Selection enables enhancement: An integrated model of object tracking

Andrew Lovett U.S. Naval Research Laboratory, Washington, DC, USA

Will Bridewell U.S. Naval Research Laboratory, Washington, DC, USA

Paul Bello U.S. Naval Research Laboratory, Washington, DC, USA

The diversity of research on visual attention and multiple-object tracking presents challenges for anyone hoping to develop a unified account. One key challenge is identifying the attentional limitations that give rise to competition among targets during tracking. To address this challenge, we present a computational model of object tracking that relies on two attentional mechanisms: serial selection and parallel enhancement. Selection picks out an object for further processing, whereas enhancement increases sensitivity to stimuli in regions where objects have been selected previously. In this model, multiple target locations can be tracked in parallel via enhancement, whereas a single target can be selected so that additional information beyond its location can be processed. In simulations of two psychological experiments, we demonstrate that spatial competition during enhancement and temporal competition for selection can explain a range of findings on multiple-object tracking, and we argue that the interaction between selection and enhancement captured in the model is critical to understanding attention more broadly.

Introduction

Attention lies at the heart of human visual processing. Our ability to selectively attend to relevant objects helps us make sense of the complex world around us. Thus, considerable effort has gone toward exploring how attention operates, when it succeeds, and when it fails. Multiple-object tracking (MOT) is a task that has proven useful for exploring attention to dynamic scenes. In this task, participants must track a set of moving objects, called *targets*, while distinguishing them from identical-looking distractors (Pylyshyn & Storm, 1988). Although MOT was not originally designed to study attention, there is ample evidence that it requires attention (e.g., Kunar, Carter, Cohen, & Horowitz, 2008; Tombu & Seiffert, 2008; for a review, see Meyerhoff, Papenmeier, & Huff, 2017), and the task has been used to study how attention is redeployed to address challenging situations (Holcombe & Chen, 2012; Meyerhoff, Papenmeier, Jahn, & Huff, 2016; Srivastava & Vul, 2016).

Despite MOT's popularity, there is widespread disagreement on how to interpret the results from object-tracking research (Alvarez & Franconeri, 2007; Franconeri, Jonathan, & Scimeca, 2010; Holcombe & Chen, 2013), much of it stemming from a lack of clarity about what visual attention is. For example, consider two key findings: Tracking performance suffers as the number of targets increases (Pylyshyn & Storm, 1988), and although four targets can be tracked effectively, some information about the targets, such as their motion histories, is unavailable when more than two targets are tracked (Luu & Howe, 2015). These findings suggest that targets compete for attention, such that only a finite number of targets can be processed fully, but what is the nature of this competition? To what extent is attention limited by space in the visual field (Franconeri et al., 2010), processing time (Holcombe & Chen, 2012), memory (Horowitz & Cohen, 2009), or other constraints? To address these questions, we need to explore the specific mechanisms that support attentional processing, and the conditions under which those mechanisms can provide sufficient information to track targets.

As an aid to developing a concrete theory of attentional mechanisms, we present the Integrated Model of Object Tracking (IMOT). IMOT is a computational model that mimics human tracking performance, operating autonomously on videos that match those shown to human participants. The model relies on two tracking mechanisms, one that can track multiple targets in parallel but is constrained by target spacing and another that serially picks out individual targets to compute further information, and thus is

Citation: Lovett, A., Bridewell, W., & Bello, P. (2019). Selection enables enhancement: An integrated model of object tracking. *Journal of Vision*, *19*(14):23, 1–31, https://doi.org/10.1167/19.14.23.

https://doi.org/10.1167/19.14.23 Received February 22, 2019; published December 23, 2019 ISSN 1534-7362

constrained by timing. This distinction is motivated by decades of research suggesting that at least two mechanisms support visual attention: enhancement and selection. Enhancement increases sensitivity to stimuli at one or more locations while decreasing sensitivity in the area surrounding each location (Egly, Driver, & Rafal, 1994; Eriksen & St. James, 1986; Posner, 1980; Tsotsos, 1990), whereas selection singles out an element for further processing (Rensink, 2000; Treisman & Gelade, 1980; Wolfe, 2007).

IMOT applies enhancement to mark each target's location, and these marks are updated in parallel to follow moving targets. Although this mechanism often is sufficient for tracking, sometimes location is not enough for distinguishing targets from distractors. For example, in a two-dimensional tracking task, it may be possible for a target to overlap a distractor in space, such that their locations briefly appear identical. To address this challenge, an individual target can be selected and its motion history computed before it overlaps a distractor. The motion history can be used to predict where the target will emerge after the overlap event ends, so that the target continues to be tracked successfully. However, because only one target can be selected at a time, participants will be less effective at processing motion histories as the number of targets competing for selection increases (Luu & Howe, 2015).

In the following sections, we demonstrate IMOT's ability to explain existing findings and generate novel, testable predictions about when targets will compete for attention. We begin by summarizing a range of MOT findings and proposing that most findings can be explained by enhancement, which is constrained spatially, and selection, which is constrained temporally. Next we describe IMOT, demonstrating that it is well grounded in visual-perception and attention research. We evaluate the model by simulating two behavioral experiments, one involving competition for space and the other involving competition for processing time. We stress that no existing model or theory provides a compelling explanation for the results across the two experiments. We close by discussing how IMOT could be expanded to explain a broader range of results and by considering the model's predictions. Notable, because enhancement and selection are general attentional mechanisms, these predictions go beyond MOT in isolation, and include the ways that object tracking and other attentionally demanding tasks should interact in complex, dual-task scenarios.

Background

From its beginning, research on MOT has both followed and informed theories of visual attention. The initial MOT research addressed an apparent paradox where theories positing that people can attend to one location at a time (Eriksen & St. James, 1986; Posner, 1980) conflicted with the capacity to reference multiple locations when evaluating a spatial relationship between two objects (e.g., *above*) and when reasoning about large-scale patterns. To account for this ability, Pylyshyn (1989) suggested that people use *visual indices*, pointers from representations in their minds to object locations in their visual fields. These visual indices are considered preattentive in the sense that they follow moving objects without requiring attention. In early investigations of MOT, Pylyshyn and Storm (1988) demonstrated that people can track five moving objects accurately. Using a mathematical model, they argued that it would be impossible to track that many targets by serially attending to each one.

In contrast with earlier findings, later research demonstrated that people can attend to multiple, distinct locations in parallel (Awh & Pashler, 2000; McMains & Somers, 2004). For example, Kramer and Hahn (1995) cued participants to expect letters at two locations and found that participants could compare the letters without being distracted by task-irrelevant letters positioned between the cued locations. Given this evidence, Cavanagh and Alvarez (2005) suggested that MOT relies on multifocal attention: People attend to and track multiple objects in parallel. If this is the case, then MOT should provide important information about the limits of multifocal attention. By exploring the conditions where tracking fails, researchers can determine when it is or is not possible to attend to moving objects.

Key findings in MOT

Here we summarize five classes of MOT findings, each of which helps to elucidate the limits of human visual attention. Two classes, spatial constraints and temporal constraints, directly address limits on attention. The third class, hemifield advantages, indicates one way of overcoming those limits. The fourth class, dynamic operation, suggests that the demands on attention can change over time during tracking. The final class, sensitivity to motion, demonstrates that processing complex visual features puts additional demands on attention beyond those required simply for tracking targets.

Spatial constraints

Several studies have found that tracking ability depends on the distance between targets, with performance declining when targets are able to move closer to each other (Carlson, Alvarez, & Cavanagh, 2007;

Holcombe, Chen, & Howe, 2014; Shim, Alvarez, & Jiang, 2008; Vater et al., 2017). These findings have given rise to a *spatial interference* account, which claims that neighboring targets interfere with each other during tracking (Franconeri et al., 2010). Such an account is consistent with the well-studied phenomenon of *crowding*, in which visual processing of a stimulus is impeded by the presence of nearby stimuli (Whitney & Levi, 2011), and indeed spatial interference appears to follow Bouma's law, which states that crowding occurs when the distance between stimuli is less than half the distance from a stimulus to the gaze fixation (Holcombe et al., 2014). Potentially, a target could be crowded either by nearby targets or by nearby distractors. However, several theories of attention posit that attending to a location results in enhanced sensitivity to stimuli at that location and suppressed sensitivity to stimuli in the surrounding area (Desimone & Duncan, 1995; Reynolds & Heeger, 2009; Tsotsos, 1990), and research has indicated that attending to a target makes it resistant to being crowded by its neighbors (Dakin, Bex, Cass, & Watt, 2009; Sundberg, Mitchell, & Reynolds, 2009; Yeshurun & Rashal, 2010). Therefore, it seems likely that targets strongly suppress processing of their neighbors, whereas distractors do so weakly.

Researchers have examined enhancement and suppression by measuring whether participants notice task-irrelevant probes (e.g., flashes of light) occurring on MOT targets, distractors, or the background (Doran & Hoffman, 2010; Drew, McCollough, Horowitz, & Vogel, 2009; Pylyshyn, 2006). In these studies, both behavioral and neural measures show greater sensitivity to probes on targets than to probes on distractors, which suggests that targets are enhanced, distractors are suppressed, or both effects occur simultaneously. Notably, it can be difficult to distinguish between enhancement and suppression without a clear baseline.

Temporal constraints

There is a well-established decline in tracking performance as object speed increases (Alvarez & Franconeri, 2007; Shim et al., 2008). However, in many studies the cause of this decline is unclear: It might result from fast-moving targets being harder to track or from fast-moving targets having more opportunities to crowd each other (Franconeri, Lin, Pylyshyn, Fisher, & Enns, 2008). Therefore, researchers have studied timing effects by employing circling stimuli, in which timing can be varied while spacing is held constant (Holcombe & Chen, 2012). In a typical display, each target travels around a circular trajectory, while one or more distractors follow each target along its circle. When these stimuli are used, it becomes easy to manipulate several key factors independently: motion speed, the distance between targets, the distance between a target and the distractors following it along the same trajectory, and the eccentricity of each target.

Using circling stimuli that were placed far enough apart to prevent crowding, Holcombe and Chen (2012) demonstrated that there was a speed limit of about 2 rotations/s for tracking a target, and that this speed limit remained constant regardless of the radius of the circle trajectory (see also Verstraten, Cavanagh, & Labianca, 2000). In following work (Holcombe & Chen, 2013), they found that there was a temporal-frequency limit of 7 Hz, meaning that the time interval between when a target occupies a location and when a distractor following it occupies the same location must be at least 1/7 s. Critically, as the number of targets increases beyond one, both the speed and temporal-frequency limit decrease, indicating that people cannot track multiple quickly circling targets in parallel. Holcombe and Chen have suggested that the key constraint here is processing time: A serial mechanism is required to track circling targets, and this mechanism must split its time among the targets.

Hemifield advantages

Interestingly, Holcombe and Chen (2012) found that two circling targets could be tracked at the same speed as a single target if the targets were displayed in separate hemifields. Several other studies have also found a strong advantage for tracking circling stimuli that are distributed across the hemifields (Alvarez & Cavanagh, 2004; Chen, Howe, & Holcombe, 2013; Shim, Alvarez, Vickery, & Jiang, 2010) or across quadrants (Carlson et al., 2007). In contrast, hemifield findings with targets that do not circle have been mixed. Tracking four targets is easier when two are in each hemifield than when all four are in the same hemifield, although the size of the advantage varies across studies (Alvarez & Cavanagh, 2004; Hudson, Howe, & Little, 2012). On the other hand, when only two targets are tracked, there is no advantage for distributing them across hemifields (Shim et al., 2008). The absence of a hemifield advantage with two targets appears to be more than just a ceiling effect, as it remains even when motion speed increases and overall performance declines.

Dynamic operation

The distribution of attention across targets appears to vary during tracking (Meyerhoff et al., 2016). This effect is most obvious when an eye tracker is used to measure overt attention shifts. Participants fixate on *threatened* targets, that is, targets that have moved near distractors and are at risk of being lost (Vater et al., 2017; Zelinsky & Todor, 2010). Shifting gaze to a

threatened target minimizes visual crowding, making it easier to distinguish the target from its surroundings (Whitney & Levi, 2011).

In other research, participants are encouraged to fixate on a central point throughout tracking, so that only covert shifts of attention are possible. In two such studies, a target disappeared at the end of the MOT task and participants were asked to click on its final location (Iordanescu, Grabowecky, & Suzuki, 2009; Srivastava & Vul, 2016). Participants showed less positional error when the target was closer to other objects, suggesting that they were focusing their attention on threatened targets. This finding failed to be replicated in a third study (Howard, Masom, & Holcombe, 2011), though there may have been differences in how the data were analyzed.

Combined, the data suggest that threatened targets draw attention. Participants fixate their gaze on a threatened target when possible, but they deploy covert attention to the target even when eye movements are not possible. However, questions remain about how attention is able to prioritize threatened targets while participants continue to track the other targets.

Sensitivity to motion history

Recently, Howe and colleagues (Howe & Holcombe, 2012; Luu & Howe, 2015) found that participants are better at tracking targets that move in straight lines and change direction only when bouncing off walls, compared to targets that randomly change direction one or two times a second. Importantly, in these studies objects could move through each other, causing them to partially or fully overlap. One explanation for the finding is that when objects move predictably, participants can use a target's motion history to extrapolate its future position, thereby better distinguishing the target from an overlapping distractor. In the reported experiments, the advantage for predictable trajectories held for two targets but not for four, suggesting that participants could process motion history only for a small number of objects.

Explaining MOT findings

Our objective is to develop a computational model that explains these findings, so that we can make concrete claims and testable predictions about the limits of attention. The core tenet underlying our model is that targets can be tracked through parallel enhancement combined with serial selection. In the following sections, we demonstrate how these two mechanisms can explain three classes of findings: spatial constraints, dynamic operation, and sensitivity to motion history. In particular, parallel enhancement is sufficient for tracking targets based on location, but it is constrained spatially because neighboring targets can crowd each other. This mechanism is complemented by serial selection, which can be deployed to a threatened target to provide access to other features, such as motion history.

We close the article by considering how our approach might explain temporal constraints on object tracking, and by discussing future work that would allow the model to address the remaining class of findings, hemifield advantages.

Enhancement and selection

The preceding section suggests that enhancement and selection are critical to explaining attentional limits during object tracking. Here we describe these mechanisms in general terms, and in the following section we describe how they support object tracking specifically.

Enhancement increases sensitivity to stimuli at locations in the visual field. For example, behavioral studies on space-based attention show that cuing participants to a particular location increases sensitivity to stimuli that appear there (Eriksen & St. James, 1986; Posner, 1980). Similarly, findings on object-based attention show that sensitivity to stimuli appearing within the bounds of salient objects is increased (Egly et al., 1994). Moreover, neuroscientists have found that both space-based and object-based attention are characterized by increased neural activity in early visual cortex that is specific to the attended location (McMains & Somers, 2004; Roelfsema, Lamme, & Spekreijse, 1998; Somers, Dale, Seiffert, & Tootell, 1999). Together, these findings suggest that attention is related to enhanced regions in the visual field, and that these regions are correlated with increased neural activity and sensitivity to stimuli.

Contrasting with enhancement, selection refers to choosing a particular stimulus to receive further processing. Common paradigms for studying selection include visual search (Egeth, Virzi, & Garbart, 1984; Wolfe, 2007), where participants look for objects that match a particular description, and change blindness (Rensink, O'Regan, & Clark, 1997; Simons, 2000), where noticing differences between visual scenes requires attending to changed elements. Findings resulting from these paradigms indicate that there is a processing bottleneck surrounding the binding of features into a cohesive representation. Specifically, research supports the view that the features from only one object can be bound together at a time (Rensink, 2000; Treisman & Gelade, 1980). Accepting the evidence that selection enables the construction of object representations, it follows that further mental operations on those objects rely on that initial, selective step.

