Cognition xxx (xxxx) xxxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

A model that adopts human fixations explains individual differences in multiple object tracking

Aditya Upadhyayula, Jonathan Flombaum*

Johns Hopkins University, Psychological & Brain Sciences, United States of America

ARTICLE INFO

Keywords:
Multiple object tracking
Eye tracking
Individual differences
Computational modelling
Kalman filter

ABSTRACT

In many settings "keep your eye on the ball" is good advice. People fixate important objects to obtain high quality information. Perhaps equally often, however, we engage with multiple important, moving, and unpredictable objects. Where should we fixate in these situations, and where do we? Do we for example appropriately center fixations to manage spatial non-uniformity in our visual system? And do we fixate empty space strategically to gain as much information as possible about multiple objects of interest? We explored these issues in the context of Multiple Object Tracking (MOT), wherein observers track several moving objects (targets) within a larger set of moving objects (nontargets), all the objects physically indistinguishable from one another. Among the features that make MOT an interesting paradigm is that it cannot be accommodated by continuous gaze to one important object, because there are multiple such objects in a given trial. Instead, it demands sustained processing of inputs from an entire display and iterated inferences about target versus nontarget identities. MOT therefore demands a strategic interaction between eye movements and cognition: the observer should seek fixation locations that minimize the aggregate probability of confusing any target with any nontarget. Individuals who meet this fixation challenge should perform the task better than those who meet the challenge less effectively. Here we describe a probabilistic model that implements the basic computations needed to do MOT, estimating the positions of targets, predicting their future positions, and inferring correspondences between new inputs and represented targets. The quality of the input received by the model depends on its fixation location at a given moment. We simulated a group of fifty participants who all performed the same MOT trials, with the model adopting each observer's fixation locations in the respective simulations. The model reliably predicted individual participant tracking performances and their relative rankings within the cohort. The results suggest that an individual's relative capability in this cognitively demanding task is in part determined by his/her utilization of eye fixations to control the quality and relevance of incoming visual input.

1. Introduction

Photoreceptors in the human eye are far more densely packed in a central region called the fovea than they are in the periphery. The early visual system then mimics this organization, with more cells and smaller receptive fields dedicated to foveated space than to peripheral space. As a result, objects in the periphery are more difficult to perceive than fixated ones. Demonstrations of poor peripheral vision can be found under the rubric of 'crowding', wherein peripheral objects can be detected, but they remain difficult to attend, recognize, or localize when other objects are nearby (Freeman & Simoncelli, 2011; Ma, McCloskey, & Flombaum, 2015; Pelli & Tillman, 2008; Whitney & Levi, 2011). A consequence for cognition is that fixating objects is a crucial step in the acquisition of visual information. To name only a few examples: invariant object recognition appears to rely on object fixation (Cox, Meier, Oertelt, & DiCarlo, 2005); reading involves the recognition of characters within a small window centered on fixation location (Binder, Pollatsek, & Rayner, 1999); and semantic parsing of complex scenes depends critically on eye movements to individual objects (Henderson & Hollingworth, 1998).

Yet certain activities and scenarios may be better-served by eye movements that do not target individual objects. During visual search in noisy environments, for example, eye movements to empty space are used nearly optimally to maximize information gained per fixation (Najemnik & Geisler, 2005). More generally, in settings with multiple important objects, when unpredictable events may take place, or with moving objects, the best place to look may not be any single object and instead, it may often be to empty locations that maximize the overall

* Corresponding author.
E-mail address: flombaum@jhu.edu (J. Flombaum).

https://doi.org/10.1016/j.cognition.2020.104418
Received 5 November 2019; Received in revised form 26 July 2020; Accepted 27 July 2020

quality and relevance of the resulting input. Choosing where in empty space to look should therefore play a causal role in how well people perform certain tasks.

We sought to investigate how fixation choices constrain performance in once such task. Specifically, we investigated the relationship between eye movements and performance in multiple object tracking (MOT) (Pylyshyn & Storm, 1988). MOT is a common laboratory task used in the study of visual cognition. The nature of the task makes it so that empty space *should* often be the best place to fixate, as opposed to individual objects.

1.1. Multiple object tracking (MOT)

In a typical trial, a set of identical objects appears on the screen. Some of the objects are then designated targets, by flashing or changing color, leaving the remaining objects designated as nontargets (sometimes also called 'distractors'). After designation, the targets and nontargets become physically identical, and all the objects begin to move independently on the screen. At the end of a trial, the objects stop moving and the task for the participant is to identify the objects that were originally designated targets (Pylyshyn & Storm, 1988). Because the targets and nontargets are identical in appearance at all times, save for the start of the trial, the participant must track the target positions continuously during the motion phase if she will be able to accurately identify them later. The requirement of sustained attention therefore distinguishes the task among many other attention-demanding paradigms.

Fig. 1 schematizes the sequence of events in a typical trial. One advantageous feature of the task is that difficulty can be manipulated intuitively and continuously. Increasing the number of targets to track, the number of confusable nontargets, or the speeds at which the objects move will each make a participant more likely to make an erroneous report at the end of a trial (Alvarez & Franconeri, 2007; Bae & Flombaum, 2012).

Our present interest in the MOT task arises from an interest in understanding how fixation choices may place inherent limits on human abilities. We expect that fixation should often favor empty locations because of the inadequacy of object-directed fixations under the constraints of the task. Consider just the difference between tracking one target in the MOT task and tracking two: when tracking one, the best place to point one's eyes is directly at the moving target, and to pursue it smoothly. A strategy that is probably not the best when simultaneously tracking two independently moving targets. What is? One option might be to serially fixate each of the individual targets. But another might be to mostly fixate the centroid of the two targets, *an empty location* that will maximize the joint quality of the target inputs. And other strategies are conceivable as well, ones that combine target and empty space fixations in response to the current proximity between individual targets and nontargets (i.e. the current risk of confusing any given target). Regardless, tracking two targets demands that one resigns himself to noisier inputs than he could otherwise acquire about a single target, were it on its own. Tracking two targets means that the best fixation positions over the course of a trial are different from what they would be if the trial included only one of the targets, and fixating empty spaces might be a necessary concession in the effort to maximize shared input quality from the targets.

Now consider a trial with three targets. Again, the best fixation strategy is unlikely to be the same as it would be if only two of the three targets were included in the trial. MOT therefore demands a strategic interaction between eye movements and cognition, as a function of the current configuration of the input: the observer should seek fixation locations that minimize the aggregate probability of confusing any target with any nontarget. Individuals who meet this fixation challenge should perform the task better than those who meet the challenge less effectively. This is a specific prediction that we examine here, in the case of MOT. It likely applies more broadly as well: spatial non-uniformity in human vision turns one's fixation choices into a lever for controlling input relevance and quality.

1.2. Eye movements and performance: many paths (could) lead to Rome

Given what should be an important role for fixation selection in determining performance, there is a surprising dearth of research on eye movements in the task. This is possibly because of the early suggestion that eye movements do not impact performance; specifically, that task performance is no different when participants are forced to fixate the center of a display or allowed to freely move their eyes (Scholl & Pylyshyn, 1999). But three relatively more recent studies have investigated eye movements in MOT with interesting results. Fehd and Seiffert (Fehd & Seiffert, 2008; Fehd & Seiffert, 2010) found that a meaningful proportion of fixations are close to the centroid of the polygon formed by the targets at a given moment (see also (Yantis, 1992)). This suggests a potentially productive strategy for performing the task, one that some observes may adopt more than others. Zelinsky

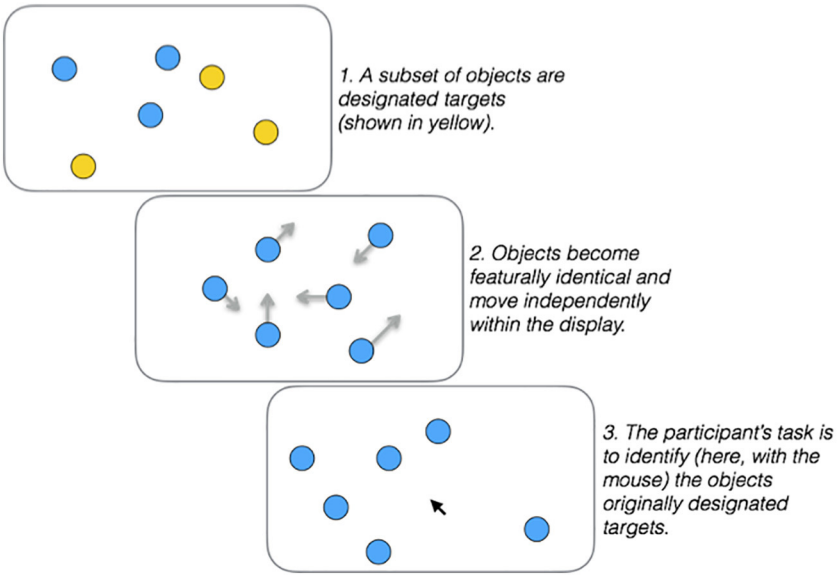


Fig. 1. A typical multiple object tracking (MOT) trial.

and Neider (Zelinsky & Neider, 2008) similarly found a large number of centroid fixations in a tracking cohort, though for target loads of only two to four. For loads greater than four the proportion of fixations to the centroid declined, replaced by an increase in shifting fixations to individual targets. The authors suggested that the right combination of centroid fixations and target-directed 'rescue saccades' could be the best strategy for more difficult MOT loads. Unfortunately, none of the three studies identified links between the proportion of target vs. centroid fixations and performance, nor did they demonstrate that target fixations occur at the right times, that is when a target is more at risk of being confused than not. And these studies are likely to have overestimated the degree to which participants look at the target centroid because they classified each fixation as exhaustively belonging to one of two categories; classifications were made competitively not in terms of goodness of absolute fit.

