Attentive and Pre-Attentive Processes in Multiple Object Tracking: A Computational Investigation

Paul Bello (paul.bello@nrl.navy.mil)
Will Bridewell (will.bridewell@nrl.navy.mil)
Christina Wasylyshyn (christina.wasylyshyn@nrl.navy.mil)
Naval Research Laboratory, 4555 Overlook Ave. S.W.
Washington, DC 20375 USA

Abstract

The rich literature on multiple object tracking (MOT) conclusively demonstrates that humans are able to visually track a small number of objects. There is considerably less agreement on what perceptual and cognitive processes are involved. While it is clear that MOT is attentionally demanding, various accounts of MOT performance centrally involve pre-attentional mechanisms as well. In this paper we present an account of object tracking in the ARCADIA cognitive system that treats MOT as dependent upon both pre-attentive and attention-bound processes. We show that with minimal addition this model replicates a variety of core phenomena in the MOT literature and provides an algorithmic explanation of human performance limitations.

Keywords: attention; visual cognition; multiple object tracking; cognitive model

Introduction

A sizeable portion of the visual cognition literature has been consumed with trying to produce a detailed story about how objects in the world are visually represented and tracked through time. Attention, broadly construed, is central to many of the explanations on offer. Insofar as object-tracking behavior is observed in human visual cognition in the absence of attention, it is precisely this absence that is striking and calls out for explanation. Nowhere is this clearer than in the substantial literature on multiple-object tracking (MOT; Pylyshyn & Storm 1988). While almost universally considered to be an attentionally demanding task, MOT has been intensely investigated because it appears that object tracking can be sustained for short periods in the absence of attention (Alvarez et al. 2005). This superficial inconsistency suggests that both attention-bound and pre-attentive processes are partially constitutive of object-tracking capacity, although perhaps do not fully exhaust it, since strategies may play a substantial role as well.

But what can performance characteristics on the MOT task tell us about the nature of object tracking and the role of attention in tracking? If attention can be divided during MOT coupled with a dual-task, what are the mechanisms that explain successful tracking performance? Finally, how would these mechanisms fit into a larger computational theory of human visual cognition? The plan for the remainder of this paper is to address these questions within a computational system. After briefly summarizing some important results from the MOT literature, the discussion will then turn to a proposal by Dawson (1991) that mechanisms involved in the perception of apparent motion may be at the heart of the pre-attentive computations that enable MOT.

Against the backdrop of these requirements, we summarize ARCADIA[1] with a particular eye on its components that contribute to object tracking. These components include a visual short-term memory (vSTM), a volatile mirror image of vSTM that stores only location as suggested by our discussion of apparent motion, and respective update mechanisms. We then show via simulation that the same proximity-based mechanism is involved in producing apparent motion suffices for explaining performance on MOT tasks and accounts for errors generated when tracked targets become crowded and when their speed limits tracking capacity.

Multiple Object Tracking

In a typical multiple object tracking experiment, subjects are shown a display of some number (usually > 8) of identical objects such as circles. A subset of these objects (the targets to be tracked) are flashed or highlighted to facilitate encoding before returning to their original state. After a brief pause each object in the display moves in a random fashion for a short period of time, after which subjects are to indicate via mouse click which objects are the targets.

In a recent review, Scimeca and Franconeri (2015) lay out a set of four core capacity limits that any adequate theory of MOT competence must explain: (1) capacity, (2) crowding, (3) hemifields, and (4) speed. In short, tracking more rather than fewer objects engenders lower accuracy. Targets that are packed closer together negatively impact accuracy. When targets are clustered in the same hemifield or quadrant of the display, accuracy drops. Finally, accuracy drops for targets that move faster. Some of these factors are not independent of one another. For example, fast-moving targets raise the probability that they will crowd with others as a function of time and distance traveled. Similarly, there may be an interaction between speed and the number of targets in a specific hemifield at any one time.

As we move on in our discussion, we turn back to these *core four* and show how a very simple proximity heuristic suggested in the literature on the perception of apparent

[1] **A**daptive **R**eflective **C**ognition in an **A**ttention-**D**riven **I**ntegrated **A**rchitecture

motion might help explain some of these limitations in tracking performance. The proximity heuristic predicts an interaction between the speed of targets and the amount of spacing between them. The heuristic fails when speed increases and spacing decreases, because objects will have more *close encounters* over time and thus more opportunities to have their identities confused in the MOT task.