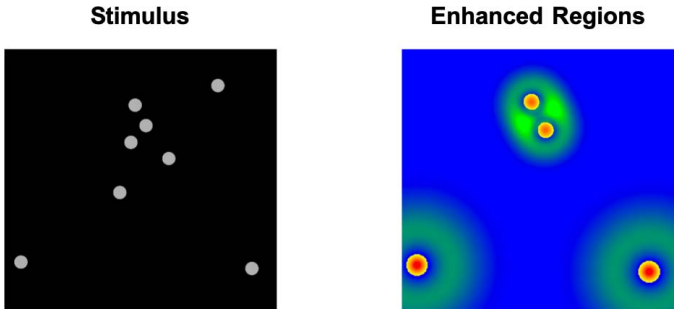

Figure 1. Enhanced (red and yellow) and suppressed (green) regions for four tracked objects.

Although they reflect separate mechanisms, selection and enhancement work in combination. Specifically, after individuals select an object as the focus of attention, they show enhanced sensitivity to its location in the visual field (Egly et al., 1994; Posner, 1980; Pylyshyn, 2006). Consequently, objects that appear in that region in the future are more likely to be selected than others. Because the mechanisms are so tightly interwoven, it can be difficult to distinguish one from the other. However, we believe a key distinguishing factor is that objects must be selected serially, whereas multiple regions associated with previously selected objects can be enhanced in parallel (Kramer & Hahn, 1995; McMains & Somers, 2004).

A theory of enhancement and selection in MOT

We introduce a theory in which enhancement and selection play complementary roles in object tracking. As targets move, a parallel updating mechanism marks their last-known locations via enhancement. At the same time, selection of a single target can support further processing of its features. Critically, each of the attentional mechanisms provides useful information about tracked targets, but each is constrained in its own way. Parallel updating via enhancement provides each target's location but is constrained spatially, because enhancing a target location results in suppressing the surrounding area. In contrast, serial selection provides access to other object features, such as a target's motion history, but is constrained temporally, because only one target can be selected at a time.

Because center–surround suppression is central to our views on object tracking, it is worth noting that there is disagreement on how this mechanism operates. Some researchers have argued that enhancing a target area results in reduced sensitivity to the surrounding area (Desimone & Duncan, 1995; Reynolds & Heeger, 2009), whereas others have argued that suppressing the surrounding area results in increased sensitivity to the target area (Tsotsos, 1990; Tsotsos et al., 1995). Unfortunately, it is difficult to distinguish these possibilities behaviorally or even neurally. Because both possibilities ultimately result in increased sensitivity in the center and reduced sensitivity in the surround, the present model performs both enhancement and suppression.

Figure 1 provides an example of enhancement and suppression. In this case, four disks are being tracked. Their locations are marked by enhanced regions, depicted in yellow and red, that are surrounded by suppressed regions, depicted in green. When two targets move close to each other, the suppressed region around one overlaps the enhanced region around another (the upper two targets in Figure 1). This interference leads to smaller enhanced regions, which increases the risk that the targets will be lost (Shim et al., 2008).

Parallel updating provides basic tracking capabilities, but sometimes information beyond target locations is needed. For example, when a target moves through an identical-looking distractor, the two briefly overlap and cannot be distinguished by their locations. However, the target's motion history can be used to predict its future location after the two objects separate. To facilitate processing motion history, a target that is threatened by nearby objects and thus in danger of overlapping them will tend to be selected (Iordanescu et al., 2009; Srivastava & Vul, 2016; Zelinsky & Todor, 2010). This strategic selection of threatened targets becomes less effective as the number of targets increases, because there is a greater probability that multiple targets will be threatened at the same time, and thus a greater risk that a target will not be selected before it overlaps another object. As a result, the processing of motion histories will be less effective when there are more targets (Howe & Holcombe, 2012; Luu & Howe, 2015).

Example

Figure 2 illustrates multiple events during a MOT trial. The left column depicts five time slices of the trial. The selected object at each time is indicated by a blue square. The right column depicts the enhanced and suppressed regions used to support parallel updating. Initially, two disks turn red to indicate that they are targets (Figure 2a). Each red disk is selected, which enables feature binding and creates an object representation (Treisman & Gelade, 1980) that is stored in visual short-term memory (VSTM). Each object representation in VSTM is assigned a visual index (Pylyshyn & Storm, 1988) that the parallel updating mechanism uses to track its location. The visual index

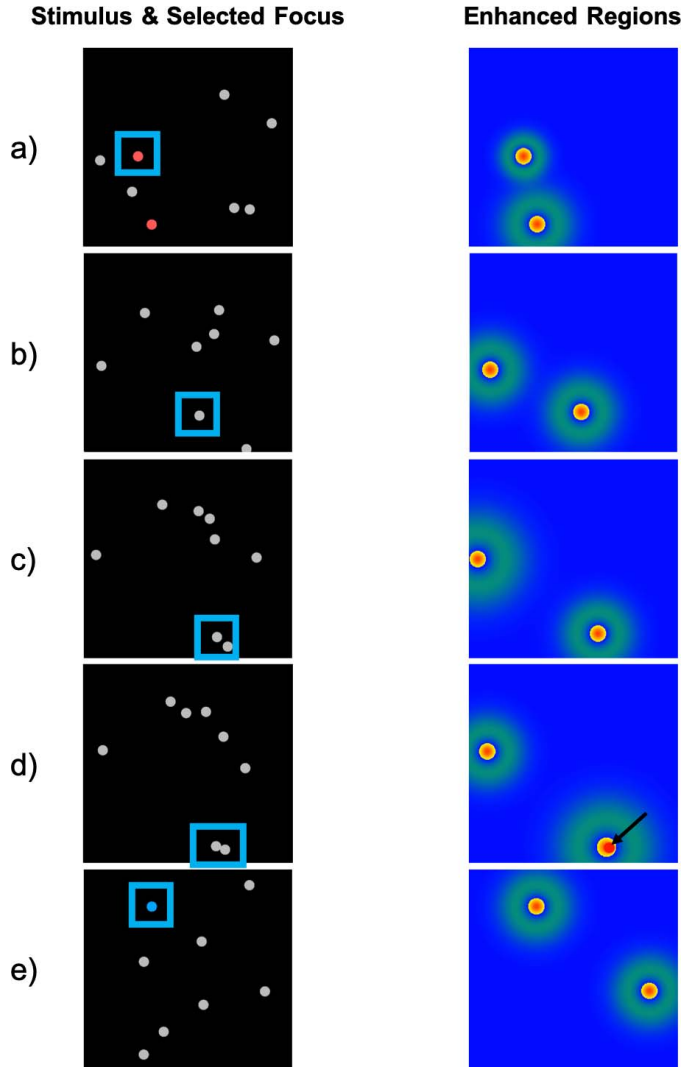

Figure 2. The selected focus and the enhanced and suppressed regions at five time slices (a–e).

marks the target's last-known location via center–surround suppression, with an inner, enhanced region (red and yellow) surrounded by an outer, suppressed region (green).

After a short interval, the red disks change to gray disks and all disks begin moving (Figure 2b). As the trial progresses, the parallel updating mechanism tracks the two targets without requiring selection (in this example one of the targets is selected, but selection is providing no added value). This mechanism addresses the *correspondence problem* (Pylyshyn, 2004), the challenge of determining which disk in the video corresponds to each target's visual index. A disk is matched to an index if two conditions are met: The disk overlaps the index's enhanced region, and the disk is the closest to the enhanced region. The correspondence problem must be solved at every time step as the objects move.

At time C (Figure 2c), one target moves near a distractor. Because this target is threatened by a nearby object, it is selected, and its motion history is computed over time. When the selected disk overlaps the distractor disk, the motion history is used to predict the selected disk's location (Figure 2d). To this end, a bias, shown by the extra red circle with an arrow pointing to it, is added to the disk's enhanced region at the predicted location. This bias enables the correct identity assignment once the disks no longer overlap.

At the end of the sequence, a disk changes to blue, which causes the disk to be selected (Figure 2e). The blue disk's location is compared to the last-known locations of the target disks to determine whether it is a target or a distractor.

The ARCADIA modeling framework

IMOT is implemented in ARCADIA (Bridewell & Bello, 2016), a computational framework that has been used to model human thought and behavior across multiple tasks. ARCADIA was built to explore the relationships among attention, perception, cognition, and action. A key claim underlying its development is that attention biases mental processing such that a selected element (which may derive directly from an external stimulus or may be an internal, "mental" representation) receives more processing than other elements. Specifically, the framework distinguishes between focus-sensitive processing that occurs only for a single, selected element and focus-independent processing that occurs regardless of what is selected. This distinction is critical for modeling selection and enhancement in MOT. Focus-sensitive processing supports integrating a selected object's features and computing its motion history, whereas focus-independent processing supports parallel enhancement of multiple target locations.

Mental processing in ARCADIA is simulated by a set of components that operate in parallel over a sequence of cycles. These components listen for broadcast information that they have the ability to process, apply computational routines to that information, and broadcast the resulting information for components to use in the subsequent cycle. On each cycle (see Figure 3), components process input and generate output. One of the outputs is then selected as the focus, which directs focus-sensitive components to process it if they are able to do so. The components may receive input from three sources: the output produced by all components on the previous cycle, the output item that was selected as the focus on the previous cycle, and sensors that retrieve data from the

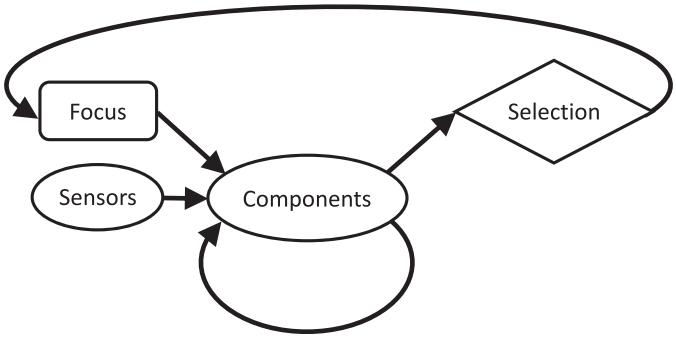

Figure 3. ARCADIA's processing cycle.

environment (e.g., by providing still frames from a video depicting a MOT trial).

Each ARCADIA model consists of a set of components and an *attentional strategy*, which provides the model's task-specific attentional priorities. For example, in a tracking task it may be important to select tracked objects that are threatened by nearby distractors. On each cycle of processing, the output item with the highest attentional priority is selected as the focus of attention.

The ARCADIA framework is largely unconstrained, as each component can, in principle, perform any operation that a function in a computer program might perform. However, the framework does rely on several key commitments: Processing is controlled through the selection of a focus of attention, and only one focus may be selected at a time; representation and processing are distributed among components; and there is a short cycle time (typically corresponding to 25 ms of real time), which bounds the amount of processing in each component. For instance, as the next section describes in detail, encoding the representation of a visual object in VSTM takes three cycles or roughly 75 ms, involves at least four components, and transforms unstructured perceptual information into a structured representation that includes nonconceptual content.

In addition to the framework constraints, model design is undertaken with reference to the literature from psychology, neuroscience, and philosophy of mind, including both experimental results and theoretical arguments. An underlying objective is to develop broad mechanisms that operate in the same or similar ways across models, rather than building isolated, task-specific models. With that objective in mind, our goal in this article is to give a mechanistic account of how general cognitive structures and processes combine to enable the execution of the object tracking task.

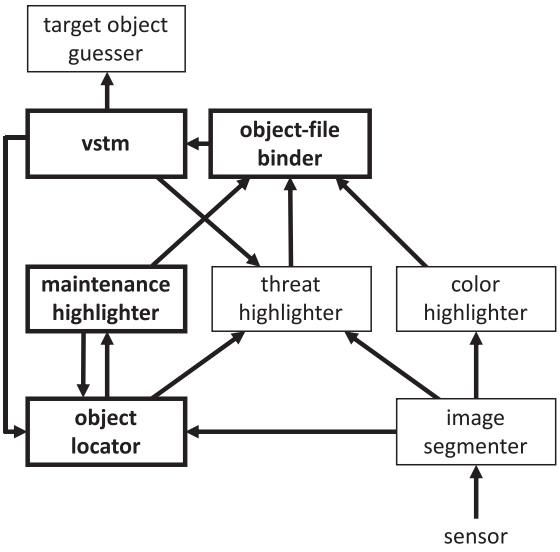

Figure 4. Flow of information in IMOT. Components in bold are sensitive to the selected focus.

Integrated model of object tracking

IMOT is an implementation of selection and enhancement during object tracking, built within the ARCADIA architecture. IMOT's source code is available for download.[1] Figure 4 illustrates IMOT's components and shows how information flows between them (those components that are sensitive to the focus of selection are indicated with bold text), and Table 1 presents the model's attentional strategy. Critically, the model is meant to capture the flow of information in the mind during tracking, but not computation at a neural level. Although we will motivate each component, we are not pursuing a direct mapping between components and brain regions or functions. Additionally, the output generated by components may be symbolic (e.g., an indication that a target currently is

Priority	Type	Component	Notes
1	Object representation	OBJECT-FILE BINDER	
2	Proto-object	COLOR HIGHLIGHTER	Preference for proto-objects that are not already tracked
3	Proto-object	MAINTENANCE HIGHLIGHTER	This proto-object overlaps another object
4	Proto-object	MAINTENANCE HIGHLIGHTER	THREAT HIGHLIGHTER indicates that this proto-object is threatened or has the closest threat
5	Proto-object	THREAT HIGHLIGHTER	THREAT HIGHLIGHTER indicates that this proto-object has the closest threat

Table 1. IMOT's attentional strategy.

threatened by nearby objects), numerical (e.g., the location and size of an object), or pictorial (e.g., the color pixel values for an object). See Appendix A for more information on how output is represented. In the next section, we present IMOT's components and show how they would process the example sequence in Figure 2. Afterward, we describe each component in greater detail.

Model walkthrough

Suppose IMOT is presented with the video summarized in Figure 2. A sensor captures a still frame from the video every 25 ms and makes it available to the components, and one cycle of processing corresponds to 25 ms of real time. On each cycle, IMAGE SEGMENTER[2] performs figure–ground segregation on the still frame provided by the sensor, partitioning the frame into *proto-objects*: short-lived representations describing possible regions of interest (Rensink, 2000). For Figure 2a, there are eight proto-objects, corresponding to the eight disks. Each proto-object encodes a disk's location and nonconceptual representations of its basic features, like size and color (i.e., proto-objects are not categorized as "red" or "large" or "round," but they do store the sensory information useful for assigning such categories).

IMAGE SEGMENTER outputs a collection of proto-objects in the visual scene, and the individual proto-objects must be selected before their features can be processed and stored in memory. To this end, a set of components called *highlighters* suggest particular proto-objects as candidates for selection—for example, COLOR HIGHLIGHTER suggests selecting brightly colored proto-objects. One proto-object is selected according to the attentional strategy (Table 1), and then on the following cycle, OBJECT-FILE BINDER constructs an object representation to describe what was found at that proto-object's location (e.g., a red disk). This description, which includes location and feature information, is then selected as the focus because the attentional strategy gives top priority to object representations.

After an object representation is selected, VSTM stores it in visual short-term memory, which maintains information about recently selected items. On the cycle after an object representation is selected and stored in memory, a new proto-object is selected and the pattern repeats. Within the context of this model, we can say that IMOT is focusing on an *object* when the model selects the proto-object corresponding to it and then selects the resulting representation, storing it in memory. Additionally, we can say that IMOT is maintaining focus on an object when the model repeatedly alternates between selecting the proto-object corresponding to a stored object representation and selecting the updated representation which refreshes properties stored in memory (e.g., if a color change has taken place, the memory of the object will be updated with the new color).