What if participants make fixations to empty space that are not intended as centroid directed? What if participants have a variety of strategies for choosing where to fixate, emanating from a variety of considerations? For instance, what if the best places to fixate include places like the centroid of only 'at-risk' targets, excluding those that are currently far from any nontargets and unlikely to be confused. Perhaps participants fixate a version of the target centroid that is weighted by risk-of-loss? What if fixations are programmed to anticipate future confusion risks, based on current trajectories?

Parametrically investigating all these possibilities (and others) is a nontrivial experimental challenge. Put more generally, we perceive two related challenges for understanding how (and whether) fixation choices constrain MOT performance. One is that there may be effective strategies that researchers fail to think of, neglecting to include them in a classification scheme. Second, it is possible that a given trial has more than one 'solution', that there are different strategies and choices that will tend to produce equally good (or equally bad) performance. More than one path could lead to Rome, in other words, with different people making idiosyncratic, although strategic and effective fixation choices.

That this could be what takes place is consistent with another recent study. Lukavskỳ (Lukavskỳ, 2013) reported that participants make correlated eye movements across repeated individual trials, that participant A's eye movements in trial one will be similar to his/her eye movements during the same trial if it appears again in the experiment session. Moreover, A's eye movements will be different from participant B's eye movements in the same repeated trials. This suggests that eye movements could be individualized responses to the evolution of events over the course of a trial. But correlated eye movements could also reflect learning through contextual cueing, which is known to take place in MOT and to improve performance over repeated trials (Ogawa, Watanabe, & Yagi, 2009). The results are also indeterminate with respect to whether performance differences at an individual level are related to eye movement differences.

At this time, the most one can say about eye movements and MOT performance is that observers execute systematic fixation choices, at least at a group (if not individual) level. Fixations are presumably localized to reduce item confusability and in response to the moment-to-moment configuration of the display. Yet there remains a lack of direct evidence connecting individual differences in fixation to differences in performance. A major challenge here is the possibility of divergent paths to the same performance outcomes. If there is more than one path to the same end, it will be difficult to diagnose through an experimental approach that seeks to classify fixations by strategy and then to correlate performance with the frequency of a given strategy.

1.3. MOT individual differences

Before moving to the current study, one may wonder whether there are systematic individual differences in MOT performance in the first place. Here too, the research is surprisingly limited, only three relevant studies that we are aware of. Drew and Vogel (Drew & Vogel, 2008) identified two groups of participants in a test cohort, one with a high tracking capacity and one with a lower capacity. This group-level difference in performance paralleled a group-level difference in the amplitudes of two ERP components, the CDA (contralateral delay activity) —thought to correlate with working memory capacity— and the N2PC —thought to reflect attentional selection. Thus, measurable ERP differences correlate with group-level differences in performance. Similarly, individual differences in MOT have been found to correlate with individual differences in other tasks that measure aspects of visual cognition, attention, and working memory (Huang, Mo, & Li, 2012; Oksama & Hyönä, 2004). The scant evidence available therefore suggests systematic as opposed to random reasons for why some participants perform better than others. Here we suggest that fixation differences are the primary determinant of systematic performance differences.

1.4. The current study

The current report takes a novel approach to confirm a link between fixations and individual task performance. We devise a model that adopts individual participant fixations so that we can observe the impacts of those fixations on the model's performance. As will become evident, this approach does not require that we classify individual fixations into predetermined categories. We discuss the model's specific implications in the case of MOT, and as a general framework for thinking about how unobservable mental computations interact with input-seeking behavior.

2. Methods

2.1. Participants

50 Johns Hopkins undergraduate students participated for course-related credit. All the participants had normal or corrected-to-normal visual acuity and completed informed consent prior to participation. The protocols of all the reported experiments were approved by the Homewood Institutional Review Board of the Johns Hopkins University. In accordance with the approved IRB, demographic information was collected anonymously and without identifiers linked to specific experimental results, used only for year-end reporting.

2.2. Trial structure and procedures

Each participant completed 120 trials of Multiple Object Tracking (MOT), equally distributed across a combination of six loads (number of targets: 3, 4, 5, 6, 7, 8) and four speeds (2.8, 5.6, 8.4, 11.2 deg./s), such that there were five trials in each condition. The total number of moving objects in a trial was always twice the number of targets (i.e. the number of nontargets equaled the number of targets).

At the start of each trial, targets were identified by turning yellow, before turning blue to match the nontargets. Targets and nontargets then began to move on independent trajectories for a duration of 10 s. They moved with constant velocity until they collided with one another or with the edges of the display, at which point they bounced while conserving momentum. At the end of motion, the participant was instructed to mark all the targets using a mouse. The participant was required to mark as many items as there were actual targets, no fewer or more. Participants were instructed to guess when uncertain.

All 120 trials in the experiment were pre-generated, so that each participant completed the same set of trials, though in a randomly distributed order across four blocks. The experiment took one hour to complete.

2.3. Display apparatus

Stimuli were presented on an LCD monitor with a refresh rate of

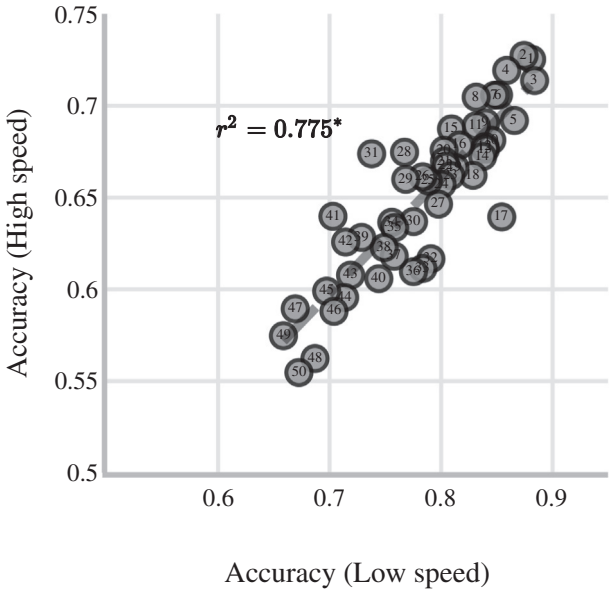


Fig. 2. Split half analysis of tracking performance comparing high speed and low speed trials. Numbers on the markers indicate the overall rank of the participant in the experiment.

60 Hz, controlled by a Mac mini (Apple Inc., Cupertino, CA). The viewing distance was 55 cm (fixed by headrest) so that the tracking area subtended 40.4° × 30.7° of visual angle. The size of each disc was 0.6° in diameter.

2.4. Eye tracking

We used an SR Research Eyelink 1000 tracker with the headrest tower. We recorded the data from the right eye for all the participants. We calibrated their eye movements at the beginning of the trial block using a 13 point calibration with the eye tracker reporting an average accuracy of 0.45° ± 0.02 (avg ± 95 % CI) in the horizontal and vertical direction. The calibration was done at the beginning of each block to avoid the accumulation of drift errors over time.

The tracker recorded at 500 Hz for the first 25 participants, and at 1000 Hz for the remaining 25 participants. (This difference was for not for any substantive reason. The 500 Hz setting leftover erroneously from a previous study). Thus the eye tracker collected 5000 or 10,000 samples in each 10 s trial. The data obtained were classified as Saccades, Fixations, and Blinks using the Eyelink native online parser, with the following criteria: Saccades were classified when eye velocity was greater than 30°/s, acceleration greater than 8000°/s, and Saccadic motion was no larger than 0.15°; blinks were classified when the pupil was not detected; the velocity threshold for saccadic motion was automatically adjusted to 60°/s by the online parser during the detection of smooth pursuit movements.

Our analysis and modelling used participant *fixation locations*, defined as follows. In this experimental setting, a fixation is defined as a period of time during which a specific part on the screen is looked at and thereby projected to a relatively constant location on the retina. This is operationalized as a relatively still gaze position in the eye-tracker signal implemented using the EyeLink algorithm with the Cognitive configuration as outlined in the EyeLink 1000 manual v 1.5.

