, the model evaluates the potential *confusability* of all targets based on the object states the low-level Kalman filters. We approximate the probability of confusion as a logistic sigmoid decreasing with the distance between the target and its nearest distractor, but critically, this distance is scaled by the spatial resolution that each tracker's allocated attention resource permits it to have. These convergent desiderata inform our formal definition of confusability for each object p as

$$c(p) = \exp(-Ka_{t-1}(p)d_t^*(p)),$$

where K is a scaling parameter, $d_t^*(p) = \min d_t(p)$, and $d_t(p)$ is the estimated distance at model iteration t between p and all distractors if p is a target or between p and all targets if p is a distractor.

If a target is easily confusable with a distractor and vice versa, the two will swap target/distractor labels with a probability determined by the magnitude of their confusability. Once all possible swaps have been resolved, the particle possesses a new list of targets (which could be the same as the old list if no swaps occurred).

Since the model's current top-level attention allocation to all trackers is based on the previous iteration's distance estimates, it now determines a new attention allocation for each of A "units" of attention. Each unit is assigned to an object p by sampling an object index from a mixture model: With probability τ, an object index is sampled from a distribution obtained by normalizing the confusability of all objects, and with probability $1-\tau$, an object index is sampled from the targets with probability proportional to their confusability. The parameter τ controls the extent of inhibition of distractor particles. A value of 1 would mean that the model treats targets and distractors equally while dividing up attention. A value of 0 would mean that the model ignores all distractors and attends only to the targets. Grid search in parameter space suggested that a useful value of τ would be 0.4; this is the value we have used throughout our experiments.

Finally, reflecting sensitivity to cognitive processing costs, we assumed the model would possess some degree of inertia to changing its attention allocation, so that

$$a_t(p) = \lambda a_{t-1}(p) + (1 - \lambda)\overline{a_t(p)},$$

where $\overline{a_t(p)}$ is the allocation computed for the present iteration as above.

2.3. Experiment design

The basic MOT task is illustrated in Fig. 1a. After initial presentation of 12 objects, $n = \{1, 2,\ldots,6\}$ of which were red (targets) and the rest (distractors) blue at the beginning of each trial, the subject pressed a key to set them in motion. The objects all turned blue and moved on the screen following modified Ornstein–Uhlenbeck dynamics, as outlined below. After 5 s, the objects stopped moving and one of them, sampled from among the set of targets and set of distractors with equal probability, turned red. The subject had to indicate, by pressing "y" or "n," if the red object was red at the beginning of the trial too.

Subjects were allowed to practice the task they were to perform until they verified that they understood the objective and were accustomed to the keyboard controls. Practice data were discarded from subsequent analysis in all cases. All experiments were IRB approved and 50 undergraduate students volunteered as subjects for course credit. Subjects viewed the MOT display within a 720 × 720 pixel window on a 19-inch PC monitor at a viewing distance of 55–60 cm and used mouse and keyboard for inputs. Head movements and eye fixations were unconstrained throughout experiments. Each pixel on screen, therefore, subtended approximately 0.029° ± 0.004° visual angle for our subjects.

The position and velocity for each object evolve independently according to a modified Ornstein–Uhlenbeck process:

$$x_t = x_{t-1} + v_t,$$

$$v_t = \lambda v_{t-1} - k x_{t-1} + w_t,$$

$$w_t \sim N(0, \sigma_w),$$

where x and v are the position and velocity at time t; λ is a friction parameter constrained to be between 0 and 1; k is a spring constant which pulls the particles mildly to the center of the screen; and w_t is random acceleration noise added at each time point which is distributed as a zero-mean Gaussian with standard deviation σ_w.

In two dimensions, this stochastic process describes a randomly moving cloud of objects; the spring constant assures that the objects will not drift off to infinity, and the friction parameter assures that they will not accelerate to infinity. Within the range of parameters we consider, this process converges to a stable distribution of positions and velocities.

3. Results

3.1. More targets become harder to track

We replicate the finding that object tracking becomes harder both with increasing velocity of the particle swarm and with increasing number of targets (Alvarez & Franconeri, 2007). Unlike the original experiment, where subjects were allowed to adjust their own speed to what they felt was subjectively comfortable, we used a 3 up–1 down staircase, varying the parameter σ_w in steps of 0.5, thereby objectively measuring a 79% psychometric accuracy threshold for subjects.

