Tracking Multiple Objects without Indexes

Shubhamkar Ayare (sbajranga21@iitk.ac.in)
Department of Cognitive Science, IIT Kanpur
Kalyanpur, Kanpur, Uttar Pradesh 208016, India

Nisheeth Srivastava (nsrivast@iitk.ac.in)
Department of Cognitive Science, IIT Kanpur
Kalyanpur, Kanpur, Uttar Pradesh 208016, India

Abstract

Computational models of multiple object tracking (MOT) presuppose the existence of non-conceptual indexes in visual perception, and as a result predict that ID (identification) performance on MOT tasks should be no worse than tracking performance for the same stimuli. However, empirical evidence suggests that ID performance is worse than tracking performance in MOT. We propose a computational model of MOT that is able to account for several empirical results related to tracking performance without the use of indexes and thus avoids yoking tracking performance to ID performance. We also test our model empirically, contrasting it with an existing index-based model, and show that an assumption that avoids indexes and instead incorporates an explicit (rather than an implicit) mechanism for identity maintenance accounts well for the variation in ID performance with increasing number of targets in MOT with visually identical objects.

Keywords: Visual Indexing Theory; Multiple Object Tracking; Computational Modeling

Introduction

Computational models of Multiple Object Tracking (MOT) of visually identical objects are often premised on the existence of non-conceptual pre-attentive indexes in the visual system (Alvarez & Franconeri, 2007; Oksama & Hyönä, 2008; Srivastava & Vul, 2016). This assumption is supported in such models by reference to Pylyshyn's FINgers of INSTantiations (Pylyshyn & Storm, 1988; Pylyshyn, 2009) which, theoretically, are meant to act as indexes, enabling models to establish a correspondence between the representation of an object that is being noticed at the current point of time to earlier ones.

While Pylyshyn's work indicated the number of FINSTs available in human visual perception to be between 3-4, Alvarez and Franconeri (2007) has subsequently shown that, MOT task participants are able to track not just 4 but even 8 objects at sufficiently low movement speeds, suggesting either that the number of FINSTs is larger than Pylyshyn found, or that MOT is possible without using FINSTs.

Further, because FINSTs provide an incorruptible mechanism to establish a correspondence between two visual elements, one being accessed currently while the other that was present at an earlier point of time, they do not offer a natural explanation for errors in identifying objects in such tasks. In particular, since incorruptible indexing makes the identification problem trivial, respondents should be able to identify correctly any objects that they are able to successfully track in MOT tasks. However, empirical evidence suggests a disparity between the two. Participant's ability to identify which particular target had which particular ID (ID performance) is known to be worse than their ability to correctly identify the targets (tracking performance) (Pylyshyn, 2004). And while it is always possible to explain the errors by positing that the indexes are fallible or corruptible, it has been argued (Scholl, 2009) that doing so deprives indexes of the critical capacity they were supposed to provide.

In contrast to the identity maintenance involved in the tracking of visually distinct objects, our discussion concerns only the nonconceptual index-based identity maintenance involved in the tracking of visually identical objects. That there is a difference in the tracking of visually distinct objects vs visually identical objects is also noted by Horowitz et al. (2007); Li, Oksama, and Hyönä (2019).

This paper then provides a computational account of what Scholl (2009) has described as "Tracking in the Present". We propose a computational model of multiple object tracking - MOTUAL - that is able to account for several empirical results related to tracking performance without using indexes, and thus avoids yoking ID performance with tracking performance.

The model's basic mechanism is straightforward. It assumes access to a fixed grid of locations on a retinotopic map indicating if an object is present in each grid, and if so how many, and maintains and updates a set of certain locations - the attended locations - indicating which particular grids supposedly contain the targets. The process of updating is assumed to consume resources from a limited pool. The greater the number of such locations, the less frequently are individual location maps updated, resulting in worse tracking performance with increasing number of targets, as seen in Alvarez and Franconeri (2007) and Vul, Alvarez, Tenenbaum, and Black (2009). We describe this model in more detail below.