We limited our analyses to the events that were classified as *fixations* by the software. Here and elsewhere in this report *fixations* refers to locations in screen-based coordinates. We further down-sampled the classified fixation events to synchronize with the display refresh rate (60 Hz). We did this by matching the time of all of the fixation periods with the 600 display video frames for the entire ten second duration of a trial. Any motion frame that fell between the ending time of the $(i)^{th}$ fixation period and ending time of the $(i + 1)^{th}$ fixation period was assigned the average gaze location of that fixation period. Further, any motion frame that contained blinks was assigned a gaze location that was last recorded before the blink occurred. After this pre-analysis, we ended up with 600 pairs of eye gaze coordinates that corresponded to the 600 motion frames. These coordinates were used in the model simulations of the task, and the related behavioral analyses.

2.5. A note on the use of the term ''fixations'

For simplicity, we use the term 'fixation(s)' throughout to refer to the eye tracking data recorded and then extracted as just described in section 2.4. Specifically, in each instance, the term refers to a pair of eye gaze coordinates (groups thereof when plural) from among the 600 pairs obtained for each participant in each trial of the experiment, each pair referenced with respect to the monitor and linked with a single monitor frame.

3. Behavioral results

3.1. Behavioral results 1: reliable individual differences in MOT

We sought to confirm that participants with overall better performance on the task were more likely to perform the task well throughout the experiment, not just during one epoch or condition. To do this, we conducted a split half correlation across participants using object speed to divide the data. For each participant we computed average tracking accuracy for the slower speed trials (2.8° and 5.6° per second —half of all trials) and for the faster speed trials (8.4° and 11.2° per second —the remaining half of trials). Fig. 2 shows the correlation in performance by participant in these two halves of the data. Each point is a participant, and the number (1–50) labeling each point reflects that participant's rank among the participants across the experiment. What the graph shows is a strong and significant correlation ($r^2 = 0.775$, $p < .001$) for performance in high and low speed conditions. Trials were interspersed in a different random order for each participant so that the correlation observed reflects consistency in performance as opposed to the vagaries of trial order and/or effort distribution.

We then performed a similar analysis, this time dividing the data between the lower target loads (3–5) and the higher loads (6–8). Those results are shown in Fig. 3. Again, performance was consistent by participant, with a significant and strong correlation between the two halves of the data ($r^2 = 0.569$, $p < .001$).

3.2. Behavioral results 2: general task performance

In addition to individual differences, we analyzed the effect of task difficulty on performance at the group level. The results were typical. Fig. 4 shows that performance decreases as a function of task difficulty, i.e. accuracy in a given trial decreases as target load increases, and also as speed increases. We also visualized these results in a less common format: Fig. 5 plots the frequency with which a given number of targets was chosen (on average) for each tracking load. The purpose is to illustrate that participants accurately chose five or more targets in fewer than 30% of the available trials.

3.3. Behavioral results 3: quantifying fixation strategies

The motivation for this project is the hypothesis that individual differences in tracking performance can be explained causally as a consequence of moment-to-moment inferences about target identities. We argue that testing the hypothesis requires a model that adopts observer fixations on a moment-to-moment basis. But perhaps there is a single strategy for selecting where to fixate, one that could produce good performance for obvious causal reasons? The primary contender is a strategy of fixating the centroid of the targets in a given display at a

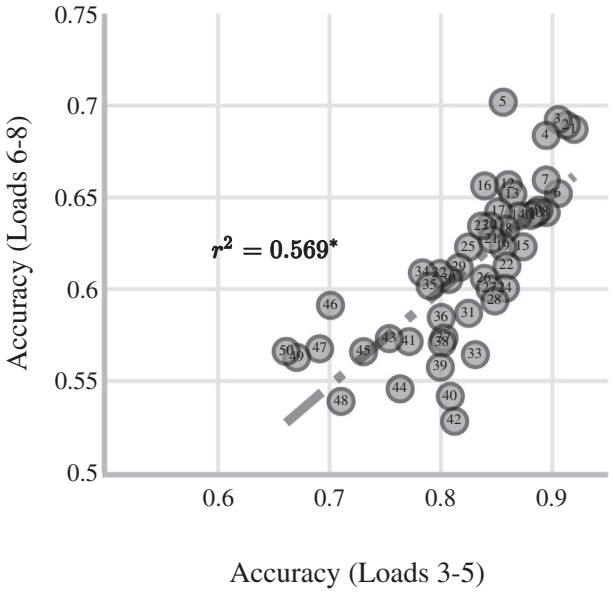


Fig. 3. Split half analysis of tracking performance comparing high load and low load trials. Numbers on the markers indicate the overall rank of the participant in the experiment.

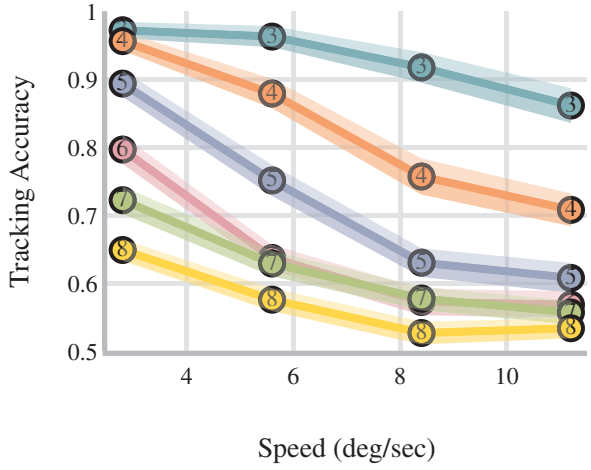


Fig. 4. Overall tracking performance (N = 50) as a function of speed (^{o}/sec) and tracking load. Tracking loads differentiated by color and numbered labels inside markers. Shading reflects 95% confidence intervals.

given moment. Previous research (Fehd & Seiffert, 2008; Fehd & Seiffert, 2010; Zelinsky & Neider, 2008) has suggested that participants do regularly fixate target centroids. We therefore sought to quantify the proportion of time that participants look at the target centroid, and then to test the hypothesis that better trackers spend more time looking at target centroids than do poor performers.

Previous studies have tried to quantify fixation strategies under the assumption that only two strategies are available: look at objects, or look at object centroids (Fehd & Seiffert, 2008; Zelinsky & Neider, 2008). If a fixation falls outside a window that reasonably includes an object, that fixation is automatically classified as centroid directed. We applied a different analysis with the expectation that a large number of fixations are directed towards empty space, but not at the centroid —fixations that could reflect strategic, moment-to-moment dependencies. We therefore classified fixations as centroid directed or object directed if they were within a 4^o radius of the target-defined centroid or any individual object, respectively. All fixations that failed to meet either criterion were classified as 'other,' i.e. non-centroid directed empty space fixations. Fig. 6 plots the resulting classifications,

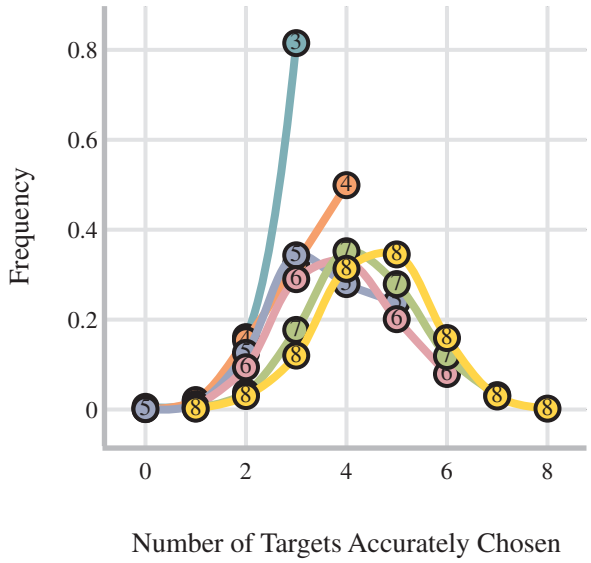


Fig. 5. Frequency with which a given number of targets was accurately chosen as a function of tracking load. Tracking loads differentiated by color and numbered labels inside markers.

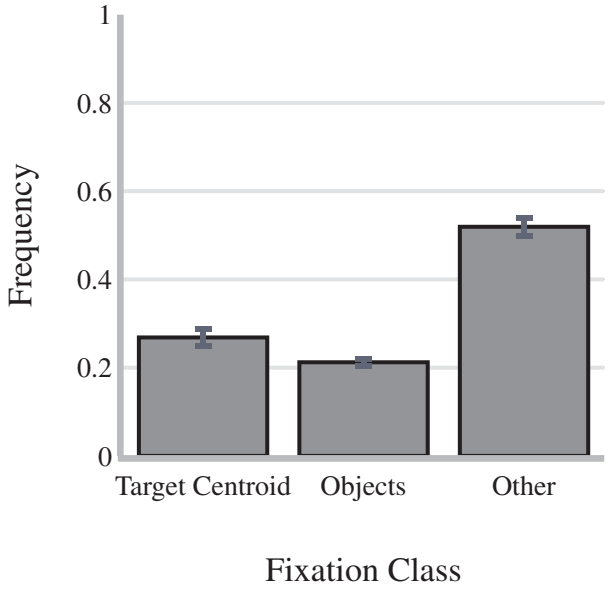


Fig. 6. Proportion of fixation time by strategy. N = 50. Error bars reflect 95% confidence intervals.

demonstrating that fewer than 50% of fixations conform to either the centroid or the object fixation strategies. More than 50% of fixations are therefore found in non-centroid empty space. (When we repeated the analysis with a 2^o radius the same was true for 80% of fixations). Given the relative infrequency with which participants direct fixations to the centroid, there should be a great deal of performance variance that remains to be explained by non-centroid, empty space fixations.

