Dynamics of target selection in multiple object tracking (MOT)

Zenon W. Pylyshyn & Vidal Annan, Jr.
Rutgers Center for Cognitive Science,
Rutgers University, New Brunswick, NJ

Suggested running head: Target selection in multiple object tracking

Key Words: Tracking, attention, selection, attention capture, multiple object tracking

Address all correspondence to:
Zenon Pylyshyn
Rutgers University,
Center for Cognitive Science
Psychology Building Annex,
152 Frelinghuysen Road
Piscataway, NJ 08854-8020

Email: zenon@ruccs.rutgers.edu
Dynamics of target selection in Multiple Object Tracking (MOT)

Zenon W. Pylyshyn & Vidal Annan, Jr.
Rutgers Center for Cognitive Science,
Rutgers University, New Brunswick

ABSTRACT

In four experiments we address the question whether several visual objects can be selected voluntarily (endogenously) and then tracked in a Multiple Object Tracking paradigm and, if so, whether the selection must be carried out serially. Experiment 1 showed that items can indeed be selected based on their labels. Experiment 2 showed that to select the complement set to a set that is automatically (exogenously) selected – e.g., to select all objects not flashed – observers require additional time and that given 1080 ms they were able to select and track them as well as those selected automatically. Experiment 3 showed that the additional time needed in the previous experiment cannot be attributed solely to time required to disengage attention from the initially automatic selections. Experiment 4 showed that the added time provides a monotonically greater benefit when there are more targets, suggesting a serial process. These results are discussed in relation to the Visual Index (FINST) theory which assumes that visual indexes are captured by a data-driven process. It is suggested that voluntarily allocated attention can be used to facilitate the automatic attention capture by objects of interest.

Introduction

Perceptual systems must be selective. What they select and under what conditions selection takes place has been the subject of extensive research, often conducted under the heading of focal attention (for a review see, Pylyshyn, 2003, Chapter 4). In the case of vision there is evidence that selection can be based on a number of different properties. For example, it has been shown that selection can be based on different spatial frequency bands (Julesz & Papathomas, 1984; Shulman & Wilson, 1987), and on such features as color (Friedman-Hill & Wolfe, 1995; Green & Anderson, 1956), local shape (Egeth, Virzi, & Garbart, 1984), motion (McLeod, Driver, Dienes, & Crisp, 1991) (particularly looming, Franconeri & Simons, 2003), stereo-defined depth (Nakayama & Silverman, 1986) or other dynamic events (Franconeri, Simons, & Junge, 2004). Such selection is typically course-grained compared to the main types

---

1 We wish to acknowledge the support of NIH Grant number 1R01-MH60924 to the first author and an NRSA post-doctoral award to the second author.
of selection studied most recently which include: (1) selection based on spatial properties such as location (often described in terms of what is referred to as a “spotlight” of attention, LaBerge, 1998; Posner, 1980) or different spatial extents (often described as involving a “zoom lens” of attention, Eriksen & St. James, 1986), and (2) selection based on individual token objects. Recent evidence has shown that in focusing limited perceptual resources, observers do indeed (perhaps even must) select individual objects (often moving objects) in their visual field (see the review in, Scholl, 2001).

From an evolutionary perspective being able to select real objects that appear in the field of view and to track them as they move would obviously be a capacity useful for survival. Tracking both predators and prey under conditions where their optical properties keep changing but their identity persists as they move is clearly critical, as is the ability to dodge unidentified moving objects of all kinds. This ability has been documented by experiments that show that observers can pick out and keep track of objects even when they have not encoded their properties. In the present study we will be concerned with the process of making selections based on individual objects, where objecthood itself is the defining category.