Although the objects represented in VSTM must be selected again for their feature information to be updated, their location information is updated each cycle by OBJECT LOCATOR, which maintains visual indices pointing to each object's last-known location. This component tracks recently selected objects via three steps: Mark the last-known locations with enhanced regions (Figure 2, right column); identify the current proto-object closest to each enhanced region; and update each visual index to point to the corresponding, current proto-object. As a result, other components always have access to the current location of a remembered object, in the form of its current proto-object.

Returning to Figure 2a, COLOR HIGHLIGHTER suggests the two proto-objects corresponding to the red disks as candidates for selection. One of these proto-objects is selected and represented in VSTM, and then the attentional strategy gives priority to the other red disk (the one whose location is not already being tracked by OBJECT LOCATOR), to ensure that both will be selected. Thus when the red disks change to gray and all disks begin moving (Figure 2b), both target disks are tracked.

Most of the time, OBJECT LOCATOR can track moving targets in parallel based on their last-known locations. However, when a target draws close to a distractor (Figure 2c), there is a risk that the two disks will overlap and their locations will become indistinguishable. To address this concern, two components—THREAT HIGHLIGHTER and MAINTENANCE HIGHLIGHTER—support selecting a target that is threatened by nearby objects and computing its motion history so that the object can be tracked successfully if it overlaps another object. Like COLOR HIGHLIGHTER, these components suggest proto-objects as candidates for selection, but they suggest proto-objects corresponding to currently tracked targets, and they associate each proto-object with additional information.

THREAT HIGHLIGHTER computes a *threat distance* for each tracked target's current proto-object, based on the distance to the nearest other proto-object. This component suggests every tracked target's proto-object as a candidate for selection, and it associates each proto-object with an indication of whether this target has the closest threat and an indication of whether this target is threatened. A target has the closest threat if its threat distance is the lowest among all targets, and it is classified as threatened when the threat distance falls below a threshold which indicates that other objects are close enough to interfere with successfully tracking that target.

MAINTENANCE HIGHLIGHTER computes motion history for the currently selected proto-object and detects overlaps with other objects based on sudden size changes (e.g., as in Figure 2d). This information can be generated only if IMOT focuses on an object over time. To facilitate the collection of information about object motion, this component supports maintaining focus on an object by suggesting the current proto-object corresponding to the last-selected object. If MAINTENANCE HIGHLIGHTER detects that the target is currently overlapping another object, it also provides that target's predicted location based on its motion history.

The output from these two highlighters is managed according to the attentional strategy (Table 1). Initially, the proto-object corresponding to the target with the closest threat is selected. Then, MAINTENANCE HIGHLIGHTER suggests continuing to select the current proto-object corresponding to that target. IMOT's attentional strategy encodes a strong bias for maintaining focus on a selected target. More specifically, the proto-object identified by MAINTENANCE HIGHLIGHTER is preferred unless both another target has the closest threat and the current target is neither threatened nor overlapping another object.

While a selected target overlaps another object, OBJECT LOCATOR uses the predicted location provided by MAINTENANCE HIGHLIGHTER, adding enhancement at the predicted location in the priority map (see the small red circle indicated by an arrow in Figure 2d). This additional enhancement ensures that after two overlapping objects separate, the correct proto-object will be matched to the target's enhanced region so that the target will continue to be tracked. Notably, this process requires that the target be selected long enough to collect its motion history before the objects overlap. If the target is not selected or is selected too late, then its motion history will be unavailable, and the model will be unable to distinguish the two overlapping objects.

At the end of a MOT trial, one of the disks changes to blue, and IMOT must indicate whether it is one of the tracked targets (Figure 2e). After the proto-object for the blue disk is picked out by COLOR HIGHLIGHTER and selected, TARGET OBJECT GUESSER generates a "match" or "mismatch" signal, based on whether the blue disk matches the current location of a tracked target. Optionally, multiple disks may change to blue in sequence, and this component generates a signal for each.

Components

With the exception of TARGET OBJECT GUESSER, IMOT's components are not specific to object tracking. These components, which constitute a general front end to visual processing, support selecting objects, processing their features, and storing their representations in memory. The same or earlier implementations of these components have been used in ARCADIA models of other visual tasks, such as enumeration (Briggs, Bridewell, & Bello, 2017) and change detection (Bridewell & Bello, 2015). In the remainder of this section, we stress the role that each component plays in visual processing, rather than its role in MOT alone.

Image segmenter

Tracking objects requires the basic ability to distinguish them from the unimportant parts of the visual field, foregrounding items of interest and backgrounding the rest. To this end, IMAGE SEGMENTER takes input from the visual sensor and carries out figure–ground separation, an important first step in visual processing that identifies regions that correspond to possible objects (Palmer & Rock, 1994). These possible objects have been referred to as proto-objects (Rensink, 2000), which are volatile representations that require attention to become stable over time. The regions of IMAGE SEGMENTER are maximally large, connected areas that differ in intensity from the background. The component generates a list of proto-objects describing each region's location, size, and color profile. Because a new set of proto-objects is produced on each cycle, they are forgotten by the model unless one is selected as the focus of attention.

Color highlighter

In MOT tasks, there are various ways to draw a person's attention to the target objects and to the queried objects at the end of a trial. Many of these methods rely on pop-out effects of visual perception, such as having the target disks blink or coloring them. In general, the pop-out effect refers to the tendency of objects that contrast with their surroundings to grab attention in a manner consistent with fast, parallel processing (Wolfe & Horowitz, 2004). As implemented for this model, COLOR HIGHLIGHTER outputs requests to select proto-objects that are not grayscale. Because the default color of the disks is gray, the targets are red at the beginning of each trial, and the queried disks are blue at the end, this component simulates the pop-out effect in IMOT.

Threat highlighter

Several MOT studies have found that people attend to targets as they draw close to other objects (Iordanescu et al., 2009; Srivastava & Vul, 2016; Zelinsky & Todor, 2010). This behavior may increase visual acuity or support the collection of information about an object's motion. In IMOT, THREAT HIGH-

LIGHTER simulates this draw on attention by computing the threat distance from each target's proto-object to the nearest other proto-object, which enables the selection of the appropriate target according to the attentional strategy. Specifically, the component generates a request to focus on the proto-object associated with each tracked target, accompanied by two pieces of information: an indication of whether this proto-object currently has the closest threat, and an indication of whether this proto-object is classified as threatened.

Classifying targets as threatened influences IMOT's behavior because its attentional strategy (Table 1) indicates that once a selected target becomes threatened, focus must be maintained on it. This classification is made based on a parameter labeled *max-threat-distance*, which is the maximum threat distance at which a target will be considered threatened. The value of this parameter controls how quickly the model switches which target is selected. If the parameter is small, targets will rarely be considered threatened, and the model will immediately focus on whichever target has the closest threat; but if the parameter is large, then targets will often be considered threatened, and the model will be highly resistant to changing the focus to a new target. We discuss this threshold and other free parameters further at the end of this section.

Maintenance highlighter

Within IMOT, tracking targets that overlap distractors requires a motion history that lets the model extrapolate the location of a currently selected object when there is perceptual uncertainty. Consequently, MAINTENANCE HIGHLIGHTER has four roles. First, to keep a target selected long enough to store that information, MAINTENANCE HIGHLIGHTER generates requests to select the proto-object corresponding to the current selection focus. Second, to enable extrapolation, while focus is maintained on a target the component computes and stores its motion history.[3] Third, to determine when to provide motion information, MAINTENANCE HIGHLIGHTER detects overlap-event onsets and offsets based on abrupt increases or decreases in the proto-object's size.[4] And fourth, when a selected target overlaps another object, this component outputs its predicted location based on the motion history and a signal indicating that the overlap is occurring.[5] This information is used by OBJECT LOCATOR to track the target, as described later.

Object-file binder

Tracking objects requires the ability to store targets over time and distinguish them from potential distractors. Proto-objects are ill suited for that role, since they are transient encodings and do not survive the presentation of new information in the visual field. Instead, IMOT incorporates the concept of *object files* from feature-integration theory (Treisman & Gelade, 1980). Object files in the model reflect current evidence that once objects receive visual attention, they are encoded into stabilized representations that integrate the visual features and enable storage in short-term memory (Lee & Chun, 2001; Luck & Vogel, 1997). Consequently, when IMOT selects a proto-object suggested by one of the highlighters, OBJECT-FILE BINDER generates a representation that consolidates the features found at that proto-object's location. These features include size information found in the proto-object itself and information from other components, such as threat information generated by THREAT HIGHLIGHTER.

Visual short-term memory

There are several memory systems that may contribute to object tracking and a variety of potential theories to explore. On this front, IMOT relies on a slot-based view of VSTM, where four slots each store a single, integrated object file (Pylyshyn & Storm, 1988; Vogel, Woodman, & Luck, 2001; but see Wilken & Ma, 2004). The VSTM component is an active storage system that monitors which item is selected on each cycle and stores any object files that become the focus of attention. If the slots in this component are full, any new representation will unseat the least recently encoded object file. This assumption follows from evidence of a recency effect in VSTM (Kool, Conway, & Turk-Browne, 2014).

Several studies have suggested that people use the same mental resources to remember object locations as they use to track targets (Bettencourt, Michalka, & Somers, 2011; Drew & Vogel, 2008; Oksama & Hyönä, 2004). In line with this research, VSTM slots in IMOT are closely associated with visual indices (Pylyshyn, 1989), which are maintained by OBJECT LOCATOR and used for tracking (discussed next). This connection is used by VSTM to determine whether a focused object file represents an unstored object or an update to an existing one: If the visual indices for the two object files point to the same location, then the object files likely represent the same object.[6] When an object file matches an object already represented in VSTM, the old representation is updated.

In ARCADIA models, short-term memory components operate by outputting their contents on each cycle. As a result, up-to-date, memorized representations are available to all components simultaneously. In IMOT this means that VSTM outputs each of its stored object files. Additionally, when this component stores an object file, it issues a signal indicating whether the

corresponding object is recognized as already in VSTM or new.

Object locator

The components listed so far enable IMOT to notice, represent, and remember target objects. However, the model still needs a way to track moving objects in parallel. To this end, OBJECT LOCATOR uses a combination of visual indices and spatial enhancement to address the motion-correspondence problem (Dawson, 1991). This visual problem is most commonly encountered in the case of apparent motion and concerns how the visual system tracks the stable identity of a moving object over time even when motion is sampled discretely (e.g., frames of a movie, psychological experiments on computer screens). In broad strokes, this component's approach comprises two steps: Mark each tracked object's last-known location with an enhanced region, and identify the current proto-objects that best match those regions.

To carry out its function, this component maintains visual indices pointing to each remembered object's last-known location. Following Pylyshyn's (1989) view, there is a finite number of visual indices: one index for each of the four VSTM slots, and an additional index for newly focused proto-objects that haven't yet been represented in VSTM. Whenever a proto-object is selected as the focus, a visual index is attached to that proto-object, and it will follow the corresponding object's motion until it is assigned to another proto-object. On each ARCADIA cycle, IMOT tracks objects by solving the correspondence problem: identifying the current proto-object that best matches each visual index. Afterward, the indices are updated to point to the newly identified proto-object locations. If no proto-object matches an index, then that index is abandoned, which in MOT results in losing track of a target.

Visual indices are matched to proto-objects using a *priority map* (Bisley & Goldberg, 2010; Fecteau & Munoz, 2006),[7] which is implemented as a two-dimensional array that can be overlaid on the visual input. Each location in the array can store a positive number, indicating how strongly that location is enhanced, or a negative number, indicating how strongly it is suppressed. Visual-index locations are marked on the priority map via center–surround suppression, with an enhanced region of positive values surrounded by a suppressed region of negative values (see the red and yellow enhanced regions and the green suppressed regions in Figure 1). If two indices point to nearby locations, one index's suppressed region may overlap another's enhanced region, resulting in smaller enhanced-region sizes (the top two regions in Figure 1).

Enhanced and suppressed regions are generated using the Marr wavelet:

$$\psi(t) = \frac{2}{\sqrt{3\sigma}\pi^{1/4}}\left(1 - \left(\frac{t}{\sigma}\right)^2\right)e^{-\frac{t^2}{2\sigma^2}}$$

This function forms a one-dimensional "Mexican hat" shape[8] whose positive region extends to $\pm\sigma$. Negative regions extend outward from these points, asymptotically approaching 0, but they are near 0 by $\pm 3\sigma$.

A visual index's enhanced and suppressed regions are computed by entering each location's distance from the index into the Marr wavelet to determine how much that location is enhanced or suppressed. The function is scaled such that the enhanced region's radius σ is the tracked proto-object's radius multiplied by 1.5, meaning the enhanced region extends 50% beyond the object's radius. In contrast, the function is scaled so that the suppressed region's radius increases for objects in the periphery—we assume here that gaze is centered in the display, as in many MOT studies.[9] The value of σ is scaled such that the distance from σ to 3σ is half the proto-object's distance from the display center. This scaling approximates Bouma's law for visual crowding (Whitney & Levi, 2011), which states that perception of an object that is x distance from the gaze center will tend to be disrupted by other objects at a distance of up to $x/2$.

Enhanced regions are matched to proto-objects in three steps. First, a score is assigned to each pairing of an enhanced region E and a proto-object P:

$$\begin{aligned}\text{Score}_{EP} = {} & \text{Radius}_E + \text{Radius}_P \\ & - \text{Distance}(\text{Center}_E, \text{Center}_P).\end{aligned}$$

This equation sums the radii of the enhanced region and the proto-object and subtracts the Euclidean distance between their centers to compute the width of overlap between them. Nonoverlapping items will receive a negative score. The equation makes the simplifying assumption that proto-objects are circular, which is reasonable for the stimuli discussed in this article. Second, enhanced regions are matched greedily to proto-objects, beginning with the highest scoring pair and enforcing a one-to-one mapping constraint: An enhanced region can match at most one proto-object, and vice versa. Only positively scored pairs can be matched. Third, the one-to-one mapping constraint is relaxed, and each unmatched enhanced region is matched to the proto-object with which it has the highest positive score. This final step lets two visual indices point to the same location, as when two targets overlap. On each cycle, OBJECT LOCATOR outputs a list of the updated visual indices.

Adding a bias: When MAINTENANCE HIGHLIGHTER provides a predicted location for a target that overlaps another object, OBJECT LOCATOR adds a bias to that target's enhanced region (see the red circle indicated by

an arrow in Figure 2d). When possible, the enhanced region is matched to a proto-object that overlaps the bias, so that when the overlapping disks split apart, the correct disk will be tracked.

When an enhanced region includes a bias B, the score for the pairing between that enhanced region E and each proto-object P is computed as follows:

$$\text{Score}_{EP} = \begin{cases} \text{Radius}_B + \text{Radius}_P - \text{Distance}(\text{Center}_B, \text{Center}_P), \\ \quad \text{if this is } >0. \\ \text{Radius}_E + \text{Radius}_P - \text{Distance}(\text{Center}_E, \text{Center}_P), \\ \quad \text{otherwise.} \end{cases}$$

The equation's top case refers to the width of overlap between an enhanced region's bias and a proto-object, whereas the bottom case refers to the width of overlap between the overall enhanced region and the proto-object (as in the simpler equation provided earlier). In comparing scores for each pairing, priority is given to pairings with a positive overlap between an enhanced region's bias and a proto-object.