Next, for each observer we computed the average amount of time he/she spent fixating within the target centroid for the entire experiment. We correlated that average time spent with the participants' ranked performance scores over the course of the experiment. Better ranked individual trackers did spend more time looking at the target-centroid than those with lower performance ranks, as shown in Fig. 7a and b. But the correlation captured only 19% of the variance. One way to summarize all this is as follows: Looking at the target centroid may be a good strategy if it is applied at the right times. But doing this task

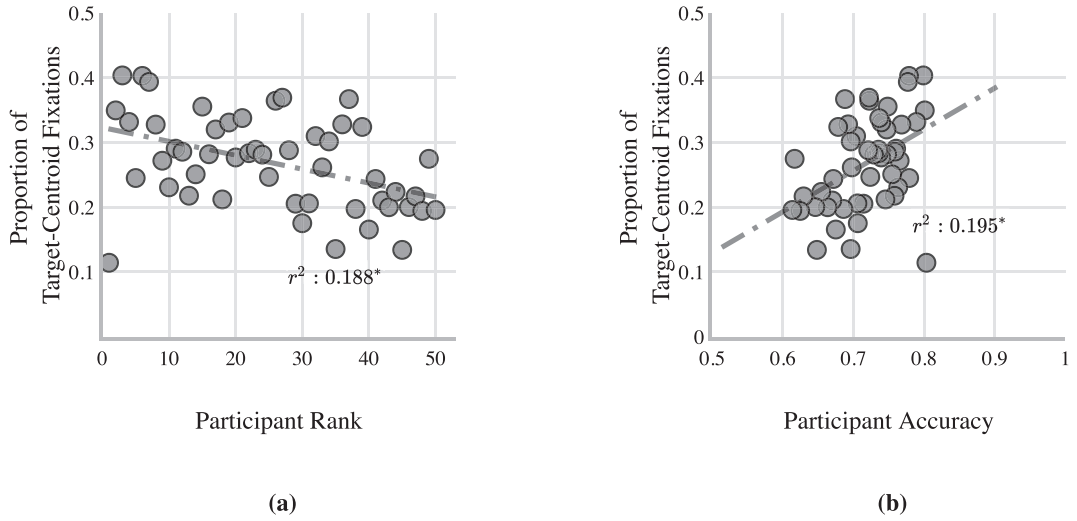


Fig. 7. Percentage of time a participant (N = 50) fixated the target-centroid correlated against the individual's performance rank (a), and accuracy (b) in the experiment overall.

well involves controlling not only where one looks, but also when.

We conducted one more analysis, in order to consider the possibility that some individuals make larger fixation changes than others, potentially causing or reflecting individual differences. We computed the average spread of fixations over the course of each trial, for each individual participant. The average spread of the fixations is operationalized as the overall variation in the fixations made by a participant relative to a reference point (the origin). We computed the distances from each fixation to the origin averaging across trials within participant. Larger numbers (in degrees) reflect fixations that deviate more from one another. Fig. 8 plots the results of this analysis against participant rank. Spread of fixations only explained about 20% of the observed individual differences. This result could reflect the fact that poor performers end up moving their eyes more because they confuse targets and distractors, not necessarily that moving one's eyes leads to confusions. The blue dot in Fig. 8 shows the average spread of the target centroid over the course of the trials. Observers by and large change their fixations to a far greater degree than the target centroid changes.

3.4. Interim summary

Our initial analyses of the behavioral and eye tracking results demonstrate that there are individual differences in tracking performance across participants, and that these individual differences cannot be explained entirely in terms of a centroid-directed fixation strategy. These results are consistent with the hypothesis that fixations constrain input quality and target inferences in a more dynamic way.

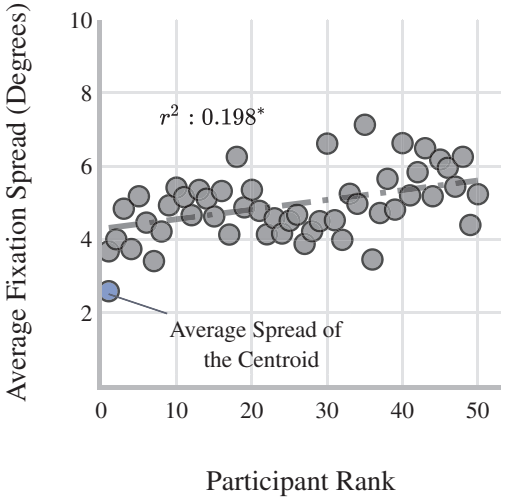


Fig. 8. Average spread of the fixations made by the participants in a trial plotted as a function of the participant rank.

4. Model methods and results

4.1. Overview: a model of fixation-dependent probabilistic tracking

Among the features that makes Multiple Object Tracking an interesting paradigm is that it cannot be accommodated by continuous fixation or attention to one important object, because there are multiple such objects in a given trial. Instead, it demands sustained processing of inputs from an entire display and iterated inferences about target versus nontarget identities.

Previous research has productively formalized the process of receiving input and making identity inferences within a probabilistic framework of state estimation (also called Kalman filtering; (Ristic, Arulampalam, & Gordon, 2004)) combined with nearest-neighbor correspondence assignments (Li, Wang, Wang, & Li, 2010; Vul, Alvarez, Tenenbaum, & Black, 2009; Zhong, Ma, Wilson, Liu, & Flombaum, 2014). The Kalman filter portions of these models estimate the positions of the tracked objects iteratively. Consider tracking a single object: the model behaves as an observer who receives noisy input about the position and velocity of the object, makes a prediction (a prior) about where it should next encounter the object, and it then updates its estimate about the object's current position (a posterior) by optimally combining its prediction with a new noisy observation, over and over until tracking is complete.

With more than one object to track a second challenge arises: the need to identify which observations come from which targets whenever new observations are obtained. Assume that at regular intervals an observer receives noisy observations from *all* the identical objects in the display. The observations provide the participant with an approximate sense of the positions of each of the objects. The trouble is that the observations are unlabeled with respect to their statuses as emanating from targets or nontargets, let alone which targets; during tracking, all the observation-yielding objects are featurally indistinguishable. The participant must therefore make inferences about which observations came from targets and which from nontargets, i.e. which observations correspond with which represented objects. In the model that we present below this is done by minimizing the collective differences in position between each observation and the prior estimate of position for

the object that the respective observation is assigned (see also (Vul et al., 2009; Zhong et al., 2014)). In other words, the model assigns correspondences by searching for a nearest neighbor for each observation and its respective prior, while minimizing the overall error in the complete set of assignments. This is similar to how models of apparent motion perception determine which motion paths are observed, in particular, by assuming that each object in one frame has a mutually exclusive and exhaustive correspondent in the next frame, and by assuming that object positions tend to change by small amounts on average (e.g. (Dawson, 1991)).

What follows is a formal description of a model that implements both Kalman filtering and correspondence assignment. Because the model receives noisy observations of position at regular intervals, it has two additional key properties, a sampling rate and eccentricity dependent spatial noise. Based on previous research describing the temporal profile of visual attention and perception we chose a sampling rate of 20 Hz (the high end of a range estimated to be between 8 and 20 Hz; e.g. (Holcombe & Chen, 2013; Landau & Fries, 2012; VanRullen & Koch, 2003; VanRullen & Macdonald, 2012; VanRullen, Reddy, & Koch, 2005)). For spatial noise we assumed that observations are derived from two dimensional Gaussian distributions centered on the positions of the observation-generating objects. As detailed further below, the variances of these distributions were computed independently for each object at each sampling point, as a function of the respective object's distance from the relevant observer's current fixation. The model therefore adopts the fixations of individual observers in the sense that when it simulates a given participant the amount of noise in each received observation depends entirely on that specific participant's fixation at the time of the observation. The function linking spatial noise to eccentricity is derived from previous research (Carrasco, Evert, Chang, & Katz, 1995; Rovamo & Virsu, 1979).