The question of what is selected by visual attention goes hand-in-hand with the question of how and under what conditions selection takes place. For example, it has been shown that selection can be automatically induced by what some have called exogenous cues that are automatic and data-driven, or by can be voluntarily allocated by symbolic or endogenous cues (e.g., Theewues, 1994). There is also abundant evidence that multiple selection can occur, with at least 4 or 5 objects being available simultaneously (Pylyshyn, 2001; Pylyshyn et al., 1994). Although the evidence shows that selected objects are available simultaneously, it is not clear whether they must be selected in parallel or whether serial selection occurs under some conditions. Data showing serial search (say of feature conjunctions, Treisman & Gelade, 1980) is silent on the question of whether the candidate items among which search is carried out are themselves selected in series or in parallel. In fact the question of what the search is carried out over (whether search scans through objects or through space) is rarely addressed. Our position has been that some selection of candidate objects must occur prior to testing whether the candidates meet the search criteria (Pylyshyn, 2001). But the question of whether candidate-selection occurs in parallel or in series and whether it is subject to any special conditions has not been tested outside a few tasks such as searching through subsets (Burkell & Pylyshyn, 1997) or subitizing (Trick & Pylyshyn, 1994).
Another closely related question concerns the *maintenance* of selections once they have been made. It would be of limited value if objects were selected and then lost as they changed their visual properties or moved from their initial locations, so the study of the continued maintenance or tracking of selected visual objects is an integral part of the study of selection. One of the experimental paradigms of choice for studying both initial and continuing object-based selection is the Multiple Object Tracking (MOT) task developed by (Pylyshyn & Storm, 1988).

The multiple-object tracking (MOT) task has been used widely in the study of attention and particularly in the study of sustained multiple-locus of attention. In MOT (see Figure 1) a set of simple identical objects (typically 8 circles) is presented on a computer screen. A subset of them (the “targets”) is made visually distinct, typically by flashing them on and off for a brief period of time. Then all objects move about in an unpredictable manner and the task is to keep track of the now-identical objects and to identify the targets at the end of a short trial (usually about 5-10 seconds). Observers can do this under a variety of conditions at better than 90% accuracy. The theory we have developed to account for these capacities is called Visual Index (or FINST) theory (Pylyshyn, 2001). While the present study is directed at the empirical questions surrounding the selection stage of tracking, we will occasionally refer to the FINST theory in discussing the hypotheses and findings.

Figure 1. In Multiple Object Tracking a set of simple objects is briefly identified as the target set (e.g., by flashing the ones shown here with shadows around them) and then they move around unpredictably among identical nontargets for 5-10 seconds. The observer must identify the targets using a computer mouse.

MOT is particularly well suited as a tool for studying object selection because only the individual identity of objects, qua selected individuals, appears to be involved in MOT performance. Thus when we select certain elements in MOT the selection is essentially of particular individual objects as opposed to a set of locations or featural properties. After the brief
initial identifying phase, objects are not only identical in appearance (or, in some cases change their properties either synchronously or asynchronously, Pylyshyn & Dennis, in preparation) but they have unpredictable constantly-changing locations. Thus MOT allows us to ask questions about the process of selecting token objects as persisting individuals. In FINST theory (as described, for example, in the original presentation in, Pylyshyn, 1989) it was assumed that selection (via the limited indexing mechanism) occurs automatically; indexes (up to the maximum available total of 4 or 5) are *captured* by one of a small number of event types, most notably the onset of new objects. However, in an updated exposition (Pylyshyn, 2001) a way was suggested whereby voluntary (endogenous) assignment of indexes could also occur despite the automatic nature of basic index capture mechanism. Whatever the merits of this particular suggestion (discussed briefly in the Summary and General Discussion section below), the basic prediction is that automatic capture can operate in parallel across the visual field whereas voluntary assignment of indexes requires that targets be visited serially. Put in terms of the theoretically more neutral notion of selection, the theory predicts that up to 4 or 5 objects can be selected at once if they have certain features (e.g., if they have what have been called “popout” features, such as onsets or color singletons, or perhaps any property that allows targets to be linearly-separated from nontargets, Bauer, Jolicur, & Cowan, 1999; Johnson, Hutchison, & Neill, 2001) whereas if such features are not present the objects can be selected only by serially visiting each one. The present studies address this general prediction and also explore the parameters that determine whether items are selected serially or in parallel.

**General Method**

Except as noted, experiments in this report used the following general procedure for presenting the Multiple Object Tracking task.

*Stimuli and Apparatus:* The experimental stimuli were generated using the Visionshell Software libraries (Comtois, 1999), and presented using an Apple iMac G3 computer. Targets and non-targets were identical except in the initial target-designation phase of each trial. Objects (targets and nontargets) consisted of white rings on a black background, subtending a visual angle of 2.7 degrees. The width of the rings was 0.11 degrees.