Adding noise: OBJECT LOCATOR will stop tracking objects when they move fast enough so that the enhanced regions from one cycle fail to overlap the proto-objects on the following cycle. Until that point, the component will track perfectly. In contrast, human accuracy gradually drops as object speeds increase (Shim et al., 2008). To account for this gradual decrease, we add noise to the model. On each cycle, a pairing of an enhanced region E and a proto-object P will be discarded if either its score is negative or the following condition is met:

$$\text{Score}_{EP} < \text{Gaussian} \times \text{noise-width} + \text{noise-center}.$$

In this equation, Gaussian returns a random variable drawn from a standard normal distribution, and noise-width is a multiplier that increases the width (and standard deviation) of the noise, whereas noise-center translates the center of the noise. For example, if noise-width = 0.1 and noise-center = 0.2, then for each pairing of E and P, a random variable is drawn from a Gaussian distribution centered at 0.2 with a standard deviation of 0.1. If the width of the overlap of E and P is less than the random variable, then the pairing is discarded.

Target object guesser

The final component in IMOT is task specific and enables us to gather data on the model's tracking accuracy, so that we can compare it to human data. TARGET OBJECT GUESSER reports whether the highlighted object at the end of a MOT trial matches one of the tracked targets.[10] This component uses the signal that VSTM generates when a new object representation is added to memory or an old object representation is updated. If a new object representation with the highlighting color is added—indicating that the object is untracked—then this component outputs a mismatch signal. If an old object representation is updated but the object's color has changed to the highlighting color, then this component outputs a match signal.

Free parameters

The behavior of IMOT's components depends on three important parameters that have been described in the previous sections. The first, THREAT HIGHLIGHTER's max-threat-distance, determines how close a threat must be for a target to be classified as threatened. Because the model will maintain focus on a threatened target, a high max-threat-distance makes it more difficult for the model to flexibly change its focus to whichever target has the closest threat.

The other parameters, OBJECT LOCATOR's noise-width and noise-center, determine how much noise is applied in evaluating a match between an enhanced region and a proto-object. As the noise increases, there is a heightened risk of failing to match an enhanced region to a proto-object, and thus losing track of a target.

In the simulations that follow, we explore the effects of varying the three parameters, including examining what happens when they are set to zero.

Simulation 1

Our first simulation explores spatial competition and its contribution to *dropping*, or losing track of targets. In IMOT, the spatial arrangement of targets is critical because each target's location is marked via center–surround suppression. When one target moves near another, its suppressive surround overlaps the other's enhanced region, resulting in a smaller enhanced region. Smaller enhanced regions are dangerous because IMOT's OBJECT LOCATOR uses the overlap between the enhanced region from one cycle and the proto-object on the following cycle. If the two fail to overlap, the model cannot identify the target's current location, and the target is dropped. An additional factor that contributes to dropping targets is the speed of the objects. If objects travel a greater distance between cycles, then there is a higher risk that the enhanced region and proto-object will not overlap. Thus, targets will be dropped more frequently either when targets are closer together or when objects move faster.

If IMOT's explanation is correct, then the model's object-tracking accuracy should match human accuracy as speed and target spacing are varied. To test this conjecture, we simulated a human MOT experiment originally conducted by Shim et al. (2008, experiment 2) that varied the two factors. In the experiment,

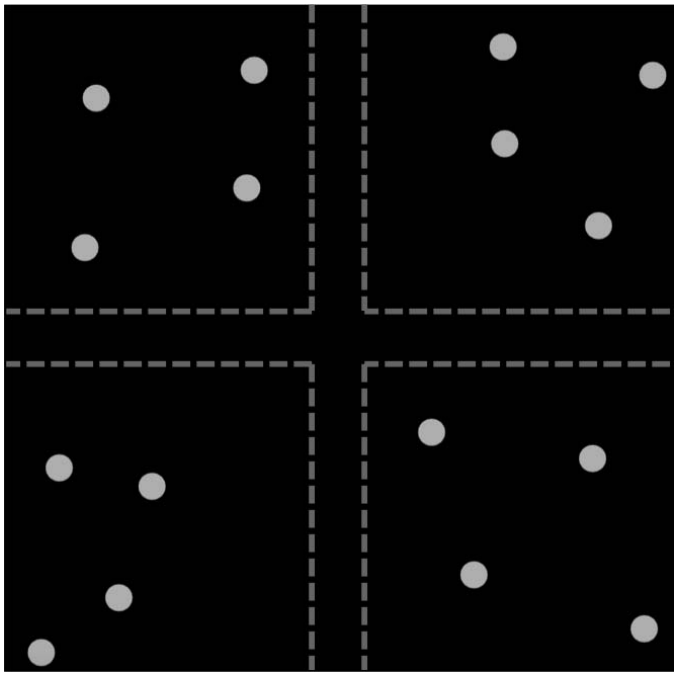

Figure 5. Example of a multiple-object tracking task in which the disks are restricted to quadrants. The dashed lines show the boundaries of each quadrant and are not visible to participants.

participants tracked one or two out of 16 moving objects. The 16 objects were evenly distributed into four quadrants, and their movement was restricted such that they could not leave their quadrant (see Figure 5). Shim et al. reported that tracking accuracy dropped as motion speed increased and when there were two targets in the same quadrant.

Experiment description

In the original experiment (Shim et al., 2008), six participants tracked black disks while maintaining center fixation. On each trial, one or two disks were highlighted in red (or green) to indicate that they were the targets. Afterward, these disks changed back to black, and the 16 disks began moving. Disks moved in straight lines, bouncing off the edges of their quadrants. They also bounced off each other when they reached a minimum between-disks distance (1.61° visual angle). After 8 s, the disks stopped moving and a randomly selected disk from one of the tracked quadrants was highlighted. Participants pressed a key to indicate whether the highlighted disk was a target or a distractor. The probability that the highlighted disk was a target was always 50%.

On separate trials, there were three tracking conditions: one target (Track-1), two targets in adjacent quadrants (Track-2-far), and two targets in the same quadrant (Track-2-near). There were five motion speeds, ranging from 5.73°/s to 11.43°/s. Each participant saw 240 videos: 4 tracking conditions × 5 motion speeds × 16.

Figure 6a depicts the human experimental results. There were three key findings:

- Accuracy dropped as speed increased.
- Accuracy was higher for the Track-2-far condition than for the Track-2-near condition, indicating that it was easier to track two targets when they were far from each other.
- Accuracy was the same for Track-2-far and Track-1, indicating that tracking two distant targets was as easy as tracking a single target.

The researchers conducted additional pairwise comparisons to test for the second and third findings at each motion-speed level. The results held at all speeds except the slowest, where there was no difference between conditions. It is likely that there was a ceiling effect at the slowest speed—that is, tracking was so easy that participants were equally accurate even when they were tracking two targets in the same quadrant.

Computational simulation

To evaluate the computational model, we randomly generated MOT videos similar to those used by Shim et al. (2008). These videos match the description in the original article as closely as possible, with three exceptions that have no bearing on the results:

- There is no fixation cue in the center of the display, but gaze is always centered in the model.
- To save simulation time, disks are highlighted for a shorter duration at the beginning and end of each video.
- We use a common color scheme across this simulation and Simulation 2, despite differences in the colors used in the two original experiments. Disks are highlighted in red at the beginning of a trial and in blue at the end.

The original display dimensions are specified in degrees of visual angle, whereas the simulation display dimensions are in pixels. However, all proportions from the original display are preserved. We report speeds in degrees per second rather than pixels per second.

Videos were generated in batches of 240: 3 tracking conditions × 5 motion speeds × 2 correct responses ("match," meaning a highlighted disk matches one of the targets, or "mismatch") × 8. Each batch corresponds to the videos viewed by one human participant. To increase statistical power, we ran the simulation on 36 total batches—equivalent to six times the number of

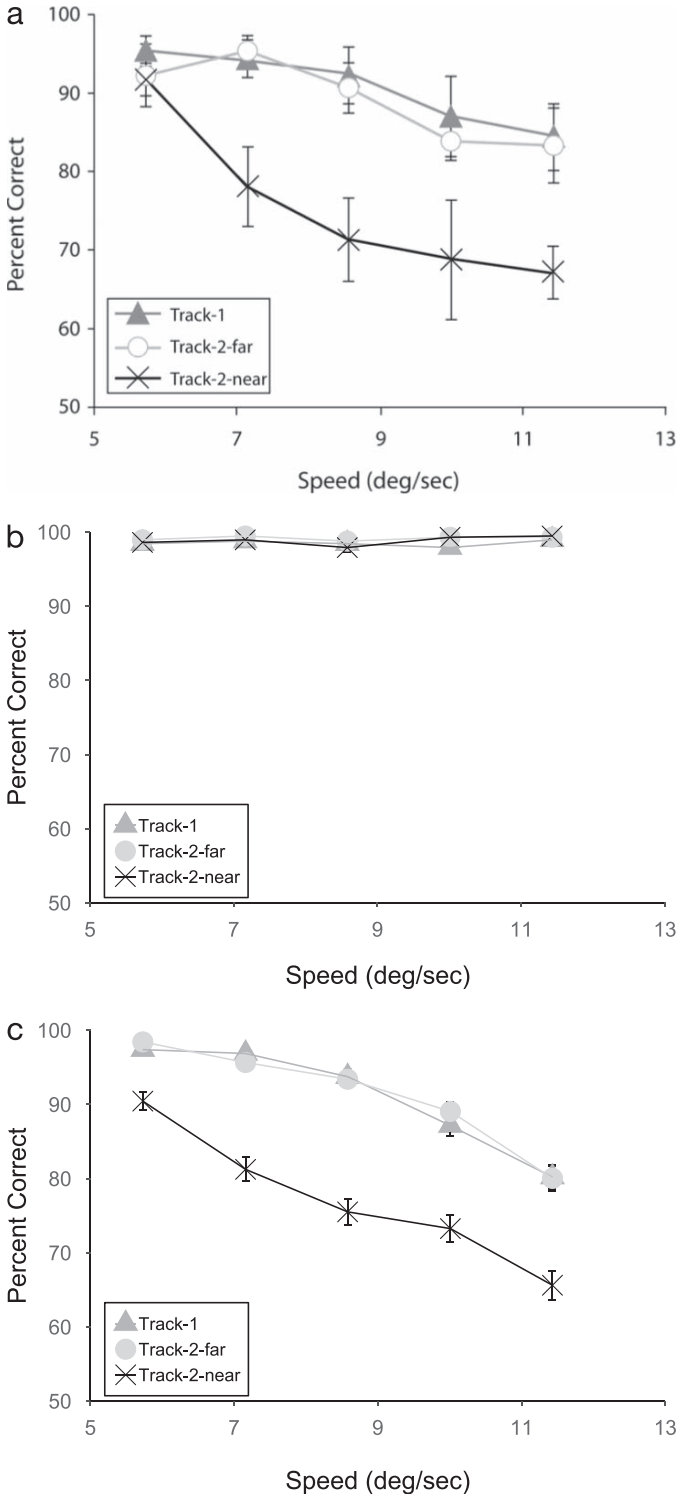

Figure 6. Behavioral results and model results. (a) Humans. (b) Model without noise. (c) Model (noise-center = 0.2, noise-width = 0.15). Error bars are ±1 standard error. Panel (a) reprinted from "Spatial Separation Between Targets Constrains Maintenance of Attention on Multiple Objects," by W. M. Shim, G. A. Alvarez, and Y. V. Jiang, 2008, *Psychological Bulletin Review*, *15*(2), p. 393. Copyright 2008, The Psychonomic Society. Reprinted with permission.

participants in the original study—and analyzed by trial.

Data, code for generating stimulus videos and analyzing results, and a subset of the stimulus videos for Simulations 1 and 2 are available for download at https://osf.io/5m9rh.

Ablation manipulation

For this simulation, we predicted that targets could be tracked via parallel enhancement, without any need to select individual targets. We predicted this because objects never overlapped, and so their locations were sufficient to distinguish targets from distractors in all cases. To test this prediction, we ran the simulation using an ablated version of IMOT, from which THREAT HIGHLIGHTER and MAINTENANCE HIGHLIGHTER were removed. In this simplified version, each target was selected at the beginning of a trial when the targets changed color, but there was no meaningful target selection during object tracking.[11]

Note that after Simulations 1 and 2 were used to calibrate all of IMOT's free parameters, we reran Simulation 1 with the complete model, confirming that the results did not change when THREAT HIGHLIGHTER and MAINTENANCE HIGHLIGHTER were included.

Results without noise

We first tested whether noise is necessary to explain human behavior. To this end, we ran the simulation with OBJECT LOCATOR's noise parameters (noise-center, noise-width) set to zero, meaning that during tracking a proto-object would always be assigned to the closest overlapping enhanced region. Unsurprisingly, the model had near-perfect accuracy, correctly classifying the highlighted disk as a match or mismatch to the targets in almost every video (Figure 6b). This result demonstrates that the model is capable of tracking targets but does not match human results.

Results with noise

To determine the appropriate noise-parameter values, we conducted a series of simulation runs across a set of possible values (see Appendix B). These runs demonstrated that IMOT replicates human results across a wide range of moderate noise values. Here, we report the results with noise-center = 0.20 and noise-width = 0.15, a representative instance which mirrors the human results closely—compare Figure 6a and 6c.

To evaluate the simulation results, we conducted a two-way analysis of variance (ANOVA) with disk

speed and tracking condition as variables. There was a significant main effect of disk speed, $F(4, 8625) = 96.82$, $p < 0.001$, indicating that accuracy dropped as speed increased. There was also a significant main effect of tracking condition, $F(2, 8625) = 174.90$, $p < 0.001$, indicating that difficulty varied across the three tracking conditions. Finally, there was a much smaller, but significant, interaction, $F(8, 8625) = 2.94$, $p = 0.002$.

Following Shim et al. (2008), we conducted t tests for pairs of tracking conditions. Track-2-far accuracy was significantly higher than Track-2-near accuracy, $t(5758) = 14.97$, $p < 0.001$, and this difference held at each speed level (all ps < 0.05). Track-2-far accuracy was not significantly different from Track-1 accuracy, $t(5758) = 0.33$, $p = 0.745$, and this result held at each speed level (all ps > 0.05).

These results replicate the three key findings from the original experiment. The only difference in the simulation analysis was that the advantage for Track-2-far over Track-2-near held even at the slowest speed. In the original experiment, this advantage was eliminated at the slowest speed, suggesting a ceiling effect.

Discussion

As conjectured, IMOT demonstrates accuracy levels similar to those of humans across a range of noise-parameter values. Tracking accuracy declines when objects move faster because there is less overlap between target locations and the enhanced regions marking each target's last-known location. Additionally, accuracy declines when two targets are close to each other because they suppress each other's enhanced regions, which leads to smaller region sizes and, again, less overlap with the target locations.

The primary difference between IMOT and humans is that the model fails to achieve ceiling-level accuracy at the slowest speed. It is possible that humans adjust their tracking strategy when objects are moving particularly slowly. Alternatively, there may be low-level differences between the model and humans in the size or strength of suppressive fields.

IMOT, like humans, tracks two distant targets as easily as a single one. This finding is expected because there is no competition among the targets. They do not compete for space, because they are too distant to suppress each other, and they do not compete for processing time, because they can be tracked via parallel updating of each target location. However, if objects are allowed to overlap each other, then location information should be insufficient for distinguishing targets from distractors. In that case, serial selection will be needed, and so the targets should compete for processing time. To examine this claim, we conducted a second simulation.

Simulation 2

Our second simulation explores competition between targets for serial selection, which assists in distinguishing targets from overlapping distractors (e.g., Figure 2d). Because overlapping objects cannot be differentiated by their locations, a target must be selected before the overlap event begins so that its motion history can be computed. This information can then be used to predict where the target will emerge after the overlap event ends. Notably, because this approach requires selecting a target before it overlaps another object, people should be able to track targets through overlap events more successfully when there are fewer targets competing for selection.