4.2. Formal model details

N_T denotes the number targets to be tracked and N_D denotes the non-targets such that the total number of objects in any display is denoted as N, where $N = N_T + N_D$, and $N_T = N_D$. At any given point in time, the state of a given item (i) is denoted as $(\mathbf{S})_t^i$. This is a vector containing the position and velocity of the item.

$$(\mathbf{S})_t^i := [(\mathbf{S}_p)_t^i; (\mathbf{S}_v)_t^i] := [x_t^i\ y_t^i\ v_{xt}^i\ v_{yt}^i]^T \tag{1}$$

Here, x_t^i and y_t^i are the x and y coordinates of the item. v_{xt}^i and v_{yt}^i are the respective components of the item (i)'s instantaneous velocity. Let $(\tilde{\mathbf{S}})_t^i$, $(\hat{\mathbf{S}})_t^i$ denote prior and posterior expectations about the true state respectively.

During a sampling frame, observations are received from each of the items in the display. Because the model does not know which item generated which observation, we denote the observations with the superscript m, so that each observation at given time t is denoted as, and includes only a noisy sample of m's true position where:

$$\begin{aligned}(\mathbf{Z}_p)_t^m &= \mathbf{B} * (\mathbf{S})_t^m + (\mathbf{R})_t^m \\ B &= \begin{bmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\end{bmatrix}\end{aligned} \tag{2}$$

$(\mathbf{R})_t^m$ is the noise in the observation for the m^{th} item, assumed to be zero mean Gaussian white noise with covariance as follows:

$$(\mathbf{R})_t^m = (\sigma^2(E))_t^m \mathbf{I}_2 \tag{3}$$

where $\mathbf{I}_2$is a [2 × 2] identity matrix. In our model, we utilized an eccentricity dependent σ, meaning that the position of the eyes control the amount of noise in each individual observation (Carrasco et al., 1995; Rovamo & Virsu, 1979) as follows:

$$\sigma(E) = c * (1 + 0.42 * E) \tag{4}$$

The value of c in our model is 0.08, based on unpublished experiments in our lab, and consistent with a value used in a previous modelling study (Vul et al., 2009). In our model, each target is tracked by its own Kalman filter with its own a-priori and a-posteriori error covariances (4 × 4), denoted as $(\tilde{\mathbf{P}})_t^i$ and $(\hat{\mathbf{P}})_t^i$ respectively. Following standard Kalman filter procedures the model employs a prediction and an update step, along with a correspondence step in-between as necessitated by the MOT paradigm.

4.2.1. Prediction step

A key feature of model is that it generates predictions about the upcoming state of the system. Given a posterior belief about where a tracked target is $(\hat{\mathbf{S}})_{t-1}^i$, the model generates a prediction —a prior $(\tilde{\mathbf{S}})_t^i$— about where it will next find the object, that is the next time that it receives observations. It does this using a state transition matrix that is governed by the Newtonian laws of motion, as follows:

$$(\tilde{\mathbf{S}})_t^i = \mathbf{A}(\hat{\mathbf{S}})_{t-1}^i \tag{5}$$

where

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix}1 & 0 & 1 & 0\\ 0 & 1 & 0 & 1\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix}\\ (\tilde{\mathbf{P}})_t^i &= \mathbf{A}(\hat{\mathbf{P}})_{t-1}^i \mathbf{A}^T + \sigma_s^2 \mathbf{I}\end{aligned} \tag{6}$$

4.2.2. Update step

A typical Kalman Filter knows the correspondence between observations and objects. We therefore describe the update step of our model before describing the correspondence procedure. It is important to understand, however, that this updating takes place after the correspondence step provides its output, a mutually exclusive and exhaustive mapping between each of the tracked targets and a new position observation.

Accordingly, updating proceeds given a set of *labeled observations* $\mathbf{Z}_t^i$ and the Kalman gain $\mathbf{K}_t^m$ [4 × 2] matrix at time step t. This Kalman gain is used to optimally weigh in the prior and the observations at time t to generate the posterior for that time frame.

$$\begin{aligned}(\mathbf{K})_t^i &= (\tilde{\mathbf{P}})_t^i * \mathbf{B}^T * (\mathbf{B} * (\tilde{\mathbf{P}})_t^i * \mathbf{B}^T + (\mathbf{R})_t^i)^{-1}\\ &= [(\mathbf{K}_p); (\mathbf{K}_v)]_t^i\\ (\hat{\mathbf{S}}_p)_t^i &= (\tilde{\mathbf{S}}_p)_t^i + (\mathbf{K}_p)_t^i((\mathbf{Z}_p)_t^i - \mathbf{B} * (\tilde{\mathbf{S}})_t^i)\\ (\hat{\mathbf{S}}_v)_t^i &= (\mathbf{Z}_p)_t^i - (\mathbf{Z}_p)_{t-1}^i\\ (\hat{P})_t^i &= (\mathbf{I}_2 - (\mathbf{K})_t^i * \mathbf{B}) * (\tilde{\mathbf{P}})_t^i\end{aligned} \tag{7}$$

4.2.3. Correspondence step

At a given time step 't', the model has access to unlabeled bag of observations $\{\mathbf{Z}_t^m\}$ where $m \in M = \{1,2,3,...N\}$, and does not know which observations correspond to which objects in the a-priori state $(\tilde{\mathbf{S}}_p)_t^i$ in order to generate the a-posteriori state estimate $(\hat{\mathbf{S}}_p)_t^i$ of the targets. The assignment vector i is obtained by maximising the probability that an observation resulted from a given a-priori state while minimizing the associated overall cost $\mathbb{C}$ for this assignment vector where

$$\mathbb{C} = \sum_{i,m} d((\tilde{\mathbf{S}}_p)_t^i, (\mathbf{Z}_p)_t^m) \tag{8}$$

and $d(.)$ is the Euclidean distance. We used the python inbuilt linear optimal assignment package (scipy) to compute the above correspondence vector.

4.3. Model simulations

To simulate the tracking performance of experiment participants, we did the following. The model performed each trial completed by

each participant. On the first frame of each simulation, the model received accurate input about the positions of each of the targets in the display. On each subsequent monitor frame the model either received input or not depending on its refresh rate (i.e. at 20 Hz, the model received input on every third monitor frame). When the model received input, it consisted of an identity-unlabeled position sample from each object in the display, both the targets and nontargets. The samples were drawn from two dimensional Gaussian distributions, each centered on the true location of the relevant source object, and with a σ value determined by Eq. (4) (Carrasco et al., 1995; Rovamo & Virsu, 1979) as described above. The fixation position of the model during each input frame was the recording-derived fixation position of the observer during that frame, as described in Section 2.4. The model completed each trial 100 times as a given participant. Noisy samples were drawn independently during each simulation. Each simulation ended by recording the items that the model had labeled as targets, such that the model could identify anywhere between 0 and N_t items correctly in one simulation. Average performance across the 100 simulations of a trial served as a prediction for how well a given participant should perform in a given trial.

4.4. Model simulation results 1: general task performance

Qualitatively, the model reproduced typical patterns of human performance in MOT, reduced tracking performance as speed and load increase. Fig. 9 plots model performance as a function of object speed and tracking load. The figure was generated by analogy to Fig. 4, by averaging together the performance of all the simulated participants at each tracking load and speed. At the smallest speeds and loads the model under performed human participants, although broadly, it outperformed. Fig. 10 (right panel) is drawn by analogy to Fig. 5 (reproduced for comparison, in the left panel). It depicts the frequency with which simulated participants correctly picked a given number of targets given some number of targets to track. Here too, the model mostly outperforms the human cohort, for example, picking 4 targets out of 6 correctly more than 60% of the time whereas the participant cohort did the same less than 40% of the time. Importantly, though, the model's performance, visualized in this way, is qualitatively similar to participant performance. Notably the model's effective tracking capacity is limited even though there are no capacity limits built-into it. For example, the model never identified eight targets correctly (given a load of eight). It was accordingly in the position of representing (and tracking) eight items in the relevant trials, some of which were always nontargets (in point of fact) by the end of each such trial.

4.5. Model simulation results 2: individual differences

The model significantly predicted human performance at an individual level ($r^2 = 0.38$, $p < 0.005$), and individual rankings within the group ($r^2 = 0.40$, $p < 0.005$). Fig. 11 plots the model's predicted accuracy for each of the 50 observers (across the entirety of the experiment) against measured accuracy and also the model predictions about observer rankings within the cohort of 50 participants. Recall that we did not fit any parameters to the model on an individual observer basis. (In fact, we did not fit any parameters at all. Instead we fixed the values of the relevant parameters based on extant research pertaining to spatial observation noise and the temporal frequency of perception). And the observers all completed the same exact trials. Adopted fixations were the only factor that differed when the model simulated one participant or another in a given trial. Fixation differences are therefore the only variable that can explain how and why the model made reliably different predictions about the performances of different observers.