MOTUAL: Multiple Object Tracking as Updates of Attended Locations

MOTUAL comprises basically of two retinotopic grid maps, one representing parallelized low-level object detection mapped to a location map, and one an isomorphic map of attended locations, and an account of their interactions. That the position representations of targets are distinct from actual target locations has also been suggested by Howard, Masom,

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

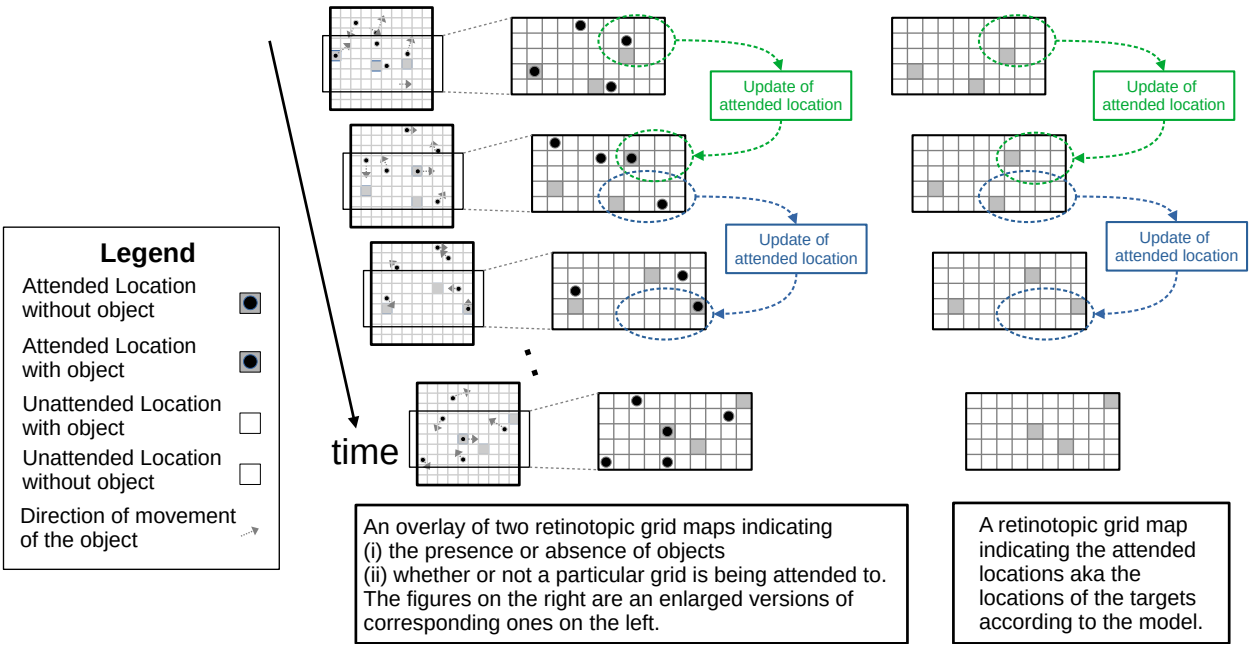

Figure 1: An illustration of our model with the shaded grids in the retinotopic grid map indicating the attended aka target locations, along with the presence / absence of objects at each location. While indexes provide a location-independent means to individuate two objects, there exist no such location-independent way to distinguish between two attended locations.

and Holcombe (2011). While we make no commitments to the actual structure of these maps, for purposes of computational modeling, we assume that the two maps comprise of equal-sized grids indexed by cartesian coordinates (Figure 1). At time t, let the matrix O_t denote the first of these, with each entry $(O_t)_{yx}$ in the matrix indicating the number of objects, and thus their presence (count > 0) or absence (count $= 0$), in that particular grid cell yx. Similarly, let the matrix A_t denote the second of these at the same time t, with each entry $(A_t)_{yx}$ indicating whether or not a particular grid cell is being attended to on the retinotopic grid map. $(A_t)_{yx}$ takes larger values when the underlying grid cell contain multiple objects at any point in time, and end up focusing attention more to that grid cell in the immediately succeeding time steps through an update process we describe below.

At the start of the trial, at time t_0, attended locations are identical to the locations of the targets. That is, for each location where $(A_{t_0})_{yx}$ is non-zero, $(O_{t_0})_{yx}$ is also non-zero. But as the trial progresses, the attended locations may or may not correspond to the exact locations of the objects. At an arbitrary point of time t during the trial, some $(A_t)_{yx}$ may be non-zero even though $(O_t)_{yx}$ is zero. This is the main difference with respect to the FINST theory, which considers locations as being accessed through the indexed objects thus disallowing attention to locations without objects.