We evaluated this claim by simulating a MOT behavioral experiment by Luu and Howe (2015, experiment 1) in which two-dimensional objects were allowed to move through each other. Participants tracked either two or four out of eight total disks. Disks moved either predictably, changing direction only when they hit the edge of the display; or unpredictably, changing direction randomly. Luu and Howe found that when there were two targets, participants were more accurate with predictable motion, suggesting that they used information about motion histories to aid in tracking. However, when there were four targets, participants were equally accurate with predictable or unpredictable motion, suggesting that they failed to benefit from motion histories.

Experiment description

In their experiment, Luu and Howe (2015) had 15 participants track black disks while maintaining center fixation. On each trial, two or four disks were highlighted in red to indicate that they were the targets. Afterwards, the targets changed back to black, and all eight disks began moving in straight lines, traveling through each other but bouncing off the edges of the display. After 5.5 s, the disks stopped moving and two were highlighted in sequence. Participants indicated whether each highlighted disk was a target or a distractor. As in Simulation 1, the probability that a highlighted disk was a target was always 50%.

On separate trials, there were two motion conditions: *unpredictable*, where each disk changed direction randomly every 300 to 600 ms, and *predictable*, where each disk changed direction only when it hit the edge of the display. Additionally, there were two target-number conditions: two and four. Separate motion speeds for two and four targets were calibrated for each participant by finding the speed at which the participant

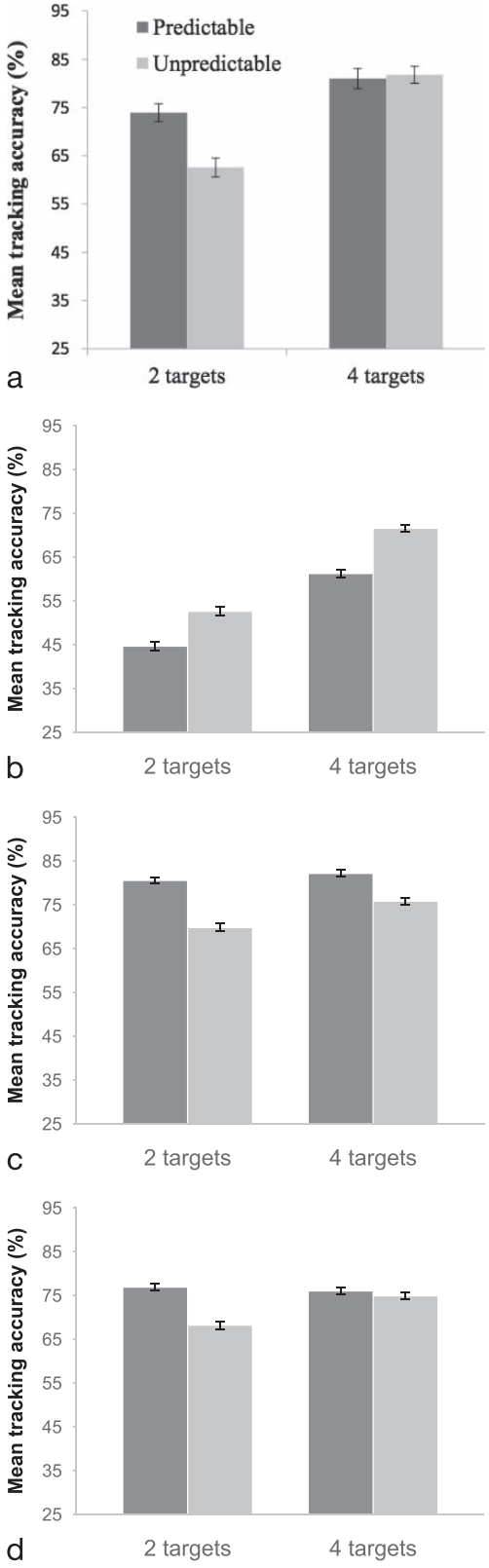

Figure 7. Behavioral results and model results. (a) Humans. (b) Model without selection. (c) Model (max-threat-distance = 0). (d) Model (max-threat-distance = 2.8). Error bars are ±1 standard error. Panel (a) reprinted from "Extrapolation Occurs in Multiple Object Tracking When Eye Movements Are Controlled," by T. Luu and P. D. L. Howe, 2015, *Attention, Perception, & Psychophysics*, 77(6), p. 1924. Copyright 2015, The Psychonomic Society. Reprinted with permission.

achieved 75% accuracy in the predictable motion condition.

Figure 7a depicts Luu and Howe's results. There were three key findings:

- Mean tracking accuracy was higher with predictable motion.
- Predictable motion increased accuracy more for two targets than for four targets.
- The improvement in accuracy was statistically significant for two targets but not for four targets.

Simulation

To evaluate IMOT, we randomly generated videos similar to those used by Luu and Howe (2015). These videos differed from the description in the original article in the three ways discussed in Simulation 1. Additionally, the videos were constrained to begin and end with all disks at least one radius apart, to ensure that there were no overlap events at the beginning or end of a trial.

We generated videos in batches of 120: 2 motion-type conditions × 2 target-number conditions × 30. Each batch corresponded to the videos viewed by one human participant. To increase statistical power, we ran the simulation on 90 total batches—equivalent to six times the number of participants in the original study—and analyzed by trial.

Following the original experiment, the simulation used speeds of 5°/s and 2.5°/s for two and four targets, respectively, because preliminary work indicated that the model achieved about 75% accuracy on predictable motion with these speeds (Lovett, Bridewell, & Bello, 2017). In addition, the simulation used the noise-parameter values from Simulation 1 (noise-center = 0.20, noise-width = 0.15).

Results without selection

We first ran the simulation with the ablated tracking model, which lacked THREAT HIGHLIGHTER and MAINTENANCE HIGHLIGHTER. Because this model does not strategically deploy selection during tracking, it should not benefit from predictable motion trajectories. Figure 7b shows the results. A two-way ANOVA with motion type and target number as variables was conducted. There was a significant main effect of motion type, $F(1, 10795) = 22.55$, $p < 0.001$, indicating that accuracy was

higher with unpredictable motion. There was also a main effect of target number, $F(1, 10795) = 85.24$, $p < 0.001$, indicating that accuracy was higher for four targets. The interaction was not significant, $F(1, 10795) = 1.64$, $p = 0.200$.

These results diverge from the human data in two important ways. First, as expected, accuracy is not higher with predictable motion trajectories. Second, accuracy is far lower for two targets than for four targets (recall that the motion speeds for two targets are twice as fast as the motion speeds for four targets).

Results without a max threat distance

We next ran the simulation with the full model (including THREAT HIGHLIGHTER and MAINTENANCE HIGHLIGHTER), setting THREAT HIGHLIGHTER's parameter max-threat-distance = 0 (Figure 7c). Recall that the model maintains focus on a target that is threatened by nearby objects. When max-threat-distance is set to zero, targets are never classified as threatened, so the model always focuses immediately on whichever target has the closest threat.

A two-way ANOVA with motion type and target number as variables was conducted. There was a significant main effect of motion type, $F(1, 10795) = 19.74$, $p < 0.001$, indicating that accuracy was higher with predictable motion, and a main effect of target number, $F(1, 10795) = 22.57$, $p < 0.001$, indicating that overall accuracy was higher with four targets. There was a significant interaction, $F(1, 10795) = 7.62$, $p = 0.006$, indicating that the advantage with predictable motion was greater for two targets than for four targets.

Following Luu and Howe's (2015) methodology, we conducted t tests comparing predictable with unpredictable motion for each target number. Accuracy was higher with predictable motion both for two targets, $t(5397) = 9.23$, $p < 0.001$, and for four targets, $t(5398) = 5.73$, $p < 0.001$.

Results with a max threat distance

The previous simulation results matched the finding that humans use predictable motion more effectively when they are tracking only two targets. However, unlike humans, the model also showed a benefit for predictable motion when tracking four targets. The model might show this benefit because it can change focus immediately to whichever target has the closest threat, catching overlap events that people might miss. To test for this possibility, we varied the max-threat-distance parameter, which determines how easily the model can change its focus. The simulation replicated the three human findings across a range of parameter values (see Appendix C). Here, we report the results with max-threat-distance = 2.8, a representative instance which closely fits the human results—compare Figure 7a and 7d. Note that the distance threshold is expressed as a multiple of the proto-object's radius, so a target is considered threatened when another object is within 2.8 × the target's radius.

A two-way ANOVA with motion type and target number as variables was conducted. There was a significant main effect of motion type, $F(1, 107955) = 33.91$, $p < 0.001$, indicating that accuracy was higher with predictable motion, and a main effect of target number, $F(1, 10795) = 11.97$, $p < 0.001$, indicating that overall accuracy was higher with four targets. There was a significant interaction, $F(1, 10795) = 20.29$, $p < 0.001$, indicating that the advantage with predictable motion was greater for two targets than for four targets.

We conducted t tests comparing predictable with unpredictable motion for each target number. Accuracy was higher with predictable motion for two targets, $t(5397) = 7.19$, $p < 0.001$. However, there was no significant difference between predictable and unpredictable motion for four targets, $t(5398) = 0.95$, $p = 0.343$.

Replication of Simulation 1

Recall that Simulation 1 was run with the ablated model, which lacked the ability to select targets and compute motion history during tracking. Including this ability should not change the results, because in the present model, selection helps only when targets cannot be distinguished by their location. To test this claim, we reran Simulation 1 using the complete model with max-threat-distance = 2.8. As expected, we replicated our previous findings, with all statistical test results coming out the same.

Discussion

As anticipated, IMOT successfully replicates the human results. When the model tracks two targets, accuracy is higher with predictable motion than with unpredictable motion. When it tracks four targets, the advantage with predictable motion is weakened, and the advantage is eliminated entirely with a sufficiently large max-threat-distance (Figure 7d).

The results demonstrate the importance of selection for explaining human results. When the model does not deploy selection to aid in tracking (Figure 7b), its accuracy does not improve with predictable motion. Interestingly, IMOT is *less* accurate with predictable motion in this case. We speculate that objects may cluster more closely when they are not randomly

changing their directions of motion, resulting in more overlap events that cannot be handled without selection.

When selection is not deployed, we find also that accuracy when tracking two targets falls far below accuracy when tracking four. This result is unsurprising, given that objects in the two-target videos moved considerably faster than objects in the four-target videos. The finding that two-target tracking catches up to four-target tracking when selection is deployed (Figure 7c) demonstrates that selection is more valuable when only two targets are competing.

When max-threat-distance = 0 (Figure 7c), four-target tracking shows some benefit for predictable motion, indicating that it is possible to track targets through overlap events even when there are four targets. However, strategic deployment of selection is less effective when there are more targets. When max-threat-distance increases to 2.8 (Figure 7d), the four-target advantage for predictable motion vanishes, whereas the two-target advantage remains.

In the present model, increasing max-threat-distance makes it more difficult to switch focus among targets—a selected target will appear threatened for longer, and so it will remain the focus for longer. We concede that a similar effect might be achieved in other ways, for example by modeling a general resistance to switch focus among targets. Thus, the results do not conclusively demonstrate that selection lingers on threatened targets; rather, they suggest only that participants do not immediately change focus to the target with the closest threat.

Lastly, we wish to stress an important point: The model may receive *some* benefit from selection even when tracking four targets. If we compare the results without selection (Figure 7b) to those with selection (Figure 7d), performance with four targets does appear to improve, especially in the predictable motion condition (61% without selection, 76% with selection). Given this finding, we believe it is possible that humans sometimes deploy selection to track targets through overlap events, even when there are four targets. However, four-target tracking performance with predictable motion fails to exceed performance with unpredictable motion due to a combination of two factors: In the absence of selection, predictable motion may be more difficult than unpredictable motion (Figure 7b); and the benefit for selection among four targets is limited, because each target competes for selection and there is a tendency to linger on a selected target.

General discussion

The diversity of Simulations 1 and 2 demonstrates the breadth of IMOT's explanatory power. By integrating parallel enhancement with serial selection, this model accounts for both spatial and temporal constraints on object tracking. Notably, it tracks moving targets in parallel via enhancement, but targets can crowd each other, resulting in smaller enhanced regions and a greater chance of dropping targets. At the same time, the model can select a target to process its motion history and predict its future location, but as the number of targets increases, there is more competition among them for selection.

In the following sections, we compare IMOT to other models of MOT and then discuss how fully the model explains the findings presented in the Background. Finally, we consider predictions about object tracking that follow from the model.

Previous models

Previous computational MOT models fall into three categories: serial tracking, parallel tracking, and hybrid. Despite their strengths, none of them can explain human performance on the two experiments simulated in this article because in these models there is no crowding between targets, and motion histories are not used to track targets through overlap events.

Serial tracking

A serial tracking model attends to each target in sequence to update that target's location. Oksama and Hyönä (2008) propose such a model for multiple-identity tracking, a task similar to MOT where each object has a distinct visual appearance (e.g., a different shape). Their MOMIT (Model of Multiple Identity Tracking), which is described but not implemented, uses a spatial index to point to each target's last-known location. As objects move, MOMIT serially reattends to each target to select the current object that best matches that target's spatial index. When MOMIT selects the wrong object as the target, it can detect and correct the mistake by noticing that the selected object's visual appearance does not match that of the target. Notably, a recent update to MOMIT incorporates parallel processing of visual features, but still requires that targets be serially reattended to refresh their representations (Li, Oksama, & Hyönä, 2019).

In contrast with MOMIT's serial updating, IMOT updates all target locations in parallel, so each location is updated more frequently and there should be fewer tracking errors. A reduction in tracking errors is important because the objects in a MOT task are visually identical, which means that IMOT cannot use mismatched visual appearances to correct errors.

Parallel tracking

A parallel tracking model applies the same tracking mechanism to each target, without any serial component. Kazanovich and Borisyuk's (2006) model uses oscillating artificial neurons to mark target locations. For each target, there is a specialized neuron that fires in synchrony with the neurons representing that target's location. Importantly, unlike IMOT, this model does not process motion histories, and hence cannot reliably distinguish a target from a distractor when the two overlap. Instead, after an overlap event ends, the model arbitrarily chooses either the target or the distractor and continues tracking that object.

Zhong, Wilson, and Flombaum (2014) evaluate the benefits of using motion histories in a Bayesian computational model. At each time step, this model is given noisy measurements of every object's location and determines which objects are targets by comparing the noisy measurements to the targets' expected locations. A target's expected location is a weighted average of two values: the noisy measurement of its location from the previous time step and its predicted location based on its past motion history. Notably, the model predicts locations for all targets in parallel—up to eight targets are used—despite evidence that humans fail to predict locations for more than two targets (Howe & Holcombe, 2012; Luu & Howe, 2015).

Zhong et al. (2014) find that their model works best when little weight is allocated to predicted locations, even when objects move predictably and can travel through each other. They conclude that motion information provides minimal value during tracking. However, we would argue that their model fails to benefit from motion information because it does not need motion histories to disambiguate overlapping objects. The model is always provided with separate location measurements for each object, even when two objects overlap. In contrast, IMOT operates directly on visual input from a video and is unable to distinguish the locations of two overlapping objects through visual appearance alone.

Hybrid tracking

Similar to IMOT, hybrid models track targets in parallel but can allocate attention to where it is needed most. Srivastava and Vul (2016) describe a Bayesian, hybrid model based on the attentional-resource account of object tracking (Alvarez & Franconeri, 2007). Their model distributes an attentional resource among the targets, dynamically allocating more of it to targets that are threatened by nearby objects. As the amount of attention assigned to a target increases, the model correspondingly lowers its uncertainty about the target's position, meaning that the target is less likely to be confused with its neighbors. However, the model is primarily descriptive of behavior, offering no account for why allocating more of the resource should decrease positional uncertainty in humans.

Critically, the Srivastava and Vul (2016) model does not carry out all the necessary steps for object tracking. Rather than determining which objects correspond to the targets, it receives this information as input and generates only predictions about the likelihood of confusing targets with distractors. In contrast, IMOT solves the correspondence problem directly.