Results of 14 subjects are shown in Fig. 2a and qualitatively match those from Alvarez and Franconeri (2007). An *in silico* replication of this experiment using a same-sized agent pool yields identical results, as shown in Fig. 2b, demonstrating that our model replicates aggregate human performance limitations arising out of both increasing velocity and target count. This overall pattern of behavior cannot be captured by a simple ideal observer without a constrained resource.

3.2. Predicting individual trial errors

While replicating aggregate predictions forms a useful baseline for assessing model validity, our model provides performance predictions for individual MOT trials, thereby providing a way to examine the limitations that humans face in doing this task at a much finer resolution. Pursuant to our interest in limitations to MOT performance, we are interested more in examining if our model gets the same trials wrong as humans. An

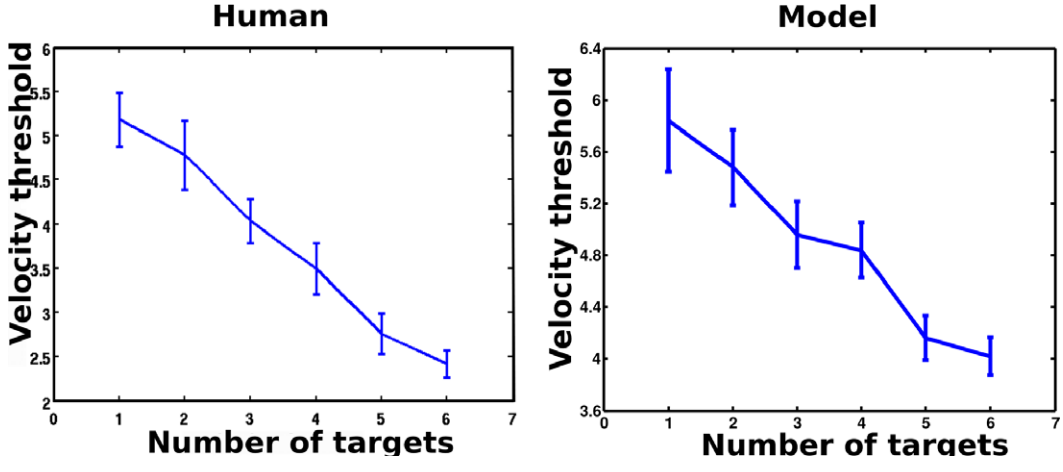

Fig. 2. The speed at which observers can maintain a particular accuracy threshold decreases as the number of objects to be tracked increases both for (*left*) 14 human subjects tested using a 3 up–1 down staircase experiment varying object velocity and (*right*) simulations of our model performing the same staircase task.

algorithm that has difficulty solving the same MOT trials that humans find difficult to solve is more interesting from a scientific standpoint than one that merely captures overall performance trends.

We conducted this analysis in the form of a binary classification study—using multiple ($N = 11$) simulations of model performance on an individual trial as a predictor for human performance. Perfect correlation between human and model predictions would be equivalent to perfect binary classification of human errors/non-errors using model predictions. As illustrated in Fig. 3a–b, our model outperforms a static spatial tracking model on two comprehensive criteria of classification performance: *F*-measure and area under the ROC curve. We obtained the ROC for our analysis by varying the threshold count of number of times the model got a trial correct (out of *N*) for us to label it positive between 1 and 11. The *F*-score reported is from the middle of the ROC, corresponding to a threshold count of 6.

In a separate experiment, we asked 22 students to perform the MOT task on 150 preset trials, with object velocity set at the average of our earlier sample. We then used classification with 20-fold cross-validation to calculate how well the performance of half the subjects on a trial predicted that of the other half, thereby obtaining a theoretical upper bound on classification performance (illustrated at the top of Fig. 3a). This upper bound places the extent of improvement in within-trial prediction performance engendered by our model in proper perspective—our model is clearly a considerable improvement over the static case, reducing nearly half the distance to the performance upper bound in terms of error classification performance.