Given these two maps, tracking comprises of maintenance of attended locations so that they keep corresponding to objects as far as possible. Such maintenance involves an update of attended locations which no longer correspond to an object to other locations which are occupied by an object.

The update from t to $t' > t$ maintain the following relation between the entries in A_t and $A_{t'}$:

$$(A_t)_{yx} = 1 \Rightarrow \begin{cases} (A_{t'})_{yx} = 1 & \text{if } (O_{t'})_{yx} \neq 0 \\ (A_{t'})_{\text{nearest-object-location}(x,y)} = 1 & \text{if } (O_{t'})_{yx} = 0 \end{cases}$$

The rest of the entries in $A_{t'}$ are 0. The above update assumes binary values of the entries in A_t for simplicity; but the model allows for larger values of both O and A grid cells arising in cases when objects overlap within grid cells.

This update is assumed to be an expensive constrained resource process characterized by a frequency of location updates f_{loc}. This denotes the total number of updates across all attended locations happening in every unit time, so that greater the number of attended locations, less frequent are the updates to each location and thus worse is the maintenance. This assumption is a specialization of more generic resource constraints found in earlier models (Alvarez & Franconeri, 2007; Srivastava & Vul, 2016).

Note that while Pylyshyn and Storm (1988) has shown that a serial tracking algorithm based on a spotlight of attention moving between the objects at finite speeds could not account for the tracking performance on the task of tracking multiple identical objects, they did not rule out the case of quantal updates, as discussed by Egeth and Yantis (1997). Serial switching theory also naturally accounts for variation in temporal resolution of tracking with the number of targets (Holcombe & Chen, 2013). Following VanRullen, Reddy, and Koch (2005), we expected the frequency of such updates to lie between 10 and 20 Hz.

Most importantly, unlike FINST (Pylyshyn, 1989, 2009) and MOT models based on FINSTs that enable individuating objects without considering their visual features or even their locations, all attended locations in our model are indistinguishable from each other except by virtue of their locations themselves.

Further, following Alvarez and Franconeri (2007), we assume that the total number of attended locations can vary within an individual based on task requirements. We expect there to be an upper limit to the number of attended locations; it is certainly above 4 given the evidence illustrating that subjects can track as many as 8 objects moving at low velocities. We speculate that the limit should be related to one's working memory capacity, but leave the exact number open as a question for future work to address.

MOTUAL Attention Updates

Here, we describe the process of updating the attention map based on object locations in more detail. For purposes of notation, let the superscripts denote one particular location on the retinotopic map, and let the subscripts correspond to the particular instance of time for which these are being considered. According to this notation, a_t^1, a_t^2, ..., a_t^n will be the non-zero grids in A_t at time point t, with n being the number of targets.

1. We note that an attended location a_t^i corresponds to the exact location of some object only immediately after an update at time t caused it to be non-zero.
2. Suppose the next update corresponding to $a_{t'}^i = a_t^i$ happens at time t', so that $a_{t'+1}^i$ once again corresponds to some object.
3. At time t', the attended location $a_{t'}^i = a_t^i$ may no longer correspond to an object, since during the time from t to t', the object would have moved from the location a_t^i to a new location l. While updating $a_{t'}^i = a_t^i$ to $a_{t'+1}^i$ at time t', the model finds a *nearest* object in the vicinity of $a_{t'}^i$. To do so, it looks for a location that is occupied by *some* object with increasing distance from $a_{t'}^i$; thus it avoids considering all the objects on display to find the nearest object. This local search is also characterized by another parameter called the nearest object bound *nob* since it is unreasonable to assume that recovering the object could work if they have moved too far away from the attended locations. The search is aborted and the location is no longer stored if no object is found within a distance of *nob* from $a_{t'}^i$.
4. Suppose this new location where *some* object is present is l'. Then, the update is performed so that $a_{t'+1}^i = l'$ holds. In general, l' may not be the same as l. For small velocities and if the time elapsed since the last update was small, l' will more likely be the same as l, and in these cases the model will not lose track of the target; but otherwise, the locations l and l' will be different and correspond to different objects. According to the model then, this is the way in which spatial interference occurs resulting in tracking errors.