Object trajectories were generated in real-time independently for each trial, producing smooth and continuous motion. Each trial consisted of 590 frames (8.55 ms each) which resulted in 5-second trials. To produce independent movement, objects were assigned random initial locations, directions, and speeds. Individual objects moved either 0 or ±1 or ±2 pixels in the x direction and y direction every other frame (corresponding to either 0 or ±0.053 degrees or ±0.11
degrees of visual angle per frame pair). When the edge of an object intercepted the edge of the screen, the x or y velocity vector was reversed, so that objects appeared to bounce off of the edge of the viewable screen. To ensure smooth motion, trajectories were characterized by an "inertia" parameter, p: objects retained their x or y velocity components except that the motion algorithm added or subtracted one pixel (0.053 degrees) after each pair of frames to either velocity component with probability p, which in our case was fixed at .10. The object speeds varied between 0 and 9.1 degrees per second with an average of 5.9 degrees per second. The objects’ motion was not restricted, except as constrained by the inertia parameter and the edges of the screen, so they could pass over each other. When this occurred, one of the circular objects, chosen at random, was designated as the near object and would gradually occlude/disocclude the other object (for this reason the area in the center of the circles was drawn as opaque, even though it was the same color as the background so the opacity only became apparent when two circles crossed). This design was chosen because it has been reported (Viswanathan & Mingolla, 2002) that T-junction depth cues tend to minimize tracking errors when objects overlap.

Procedure & Design. Observers were seated in a darkened room about 45 cm from the monitor, creating a viewable screen that subtended an angle of about 34 by 26 degrees. Except in Experiment 4, each trial began with a display of eight static circles (in Experiment 4 different numbers of circles were used). A subset of these circles was cued as targets for a specified cue duration, which varied with the experimental condition. After termination of the cue, all the circles moved as described above for 5 seconds. At the end of the trial, the circles stopped moving and observers used a mouse to select the previously designated target objects. After the correct number of circles had been selected in this way, a new trial began.

Experiment 1

This experiment examined whether multiple targets could be selected for tracking based on their symbolic properties, and in particular under conditions where the selection requires voluntary focal attention. We also examined whether such selection requires more time than selections based on exogenous (automatic) cues, such as sudden onsets. To do this we compared the performance in the baseline MOT condition, where targets were indicated by flashing them, with a condition in which targets were specified according to the label (in this case a digit) displayed on them (the “number” condition). Although it generally has been assumed that items can be selected voluntarily for purposes of tracking, this has not been explicitly tested. Moreover, if observers can select based on symbolic (endogenous) cues but in order to do so
must scan attention among the objects, then we would expect poorer performance in the number condition than in the standard (baseline) tracking condition.

**Method**

*Subjects:* Fifteen Rutgers University students participated in one 1-hour session to fulfill a psychology course requirement. All subjects had normal or corrected to normal vision.

*Stimuli and Apparatus:* The experimental stimuli and procedure were as described above. The selection cues in this experiment were provided by numerals (ranging from 1 to 8) printed inside the circles. Observers were instructed to select and then to track the objects labeled 1-4 (or in half the cases, 5-8). After the initial display of numerals on the 8 static circles, which lasted for 1.08 seconds the digits disappeared and all the circles moved as described above for 5 seconds. At the end of the trial, the circles stopped moving and observers used a mouse to select the previously designated target objects. After the fourth target was selected in this way, a new trial began.

In the baseline condition (the one depicted in Figure 1) there were no numbers in the circles. Instead, the designated targets were flashed on and off three times for a total of 1.08 seconds before all objects began to move. These two conditions were blocked. There were 128 trials in each of two blocks, which were presented in alternating orders for half the observers.