Explaining key findings in MOT

The Background section introduced five classes of MOT findings that any model would ideally address. Simulations 1 and 2 demonstrate IMOT's ability to explain three of these: spatial constraints, dynamic operation, and sensitivity to motion history. Here, we discuss each of these and then consider how IMOT could help explain a fourth class, temporal constraints.

Spatial constraints

IMOT's tracking accuracy decreases when targets can move near each other due to crowding (Holcombe et al., 2014; Shim et al., 2008). More specifically, each target's location is marked with an enhanced region, and enhancing a target's location results in suppressing the surrounding area. Neighboring enhanced regions will mutually suppress each other, resulting in smaller regions and a greater chance that targets—especially fast-moving ones—will be dropped.

Dynamic operation

Although IMOT enhances target locations in parallel, it can select a single target as the focus of attention. Notably, serial selection is strategic, picking out targets that are threatened by nearby objects and in danger of becoming lost (Iordanescu et al., 2009; Srivastava & Vul, 2016).

Sensitivity to motion history

Selecting a threatened target is particularly useful when objects move predictably, such that the target's motion history provides reliable information about where the target will be in the near future. However, IMOT shows a diminished advantage for predictable motion when the number of targets increases (Howe & Holcombe, 2012; Luu & Howe, 2015), because there is more competition for selection, and thus a threatened target is less likely to be selected in a timely manner.

Temporal constraints

The two experiments simulated in this article involve targets moving in straight lines, occasionally changing directions. In contrast, temporal constraints often have been studied using targets that follow circular trajectories. Recall that there are three key findings from this research: There is a maximum speed of about 2 rotations/s for tracking circling targets (Holcombe & Chen, 2012; Verstraten et al., 2000), there is a maximum temporal frequency of about 7 Hz (Holcombe & Chen, 2013), and the maximum speed and temporal frequency both decrease as the number of targets increases.

In explaining the circling-target research, we wish to highlight two key differences between this research and our Simulation 1 (Shim et al., 2008). Firstly, with circling targets there is a cost for increasing the number of targets from one to two, whereas in Simulation 1 there is no cost, provided the targets are far apart. Secondly, circling targets follow a highly predictable trajectory, whereas in Simulation 1 the targets move unpredictably. Note that the circling targets change direction at unpredictable intervals, but there are only two possible directions along the circle, and the act of changing direction itself may draw attention to a target. In contrast, in Simulation 1 the objects are constantly changing direction in unpredictable ways to avoid collisions.

In light of these differences—and considering Simulation 2, wherein serial selection provides more value when targets moved predictably—we propose that participants serially select circling targets, using their predictable trajectories to anticipate where they will go next. In this case, selection does not aid in distinguishing targets from overlapping distractors; instead, it aids in distinguishing targets from distractors that occupy positions recently occupied by the targets. Critically, because viewers are attempting to predict a target's future location along a circular trajectory, tracking is limited by rotational speed. At above 2 rotations/s, prediction may become impossible either because the speed is too great to be processed or because targets outpace the fastest possible predictions. Finally, because selection must shift serially between targets, tracking becomes more difficult as the number of circling targets increases.

Future work

Despite IMOT's explanatory breadth, the model does not address a key class of MOT findings: hemifield advantages. Here we consider how the model could be refined to explain these advantages and other phenomena that lie outside of the scope of the current work.

Hemifield advantages

IMOT is insensitive to the hemifield in which objects are located, but research suggests there is a considerable advantage to tracking targets that are split across hemifields (Alvarez & Cavanagh, 2004; Chen et al., 2013; Shim et al., 2010). Hemifield advantages could be addressed in several ways, including limiting suppressive surrounds to affect only other targets in the same hemifield (Chelazzi, Miller, Duncan, & Desimone, 1993), processing some information such as motion histories in parallel across the two hemifields, or even selecting separate foci of attention in each hemifield (Holcombe & Chen, 2012). Additional simulation studies would be required to determine how well these manipulations capture human hemifield advantages.

Target–distractor proximity

Although IMOT explains why tracking becomes difficult when targets are close to each other, it does not explain the related finding that tracking becomes difficult when targets are close to distractors (Shim et al., 2008; Vater et al., 2017). There are at least two possible ways that nearby distractors may disrupt tracking. First, distractors may crowd targets with their own suppressive surrounds, resulting in a smaller enhanced region around a target and a greater chance that the target will be dropped. Second, distractors may be swapped accidentally with a target. Further research is needed to evaluate to what extent each of these occurs, as most MOT studies cannot distinguish between dropping errors and swapping errors (but see Drew, Horowitz, & Vogel, 2013; Pylyshyn & Storm, 1988). In the meantime, each effect could be modeled via changes in OBJECT LOCATOR. To model drops resulting from nearby distractors, suppressed regions could be placed around distractors in the model's priority map; note that the suppressed region around a distractor should likely have a lower amplitude than the suppressed region around a target (Desimone & Duncan, 1995; Reynolds & Heeger, 2009; Tsotsos, 1990). To model swaps, noise could be added to the process that matches enhanced regions to proto-objects, such that any distractor that is close enough to overlap a target's enhanced region has a chance of being treated as the target.

Spatial precision in the periphery

IMOT's suppressed regions become wider as objects move farther from the center of gaze, simulating the increased distance over which objects can crowd each other (Whitney & Levi, 2011). However, humans exhibit a general lack of spatial precision when perceiving objects far from the center of gaze, even when those objects are not crowded (DeValois &

DeValois, 1990; Levi, Klein, & Yap, 1987). This phenomenon could be modeled by scaling the noise applied when OBJECT LOCATOR matches enhanced regions to proto-objects, such that the noise increases with the proto-object's distance from the center of gaze.

VSTM capacity

IMOT's VSTM has slots for four objects, consistent with classic views on VSTM capacity (Pylyshyn & Storm, 1988; Vogel et al., 2001). However, more recent work challenges these views, suggesting that VSTM capacity may vary depending on the amount of information stored for each object (Bays & Husain, 2008; van den Berg, Shin, Chou, George, & Ma, 2012; Wilken & Ma, 2004). As further evidence against hard capacity limits, Alvarez and Franconeri (2007) have shown that people can track up to eight targets when they move slowly and are positioned far apart from each other. Presently, we take a noncommittal position on VSTM capacity. Because tracking four targets is sufficient for the simulations reported here, we prefer to use a fixed-slot model that lacks unnecessary complexity. We may develop a variable-capacity model in the future, should new simulations require it.

Predictions

A key strength of our computational model is that we can make testable predictions about object tracking that go beyond the claims tested in this article's simulations. These predictions follow from the underlying theory, the model's implementation details, and the findings from the two simulations. Here we present five predictions, two about the MOT task in isolation and three about the interactions between object tracking and other attentionally demanding tasks. By exploring such interactions, we hope to come closer to understanding how attention is deployed during the complex, dynamic, multitask situations that humans face in their everyday lives.

Two targets will compete for selection when they are threatened simultaneously

The results from Simulation 2 suggest that serial selection can be strategically deployed to a threatened target, but that shifting the focus to a new target takes time. Thus, even when only two targets are tracked, tracking should suffer if one target draws selection shortly before the other target requires selection to distinguish it from an overlapping distractor. The first target might draw selection because it is threatened or for some other reason, for example because there is a salient flash of light at the target's location.

Note that the example of circling targets suggests two targets may compete for selection even when neither is threatened. In this case, selection—which facilitates tracking fast-moving targets along predictable trajectories—may cycle between the targets.

Motion speed and target–target distance will interact

In the present model, targets will be dropped more frequently if either speed increases or target–target distance decreases; in either case, there will be less overlap between an enhanced region at a target's previous location and a proto-object at the target's current location. The risk of losing targets is greatest when speeds are high and, at the same time, target–target distance is small. Therefore, we can predict that tracking performance should be higher if targets move slowly when they are near each other and quickly when far apart, compared to the reverse case where they move quickly when they are near each other and slowly when they are far apart.

Targets will compete with a nonvisual dual task for selection

We now consider how object tracking interacts with other tasks. IMOT and the ARCADIA framework more broadly suggest that all attentionally demanding tasks should compete for selection because there can be only one focus at a time. Therefore, just as one might fail to select a target before an overlap event because another target draws selection, one might similarly fail because a simultaneously performed, nonvisual task draws selection.

Competition from a nonvisual dual task was previously explored by Tombu and Seiffert (2008), who asked participants to judge the pitch of a tone played during a MOT task. As a second manipulation, at some time during the MOT task, tracking would briefly be made more difficult by speeding up the objects or moving them closer together. The key finding was that if the tone was played immediately before tracking difficulty increased, tracking performance suffered, compared to a control where the tone was played long before the MOT task became more difficult. The experimenters concluded that judging the pitch of the tone required attentional resources that otherwise could be devoted to tracking the targets during the more difficult period.

IMOT is able to make a more fine-grained prediction than the findings in the Tombu and Seiffert (2008) experiments: A nonvisual dual task should hamper tracking performance if it occurs specifically when one of the targets requires selection. At the same time, the model is unable to fully explain Tombu and Seiffert's results because its selection mechanism does not aid in

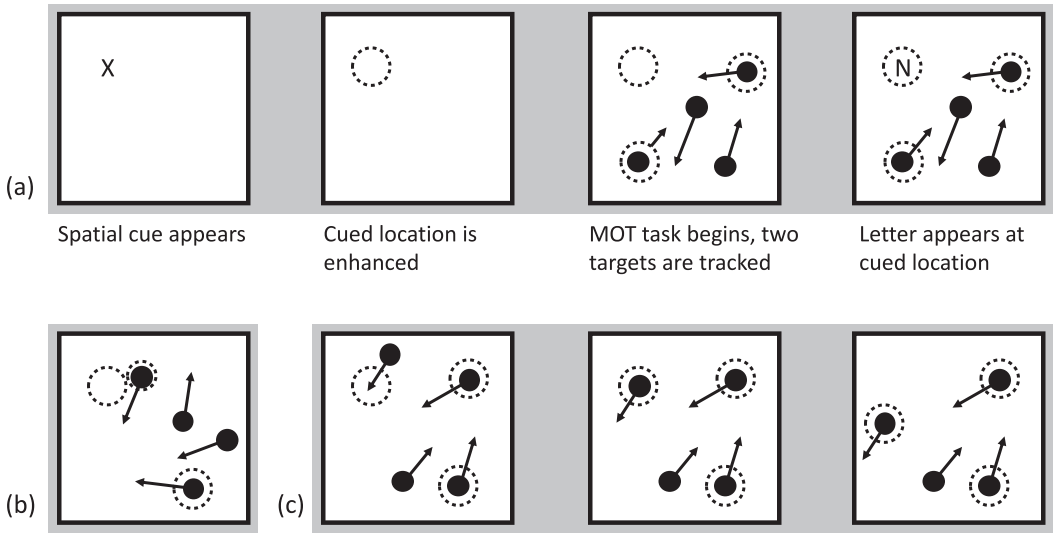

Figure 8. (a) A hypothetical spatial dual task, in which participants must remember a cued location and respond when a letter appears at that location. Dotted circles indicate enhanced regions. (b) When a target moves near the cued location, its enhanced region is suppressed. (c) When a distractor moves through the cued location, the enhanced region becomes attached to the distractor.

tracking quickly moving targets or in distinguishing targets from neighboring distractors. The example of circling targets suggests that, at a minimum, selection should aid in tracking fast-moving targets that follow predictable trajectories.

A visual dual task can cause targets to be suppressed

Because object tracking relies on enhancement, a general attentional mechanism, other visual tasks can disrupt tracking. For example, suppose participants perform a spatial-attention task, in which they are cued to expect a stimulus at a particular location and they must respond when the appropriate stimulus appears at that location. While they are waiting for the stimulus to appear, they must perform a MOT task (Figure 8a; dotted circles indicate enhanced regions). Performing these two tasks in parallel should be possible—one enhanced region can be assigned to the cued location and other enhanced regions can be assigned to the tracked targets. However, because there is a suppressive surround around each enhanced region, targets that move near the cued location would have their enhanced region suppressed, just as occurs when targets move near each other (Figure 8b). These targets would be dropped more frequently than targets that did not move near the cued location.

A visual dual task can cause distractors to be enhanced

Consider again a spatial-attention dual task, in which the region around a cued location is enhanced. If a distractor moves through this enhanced region, then the region may become associated with that distractor—this can occur because the parallel updating mechanism associates enhanced regions with matching proto-objects. If the enhanced region becomes associated with the distractor, then it should begin following the distractor as that object moves (Figure 8c). Essentially, the distractor would become a tracked object, just like the targets.

If a distractor becomes a tracked object, there should be two measurable effects. First, within the MOT task, participants should be unable to distinguish the distractor from the targets. Second, within the spatial-attention task, participants should respond slowly when a stimulus appears at the cued location, because the cued location is no longer being enhanced.

Conclusion

IMOT demonstrates the importance of both parallel and serial mechanisms in object tracking. Parallel updating tracks multiple targets' positions as they move, and serial selection provides information about a single target's motion history, letting the model track that target when it overlaps another object. With these mechanisms integrated, the model can account for the key findings from two MOT experiments, neither of which can be explained by existing computational models of object tracking.

First, tracking accuracy worsens when two targets are near each other or when motion speed increases (Shim et al., 2008). In Simulation 1, both manipulations cause targets to be dropped because they decrease the overlap between a target and the enhanced region marking its last-known location. Neighboring targets

suppress each other's enhanced regions, whereas faster motion speeds increase the distance a target moves from its last-known location each cycle. Additionally, faster motion speeds provide more opportunities for targets to crowd each other in videos with fixed durations (Franconeri et al., 2008).

Second, tracking accuracy improves when objects move predictably, but only when two targets are tracked (Luu & Howe, 2015). In Simulation 2, selection enables the processing of a target's motion history and the use of this information to track the target through an overlap event. Importantly, targets that are at risk of overlapping a nearby object can be prioritized for selection in the model. But when more than two targets are tracked, there is heavy competition among the targets, which causes many target overlap events to go unnoticed.

Because IMOT's serial and parallel tracking mechanisms are based on selection and enhancement, the model allows us to make explicit claims about attention's role in object tracking and testable predictions about when and how targets will compete for attention. When targets can be tracked using only their locations, they will compete for space because neighboring targets crowd each other. When additional information is needed to track targets, the targets will also compete for processing time, because serial selection is needed to process this information. The nature of this second competition will depend on the tracking demands—if only one target needs to be processed at a time, then competition will be minimal because selection can be strategically deployed to where it is most needed, but if multiple targets must be processed simultaneously, or if it is unclear which target needs to be processed, then competition will be heavy, and tracking accuracy will suffer.

In addition to showing how selection and enhancement support object tracking, IMOT demonstrates more generally how these mechanisms interact. After an object is selected for further processing, the region around that object is enhanced, which increases the chance that it or a nearby object will be selected in the future. We conjecture that this interaction lies at the heart of visual attention, across its many forms and the myriad tasks that have been associated with it. For example, just as spatial regions can be enhanced, visual features like color can be enhanced, such that individuals show greater sensitivity to objects that possess those features (Bichot, Rossi, & Desimone, 2005; Maunsell & Treue, 2006; Yu, Levinthal, & Franconeri, 2017). Interestingly, Huang (2010) discovered a brief delay between the times when an object is selected and when its features are enhanced, suggesting that featural enhancement, like spatial enhancement, follows selection. Future efforts will explore the limits of our theoretical account, both in vision and across cognition more broadly.