Reproducing earlier MOT results

As Srivastava and Vul (2016) point out, explaining the degradation of accuracy with an increase in the number of targets is the stiffest challenge for computational MOT models. In this section, we show how MOTUAL successfully reproduces this trend across *in silico* reproductions of four different experiments, PS1988 (Pylyshyn & Storm, 1988), AF2007 (Alvarez & Franconeri, 2007), FR2008 (Franconeri, Lin, Enns, Pylyshyn, & Fisher, 2008) and SV2016 (Srivastava & Vul, 2016).

Methods

For all simulations reported here, as well as for the design of stimuli used in the experiment described in the next section, MOT objects follow the exact same dynamics as described in Vul et al. (2009)[1].

Also, following Vul et al. (2009), tracking accuracy is defined as the percentage of the tracked objects that are targets at the end of the trial. Because we assume that the number of targets is the same as the number of tracked objects, this is also the same as the percentage of targets that are tracked.

For purposes of tracking, f_{loc} and *nob* constitute the free parameters of the model. In contrast, the free parameters of the environment and simulations include the grid size, σ, MOT simulation update rate, and the total number of simulation updates carried out (or equivalently the number of time steps).

By equating the average duration taken by an object to travel from one end of the MOT window to another, one obtains the following relationship between the parameters that are usually reported in the MOT literature, and the environment-simulation parameters that our formulation requires[2]:

$$\frac{\theta}{d} = \frac{D}{1.8\eta\sigma\frac{D}{s}}$$

Here,

- d degrees per second is the average speed of the object
- θ degrees is the angle subtended by the MOT window
- η is the frequency of simulation updates
- s is the side of the grid (in pixels)
- L is the actual distance of the screen from the participant
- D is the actual width of the grid in units identical to L

Solving for σ, one obtains $\sigma = 0.555 \times \frac{s \cdot d}{\eta \cdot \theta}$.

In our simulations, the size of the retinotopic grid map of the model equals the size of the MOT window on the display (in pixels). This is reasonable because the amount of information available to a model is indeed limited by the MOT display.

[1] Following the same notation as (Vul et al., 2009), we set the spring constant parameter to $k = 0.0005$ and the inertia parameter $\lambda = 0.9$.

[2] For $k = 0.0005$ and $\lambda = 0.9$ one simulation update covers an average of 1.8σ pixels.

Table 1: MOTUAL simulation parameters corresponding to data from previous empirical results.

Parameters \ Paper	PS1988	AF2007	SV2016	FR2008-small	FR2008-large
Visual Angle of MOT Window	21.5°	30° x 24°	16°	20.5° x 9.1°	82° x 36.4°
- display resolution	not given	not given	720x720	175x078	700x310
- grid-size for MOTUAL	360x360	720x720	720x720	180x180	720x720
Retinal Speed	1.25 - 9.4°/s	0.1 - 16°/s	not given in °/sec	5 - 25°/sec	20-100°/sec
- sigma for simulations	1.5	1.1 - 6	0.9 - 2.3	0.73 - 3.65	2.9 - 14.6
No. of OU updates (trial duration)	300 (10 sec)	150 (5 sec)	150 (5 sec)	180 (6 sec)	180 (6 sec)
Minimum Object Distance	15 (0.75°)	80 (4°)	0 (0°)	24 (2.8°)	96 (11.3°)
- actual value used for simulations	60	80	0	24	96
MOTUAL Free Parameters (f_{loc}, nob)	(20Hz, 59)	(20Hz, 60)	(20Hz, 60)	(20Hz, 58)	(20Hz, 145)

Results

Figure 2 summarizes MOTUAL's behavior vis-a-vis data from all the four experiments we evaluated. Table 1 enumerates the experiment specific parameters and free parameter values used to produce these results. MOTUAL, with only two free parameters, successfully reproduces the qualitative trends seen in all the four experiments, with some interesting exceptions, which we detail below. We imposed an upper limit of $f_{loc} = 20Hz$ following VanRullen et al. (2005) which suggests an attention-driven sampling frequency to be between 10 and 20Hz.

In Pylyshyn and Storm (1988), while the empirical data indicated a minimum separation of 0.75 degrees between any two objects on a display subtending a visual angle 21.5 degrees, that is 15 units on a 360x360 units, our model required about 60 units to obtain comparable error rates.