Finally, since our model simulates objects' movements throughout the trial, it generates predictions for which objects it considers to be targets at the end of each simulation run. By measuring the congruence of these final target sets predicted by the model with the frequency with which humans made mistakes on probed objects, we can get a sense for whether the model *makes the same mistakes* the humans did, not just mistakes on the same trials the humans did.

Panels (c)–(d) in Fig. 3 present quantitative evidence for congruence between human and model errors even at this fine-grained resolution. In both figures, the *x*-axis plots the probability rank with which the model assigns a probed object to the target set, measured across 11 simulation runs; the *y*-axis counts the number of times the probed object occurred in all error trials across 30 subjects. For false-negative trials, where humans, when probed with a target, said that it was not, panel (c) shows that the targets that humans mistook for distractors were less likely to be members of the model's target set. For false-positive trials, where humans, when probed with a distractor, said it was a target, panel (d) shows that such distractors were more likely to be members of the target set in our simulation runs.

3.3. Assessing metacognitive attention control

The model we have proposed augments flexibility in spatial resolution to existing Bayesian accounts of multiple-object tracking, and our simulation experiments show that

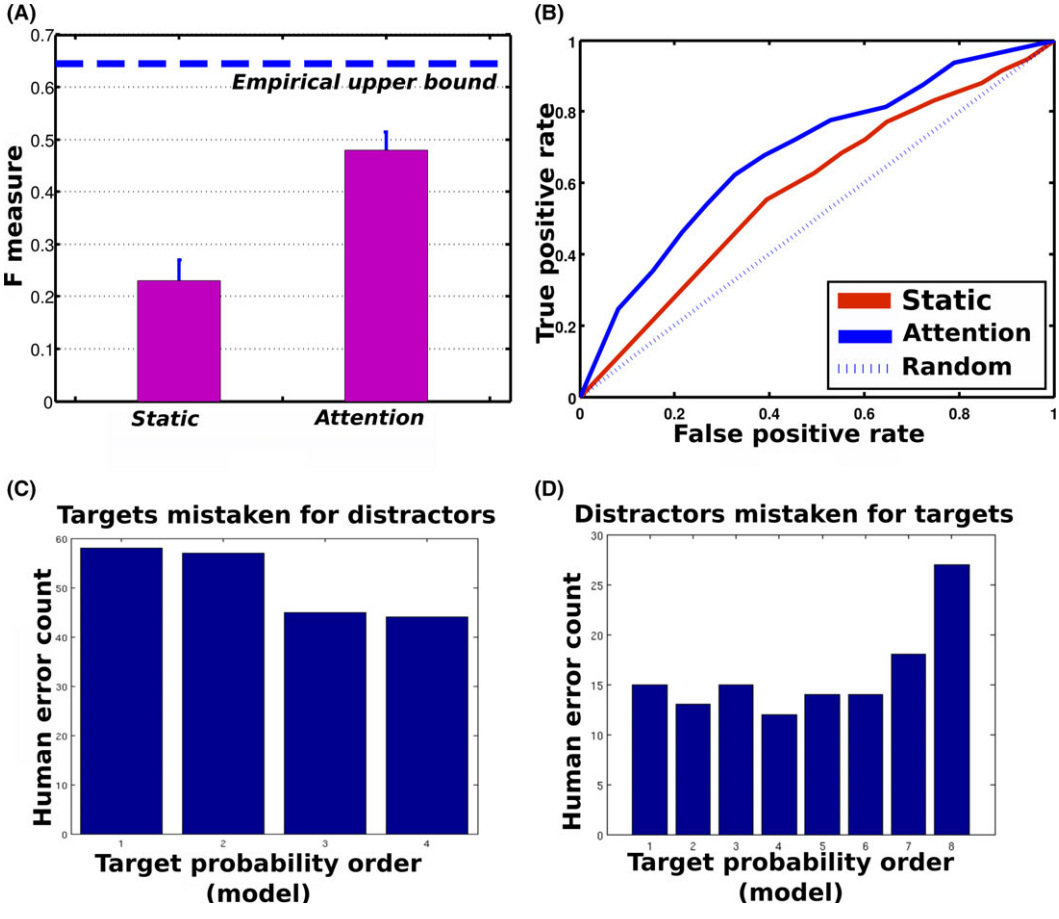

Fig. 3. Treating model performance (correct/incorrect) per trial as a binary classifier of human performance shows that attention-gated spatial tracking predicts trial-level human accuracy better than static spatial interference–based models, (A) with a considerably greater F-measure and (B) higher area under the ROC curve. Not only does our model make mistakes on the same trials as humans do, it also makes substantially the same mistakes that humans do, both for (C) trials where humans mistakenly identify a distractor is a target and (D) trials where they mistakenly identify a target as a distractor.

it does indeed improve trial-level predictions. Here, we further test some more specific predictions of the strategic-allocation MOT model: Does precision of tracking follow the predictions our model makes about allocated resources? And does the model distinguish between qualitatively different types of target-identification errors?

3.3.1. Crowded locations are tracked better

The key non-trivial prediction of our strategic-allocation model of multiple-object tracking is that subjects will localize easily confusable objects with greater precision because they will selectively attend to them more to resolve the possible ambiguity. In

contrast, a bottom-up theory of tracking would predict no relationship between crowding and localization error—location errors in such models would reflect either constant perceptual uncertainty or might even increase for more confusable objects due to crowding (Whitney & Levi, 2011).