**Results**

A t-test showed a significantly poorer tracking performance in the condition in which targets were identified by number, compared with the baseline condition in which targets were designated by flashing (t = 4.26, df = 14, p < .001). Notwithstanding this difference, however, performance in both conditions was high. Tracking in the number condition was 87.8% and in the baseline flash condition it was 93.8% (the corresponding number of targets tracked in these 2 conditions was 3.51 and 3.75 out of 4). Expressed in terms of the “effective number of targets tracked”, a measure $m$ which corrects for guessing, the performance in the number condition was 3.44 while in the baseline condition it was 3.74. (The $m$ score is required in order to take account of different probabilities of correctly guessing targets when comparing performance on conditions with different numbers of targets, as in Experiment 4. Its derivation is given in the Appendix)
Discussion

This experiment demonstrates what many have already assumed – items can be voluntarily selected for the purpose of tracking them in an MOT paradigm. Finding items scattered throughout a display by their digit identifiers is clearly a process that requires serial attention – finding items numbered 1-4 (or 5-8) is not a “preattentive” operation in the sense of (Treisman & Gelade, 1980). Perhaps not too surprising, the ability to select and then to track items is somewhat poorer in this case, since the more difficult selection may lead to errors of selection under time constraints. Even though the four cued digits can be easily located in the available 1.08 seconds, this process involves scanning the display and therefore some targets may be missed or misread. The question thus arises; if more time is available for making the selection will this eliminate the difference between a selection that requires voluntary attention and one that does not.

This question is pursued in experiment 2. Rather than use the digit task, we adopted a different pair of selection cues that had the property that detecting one is “preattentive” (i.e. it is detected automatically and in parallel) while the other is not. The preattentive selection cue used is the one that we have called the baseline condition: Items to be tracked are designated by flashing them. Since both onsets and offset transients have been shown to capture attention automatically (Atchley, Kramer, & Hillstrom, 2000), flashing seems to be a most likely candidate for being an automatic preattentive selection cue that is processed in parallel. But if flashing causes the automatic parallel-selection of targets, then a particularly well-suited candidate for being a voluntary selection cue is being one of the objects that was not flashed. Thus if observers were asked to select and track all and only the items that did not flash, they would have to first voluntarily find them by ignoring the ones that had been automatically selected. This idea formed the basis for the second experiment in which we compared the task of tracking flashed items (the “Track Flash” condition) with the task of tracking the items that had not been flashed (the “Track Nonflashed” condition). To test whether the voluntary cue requires additional time to make the selection, the comparison between these two conditions was carried out for both short and long cueing durations. This test was meant to assess whether providing enough time can eliminate the difference in the relative effectiveness of automatic and voluntary selection.
Experiment 2

If, as assumed in discussing the results of experiment 1 above, endogenous (voluntary) selection tends to be serial and therefore takes more time than exogenous (automatic, parallel, preattentive) selection, then displaying the selection cue for a longer period of time should benefit the Track Nonflashed condition more than the standard Track Flashed condition. In particular, if sufficient time is available for making a selection, the difference between automatic selection and voluntary selection should disappear, or at least substantially decrease. Assuming that the tracking performance reflects how many targets were correctly selected in the available time, this pattern should manifest itself in the performance measures for tracking when different types and durations of selection cues are used.

Method

Subjects: Fourteen Rutgers University students participated in a single session lasting approximately 75 minutes to fulfill a psychology course requirement. All subjects had normal or corrected to normal vision.

Stimuli and Apparatus: The stimuli and procedure were similar to those of Experiment 1 except it did not use numbers as selection cues. In each trial four of the objects, randomly chosen, were flashed on and off either once for a duration of 360 ms or three times for a total of 1080 ms. There were two sets of instructions: The observer’s task was either to track the flashed objects (the baseline Track Flashed condition) or to ignore the flashed objects and track the four objects that had not flashed (the Track-Nonflashed condition). These two conditions were presented in blocks of 128 trials, with order of blocks counterbalanced across observers.

Results

A repeated-measures analysis of variance was performed on 2 selection conditions (Track Flashed and Track Nonflashed) and 2 cue durations (360 vs 1080 ms). All effects were significant. Tracking in the condition in which targets were the flashed items (93.1%) was significantly better than in the condition where targets were the non-flashed items (89.9%); F=(1,13) = 17.8, MS = 136.7, p < .001. Tracking when the cue was present for 1080 ms (93.3%) was significantly better than when it was present for 360 ms (89.7%); F(1,13) = 9.1 MS = 187.8, p < .01. In addition, as predicted, the interaction between these two was also significant, F=(1,13) = 5.67, MS = 89.6, p < .03. When the difference between short (one-flash) and long (3 flash) cues was compared using a post-hoc t-test, the difference was not significant for the Track-Flashed condition (t=1.0, df=13, p>.34) whereas it was significant for the Track Non-Flashed
condition \((t=3.1, \, df=13, \, p<.01)\). Examining only the Tracked Flashed condition, there was no difference between the 360 ms cue duration and the 1080 ms cue duration, so additional time made no difference for the automatic selection case. These results are shown in Figure 2.