Also, the model continued to perform poorly compared to the human empirical data from Srivastava and Vul (2016), whereas it was able to match the human performance reported in Alvarez and Franconeri (2007). Interestingly, the experiments in Alvarez and Franconeri (2007) never allowed objects to cross each other and always maintained a separation of 4 degrees between them, while the experiments in Srivastava and Vul (2016) allowed objects to overlap freely. We suspect our model's inability to handle object overlaps may explain its inability to jointly explain both datasets.

Similarly, our model is able to reproduce the general trend from Franconeri et al. (2008) that tracking accuracy can be high even for higher retinal speeds if the space available for objects to move is large. However, relying purely on instantaneous retinotopic locations of the objects for the updates results in MOTUAL performance being worse than humans at higher velocities. It might be possible to augment this with a retinotopic velocity map with velocities in each grid cell being computed based only on local information.

Multiple identity tracking with MOTUAL

Given that at time t, the system only has access to the locations $a_t^1, ..., a_t^n$ but not the locations at other points of time, the information it has so far is insufficient to make conclusions about the target IDs.

To keep track of IDs, we therefore propose that at time t, there also exists a separate sequence $p_t^1, ..., p_t^n$ of IDs. One possible strategy for maintaining such a sequence is to keep reciting the sequence. With this, in order to match objects to IDs, one needs to go over the attended objects in some spatial sequence while reciting their IDs in that same sequence.

One particular sort order (but not the only one) could be to sort the locations in the non-decreasing order of x-and-y coordinates. With this, at the start of the trial at time t_0, the sequence of attended locations is monotonic in their x-and-y coordinates, so that for $m < n$, $a_{t_0}^m \equiv (x_{t_0}^m, y_{t_0}^m)$ and $a_{t_0}^n \equiv (x_{t_0}^n, y_{t_0}^n)$ are such that $(x_{t_0}^m < x_{t_0}^n)$ or $(x_{t_0}^m = x_{t_0}^n$ and $y_{t_0}^m \leq y_{t_0}^n)$. Given this order of $a_{t_0}^1, ..., a_{t_0}^n$, the system now has the ID of the object at $a_{t_0}^k$ in $p_{t_0}^k$ for $k \in 1, ..., n$, and it is through this correspondence that the system can infer the IDs of the objects.

At the end of the trial at time t_e, the system again sorts the sequence of attended locations $a_{t_e}^1, ..., a_{t_e}^n$ using the same sorting-rule that it had used at the start of the trial, in our particular example, this is non-decreasing order of x-and-y coordinates. It then assigns the ID $p_{t_e}^k$ to the object that is at or nearest the location $a_{t_e}^k$.

The strong assumption is that the sequence of IDs is *never* updated. In the general case, and as one's intuition would suggest, it should be possible to update the ID sequence aka do a "correspondence update" especially if the objects move sufficiently slowly. This can be characterized by an additional parameter for the model, which we call the frequency of correspondence updates aka f_{corr}. So, as opposed to the frequency of location updates f_{loc}, f_{corr} can be understood as a frequency relative to f_{loc} and will be such that $0 \leq f_{corr} \leq 1$.

Assuming $f_{corr} = 1$, aka correspondence updates occur whenever location updates occur, MOTUAL (like FINST) predicts ID performance being identical to tracking performance across any trial duration (see Figure 3 (right)). Setting this correlation to less extreme values produces a disparity between tracking and ID performance, as anticipated in the literature (Pylyshyn, 2004). For example, assuming $f_{corr} = 0$ reproduces the empirical results seen in Pylyshyn (2004) (see Figure 3 (center)) very precisely (see Figure 3 (left)).