However, a proximity-based update mechanism alone is insufficient for explaining the *core four* among other results in the MOT literature. We need to locate this heuristic in a larger framework and make choices about the number, type, and status of interacting mechanisms involved in tracking. This turns us to the discussion of object construction and tracking in ARCADIA.

The ARCADIA Cognitive System

ARCADIA as described in (Bridewell & Bello 2015) is at its heart a framework for integrating psychological and neuropsychological theories. ARCADIA consists primarily of *components*. They are the medium by which theories are implemented in ARCADIA, and insofar as ARCADIA makes any of its own commitments about how theories are realized, it is in the number and type of components used to implement it. The only restriction on components is that they are able to read and write to a common representational schema called *interlingua*. The particulars of the data structures and algorithms contained in each component are either inherited from the theories that they implement or are at the discretion of the modeler. This is one of the major features differentiating ARCADIA from other cognitive architectures.

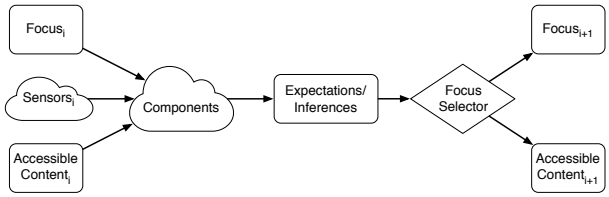

Figure 2: ARCADIA's processing loop.

Components read and write interlingua elements from *accessible content*, which is populated by the ephemeral results of system-wide behavior every cycle. Accessible content is flushed and re-populated on every cycle, making attention, both exogenous and endogenous, a critical enabler for encoding and active maintenance of mental representations (i.e., interlingua elements) across contiguous cycles.

As shown in Figure 1, on each cycle, a privileged item is selected from accessible content and broadcast system-wide to all components. The selected item serves as the *focus of attention* for that cycle, and once broadcast, all focus-responsive components take the focus and the current set of accessible content and compute their results, which are then added to the subsequent set of accessible content. The focus

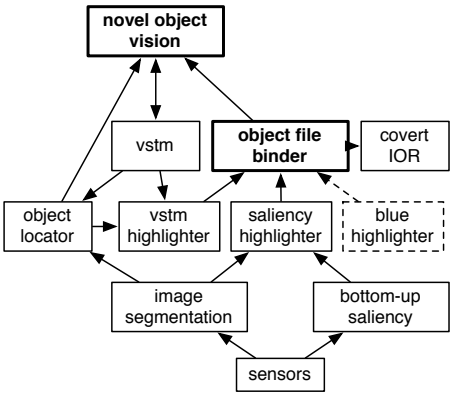

Figure 1: Informational exchange between ARCADIA components during the basic MOT task. Bolding of text indicates that the respective component is responsive to the focus of attention, and therefore attention-bound. Components having dashed borders are task-specific.

of attention is chosen by ARCADIA's focus selector, which is loaded with an *attentional strategy* for whatever task is currently being performed. Attentional strategies in ARCADIA are control knowledge, and establish selectional preferences over items in accessible content for what to focus on during each cycle. A detailed explication of ARCADIA's interlingua, processing loop, and focus selection was given by Bridewell & Bello (2015).

Modeling Object Construction and Tracking

Visual processing in ARCADIA is divisible into pre-attentive and attentive computations that can occur simultaneously. The set of computations underlying object construction can be seen in Figure 2, which may be a helpful roadmap for navigating the subsequent description. Pre-attentively, ARCADIA employs components that compute salience maps via methods described by Itti and colleagues (1998) and proto-object representations via image segmentation. This latter component produces interlingua elements that encode basic color histogram and region information wherever closed contours are found in the image. This serves as a rough and ready approximation to a high-speed, high-capacity iconic memory. The image segmenter also provides proto-object regions to the object locator component, which will be described in detail shortly and also works pre-attentively.

The next set of components in ARCADIA's visual system is responsible for pre-attentively producing requests for orientation. As shown in Figure 2, ARCADIA's saliency highlighter looks at accessible content for interlingua elements having saliency maps and others containing proto-objects. Each region containing a proto-object is co-registered back onto the saliency map and checked for salience value. The saliency highlighter outputs the N regions (with N <= 4; see Xu & Chun 2006) containing salient proto-objects, which become candidates for orientation. The vSTM highlighter produces top-down requests for orientation on objects that have been encoded into visual short-term memory.