**Tracking under different selection conditions**

![Graph showing tracking under different selection conditions](image)

**Figure 2.** Comparison of exogenous, automatic (track flashed) and endogenous, voluntary (track nonflashed) target selection at different cue durations

**Discussion**

Experiment 2 confirmed the hypothesis that items can be selected voluntarily as long as the observer has sufficient time. The finding that with the 1080 ms cue duration the difference between tracking nonflashed items and tracking flashed items disappears suggests that this provided sufficient time for voluntary selection to occur. Our interpretation is that the additional time is required in order to visit the selected items serially.

But there are alternative explanations for these results. For example, it may be that after the initial automatic assignment of attention, the subsequent disengagement and re-engagement of attention (in order to attend to the nonflashed circles) requires additional time (some reports place the time to disengage and re-engage at 300–600 ms, Duncan, Ward, & Shapiro, 1994;
Mueller, Teder-Saelejaervi, & Hillyard, 1998). If that is the case in Experiment 2, then we are not justified in concluding that the additional time is due to the involvement of a serial process. To control for this possibility Experiment 3 was designed so that no disengagement is required during voluntary selection.

**Experiment 3**

To test whether the increase time needed to select objects for tracking might be due to the extra time required to disengage the automatically captured attention, Experiment 3 incorporated several types of selection cues, including cues that did not require prior disengagement of attention from exogenously cued objects. We used items selected by the sudden appearance of vertical line segment superimposed on target circles (which presumably captured attention exogenously), as well as objects selected by horizontal and vertical line segments, described below.

**Method**

**Subjects.** Eighteen Rutgers University students participated in a session lasting approximately 90 minutes to fulfill a psychology course requirement. All subjects had normal or corrected to normal vision.

**Stimuli and procedure.** There were four conditions in this study, with one block per condition. In one block, observers were instructed to track circles upon which a vertical line was briefly flashed. In a second block they were instructed to ignore the circles with vertical lines and to track circles without any lines. In the remaining two blocks all circles had either a vertical line or a horizontal line superimposed for a brief period at the beginning of the trial. In one of these blocks observers were told to select and to track the circles with horizontal lines and in the other block they were told to track the circles with vertical lines. Other than these four types of cues, presented for varying amounts of time, the procedure was identical to that of the previous two experiments. Targets were designated as one of the following (1) circles with a vertical line presented among circles with no line (2) circles with no line presented among circles with vertical lines, (3) circles with a vertical line presented among circles with horizontal lines or (4) circles with a horizontal line presented among circles with vertical lines. Comparing conditions (1) and (2) replicates experiment 1 since it compares tracking with automatically selected targets and tracking of the complement set – circles that were not automatically selected by an onset cue. Comparing conditions (3) and (4) tested the selection of objects using a cue that requires voluntary attention (since selecting items with vertical/horizontal lines among items with
horizontal/vertical lines is not automatic – see Treisman & Gelade, 1980), yet does not require disengaging previously captured attention. The cues designating targets appeared for one of three durations; 300 ms, 600 ms or 900 ms. After the cue disappeared, observers were required to track the targets during the 5 second trial and then to indicate the targets by selecting them using a mouse pointer. The four conditions were presented in separate blocks with 64 trials per block. The order of blocks was counterbalanced across subjects.

Figure 3. Displays used in Experiment 3, showing the four types of selection cues (only one type of cue appeared in each block).