We directly tested this prediction by making a simple manipulation to the basic design. We interleaved trials probing the identity of one of four targets with ones wherein once the dots stop moving, one of them disappears, and participants were prompted to click on its latest position using a mouse. Participants were instructed to focus on getting the probe trials correct and respond on the location trials as best they could. This was done to ensure that subjects did not stop attending to targets in order to focus more generally on the entire viewing area to better minimize location errors. We further expect that the randomly interleaved presentation of both types of trials (controlled by a Bernoulli parameter $p = .5$) also dissuaded such task switching.

Unlike in the other experiments, where trials were generated de novo for each participant, all 29 participants saw the same 150 pre-selected trials in this experiment. These trials were selected to hold the distance between the probed/disappearing particle and its nearest neighbor fixed at five separate values, 30 trials per distance value.

The results shown in Fig. 4a, plotting the localization error (in pixels) that subjects make against the category of trial (sorted by distance to nearest neighbor), show a clear trend favoring our hypothesis ($\rho = 0.91$, $p = .03$), and supporting related observations from Iordanescu, Grabowecky, and Suzuki (2009). Objects that disappeared in crowded locations were localized with greater precision than objects in less crowded locations, and this effect appeared to saturate for crowding radii subtending a greater than 1.8° visual angle, potentially reflecting the limited radius of the attention spotlight. This empirical result supports our work's basic assumption—that rational attention allocation influences MOT performance via flexibility in spatial resolution.

3.3.2. Model identifies drop versus swap errors made by humans

People do not always track all of the objects they were asked to, with errors arising from swapped labels between targets and distractors; instead, they sometimes simply drop a target and stop tracking it (Drew et al., 2013). For our purposes, swaps are erroneous identifications of target-distractor labels, as uncovered in the probe trials. Drops are erroneous identifications that participants knew would likely be erroneous before responding because they knew they had dropped a target. Therefore, we can estimate whether a given error was a swap or a drop by asking participants if they were surprised by the error. When an error arises from a participant swapping the probed target for a distractor, or vice versa, they would be surprised when told that they are wrong. Conversely, participants who knew that they had dropped one or more targets would express little surprise at being wrong.