The model also can be used to make predictions at a more granular level. Fig. 12 (left panel) shows the model's average performance over 100 simulations of a given trial as a given participant, correlated against each participant's actual performance in the relevant trial. There are 6000 points in the graph, each reflecting one trial and a single participant. The strong correlation ($r^2 = 0.824$, $p < 0.05$) demonstrates how the difficulty of a given trial was similar for observers and their simulations. The right panel of Fig. 12 makes a similar point ($r^2 = 0.825$, $p < 0.05$), by reducing the number of points to the 120 trials in the experiment. Each observer completed the same set of trials. Accordingly, the graph plots average group performance on a given trial correlated with average performance of the simulated group. The trials are further coded by the tracking load (using colors and numbers). The figure demonstrates how the model performed more poorly at higher tracking loads, like the observers. More interestingly, within a load, the model found specific trials more or less challenging similarly to observers. Note for example that for loads of six and seven targets there was a wide range of performance variance across individual trials, both for human observers and for the model. This underscores the extent to which task performance depends not just on tracking load (or speed), but also on the incidental and specific dynamics of a given trial, combined with how an observer directs fixations in the respective trial.

Finally, we sought to determine whether fixation locations and *time* explain performance better than just the fixation locations. Maybe good trackers just pick good positions to fixate? Maybe there are one or more good places to point one's eyes in a trial at any time, independent of the particular constituents of the trial at a given moment or its evolution? The results presented so far all suggest that this is unlikely. We sought

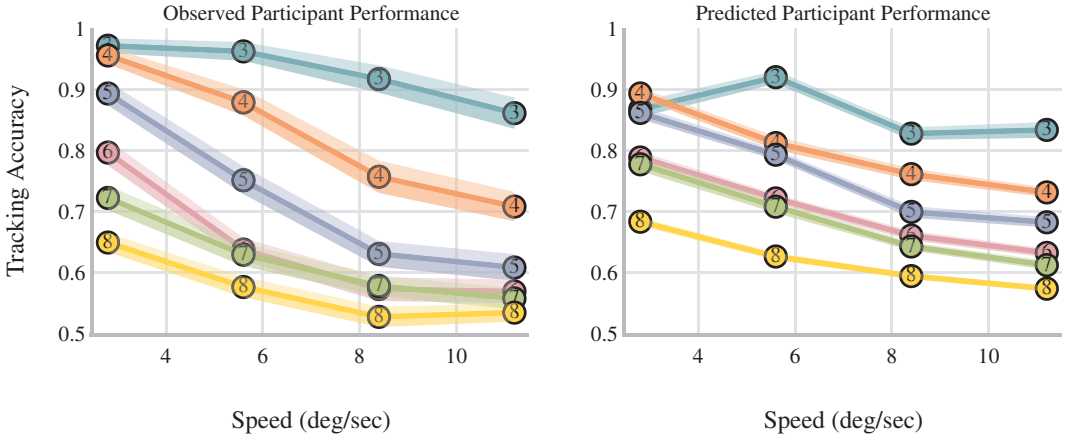
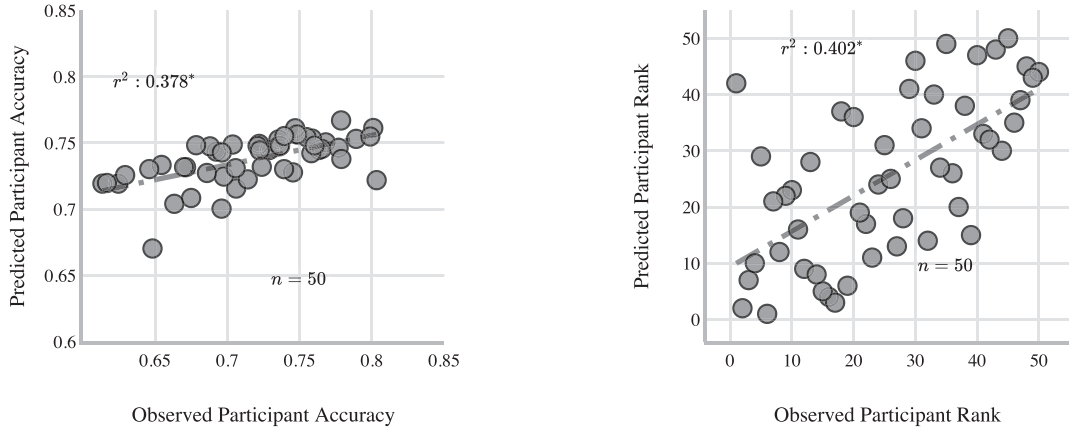


Fig. 9. Overall model (right panel) tracking performance as a function of speed (°/sec) and tracking load. Tracking loads differentiated by color and numbered labels inside markers. Shading reflects 95% confidence intervals. Model performance juxtaposed with human participant performance (left panel) also shown in Fig. 4.

Fig. 10. Frequency with which a given number of targets was accurately chosen as a function of tracking load. Tracking loads differentiated by color and numbered labels inside markers. Left panel shows human performance (reproducing Fig. 5), while right panel shows model performance.



(a) **(b)**

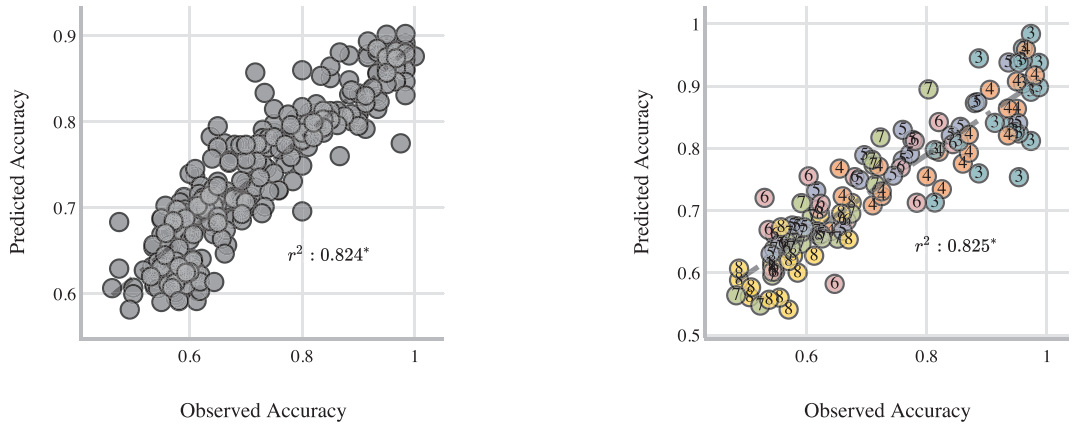
Fig. 11. Predicted accuracy vs observed accuracy by participant (left panel) and predicted observer rank vs. observed ranking (right panel). Each circle is one of the 50 participants.

further confirmation by re-simulating participants, with fixations in reversed order for a given trial. The final fixation made by the participant was treated as the first fixation input to the model (and vice versa); the penultimate fixation was treated as the second fixation input, vice versa, and so on and so forth. This of course preserves a good deal the temporal structure relating fixations and display configuration in the middle of a trial. Even so, the model only captured about 20% of performance variance by participant, as shown in Fig. 13, emphasizing the importance of the where and also the when of fixations in MOT.

5. General discussion

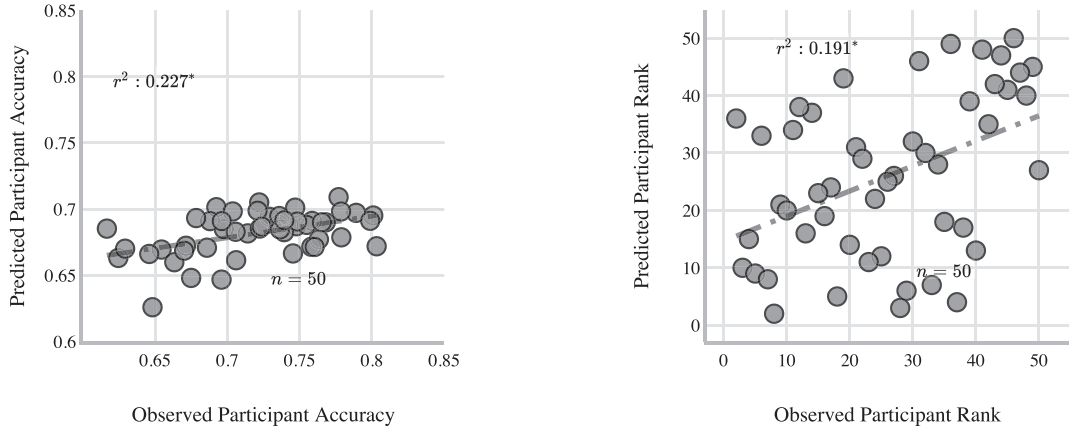
We sought to investigate the relationship between fixation choices and cognitive performance in the Multiple Object Tracking (MOT) task. The task was chosen for a number of specific reasons. It demands sustained effort over an extended period of time. Performance is limited —usually between three and five targets— and performance varies across participants. It is a task where we should expect choices about where to fixate to constrain performance, although research has not previously identified a causal link in this context. And it is a task in which good fixations should often select empty space, distinguishing it from many other research contexts, which emphasize object fixations. Finally, we anticipated that causally linking fixation locations and performance would require a moment-to-moment model of the interaction between fixation-determined inputs and mental computation. Ongoing or iterated computation that interacts with changing input is a likely aspect of a great deal of human real-world interaction. With the possible exception of visual search (Eckstein, 1998; Eckstein, 2011; Eckstein, Thomas, Palmer, & Shimozaki, 2000; Najemnik & Geisler, 2005) however, research in visual cognition tends to focus on tasks with short, punctate deployments of mental effort over the course of one or only a few fixations. Here, we saw an opportunity to leverage computational progress (Vul et al., 2009; Zhong et al., 2014) in order to investigate a sustained, interactive loop between fixations, input, and efficient mental computation in the case of MOT.