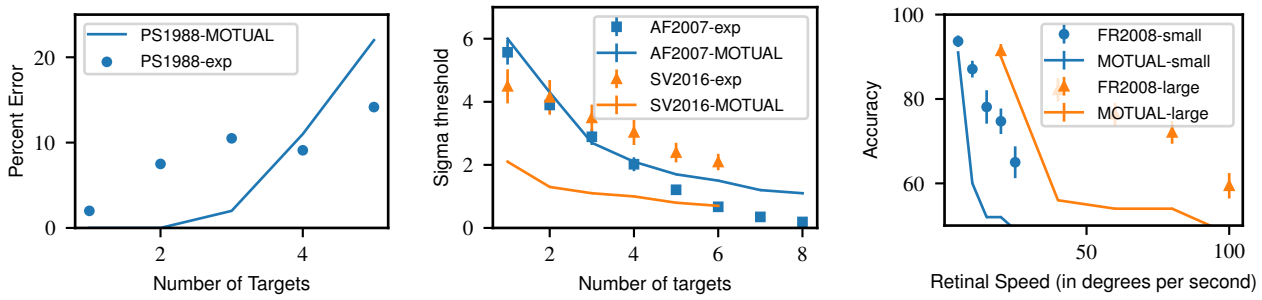

Figure 2: Reproducing previous results in Multiple Object Tracking. Left: Comparison of percentage errors with increasing number of targets corresponding with a fixed average speed of MOTUAL plotted against data from Pylyshyn and Storm (1988). Center: Velocity threshold vs number of targets patterns for MOTUAL plotted against data from AF2007 (Alvarez & Franconeri, 2007) and NS2016 (Srivastava & Vul, 2016). Right: Comparing our model against the human data in experiment 1 of Franconeri et al. (2008). MOTUAL gets the general trend, but needs to be augmented with, say the retinotopic velocity information, in order to perform as well as humans at higher velocities.

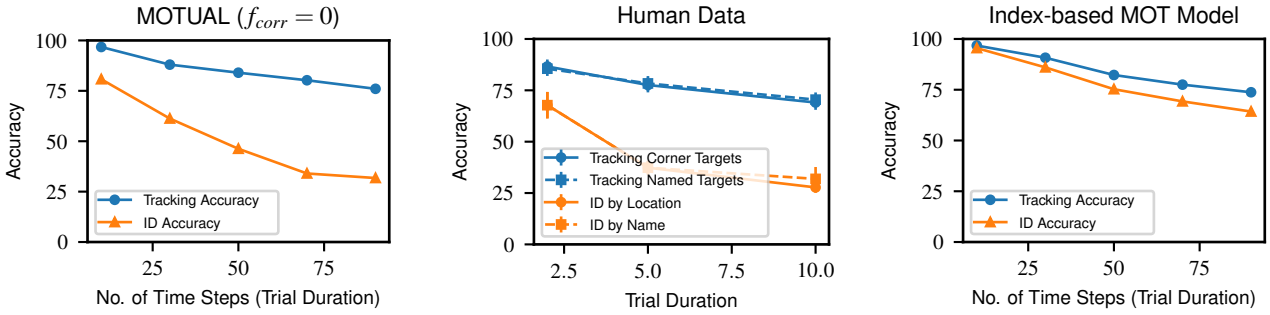

Figure 3: Tracking and ID accuracy in a MOT task with 4 targets and 8 objects (4 targets, $\sigma = 1.25$) with increasing trial duration. Left: Predictions from our modeling assuming $f_{corr} = 0$. Center: Replotted results from Pylyshyn (2004). Right: Predictions from a model based on indexes aka $f_{corr} = 1$. Our model currently only captures the notion of an ID without making a distinction between Corners vs Names.