The next layer of the visual system is responsible for object-construction, maintenance and tracking. Once orientation requests have been generated, whichever attentional strategy is currently loaded into ARCADIA's focus selector will be used to select from available orientation requests based on their relative prioritization in the strategy. Because this stage of vision is attentive, the selected region is broadcast system-wide, and any components that detect property information, such as shape or color, and can produce inferences or judgments about the content of the region now do so. The resultant judgments are passed by the property detectors up to accessible content as interlingua elements. A binding component then binds all the features that are detected in the region into an object file, which corresponds to a fully formed visual-object and a list of its properties. If focused on, new objects are tested against object representations in vSTM by the novel object vision component to determine whether they are actually new objects or should be treated as an update to a sufficiently similar object encoded in vSTM. ARCADIA assumes a fixed-slot four-element capacity for vSTM with a queue structure, so that when at capacity, new objects encoded in vSTM displace the oldest object in memory. On each cycle, vSTM pushes a list of its elements into accessible content, which are used both by the vSTM highlighter component and the object locator component.

Pre-Attentive Location Update: Object Locator

So far, we have described the normal course of processing for object construction, encoding and attention-dependent vSTM update. But what vision components are unique to MOT performance? Surprisingly, on our account there are none. Instead, the tracking mechanism most often implicated in MOT performance is motivated by other concerns.

Given two temporally contiguous visual snapshots, the human visual system is faced with the problem of re-identifying objects residing within the first frame with objects residing in the second frame, sometimes called *the correspondence problem*. Moreover, this problem is made difficult if the objects in question are in motion and change locations between frames, as would be the case for both the objects in tracking tasks. Dawson (1991) identified and computationally explored a potential solution as a corollary to his work on characterizing the mechanisms underwriting the phenomena of apparent motion.

In summary, Dawson finds that the correspondence problem is solved in the human visual system through the mutual satisfaction of three soft constraints, only two of which we will concern ourselves with in this work. The first of these constraints ensures a one-to-one mapping of each object in the first frame to a corresponding object in the second frame. The second constraint embodies a proximity-based principle such that each object in the first frame is assigned to the nearest object in the second frame in terms of Euclidean distance. The solution to the correspondence problem (1) is insensitive to object features other than spatial location and relative velocity and (2) operates on timescales well beneath those required to solve the problem attentively or deliberatively.

In ARCADIA, the object locator component serves as a bank of visual indices (Pylyshyn & Storm 1988, Alvarez & Franconeri 2007) that reference object locations. Object locator mirrors the internal structure of vSTM and stores a set of locations associated with each object encoded in it. There are two critical differences between vSTM and object locator. The first is that while vSTM stores conceptual representations of objects as collections of properties, the representations in object locator only store object location. Secondly, while vSTM requires attention to update location information for the objects it contains, object locator performs updates pre-attentively on each cycle, using Dawson's nearest neighbor proximity-based heuristic.

Object locator uses proto-object information deposited in accessible content by the image segmenter along with the contents of vSTM in updating its location information. To this end, object locator computes the N nearest proto-object neighbors in terms of Euclidean distance for each element of vSTM, and updates its location values (which correspond to their respective counterparts in vSTM) with new location information from their nearest proto-object neighbor. This basic computation immediately entails that there will be interactions between speed, crowding, and performance in MOT, since fast-moving objects will generate more instances of crowding over time and generate more opportunities for object locator to incorrectly identify vSTM elements with the wrong proto-object.

MOT: Task Specific Components and Strategy

One of the defining features of our account of MOT is just how little must be added to our model of object construction and tracking in order to simulate the MOT task. We only add a single task-specific component: a "blue highlighter." Our MOT simulation highlights the initial target set in blue before changing them back to their initial color. Blue highlighter detects proto-objects from the image segmenter with blue color histograms, and produces fixation requests on those regions of the image. We assume that the experimental instructions given to ARCADIA *qua* human subject indicate that targets will initially flash blue so that ARCADIA's attentional strategy reflects prioritization of blue-directed orientation requests over any others.