We tested fifty participants in a set of identical tracking trials across a range of speed and tracking loads, while we also recorded where on the screen they fixated. We found typical effects of speed and load on performance at a group level, and also consistent differences in overall performance across individuals. We then implemented a computational



(a) **(b)**

Fig. 12. Left panel: Average accuracy (Model vs. Observers) in each individual trial, for each individual observer. There are 6000 points (50 participants each completing 120 trials). Right panel: Average accuracy (Model vs. Observers) in each individual trial, averaged across observes. Each point is one of the 120 experimental trials, coded for target load by color (and number labels).



(a) **(b)**

Fig. 13. Left panel: Average accuracy (Model vs. Observers) in each individual trial, for each individual observer. There are 6000 points (50 participants each completing 120 trials). Right panel: Average accuracy (Model vs. Observers) in each individual trial, averaged across observes. Each point is one of the 120 experimental trials, coded for target load by color (and number labels).

model of MOT, for the first time providing it with noisy inputs that depend on where on the screen a simulated observer fixates at each moment in the trial. This effectively allowed the model to simulate each individual observer by adopting their individual fixations, allowing us to observe the effects of different fixations without classifying them in strategic terms. Without fitted parameters the model made reliable predictions about individual participant performances and it reliably ranked observers relative to one another. The results suggest a causal link between fixations and performance at an individual level because fixations were the only variable that we manipulated when the model simulated one participant or another in a given trial. The results also demonstrate how some of the general performance limitations associated with this task emerge from inherent computational uncertainty.

5.1. What we used to think about eye movements, individual differences, and MOT

We began this project with an expectation derived from first principles, that eye fixations should constrain MOT performance. We therefore hope that the results appear obvious, at least in retrospect. Yet there is surprisingly little research on fixation during MOT. In fact, most studies with the paradigm do not mention whether observers were asked to fixate or allowed to freely move their eyes, presumably because the prevailing assumption is that eye movements do not have much of an effect on performance.

Why would this be the prevailing assumption? We suspect that it owes to the assumed importance of attention as a constraint on task performance, and also to the early suggestion that eye movements do not impact performance; specifically, that task performance is no different when participants are forced to fixate the center of a display or allowed to freely move their eyes (Scholl & Pylyshyn, 1999). But there are reasons to doubt this claim beyond our results. For one, it was made in the context of experiments with relatively smaller tracking loads (four or fewer), whereas fixation selection may account for more variability as task difficulty increases. Second, the claim was made with reference to task performance on average, meaning that eye movement differences may account for individual differences that average out in a group-level comparison between free movement and fixation

conditions. It may also have to do with the fact that understanding the full causal relationship between fixations and performance requires an analysis on a moment-to-moment basis, where computational confusions do or do not occur only to manifest in performance at the end of a ten second (or longer) trial. The extended and sustained nature of the task makes it especially difficult to explain without a computational model that performs the task, an opportunity that only arose recently.

As a consequence, only four previous studies that we are aware of supply evidence linking performance and fixations. Consistent with the idea that fixation choices reflect moment-to-moment responses to the display, Lukavskỳ (Lukavskỳ, 2013) reported that participants make correlated eye movements across repeated individual trials. Fehd and Seiffert (Fehd & Seiffert, 2008; Fehd & Seiffert, 2010) found that a meaningful proportion of fixations are close to the centroid of the polygon formed by the targets at a given moment (see also (Yantis, 1992)). Zelinsky and Neider (Zelinsky & Neider, 2008) found the same, though for target loads of only two to four, whereas for loads greater than four the proportion of fixations to the centroid declined, replaced by an increase in shifting fixations to individual targets. It is important to note that in these three studies fixations were classified competitively: each screen-referenced pair of fixation coordinates was classified as either directed at an object or directed at the centroid. As a result, if a fixation to empty space was too far from an object to classify it as object-directed, then it was classified as centroid directed.

We analyzed fixations differently, only classifying them as centroid-directed when they were within a 4^o radius. And rather than correlate classes of fixations with overall performance, we evaluated the cumulative effects of serial fixations by using those fixations to determine the momentary inputs to a model that performed the task; when we reversed the order of the inputs to the model, it explained a great deal less of the performance variance (Fig. 12). The results are consistent with previous findings in the sense that they suggest systematic fixation patterns and they implicate fixations as a key determinant of performance. Our analyses further suggest that most fixations are directed neither to individual targets nor to centroids, and that those empty space fixations vary between individuals in a way that explains performance differences. Recall that the amount of time an observer spends fixating the centroid explained only a small amount of performance variance (Fig. 7) and also that participants changed fixation positions far more than the centroid position actually changes (Fig. 8). These results suggest that (especially good) trackers choose fixation locations that are not always the centroid nor individual items in order to maximize input quality in relation to the evolving risk of confusing different targets. The key determinants of performance are where one chooses to look, *and also when*.

5.2. What limits performance in the first place?

Understanding the capabilities of visual cognition requires an explanation of its limitations. Not surprisingly, examples of processing limitations often motivate large research foci. Think, for instance, of inattentional blindness, the attentional blink, and change blindness: classic case studies that elucidate how visual perception, attention and memory work by pointing to their limits. Multiple object tracking has long been deployed in experiments with a similar logic. We can perhaps understand how discrete objects are segmented and represented by asking what kinds of things we can track, what we cannot, and by asking how many is too many (e.g. (Scholl, Pylyshyn, & Feldman, 2001; VanMarle & Scholl, 2003)). In the earliest research on tracking, the apparent limit of about three to five simultaneous objects was taken as evidence of a discrete imposed limit on the underlying token representations used to index and update one's knowledge of objects in the world (Pylyshyn & Storm, 1988). In step with ongoing debates about the nature of visual working memory, later experiments were taken as evidence of more continuous resources that impose tradeoffs on representational resolution when those resources must be shared among larger sets of objects (Alvarez & Franconeri, 2007; Horowitz & Cohen, 2010; Ma & Huang, 2009). Both kinds of theories share the assumption that limitations in performance are imposed, limitations on the input to tracking which arise because some kind of vehicle needs to be consumed in order to deliver the input.

We intentionally did not impose limits of any kind upon the model in our study. It does not possess discrete limits in the sense that it can represent as many as eight targets at once (even if some of the eight represented-targets are actually nontargets; see Fig. 10). Similarly, we did not impose any resource-like limits on the model, which could manifest as increased noise and/or a reduced sampling rate as tracking load increases. The reason for avoiding imposed limits was a concern that they would make it difficult to identify the causes of individual differences. In particular, if we fit parameter values that could vary by participant we would presumably end up fitting the individual differences that we were trying to explain. The result is a model that mostly outperformed participants, but one that makes it possible to isolate the variability between participants that must be caused by differences in fixation, those being the only changing variable and one we recorded rather than fit.

Even without imposed limits, the model did show qualitatively similar effects of speed and load to the the human participants. Fig. 10 juxtaposed model performance at the group level with human performance in terms of number of items selected by load. Salient features of human performance include a modal tendency to report only four items correctly, regardless of the assigned target load, and the near absence of trials in which seven or eight targets were accurately reported. These are the kinds of effects that may seem to point to discrete, imposed capacity limits. But note that there are similarities in model performance. The model also never identified eight targets correctly even though the model did always represent eight targets when so assigned. The results therefore demonstrate how some of the costs associated with load and speed reflect inherent computational uncertainty rather than consumption limits somewhere in the hardware of the visual system. Two potentially generalizable implications with respect to how to investigate the sources of limitations in visual processing: there is something to gain by looking through the lens of individual differences, and it is useful to have a model of the mental computations that need to be performed in order to accomplish the task.

5.3. Improving the model

Although the model presented had effective limitations on its performance of the task, it outperformed human observers overall. Similarly, while the model produced very high correlations with human performance on a trial-by-trial basis (Fig. 10), correlations with individual participant accuracy and rank were a little more than half the value of the split-half correlations between participants and themselves (Fig. 11 compared with Figs. 2 and 3). A number of technical and theoretical improvements could serve to close this gap.

On the technical side, model predictions likely suffer from necessities that arose in the processing of eye tracker data. In particular, relating eye tracker outputs to individual monitor frames required significant down sampling (with the monitor running at 60 Hz). This meant that the coordinates used as a fixation for a given frame were actually an average of coordinates recorded over the period of time that a monitor frame lasted. Portions of these longer periods may not have been fixations at all, including eye movements and blinks. These limitations should only add noise to the model predictions. That we found significant effects regardless demonstrates how much individual eye position choices have an impact on individual performance. Future studies may find stronger effects by better synchronizing eye tracker recordings with screen-based events.