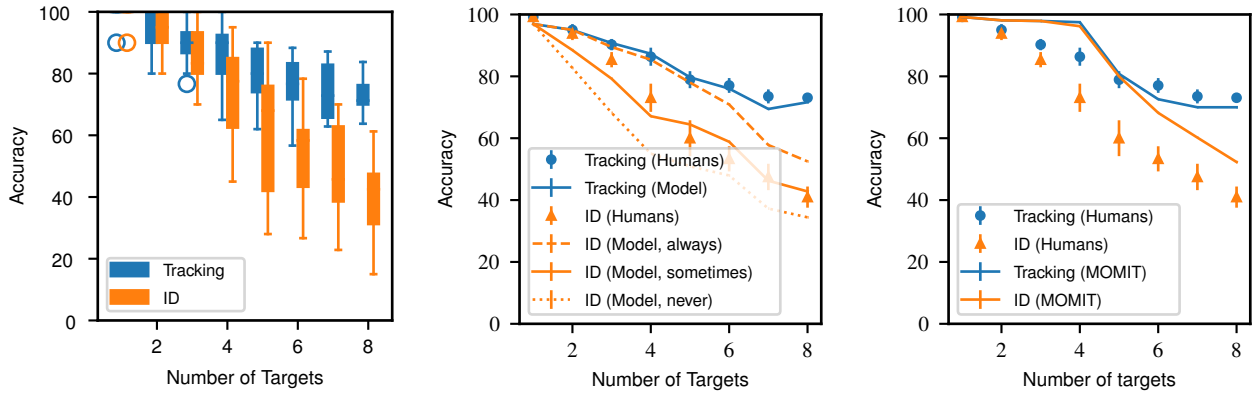

Figure 4: Left: Tracking and ID accuracy of participants with increasing number of targets. Center: Comparison of Model Tracking and ID Performance against Human Data after fitting the model's tracking performance to the human data, with differing assumptions about the ID updates (i) ID updates taking place always whenever location updates take place (ii) ID updates taking place some of the times (iii) no ID updates taking place ever as with the Center figure. Errorbars represent 1 SEM. Right: Comparison of MOMIT's Tracking and ID Performance against human data for the lowest MSE run.

An experimental test

While Pylyshyn (2004) have previously shown an ID-tracking performance disparity with respect to tracking duration, no earlier work has examined how this relationship is affected by changing the number of targets in the experiment. MOTUAL makes a non-trivial prediction that because the participants will not be able to do correspondence updates as rapidly as required by the changing spatial sequence of targets, ID performance would degrade more rapidly than tracking performance with increasing number of targets. As a strong test of its validity, we decided to test both these predictions against data from human participants performing both ID and tracking tasks in an MOT experiment, varying the number of targets across the trials.

Based on MOTUAL's predictions, with task (tracking vs ID) and number of targets as factors of a within-subjects ANOVA, we expected a significant interaction effect, as well as significant main effects of task as well as number of targets.

Participants

13 participants (5F) participated in the experiment. All had normal or corrected to normal vision, and none were color-blind. An IRB approved the protocol for the experiment.

Procedure

10 practice trials, followed by 80 main trials were employed. The number of targets varied from 1 to 8 across the trials. 10 trials for each case of number of targets from 1 to 8 accounted for the 80 trials. Each trial had 14 objects. The tracking duration in each trial was 5 seconds but the trials were self-paced and randomized. In each trial, the participant had to select all the targets and indicate the ID number for each of the target. This procedure was similar to that employed in Experiment 4 of Pylyshyn (2004).

Materials

For the purposes of the experiment, the visually identical objects comprised of small circles with a diameter of 10 pixels. The participants sat at a distance of about 60 cm from the display. Thus, each object subtended an angle of 0.23° at the retina, and the objects were allowed to move in a 720x720 pixels square area whose diagonal subtended an angle of 23.5° at the retina.

Results

Repeated measures two-way ANOVA was conducted with tracking-vs-ID accuracy as one factor, and the number of targets as the second factor. The results were as per the expectations discussed above (Figure 4, left) - there was a significant interaction effect $[F(3.46, 41.47) = 37.638, p < 0.001]$ as well as significant main effects for number of targets $[F(7, 84) = 70.845, p < 0.001]$ and task $[F(1, 12) = 113.476, p < 0.001]$.

Model-based Analysis

We minimized the mean-square error (MSE) between the tracking performance of humans and MOTUAL for f_{loc} ranging from 0 to 60Hz, and for *nob* varying from 0 to 120.

For lowest MSE (Figure 4, center), $f_{loc} = 19Hz$, $nob = 27$, $MSE = 2.9 \times 10^{-4}$, $r^2 = 0.989$. However, we note that f_{loc} and *nob* trade-off against each other around 10 to 20Hz range, and we also obtained comparable fits with $f_{loc} = 10Hz$: $nob = 52$, $MSE = 7.5 \times 10^{-4}$, $r^2 = 0.983$.

Comparing the model's ID performance with human ID performance (Figure 4, center-dotted), one notes that assuming $f_{corr} = 0$ results in the model's ID performance being worse than the human ID performance ($r^2 = 0.947$). Calculating the MSE scores between the model's ID performance and human ID performance for different f_{corr} yielded a lowest MSE for $f_{corr} = 0.7$ ($r^2 = 0.964$) (Figure 4, center-solid).