While not mentioned in the last section, ARCADIA's attentional strategy for object construction and tracking is given below:

1. If a new object file is available, make it the focus of attention so that it can be compared to and/or encoded in vSTM.

2. Otherwise, if one of the highlighters requests moving covert visual attention and there are no inhibitors

preventing the movement,[2] attend to the specified proto-object.

3. If neither of these options exists, attend to an arbitrarily selected interlingua element.

For MOT, the only change we make to the strategy above is to induce a preference ordering over highlighters such that requests from blue highlighter are prioritized over requests from vSTM highlighter, which are prioritized over requests from saliency highlighter. This has the effect of ARCADIA encoding blue targets when they first flash and serially revisiting each vSTM-encoded object over the course of tracking. Since object locator updates a mirror image of whatever is encoded in vSTM, it is unaffected by the attentional strategy above.

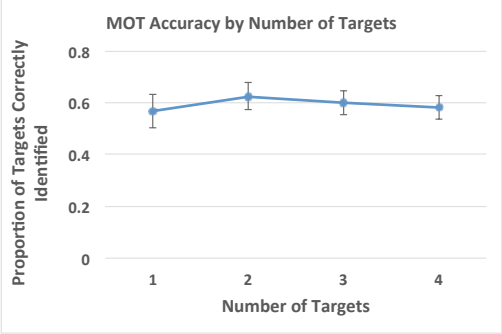

Figure 3: Results when varying the number of targets.

Computational Simulation

To test the set of predictions we have made thus far, we ran the model of ARCADIA shown in Figure 2 in a standard MOT task with sixteen total objects for five seconds per trial. We varied speed, spacing, and the number of initial targets to be tracked. Spacing between objects (circles of diameter D) varied at three levels: 0, 0.5D, and D. Effectively, these values amount to either allowing collision or requiring one or two full diameters of space between objects as they moved. Speed was normed by determining the value at which ARCADIA consistently failed to correctly identify any targets, even at D spacing. We divided this value by four and determined slow, regular, medium, and fast speed levels for the targets. Finally either 1, 2, 3, or 4 targets could be tracked, leading to a 4 x 4 x 3 configuration. We ran each configuration five times for a total of 240 system runs. Each run began with randomly selected targets in random locations with randomly selected initial trajectories. After each run, we computed the proportion of targets that ARCADIA successfully tracked.

[2] We do not discuss inhibition here since it plays no role in MOT, but other ARCADIA models utilize both task-related inhibition and covert inhibition of return (Bridewell & Bello, 2016)

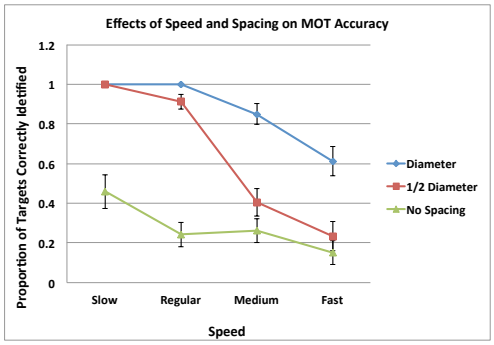

Figure 4: The interaction between speed and spacing in ARCADIA's performance on the MOT task.

Simulation Results

A 4 (number of targets) x 4 (speed) x 3 (spacing) repeated measures ANOVA was conducted. The proportion of targets correctly identified served as the dependent variable. There was no main effect for the number of targets tracked, $F(3, 12) = 0.68$, $p = 0.58$, indicating that up to 4 targets can be tracked robustly. Means and standard errors for number of targets tracked can be seen in Figure 3. There was a main effect of speed, $F(3, 12) = 44.52$, $p < 0.001$, with the proportion of targets correctly identified decreasing as speed increased ($M_{slow} = 0.82$, $M_{regular} = 0.72$, $M_{medium} = 0.51$, $M_{fast} = 0.33$). There was a main effect of spacing, $F(2, 8) = 158.81$, $p < 0.001$, with the proportion of correctly identified targets increasing as spacing decreased ($M_{0\ spacing} = 0.28$, $M_{.5D\ spacing} = 0.64$, $M_{D\ spacing} = 0.87$). These two significant main effects were qualified by a significant speed by spacing interaction, $F(6, 24) = 8.70$, $p < 0.001$. This interaction can be seen in Figure 4.