We attempted to elicit precisely this information in a third experiment. The protocol for this study followed the same staircase design used in the first experiment; we collected data only for 14 subjects and with trials involving four targets amid eight distractors. Also, every time a subject responded incorrectly to the probed particle, they were

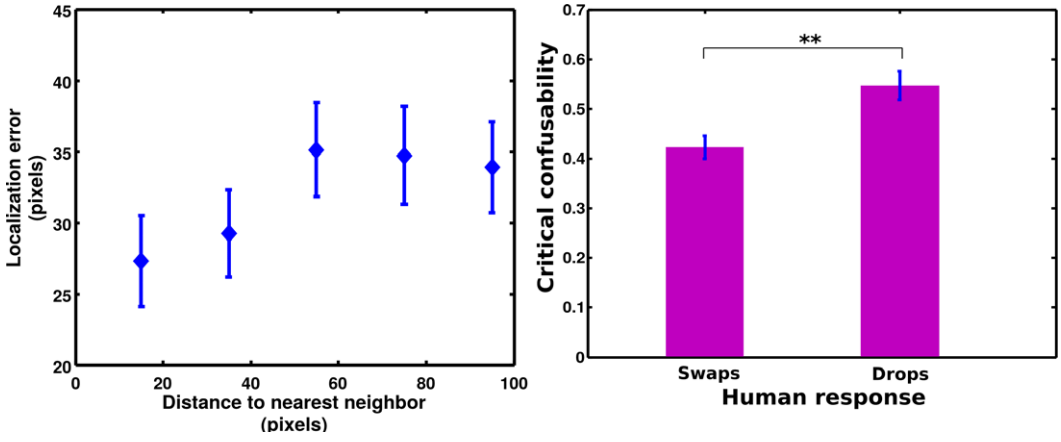

Fig. 4. Indirect measurements supporting the role of metacognitive attention dynamics in the MOT task. (*Left*) Subjects were more precise in localizing objects that were in crowded locations than those that were more isolated. (*Right*) The model's overall confusability load was significantly (p = .0017) higher at peak confusability in "drop" trials than in "swap" trials as measured behaviorally, suggesting that "dropping" could be a rational response in such situations. All error bars represent ± 1 *SEM*.

required to indicate with a keypress whether they were surprised at being wrong before proceeding to the next trial.

Even though our model does not include an explicit mechanism for dropping targets, it is possible to construct hypotheses about situations within trials that would promote drops and operationalize them in a testable manner. In particular, we expect that subjects would drop objects from the target list if their attention resources were overstretched, causing irreducible confusability among objects. In our model, the overall demand for attention might exceed capacity if there are many potentially confusable targets. Therefore, sensitivity to the drop-swap distinction in our model would predict that the cumulative confusability of all the targets would be larger at critical points in the trial for instances where errors would occur due to drops than for instances where errors would occur due to swaps. This prediction is borne out in our data, as shown in Fig. 4b, where we show that the critical confusability for trials labeled as *drop* errors from our behavioral characterization is consistently higher than for trials labeled as *swap* errors. Since errors in MOT happen at critical junctures and cannot be characterized by statistics averaged across the trial, we used the largest value of confusability obtained within a trial as our definition of critical confusability.

4. Discussion

Previous work has shown that patterns of aggregate behavior in multiple-object tracking as a function of the average speed and spacing of objects, the duration of tracking, and the number of distractors can be explained by an ideal observer iteratively solving

the correspondence problem (Vul et al., 2009). However, such ideal observers cannot capture the critical effects of tracking load—how many targets must be tracked—indicating that some sort of cognitive resource constraints limit human performance. We combined these two features to model human object tracking performance as Bayesian ideal tracking with a resource constraint and showed that such an agent exhibits the same tradeoffs between speed and number of targets tracked as people. We go further to show that this limited resource is not allocated to targets according to a fixed, static division, but is instead allocated strategically depending on the prospective costs and benefits of possible allocations.

Strategic, dynamic allocation of cognitive resources can better predict trial-level variations in performance across subjects in MOT performance. Furthermore, such strategic metacognitive allocation accounts for differences between trials when targets are dropped from consideration, rather than merely misassociated and swapped with distractors, differences that we were able to behaviorally elicit using a novel experimental manipulation. Finally, the specific combination of our presumed resource (spatial resolution—potentially mediated by attention) and our dynamic allocation policy predicts a specific pattern of variation in the precision of position estimates for individual targets (localization errors increase for less crowded objects), and we show that this holds for human observers. Together, these results represent proof-of-concept for how we can capture the interaction between bottom-up uncertainty and human cognitive resources using task-sensitive metacognitive policies: In multiple-object tracking, spatial resolution is allocated to reduce uncertainty for the correspondence problem.