Results

Performance on all four conditions is shown in Figure 4. A repeated-measures analysis of variance was performed on all 4 conditions as well as separately on portions of the data. The overall ANOVA revealed that all independent variables and their interactions were significant. The main effect of condition was significant with MS=5791, F(3,51)=165.6, p<.000), the effect of cue duration was significant with MS=1653, F(2,34)=131.9, p<.000, and the interaction of these two was significant with MS=175, F(6,102)=11.3, p<.000. Examining only the Track Line and Track Noline conditions (which are comparable to the conditions of experiment 2) showed that the main effect of condition was significant, with MS=232.8, F(1,17)=28.7, p<.000, the effect of cue duration was significant, with MS=182.6, F(2.34)=11.5, p<.000, and the interaction was also significant with MS=41, F(2,34)=4.1, p<.02, thus replicating the finding of experiment 2. When we compared the Track Horizontal and Track Vertical conditions we found that the difference between them was not significant F(2,34)=1.43, p>.23 so these two were combined for subsequent analysis. Comparing the baseline Track Line condition with the Track Vertical/Horizontal conditions showed all effects to be significant; the difference between the two conditions with MS=9862, F(1,17)=300, p<000, the effect of cue duration with MS=600,
F(2,34) = 101, and P < .000 and the interaction with MS = 342, F(2,34) = 29.1, p < .000. This finding replicates our earlier finding that a cue that requires voluntary allocation of attention produces poorer tracking performance but the performance decrement decreases markedly with longer cue durations.

Also, as in Experiment 2, automatic exogenously cued tracking did not benefit significantly from the additional cue duration. The effect of cue duration on the Track Line condition was not significant, with F(2,34) = 2.9, MS = 26, p > .07. In contrast, the added cue time markedly improved tracking in all the endogenously cued conditions (for the Track Noline condition the repeated measures ANOVA yielded F(2,34) = 11.5, MS = 197, p < .000, and for the combined track horizontal/track vertical it was F(2,34) = 104, MS = 976, p < .000).

In Experiment 2 we found that when cues were available for 1080 ms the gap between the automatic and the attentive selection disappeared. In the present experiment the gap did not disappear even at the longest duration tested (900 ms). However it did decrease significantly as cue duration increased, as shown by the significant interaction between the effect of condition (Track Line vs Track Horizontal/Vertical) and cue duration. To examine this interaction in more detail we computed the gap between baseline and the endogenous (Track Horizontal/Vertical) cue condition for each cue duration, and then compared the difference in this gap as cue duration increased. A post-hoc t-test showed that the difference in the gap size as we go from 300 ms to 600 ms was significant (t = 4.2, df = 17, p < .001) as was the difference in the gap size as we go from 600 ms to 900 ms (t = 3.0, df = 17, p < .01), showing the gap kept getting significantly smaller as cue duration increased. Perhaps selecting horizontal/vertical bars among vertical/horizontal bars involved a slower search that would require a cue duration longer than 900 ms to fully eliminate.
Figure 4. Tracking performance when the task was to track the circles cued for three different durations by (1) the presence of a vertical line among circles without any line (2) the absence of a vertical line among circles with vertical lines (3) the presence of a vertical line among circles with horizontal lines or (4) the presence of a horizontal line among circles with vertical lines (for m score see text).

Discussion

The results of Experiment 4 support the hypothesis that voluntary selection of objects for tracking takes more time than automatic selection, even if we control for the effect of any additional time it takes to disengage attention once it has been automatically captured by exogenous cues such as onsets. Comparing the two conditions assumed to involve endogenous cueing (Track Noline and Track Vertical/Horizontal) we see that performance in the former case, which presumably involves a disengagement stage, is even higher than in the latter case, which does not involve disengagement. More important, however, the difference between the baseline condition (Track Line) and the conditions involving endogenous cues but not requiring an attentional disengagement stage (i.e., selecting horizontal/vertical lines among vertical/horizontal lines) continued to show a marked decrease when more time was available. This finding is consistent with the assumption that voluntary selection of objects in MOT requires additional time not attributable to a disengagement operation, and therefore that it may involve serial
allocation of attention. A more direct test of this hypothesis would require varying the number of targets that have to be selected. This was the purpose of Experiment 4.

**Experiment 4**

In this experiment we examined whether the relative effectiveness of the increased cue duration, found in the previous experiments, varied with the numbers of targets selected.

*Method*

This experiment used the same procedure as in the previous experiments except that the number of targets varied between 3 and 5 while the number of nontargets was always 5. Cue duration was either 200 ms or 1200 ms. The cue used to indicate which objects to track was the absence of a vertical line on the target objects. This is the same cue as used in the Track Noline condition of Experiment 3, which we assume provides an endogenous (non-automatic) cue. The number of targets (3 vs 4 vs 5) was blocked, with 128 trials per block. The order of blocks balanced across subjects and the cue duration was assigned randomly.