General Discussion

In general, the results of the simulation study are consistent with the vast majority of literature on limitations in MOT performance. The lack of a main effect of target reflects robust tracking of one to four objects via visual indices. The main effects of speed and spacing found here are in accordance with previous findings and correspond to two elements of Scimeca and Franconeri's *core four* signature performance limits on MOT. Crucially, we also found a significant interaction between speed and spacing, which is both predicted by any nearest-neighbor type model. While not conclusive with respect to the prediction made by Franconeri and colleagues (2010) that takes spacing to be the only theoretical boundary on MOT performance, our results do provide evidence that spacing plays a considerable role in offsetting the performance-reducing effects of high-speed object movement. To more fully pursue Franconeri's hypothesis, we would need to decrease target size and increase the number of possible spacing conditions, but we would expect to find performance dropping off much more slowly as a function of speed, given extra space.

We do not report results here related to hemifield and quadrant effects nor do we report capacity effects. The former, while interesting, are contentious (see Hudson et al. 2012 for details). Capacity effects for tracking loads greater than four are not reported here since ARCADIA's vSTM is only four objects deep. Explanations for how subjects manage to track more than four objects are open to many different interpretations that invoke strategies and mechanisms beyond the simple pre-attentive updating mechanism described here.

So-called "flexible resource theories" have been proposed to explain human tracking of more than four targets (Alvarez & Franconeri 2007). One of the nagging problems about resource theories is that it remains unclear what a "resource" could be (Franconeri et al. 2010). Similarly, many resource-theory explanations fail to invoke the distinction between the attentive and pre-attentive components of tracking, even though the weight of evidence points to the existence of a pre-attentive basis for MOT performance. Finally, others have argued that tracking perceptual groups of objects can give the appearance of tracking more than four targets because one or more of the four are actually sets of targets rather than individuals (Yantis 1992). This is just one example of a potential strategy that could be employed to explain tracking performance past four objects.

These factors and open possibilities have persuaded us to be methodologically conservative in the work we report here. We assume that the storehouse for visual indices in our approach mirrors the structure of vSTM, which we are conservatively assuming is four objects in capacity. One of the features of our theory is that it is insensitive to the internal structure of vSTM. If, for example, evidence persuades us to implement a resource-based account of vSTM that allows for a larger number of objects to be represented at coarser resolution, our pre-attentive update mechanism will mirror this structure and behave accordingly.

Comparison to Other Computational Models

The approach we have taken with implementing a simple nearest-neighbor pre-attentive update within a larger vision framework in ARCADIA has several explanatory advantages. The most important feature of our proposal is to link visual index updating to a known psychological process involved in other parts of visual cognition. Alternative computational models of MOT use Kalman filters and thus share an important similarity: location updating is a function of prediction (Vul et al. 2009, Zhong et al. 2014, Srivastava & Vul. 2015). However, a number of studies have demonstrated that human subjects seem to not rely on extrapolation of object trajectories during MOT, which may appear to rule out Kalman filters as lacking face validity for modeling human tracking performance (Keane & Pylyshyn 2006, Franconeri et al. 2012). A small number of studies demonstrate trajectory extrapolation in MOT under highly circumscribed conditions (Fencsik 2007, Howe & Holcombe 2012). In these latter two studies, extrapolation was only observed for tasks having a tracking load of two or less. Somewhat more disturbingly, Howard and colleagues (2011) find that some of the aforementioned results indicating trajectory extrapolation involved verbal instructions to subjects that may have unintentionally primed subjects to use explicitly extrapolative strategies.

The sharp limitations on tracking load in studies that implicate extrapolation in MOT are suggestive of different mechanisms at play, or perhaps some difference at the tracking-strategy level having to do with allocation of attentional resources. Even if extrapolation is happening for loads of two or less, it very well may be that this process is purely attention-bound and not reflective of the pre-attentive location updating mechanism under discussion in this paper. In any case, ambiguity of this sort compels us to take care as modelers to distinguish between pre-attentive and attention-bound processes, a distinction that is central in our own work.