Perhaps more importantly, the model includes likely oversimplifications of processing reality in the human mind and brain. Specifically, we employed a refresh rate of 20 Hz in our simulations.

The number itself was derived from related research concerning the rhythms of attention and perception. But frames and their respective rates are probably no more than useful analogies for more continuous processing in the visual system, in this case analogies that also provide a useful operational approach to supplying the model with inputs. Similarly, our model was handicapped by eccentricity-dependent input noise. Again, useful for operational purposes and as an analogy to visual processing, but not necessarily a perfect characterization of how unit density in the retina and in visual cortex affect perception.

Finally, the model implemented what we view as the bare-bones of the computation *necessary* to accomplish this tracking task. This meant that the model always 'knows' exactly how many targets it is meant to track in a trial, something that human observers have shown uncertainty about (Ma & Flombaum, 2013). The model also always continues to track the full expected load. It is known that human observers sometimes 'drop' targets. They give up, lose confidence in the status of certain items, or otherwise become confused about the number they should be tracking in the first place (Drew, Horowitz, & Vogel, 2013). The dynamics of stopping or failing to track a target are not well understood, and this makes it difficult to add such inflection points into the model. That human observers drop items and guess at the end of the trial is a key difference with the model we presented, a model that never guesses. This may well account for some of the model's tendency to over perform, as well as some of the individual participant differences that the model could not account for. A better understanding of how, when, and why participants track fewer (or more) items than assigned will be important for future progress.

6. Conclusion

We demonstrated that differences between participants in terms of where they locate fixations can be causally linked to performance differences in the specific case of multiple object tracking. The MOT task, like many others, has attracted research attention because it provides clear evidence for processing limitations in human perception and cognition. These limitations can be observed at a group level —in so far as everyone makes mistakes when asked to track more than four objects— and they can be observed at the level of individual differences —in so far as some people seem to have more tracking capacity than others. What causes these limitations? At both the group and individual level, the traditional answer appeals to a mental or neural resource of some kind, that performed limits reveal a limited internal capacity. We have shown that some performance limits emerge organically from an interaction between fixation-dependent inputs and the mental computations that they supply. Because the model implemented was identical across participants, except through adoption of unique fixations, it provides a computational account of individual differences in a task where those differences have been assumed to reflect reserves of something commodity-like, such as attention or working memory. Appreciating that some person performed MOT poorly because they looked at the wrong places is importantly different from ascribing their performance to inherent limits or the consumption of internal resources. Most broadly, the results therefore emphasize what may be a general property of cognitive performance: the important role of cognitive and even motor control for effectively utilizing efficient mental computation.

CRediT authorship contribution statement

Aditya Upadhyayula:Conceptualization, Writing - original draft, Formal analysis, Investigation, Methodology.**Jonathan Flombaum**:Conceptualization, Writing - original draft.

Acknowledgments

This research was funded by NSF PAC#1534568 to JIF.

Appendix A. Supplemental information

The dataset used in this research is available on http://dx.doi.org/10.17632/h5zgzrfkxr.2. Please use 'Upadhyayula, Aditya; Flombaum, Jonathan (2020), "MOT individual differences", Mendeley Data, v2 http://dx.doi.org/10.17632/h5zgzrfkxr.2' to cite this dataset.

References

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*, 1–10.
Bae, G. Y., & Flombaum, J. I. (2012). Close encounters of the distracting kind: Identifying the cause of visual tracking errors. *Attention, Perception, & Psychophysics, 74*, 703–715.
Binder, K. S., Pollatsek, A., & Rayner, K. (1999). Extraction of information to the left of the fixated word in reading. *Journal of Experimental Psychology: Human Perception and Performance, 25*, 1162.
Carrasco, M., Evert, D. L., Chang, I., & Katz, S. M. (1995). The eccentricity effect: Target eccentricity affects performance on conjunction searches. *Perception & Psychophysics, 57*, 1241–1261.
Cox, D. D., Meier, P., Oertelt, N., & DiCarlo, J. J. (2005). "breaking" position-invariant object recognition. *Nature neuroscience, 8*, 1145.
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*, 569.
Drew, T., & Vogel, E. K. (2008). Neural measures of individual differences in selecting and tracking multiple moving objects. *Journal of Neuroscience, 28*, 4183–4191.
Drew, T., Horowitz, T. S., & Vogel, E. K. (2013). Swapping or dropping? Electrophysiological measures of difficulty during multiple object tracking. *Cognition, 126*, 213–223.
Eckstein, M. P. (1998). The lower visual search efficiency for conjunctions is due to noise and not serial attentional processing. *Psychological Science, 9*, 111–118.
Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*, 1–36.
Eckstein, M. P., Thomas, J. P., Palmer, J., & Shimozaki, S. S. (2000). A signal detection model predicts the effects of set size on visual search accuracy for feature, conjunction, triple conjunction, and disjunction displays. *Perception & Psychophysics, 62*, 425–451.
Fehd, H. M., & Seiffert, A. E. (2008). Eye movements during multiple object tracking: Where do participants look? *Cognition, 108*, 201–209.
Fehd, H. M., & Seiffert, A. E. (2010). Looking at the center of the targets helps multiple object tracking. *Journal of Vision, 10*, 1–13.
Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature Neuroscience, 14*, 1195.
Henderson, J. M., & Hollingworth, A. (1998). Eye movements during scene viewing: An overview. *Eye guidance in reading and scene perception* (pp. 269–293). Elsevier.
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*, 1–19.
Horowitz, T. S., & Cohen, M. A. (2010). Direction information in multiple object tracking is limited by a graded resource. *Attention, Perception, & Psychophysics, 72*, 1765–1775.
Huang, L., Mo, L., & Li, Y. (2012). Measuring the interrelations among multiple paradigms of visual attention: An individual differences approach. *Journal of Experimental Psychology: Human Perception and Performance, 38*, 414.
Landau, A. N., & Fries, P. (2012). Attention samples stimuli rhythmically. *Current Biology, 22*, 1000–1004.
Li, X., Wang, K., Wang, W., & Li, Y. (2010). A multiple object tracking method using kalman filter. *The 2010 IEEE international conference on information and automation* (pp. 1862–1866). IEEE.
Lukavskỳ, J. (2013). Eye movements in repeated multiple object tracking. *Journal of Vision, 13*, 1–16.
Ma, W. J., & Huang, W. (2009). No capacity limit in attentional tracking: Evidence for probabilistic inference under a resource constraint. *Journal of Vision, 9*, 1–30.
Ma, Z., & Flombaum, J. I. (2013). Off to a bad start: Uncertainty about the number of targets at the onset of multiple object tracking. *Journal of Experimental Psychology: Human Perception and Performance, 39*, 1421.
Ma, Z., McCloskey, M., & Flombaum, J. I. (2015). A deficit perceiving slow motion after brain damage and a parallel deficit induced by crowding. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 1365.
Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature, 434*, 387.
Ogawa, H., Watanabe, K., & Yagi, A. (2009). Contextual cueing in multiple object tracking. *Visual Cognition, 17*, 1244–1258.
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*, 631–671.
Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience, 11*, 1129.
Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*, 179–197.
Ristic, B., Arulampalam, S., & Gordon, N. (2004). Beyond the kalman filter. *IEEE Aerospace and Electronic Systems Magazine, 19*, 37–38.
Rovamo, J., & Virsu, V. (1979). An estimation and application of the human cortical magnification factor. *Experimental Brain Research, 37*, 495–510.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking multiple items through occlusion: Clues to visual objecthood. *Cognitive Psychology, 38*, 259–290.
Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition, 80*, 159–177.
VanMarle, K., & Scholl, B. J. (2003). Attentive tracking of objects versus substances. *Psychological Science, 14*, 498–504.
VanRullen, R., & Koch, C. (2003). Is perception discrete or continuous? *Trends in Cognitive Sciences, 7*, 207–213.
VanRullen, R., & Macdonald, J. S. (2012). Perceptual echoes at 10 hz in the human brain. *Current Biology, 22*, 995–999.
VanRullen, R., Reddy, L., & Koch, C. (2005). Attention-driven discrete sampling of motion perception. *Proceedings of the National Academy of Sciences, 102*, 5291–5296.
Vul, E., Alvarez, G., Tenenbaum, J. B., & Black, M. J. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems* (pp. 1955–1963). .
Whitney, D., & Levi, D. M. (2011). Visual crowding: A fundamental limit on conscious perception and object recognition. *Trends in Cognitive Sciences, 15*, 160–168.
Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology, 24*, 295–340.
Zelinsky, G. J., & Neider, M. B. (2008). An eye movement analysis of multiple object tracking in a realistic environment. *Visual Cognition, 16*, 553–566.
Zhong, S.-h., Ma, Z., Wilson, C., Liu, Y., & Flombaum, J. I. (2014). Why do people appear not to extrapolate trajectories during multiple object tracking? A computational investigation. *Journal of Vision, 14*, 1–30.