Since our computational model is strongly predicated upon the ability of observers to consistently index objects, it fails in the same directions as indexing theory; for example, it cannot explain why humans find it easier to differentiate targets from distractors than to identify which target is which (Pylyshyn, 2006). Future work could replace our indexing assumption with more realistic models of generating attention foci given visual stimulus to accommodate these results (Trommershäuser, Maloney, & Landy, 2003). In particular, a more flexible representation could be developed where attention is split not between particles, but between retinotopic attention foci, which track areas in the visual field that contain varying numbers of objects.

Finally, classical MOT studies have tended to ask subjects to foveate a central location and maintain this fixation throughout MOT trials as a way of removing sources of perceptual uncertainty arising from body or eye movements. This constraint reduces the ecological validity of MOT performance as predictors of multiple-object tracking performance in the real world. We allowed subjects to fixate ad libitum during our experiment, which means that our present model lumps together the influence of overt and covert attention modalities. Future work could track eye fixations using eye trackers, with a richer attention model that includes a flexible covert element of the form we describe in this article as well as a more restrictive overt element, which would attenuate spatial resolution symmetrically as a function of distance from the point of present fixation. Rational predictions for such a model would include predictions of eye fixation positions as a function of target locations.

Acknowledgments

NS and EV were both supported by NSF grant 1239323. A preliminary version of this article was presented at the Annual Meeting of the Cognitive Science Society (Srivastava & Vul, 2015).

References

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you attentively track? Evidence for a resource-limited tracking mechanism. *Journal of Vision*, *7*(13), 1–10.

Drew, T., Horowitz, T., & Vogel, E. (2013). Swapping or dropping? Electrophysiological measures of difficulty during multiple object tracking. *Cognition*, *126*, 213–223.

Franconeri, S., Jonathan, S., & Scimeca, J. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science*, *21*(7), 920–925.

Franconeri, S., Lin, J., Pylyshyn, Z., Fisher, B., & Enns, J. (2008). Evidence against a speed limit in multiple object tracking. *Psychonomic Bulletin & Review*, *15*, 802–808.

Ghahramani, Z., & Hinton, G. E. (1996). Parameter estimation for linear dynamical systems. Technical Report CRG-TR-96-2, University of Toronto, Dept. of Computer Science.

Holcombe, A. O., & Chen, W. Y. (2012). Exhausting attentional tracking resources with a single fast moving object. *Cognition*, *123*(2), 218–228.

Iordanescu, L., Grabowecky, M., & Suzuki, S. (2009). Demand-based dynamic distribution of attention and monitoring of velocities during multiple-object tracking. *Journal of Vision*, *9*(4), 1.

Knill, D., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, MA: MIT Press.

Oksama, L., & Hyona, J. (2004). Is multiple object tracking carried out automatically by an early vision mechanism independent of higher-order cognition? An individual difference approach. *Visual Cognition*, *11*(5), 631–671.

Pylyshyn, Z. W. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial index model. *Cognition*, *32*, 65–97.

Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking (MOT): I. Tracking without keeping track of object identities. *Visual Cognition*, *11*, 801–822.

Pylyshyn, Z. W. (2006). Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving non-targets. *Visual Cognition*, *14*(2), 175–198.

Pylyshyn, Z., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179–197.

Scholl, B. J. (2009). What have we learned about attention from multiple object tracking (and vice versa)? In D. Dedrick & L. Trick (Eds.), *Computation, Cognition, and Pylyshyn* (pp. 49–78). Cambridge, MA: MIT Press.

Srivastava, N., & Vul, E. (2015) Attention dynamics in multiple object tracking. Proceedings of the 37th Annual Conference of the Cognitive Science Society. Pasadena, CA.

Trommershäuser, J., Maloney, L. T., & Landy, M. S. (2003). Statistical decision theory and the selection of rapid, goal-directed movements. *Journal of the Optical Society of America*, *20*(7), 1419–1433.

Vul, E., Frank, M., Alvarez, G. A., & Tenenbaum, J. B. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in Neural Information Processing Systems*, *22*, 1955–1963.

Whitney, D., & Levi, D. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*, 160–168.