Future Work

Perhaps the lowest hanging fruit involves modeling results that show subjects are capable of MOT with occlusions using a proximity heuristic such as the one implemented in ARCADIA's object locator (Franconeri et al. 2012). Because object property information other than location is updated attentively in ARCADIA's vSTM, identifying information about targets would be lost over the course of tracking, which is primarily served by pre-attentive mechanisms (Pylyshyn 2004). Pylyshyn (2006) has suggested that distractor inhibition plays a role in explaining why target information is recalled poorly. ARCADIA's attentional strategy for MOT prefers fixating on objects encoded in vSTM over anything driven by salience or other bottom-up processes. In this way, inhibition is built in as a function of being task focused.

Because ARCADIA is driven by attentional strategies it can be used to capture a variety of plausible strategy-driven features of MOT. Recent results are suggestive of better tracking performance for targets in crowded parts of the MOT display (Srinistava & Vul 2015) due to strategic deployment of attention to minimize uncertainty. Attentional strategies along with other components to detect relations could also be used to encode targets as the vertices of a polygon to be tracked (Yantis 1992). ARCADIA would need to be outfitted with a theory of overtly deployed visual attention and an accompanying model of eye movements to begin attacking any of the above in earnest. Work on these additions to ARCADIA has begun, allowing for a much richer and fuller exploration of the range of human object tracking capacity.

Acknowledgments

The authors would like to acknowledge generous support from the Office of Naval Research under grants N0001414WX20179 and N0001415WX01339. The views expressed in this paper 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.

References

Alvarez, G., Horowitz, T., Arsenio, H., DiMase, J., & Wolfe, J. (2005). Do multielement visual tracking and visual search draw continuously on the same visual attention resources? *Journal of Experimental Psychology: Human Perception and Performance, 31*, 643–667.

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

Bridewell, W., & Bello, P. (2015). Incremental object perception in an attention-driven cognitive architecture. In *Proceedings of the thirty-seventh annual conference of the cognitive science society* (pp. 279-284). Austin, TX: Cognitive Science Society.

Bridewell, W., & Bello, P. (2016). Inattentional blindness in a coupled perceptual-cognitive system. In *Proceedings of the thirty-eighth annual conference of the cognitive science society*. Austin, TX: Cognitive Science Society.

Dawson, M. (1991). The how and why of what went where in apparent motion: Modeling solutions to the motion correspondence problem. *Psychological Review, 98*, 561–603.

Fencsik D., Klieger S., & Horowitz T. (2007). The role of location and motion information in the tracking and recovery of moving objects. *Perception & Psycho-physics , 69,* 567–577.

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

Franconeri S., Pylyshyn Z., & Scholl B. (2012). A simple proximity heuristic allows tracking of multiple objects through occlusion. *Attention, Perception and Psychophysics*, *74*, 691–702.

Howard, C., Masom, D., & Holcombe, A. (2011). Position representations lag behind targets in multiple object tracking. *Vision Research*, *51*, 1907-1919.

Howe P., & Holcombe A. (2012). Motion information is sometimes used as an aid to the visual tracking of objects. *Journal of Vision*, *12*, 1–10.

Hudson C., Howe P., & Little D. (2012). Hemifield effects in multiple identity tracking. *PLoS One 7*: e43796.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*, 1254–1259.

Keane, B., & Pylyshyn, Z. (2006). Is motion extrapolation employed in multiple object tracking? Tracking as a low-level, non-predictive function. *Cognitive Psychology*, *52*, 346–368.

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

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

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

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

Scimeca, J., & Franconeri, S. (2015). Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science*. Advance online publication.

Srivastava, N., & Vul, E. (2015). Attention dynamics in multiple object tracking. *Proceedings of the thirty-seventh annual conference of the cognitive science society* (pp. 2266-2271). Austin, TX: Cognitive Science Society.

Vul, E., Frank, M., Alvarez, G., & Tenenbaum, J. (2009). Explaining human multiple object tracking as resource constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems*, *22*, (pp. 1955-1963).

Xu, Y., & Chun, M. (2009). Selecting and perceiving multiple visual objects. *Trends in Cognitive Science, 13*, 167–174.

Yantis, S. (1992). Multielement visual tracking: attention and perceptual organization. *Cognitive Psychology, 24,* 295–340.

Zhong, S., Ma, Z., Wilson, C., Liu, Y., & Flombaum, J. (2014). Why do people appear not to extrapolate trajectories during multiple object tracking? A computational investigation. *Journal of Vision*, *14*, 1–30.