Keywords: attention, multiple-object tracking, computational modeling

Acknowledgments

This research was performed while AL held an NRC Research Associateship award at the U.S. Naval Research Laboratory. The authors acknowledge support from the Office of Naval Research under Grants N0001417WX00153, N0001417WX00904, and N0001417WX01804. The views expressed in this article are solely the authors' and should not be taken to reflect any official policy or position of the United States Government or the Department of Defense.

The authors thank Steve Franconeri and Alex Holcombe for their helpful suggestions and feedback.

Commercial relationships: none.
Corresponding author: Andrew Lovett.
Email: andrew.lovett@nrl.navy.mil.
Address: U.S. Naval Research Laboratory, Washington, DC, USA.

Footnotes

[1] https://osf.io/5m9rh.

[2] For clarity, we use small caps whenever we refer to a component's name.

[3] For simplicity, motion history is computed as a change in location from one cycle to the next. Note that we do not claim selection is required to compute this information in humans, only that selection makes this information available for extrapolation.

[4] This approach is task specific. Cues other than size might be used to detect overlapping objects or three-dimensional occlusion in other tasks. For example, if there is an outline around each object, then T-junctions are a good indicator that one object is occluding another (Viswanathan & Mingolla, 2002).

[5] Overlap events are treated as ongoing for two additional cycles after they appear to have ended. This step is necessary because sometimes an overlap event appears to end early when two objects overlap perfectly, such that they produce a proto-object the size of just one object.

[6] For example, imagine tracking birds in flight while their shapes in the visual field change based on direction of flight and bodily movement. As a result,

location will play a key role in differentiating the birds from each other and potential distractors.

[7] The term *priority map* is sometimes seen as guiding where a person will move their eyes in a scene. However, it is well understood that overt attention (eye movements) typically follows covert attention (Deubel & Schneider, 1996; Peterson, Kramer, & Irwin, 2004). If we take the view that enhancing regions of interest is a form of covert attention, then the term *priority map* is appropriate.

[8] Mexican-hat functions have often been used to approximate center–surround suppression in vision (e.g., Kang, Shelley, & Sompolinsky, 2003; Müller, Mollenhauer, Rösler, & Kleinschmidt, 2005).

[9] A central gaze fixation is encouraged in many MOT experiments, including the experiment modeled in this article's Simulation 1. In contrast, it is directly enforced with an eye tracker in other experiments, including the experiment modeled in Simulation 2. Note that when the fixation is merely encouraged, participants might move their eyes, for example to follow the centroid of the group of targets (Fehd & Seiffert, 2008).

[10] A highlighting color can be specified when the component is instantiated. In the present work, objects at the end of MOT trials are highlighted in blue.

[11] The simplified model did select targets at random during object tracking, but selection incurred no benefit to tracking.

[12] For the sake of clarity, some items have been given different names from the names they take in the actual code.

References

Alvarez, G. A., & Cavanagh, P. (2004). Independent attention resources for the left and right visual hemifields. *Journal of Vision*, *4*(8): 29, https://doi.org/10.1167/4.8.29. [Abstract]

Alvarez, G. A., & Franconeri, S. L. (2007). How many objects can you track? Evidence for a resource-limited attentive tracking mechanism. *Journal of Vision*, *7*(13):14, 1–10, https://doi.org/10.1167/7.13.14. [PubMed] [Article]

Awh, E., & Pashler, H. (2000). Evidence for split attentional foci. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 834–846, https://doi.org/10.1037/0096-1523.26.2.834.

Bays, P. M., & Husain, M. (2008, August 8). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851–854, https://doi.org/10.1126/science.1158023.

Bettencourt, K. C., Michalka, S. W., & Somers, D. C. (2011). Shared filtering processes link attentional and visual short-term memory capacity limits. *Journal of Vision*, *11*(10):22, 1–9, https://doi.org/10.1167/11.10.22. [PubMed] [Article]

Bichot, N. P., Rossi, A. F., & Desimone, R. (2005, April 22). Parallel and serial neural mechanisms for visual search in macaque area v4. *Science*, *308*(5721), 529–534, https://doi.org/10.1126/science.1109676.

Bisley, J. W., & Goldberg, M. E. (2010). Attention, intention, and priority in the parietal lobe. *Annual Review of Neuroscience*, *33*, 1–21, https://doi.org/10.1146/annurev-neuro-060909-152823.

Bridewell, W., & Bello, P. F. (2015). Incremental object perception in an attention-driven cognitive architecture. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 279–284). Pasadena, CA: Cognitive Science Society.

Bridewell, W., & Bello, P. (2016). A theory of attention for cognitive systems. In K. Forbus, T. Hinrichs, & C. Ost (Eds.), *Fourth Annual Conference on Advances in Cognitive Systems* (Vol. 4, pp. 1–16). Evanston, IL: Cognitive Systems Foundation.

Briggs, G., Bridewell, W., & Bello, P. (2017). A computational model of the role of attention in subitizing and enumeration. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 1672–1677). London, UK: Cognitive Science Society.

Carlson, T. A., Alvarez, G. A., & Cavanagh, P. (2007). Quadrantic deficit reveals anatomical constraints on selection. *Proceedings of the National Academy of Sciences, USA*, *104*(33), 13496–13500, https://doi.org/10.1073/pnas.0702685104.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, *9*(7), 349–354, https://doi.org/10.1016/j.tics.2005.05.009.

Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, *363*(6427), 345–347, https://doi.org/10.1038/363345a0.

Chen, W.-Y., Howe, P. D., & Holcombe, A. O. (2013). Resource demands of object tracking and differential allocation of the resource. *Attention, Perception, & Psychophysics*, *75*(4), 710–725, https://doi.org/10.3758/s13414-013-0425-1.

Dakin, S. C., Bex, P. J., Cass, J. R., & Watt, R. J. (2009). Dissociable effects of attention and crowding on orientation averaging. *Journal of Vision*,

9(11):28, 1–16, https://doi.org/10.1167/9.11.28. [PubMed] [Article]

Dawson, M. R. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review*, *98*(4), 569–603, https://doi.org/10.1037/0033-295X.98.4.569.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*(1), 193–222, https://doi.org/10.1146/annurev.ne.18.030195.001205.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837, https://doi.org/10.1016/0042-6989(95)00294-4.

DeValois, R. L., & DeValois, K. K. (1990). *Spatial vision*. New York: Oxford University Press.

Doran, M. M., & Hoffman, J. E. (2010). The role of visual attention in multiple object tracking: Evidence from ERPs. *Attention, Perception, & Psychophysics*, *72*(1), 33–52, https://doi.org/10.3758/APP.72.1.33.

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, https://doi.org/10.1016/j.cognition.2012.10.003.Swapping.

Drew, T., McCollough, A. W., Horowitz, T. S., & Vogel, E. K. (2009). Attentional enhancement during multiple-object tracking. *Psychonomic Bulletin & Review*, *16*(2), 411–417, https://doi.org/10.3758/PBR.16.2.411.

Drew, T., & Vogel, E. K. (2008). Neural measures of individual differences in selecting and tracking multiple moving objects. *The Journal of Neuroscience*, *28*(16), 4183–4191, https://doi.org/10.1523/JNEUROSCI.0556-08.2008.

Egeth, H., Virzi, R., & Garbart, H. (1984). Searching for conjunctively defined targets. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(1), 32–39, https://doi.org/10.1037/0096-1523.10.1.32.

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting visual attention between objects and locations: Evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*(2), 161–177, https://doi.org/10.1037/0096-3445.123.2.161.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*(4), 225–240, https://doi.org/10.3758/BF03211502.

Fecteau, J. H., & Munoz, D. P. (2006). Salience, relevance, and firing: A priority map for target selection. *Trends in Cognitive Sciences*, *10*(8), 382–390, https://doi.org/10.1016/j.tics.2006.06.011.

Fehd, H. M., & Seiffert, A. E. (2008). Eye movements during multiple object tracking: Where do participants look? *Cognition*, *108*(1), 201–209, https://doi.org/10.1016/j.cognition.2007.11.008.

Franconeri, S. L., Jonathan, S. V., & Scimeca, J. M. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science*, *21*(7), 920–925, https://doi.org/10.1177/0956797610373935.

Franconeri, S. L., Lin, J. Y., Pylyshyn, Z. W., Fisher, B., & Enns, J. T. (2008). Evidence against a speed limit in multiple-object tracking. *Psychonomic Bulletin & Review*, *15*(4), 802–808, https://doi.org/10.3758/PBR.15.4.802.

Holcombe, A. O., & Chen, W.-Y. (2012). Exhausting attentional tracking resources with a single fast-moving object. *Cognition*, *123*(2), 218–228, https://doi.org/10.1016/j.cognition.2011.10.003.

Holcombe, A. O., & Chen, W.-Y. (2013). Splitting attention reduces temporal resolution from 7 Hz for tracking one object to <3 Hz when tracking three. *Journal of Vision*, *13*(1):12, 1–19, https://doi.org/10.1167/13.1.12. [PubMed] [Article]

Holcombe, A. O., Chen, W.-Y., & Howe, P. D. L. (2014). Object tracking: Absence of long-range spatial interference supports resource theories. *Journal of Vision*, *14*(6):1, 1–39, https://doi.org/10.1167/14.6.1. [PubMed] [Article]

Horowitz, T. S., & Cohen, M. A. (2009). Direction information in multiple object tracking is limited by a graded resource. *Attention, Perception & Psychophysics*, *72*(7), 1765–1775, https://doi.org/10.3758/APP.72.7.1765.

Howard, C. J., Masom, D., & Holcombe, A. O. (2011). Position representations lag behind targets in multiple object tracking. *Vision Research*, *51*(17), 1907–1919, https://doi.org/10.1016/j.visres.2011.07.001.

Howe, P. D. L., & Holcombe, A. O. (2012). Motion information is sometimes used as an aid to the visual tracking of objects. *Journal of Vision*, *12*(13):10, 1–10, https://doi.org/10.1167/12.13.10. [PubMed] [Article]

Huang, L. (2010). The speed of feature-based attention: Attentional advantage is slow, but selection is fast. *Journal of Experimental Psychology: Human Per-*

ception and Performance, *36*(6), 1382–1390, https://doi.org/10.1037/a0018736.

Hudson, C., Howe, P. D. L., & Little, D. R. (2012). Hemifield effects in multiple identity tracking. *PLoS One*, *7*(8), 1–8, https://doi.org/10.1371/journal.pone.0043796.

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, 1–12, https://doi.org/10.1167/9.4.1. [PubMed] [Article]

Kang, K., Shelley, M., & Sompolinsky, H. (2003). Mexican hats and pinwheels in visual cortex. *Proceedings of the National Academy of Sciences, USA*, *100*(5), 2848–2853, https://doi.org/10.1073/pnas.0138051100.

Kazanovich, Y., & Borisyuk, R. (2006). An oscillatory neural model of multiple object tracking. *Neural Computation*, *18*(6), 1413–1440, https://doi.org/10.1162/neco.2006.18.6.1413.

Kool, W., Conway, A. R. A., & Turk-Browne, N. B. (2014). Sequential dynamics in visual short-term memory. *Attention, Perception, & Psychophysics*, *76*(7), 1885–1901, https://doi.org/10.3758/s13414-014-0755-7.

Kramer, A. F., & Hahn, S. (1995). Splitting the beam: Distribution of attention over noncontiguous regions of the visual field. *Psychological Science*, *6*(6), 381–386, https://doi.org/10.1111/j.1467-9280.1995.tb00530.x.

Kunar, M. A., Carter, R., Cohen, M., & Horowitz, T. S. (2008). Telephone conversation impairs sustained visual attention via a central bottleneck. *Psychonomic Bulletin & Review*, *15*(6), 1135–1140, https://doi.org/10.3758/PBR.15.6.1135.

Lee, D., & Chun, M. M. (2001). What are the units of visual short-term memory, objects or spatial locations? *Perception & Psychophysics*, *63*, 253–257.

Levi, D. M., Klein, S. A., & Yap, Y. L. (1987). Positional uncertainty in peripheral and amblyopic vision. *Vision Research*, *27*(4), 581–597, https://doi.org/10.1016/0042-6989(87)90044-7.

Li, J., Oksama, L., & Hyönä, J. (2019). Model of multiple identity tracking (MOMIT) 2.0: Resolving the serial vs. parallel controversy in tracking. *Cognition*, *182*, 260–274, https://doi.org/10.1016/j.cognition.2018.10.016.

Lovett, A., Bridewell, W., & Bello, P. (2017). Goal-directed deployment of attention in a computational model: A study in multiple-object tracking. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society* (pp. 2640–2645). London, UK: Cognitive Science Society.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Luu, T., & Howe, P. D. L. (2015). Extrapolation occurs in multiple object tracking when eye movements are controlled. *Attention, Perception, & Psychophysics*, *77*(6), 1919–1929, https://doi.org/10.3758/s13414-015-0891-8.

Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322, https://doi.org/10.1016/j.tins.2006.04.001.

McMains, S. A., & Somers, D. C. (2004). Multiple spotlights of attentional selection in human visual cortex. *Neuron*, *42*(4), 677–686, https://doi.org/10.1016/S0896-6273(04)00263-6.

Meyerhoff, H. S., Papenmeier, F., & Huff, M. (2017). Studying visual attention using the multiple object tracking paradigm: A tutorial review. *Attention, Perception, & Psychophysics*, *79*(5), 1255–1274, https://doi.org/10.3758/s13414-017-1338-1.

Meyerhoff, H. S., Papenmeier, F., Jahn, G., & Huff, M. (2016). Not flexible enough: Exploring the temporal dynamics of attentional reallocations with the multiple object tracking paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(6), 776–787, https://doi.org/10.1037/0096-1523.10.5.601.

Müller, N. G., Mollenhauer, M., Rösler, A., & Kleinschmidt, A. (2005). The attentional field has a Mexican hat distribution. *Vision Research*, *45*(9), 1129–1137, https://doi.org/10.1016/j.visres.2004.11.003.

Oksama, L., & Hyönä, 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, https://doi.org/10.1080/13506280344000473.

Oksama, L., & Hyönä, J. (2008). Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive Psychology*, *56*(4), 237–283, https://doi.org/10.1016/j.cogpsych.2007.03.001.

Palmer, S., & Rock, I. (1994). Rethinking perceptual organization: The role of uniform connectedness. *Psychonomic Bulletin & Review*, *1*(1), 29–55, https://doi.org/10.3758/BF03200760.

Peterson, M. S., Kramer, A. F., & Irwin, D. E. (2004). Covert shifts of attention precede involuntary eye

movements. *Perception & Psychophysics*, *66*(3), 398–405, https://doi.org/10.3758/BF03194888.

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25, https://doi.org/10.1080/00335558008248231.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97, https://doi.org/10.1016/0010-0277(89)90014-0.

Pylyshyn, Z. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition*, *11*(7), 801–822, https://doi.org/10.1080/13506280344000518.

Pylyshyn, Z. (2006). Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving nontargets. *Visual Cognition*, *14*(2), 175–198, https://doi.org/10.1080/13506280544000200.

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

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*(1–3), 17–42, https://doi.org/10.1080/135062800394667.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science*, *8*(5), 368–373.

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185, https://doi.org/0.1016/j.neuron.2009.01.002.

Roelfsema, P. R., Lamme, V. A., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*(6700), 376–381, https://doi.org/10.1038/26475.

Shim, W. M., Alvarez, G. A., & Jiang, Y. V. (2008). Spatial separation between targets constrains maintenance of attention on multiple objects. *Psychological Bulletin Review*, *15*(2), 390–397, https://doi.org/10.3758/PBR.15.2.390.

Shim, W. M., Alvarez, G. A., Vickery, T. J., & Jiang, Y. V. (2010). The number of attentional foci and their precision are dissociated in the posterior parietal cortex. *Cerebral Cortex*, *20*(6), 1341–1349, https://doi.org/10.1093/cercor/bhp197.