However, even with $f_{corr} = 0.7$, one notes that human ID performance exceeds model's ID performance for number of targets up to four, and it is worse than model's ID performance for number of targets more than four. We discuss this very interesting discrepancy further below.

MOMIT (Oksama & Hyönä, 2008) is an index-based model designed for tracking visually distinct objects. We adapted it for tracking visually identical objects by letting go of its corrective attention shift. Figure 4 (right) shows this model's fit to our data. Despite a still-reasonable ($r^2 = 0.877$) fit for tracking accuracy, we note that MOMIT is intrinsically predisposed to align ID accuracy with tracking accuracy as with the $f_{corr} = 1$ case (Figure 4, center-dashed).

Conclusion

The empirical success of MOTUAL demonstrates that it is possible to explain human MOT behavior without postulating incorruptible pre-attentive indexes. Indexes are needed if we assume that solving the correspondence problem is a prerequisite to performing the MOT task (Pylyshyn, 2004; Luo et al., 2021). We show that MOT is possible without having to computationally maintain one tracker per object. This opens up the possibility of producing fast multiple object tracking algorithms scalable to large sets of objects, particularly in combination with neuromorphic vision sensors (van De Burgt, Melianas, Keene, Malliaras, & Salleo, 2018; Pantho, Bhowmik, & Bobda, 2018), as well as the possibility of entertaining embodied cognition proposals requiring substantial metacognitive insight into attention allocation at the level of saccadic planning (Chandrasekharan et al., 2015).

In light of the discrepancy observed in our model-based analysis, where we found that fewer targets than 4 were ID'd better by humans than MOTUAL, it would be premature to conclude that FINSTs play no role in MOT. It may be that fewer targets than 4 (Pylyshyn's estimate of FINST count) could be tracked using FINSTs, and larger targets tracked by the MOTUAL mechanism. It is also possible that MOTUAL reduces to FINSTs while tracking fewer targets. Investigating this relationship presents an exciting avenue for future work.

References

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

Chandrasekharan, S., Date, G., Pande, P., Rahaman, J., Shaikh, R., Srivastava, A., ... Agrawal, H. (2015). Eye to i: Males recognize own eye movements, females inhibit recognition. In *Cogsci.*

Egeth, H. E., & Yantis, S. (1997). Visual attention: Control, representation, and time course. *Annual review of psychology*, *48*(1), 269–297.

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

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

Horowitz, T. S., Klieger, S. B., Fencsik, D. E., Yang, K. K., Alvarez, G. A., & Wolfe, J. M. (2007, February). Tracking unique objects. *Perception & Psychophysics*, *69*(2), 172–184. Retrieved from https://doi.org/10.3758/bf03193740

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

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

Luo, W., Xing, J., Milan, M., Zhang, X., Liu, W., & Kim, T.-K. (2021). Multiple object tracking: A literature review. *Artificial Intelligence*, *293*, 103448. Retrieved from https://doi.org/10.1016/j.artint.2020.103448

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

Pantho, M. J. H., Bhowmik, P., & Bobda, C. (2018). Pixel-parallel architecture for neuromorphic smart image sensor with visual attention. In *2018 ieee computer society annual symposium on vlsi (isvlsi)* (pp. 245–250).

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

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

Pylyshyn. (2009). Perception, representation, and the world: The finst that binds. *Computation, cognition, and Pylyshyn*, 3–48.

Pylyshyn, & Storm. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. Retrieved from https://doi.org/10.1163/156856888x00122

Scholl, B. J. (2009). What have we learned about attention from multiple object tracking (and vice versa). *Computation, cognition, and Pylyshyn*, 49–78.

Srivastava, N., & Vul, E. (2016, January). Attention modulates spatial precision in multiple-object tracking. *Topics in Cognitive Science*, *8*(1), 335–348. Retrieved from https://doi.org/10.1111/tops.12189

van De Burgt, Y., Melianas, A., Keene, S. T., Malliaras, G., & Salleo, A. (2018). Organic electronics for neuromorphic computing. *Nature Electronics*, *1*(7), 386–397.

VanRullen, R., Reddy, L., & Koch, C. (2005, March). Attention-driven discrete sampling of motion perception. *Proceedings of the National Academy of Sciences*, *102*(14), 5291–5296. Retrieved from https://doi.org/10.1073/pnas.0409172102

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