Simons, D. J. (2000). Current approaches to change blindness. *Visual Cognition*, *7*(1–3), 1–15, https://doi.org/10.1080/135062800394658.

Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences, USA*, *96*(4), 1663–1668, https://doi.org/10.1073/pnas.96.4.1663.

Srivastava, N., & Vul, E. (2016). Attention modulates spatial precision in multiple-object tracking. *Topics in Cognitive Science*, *8*, 335–348, https://doi.org/10.1111/tops.12189.

Sundberg, K. A., Mitchell, J. F., & Reynolds, J. H. (2009). Spatial attention modulates center-surround interactions in macaque visual area V4. *Neuron*, *61*(6), 952–963, https://doi.org/10.1016/j.neuron.2009.02.023.

Tombu, M., & Seiffert, A. E. (2008). Attentional costs in multiple-object tracking. *Cognition*, *108*(1), 1–25, https://doi.org/10.1016/j.cognition.2007.12.014.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136, https://doi.org/10.1016/0010-0285(80)90005-5.

Tsotsos, J. K. (1990). Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, *13*(3), 449–450, https://doi.org/10.1017/S0140525X00079644.

Tsotsos, J. K., Culhane, S. M., W. Y., Kei Wai, Y., Lai, N., Davis, & F. Nuflo, (1995). Modeling visual attention via selective tuning. *Artificial Intelligence*, *78*(1–2), 507–545, https://doi.org/10.1016/0004-3702(95)00025-9.

van den Berg, R., Shin, H., Chou, W.-C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, USA*, *109*(22), 8780–8785, https://doi.org/10.1073/pnas.1117465109.

Vater, C., Kredel, R., Hossner, E.-J., Reine, B., Michel, K., & Hossner, E. J. (2017). Disentangling vision and attention in multiple-object tracking: How crowding and collisions affect gaze anchoring and dual-task performance. *Journal of Vision*, *17*(5):21, 1–13, https://doi.org/10.1167/17.5.21. [PubMed] [Article]

Verstraten, F. A., Cavanagh, P., & Labianca, A. T. (2000). Limits of attentive tracking reveal temporal properties of attention. *Vision Research*, *40*(26), 3651–3664, https://doi.org/10.1016/S0042-6989(00)00213-3.

Viswanathan, L., & Mingolla, E. (2002). Dynamics of attention in depth: Evidence from multi-element tracking. *Perception*, *31*(12), 1415–1437, https://doi.org/10.1068/p3432.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in

visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92–114, https://doi.org/10.1037//0096-1523.27.1.92.

Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences*, *15*(4), 160–168, https://doi.org/10.1016/j.tics.2011.02.005.

Wilken, P., & Ma, W. (2004). A detection theory account of change detection. *Journal of Vision*, *4*(12):11, 1120–1135, https://doi.org/10.1167/4.12.11. [PubMed] [Article]

Wolfe, J. M. (2007). Guided search 4.0: Current progress with a model of visual search. In W. Gray (Ed.), *Integrated models of cognitive systems* (pp. 99–119). New York: Oxford University Press, https://doi.org/10.1093/acprof:oso/9780195189193.003.0008.

Wolfe, J. M., & Horowitz, T. S. (2004). Opinion: What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, *5*(6), 495–501, https://doi.org/10.1038/nrn1411.

Yeshurun, Y., & Rashal, E. (2010). Precueing attention to the target location diminishes crowding and reduces the critical distance. *Journal of Vision*, *10*(10):16, 1–12, https://doi.org/10.1167/10.10.16. [PubMed] [Article]

Yu, D., Levinthal, B., & Franconeri, S. L. (2017). Feature-based attention resolves depth ambiguity. *Psychonomic Bulletin & Review*, *24*(3), 804–809, https://doi.org/10.3758/s13423-016-1155-x.

Zelinsky, G. J., & Todor, A. (2010). The role of "rescue saccades" in tracking objects through occlusions. *Journal of Vision*, *10*(14):29, 1–13, https://doi.org/10.1167/10.14.29. [PubMed] [Article]

Zhong, S., Wilson, C., & Flombaum, J. I. (2014). Why do people appear not to extrapolate trajectories during multiple object tracking? A computational investigation. *Journal of Vision*, *14*(12):12, 1–30, https://doi.org/10.1167/14.12.12. [PubMed] [Article]

Appendix A: Communication among ARCADIA components

IMOT, like other models implemented in the ARCADIA cognitive system, includes a set of components that operate in parallel on each cycle, processing input from other components and generating output.

Keyword	Value
ID	4651
Name	"selection-candidate"
World	"reality"
Source	COLOR-HIGHLIGHTER
Type	"instance"
Arguments	
Candidate	<proto-object candidate>
Reason	"color"
Color-label	"red"

Table A1. Example of an interlingua element.

Although computations within each ARCADIA component may in principle take any form, the output items generated by components must be in a common language, so that components can communicate. Therefore, components generate *interlingua elements*, which are tables that associate symbolic keywords with values.

Table A1 illustrates an interlingua element that might be generated by COLOR HIGHLIGHTER.[12] **ID** is a unique numeric identifier, **name** describes what this element is (in this case, a candidate for selection), **world** is the context in which this element exists, **source** is the component that generated this element, and **type** is the general type of element. Finally, **arguments** is itself a table of keyword/value pairings that can include any additional information generated by the component. In this example, COLOR HIGHLIGHTER has provided three additional pieces of information: a data structure describing the proto-object that is the candidate for selection, the reason this proto-object is a candidate (it has an interesting color), and a textual color label for the proto-object.

Appendix B: Simulation 1 parameters

In Simulation 1, IMOT tracks the targets nearly perfectly. To match human error patterns we introduce noise, which is controlled by two parameters: noise-center and noise-width. Recall that the model's OBJECT LOCATOR can match an enhanced region to a proto-object only if the width of their overlap, in visual degrees, is greater than a random value sampled from a Gaussian distribution centered at noise-center with standard deviation of noise-width.

To determine the appropriate values for the noise parameters, we adjusted the parameters by increments of 0.1 for noise-center and 0.05 for noise-width, running the full simulation for each pair of possible values. We classified a simulation run as a success if it

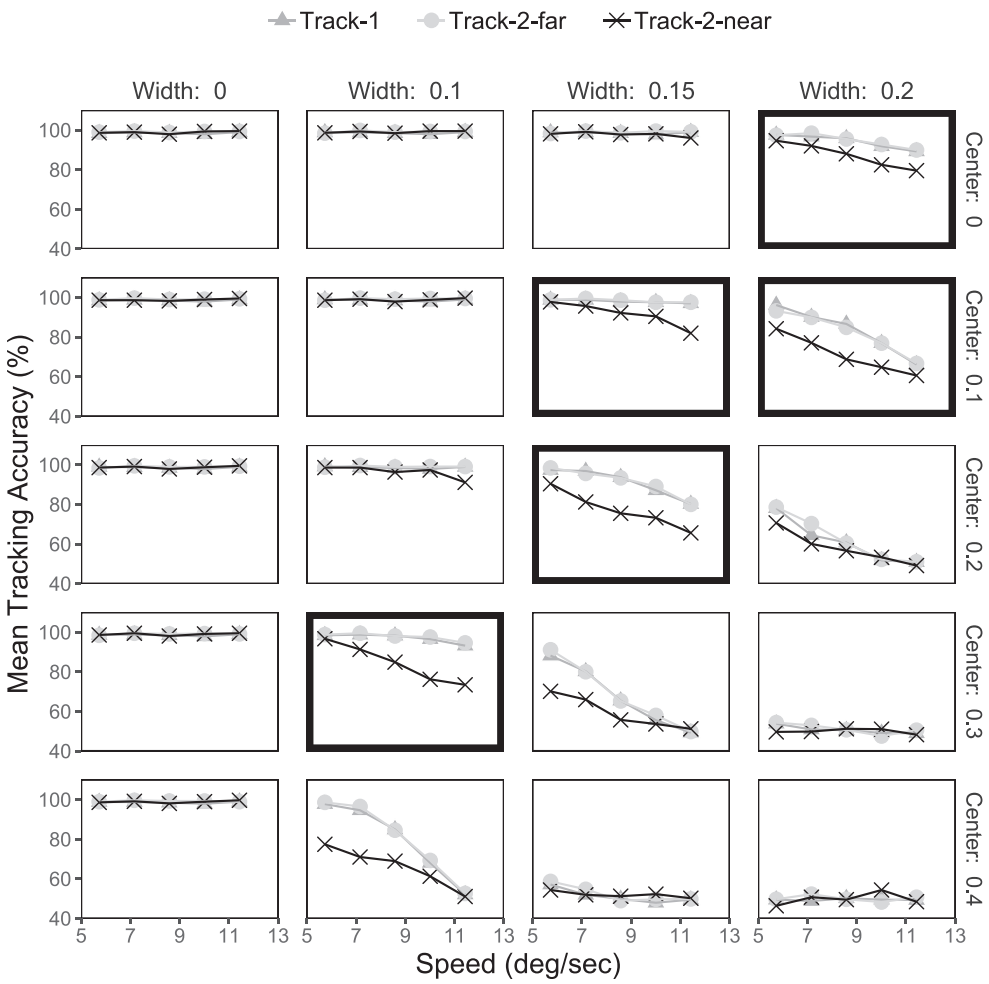

Figure B1. Simulation 1 results across a range of noise-center and noise-width values. The graphs highlighted in bold show simulation runs that replicated the human results.

replicated three results from the original experiment (Shim et al., 2008):

- Tracking accuracy drops as motion speed increases.
- Accuracy for tracking two targets in the same quadrant is lower than accuracy for tracking two targets in separate quadrants.
- Accuracy for tracking two targets in separate quadrants is the same as accuracy for tracking a single target.

The second result was evaluated both overall and at each motion-speed level except the slowest speed—recall that Shim et al. found no difference between tracking conditions at the slowest speed. The third result was evaluated overall only—we anticipated that the two conditions would be identical, but making comparisons at every motion-speed level increases the risk that a spurious difference will be detected occasionally.

Figure B1 depicts the results across all simulation runs (compare to Figure 6a). A subset of the runs (highlighted in bold) fully replicated the human results, and other runs partially replicated them. Generally, accuracy for tracking two targets in separate quadrants (Track-2-far) was the same as accuracy for tracking one target (Track-1), whereas accuracy for tracking two targets in the same quadrant (Track-2-near) was lower. But there was a ceiling effect with low noise and a floor effect with high noise.

Appendix C: Simulation 2 parameters

In Simulation 2, IMOT benefits from predictable motion both with two and with four targets. To match human error patterns, we introduce the max-threat-distance parameter, which determines whether a target will be classified as threatened. Recall that focus will

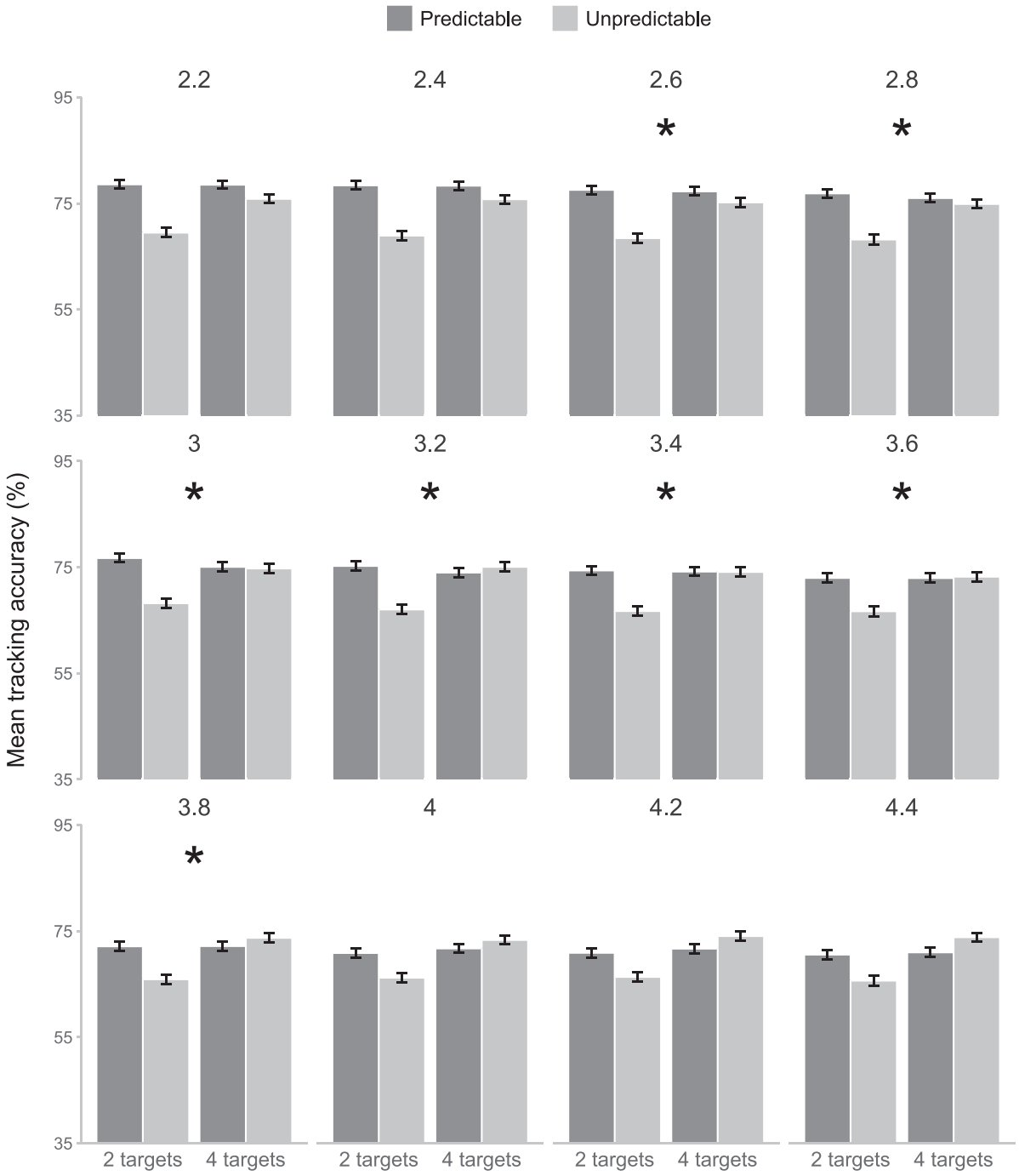

Figure C1. Simulation 2 results across a range of max-threat-distance values. The graphs with asterisks show simulation runs that replicated the human results. Error bars are ±1 standard error.

linger on a target until its threat distance rises above this value.

To determine the appropriate max-threat-distance, we adjusted the parameter by increments of 0.2, running the full simulation for each possible value. We classified a run as a success if it replicated four results from the original experiment (Luu & Howe, 2015):

- Tracking accuracy is higher with predictable motion.
- There is a significant interaction between motion type and target number, reflecting a greater benefit for predictable motion when two targets are tracked.
- When two targets are tracked, accuracy is higher with predictable motion.

- When four targets are tracked, accuracy is the same with predictable or unpredictable motion.

Figure C1 depicts the results across all simulation runs (compare to Figure 7a). Runs with a max-threat-distance between 2.6 and 3.8 (marked with an asterisk in the figure) replicate the human results. When the parameter is lower than 2.6, the runs fail to replicate the fourth result: There is a benefit for predictable motion with four targets, suggesting that the model switches its selection focus between targets more easily than humans. When the parameter is higher than 3.8, the runs fail to replicate the first result: There is no overall benefit for predictable motion, suggesting that the model switches its focus less easily than humans. Notably, all runs replicate the second result: The benefit for predictable motion is always greater with two targets than it is with four targets.