This experiment requires comparing performance for different numbers of targets. Since observers were forced to make \( n \) responses when there were \( n \) targets, the probability of correctly guessing targets increases with increasing numbers of targets. Thus rather than use the percent of targets correctly identified as the measure of performance, as was done in the first three experiments, we used an estimate of the number of targets correctly tracked that took into account the apparent performance inflation that was attributable to guessing in the forced-choice method. This estimate, which we refer to as the *effective number of targets tracked* \((m)\) is derived in the Appendix (for purposes of comparison with previous experiments the \( m \) score is also shown in earlier graphs).

*Subjects.* A total of 16 Rutgers University students participated in sessions lasting approximately 90 minutes to fulfill a psychology course requirement. All subjects had normal or corrected to normal vision. The data from two of the subjects was discarded because their overall tracking scores were low (below 50%).

*Results*

The results are shown in Figure 5. Tracking performance measured in terms of the \( m \) score increased across number of targets, reaching a maximum score of 3.97 for 1200 ms cue presentation of 5 targets. The effect of number of targets was significant (\( MS=9.6, F(2,26)=84.8, \)
p<.000), as was the effect of cue duration (MS=13.4, F(1,13)=57.7, p<.000), as well as the interaction of these two variables (MS=.22, F(2,26)=5.1, p<.01).

**Figure 5.** Differential effect of cue duration on endogenous selection of different numbers of targets (expressed in terms of the $m$ score – see Appendix).

**Discussion**

These results show that the number of targets that can be tracked increases as more targets are presented, up to the previously estimated maximum of just over 4 (Pylyshyn & Storm, 1988). More relevant to our hypothesis, however, the data also show that the benefit of the longer cue duration increased significantly when there are more endogenously-cued targets. To put it another way, for selection based on exogenous cues, the improvement in tracking resulting from longer cue times was greater when there were more targets, as would be expected if targets had to be visited serially in order to be selected for tracking.

**Summary and General Discussion**

These four experiments used the Multiple Object Tracking paradigm to address the question whether objects can be voluntarily selected for tracking and if so, whether endogenously (i.e. voluntarily) selected items are selected differently than items that capture attention in an exogenous (automatic, data-driven) manner. Experiment 1 showed that observers can easily select (and subsequently track) objects identified by particular numerals. Although they made more errors when using such symbolic cues, their tracking performance was still very high
(87.8% compared with the baseline performance of 93.8%). Experiment 2 showed that if the cue was available for longer (i.e., 1080 ms as opposed to 360 ms) the difference between tracking voluntarily selected items and tracking automatically selected items disappears. This shows that voluntary selection of the sort used in this experiment (i.e., the complement set of those that were flashed) can be successfully accomplished in 1080 ms. It also suggests that once selected, items can be tracked independent of how the selection had been accomplished.

We examined a possible confound to the interpretation that more time is required to select endogenously cued objects (e.g., nonflashed objects) because selecting them requires that they be visited serially. The alternative explanation is that it takes longer to select the nonflashed items because it takes additional time to disengage attention from the flashed objects and re-engage it on the nonflashed objects. Experiment 3 used a number of different cue types and showed that the cue that required an attentional disengagement stage (Track Noline) led to even better performance than other endogenous cues, thus suggesting that the difference found in Experiment 2 was not due to the time required for disengagement (or at least not entirely). Moreover, the pattern of diminishing difference between exogenously selected items and endogenously selected items with increasing cue duration still occurs when selection does not involve disengaging attention from one set of items and re-engaging it to the another set (since selecting vertical bars from among horizontal ones or vice versa does not involve any disengagement). Finally, Experiment 4 showed that the difference between tracking performance given a short (200 ms) endogenous cue and a long (1200 ms) endogenous cue increases as the number of targets increases. This is consistent with the hypothesis that the difference is attributable to serially attending the individual targets since the more targets there are the longer one would expect the serial process to take and therefore the more the selection process would benefit from a longer due duration.

The MOT task was initially developed to explore an independently motivated theory called Visual Index Theory (or FINST Theory). The FINST mechanism was hypothesized because of a need that arose in designing a computational system that drew diagrams and made conjectures on the basis of what it noticed (Pylyshyn, Elcock, Marmor, & Sander, 1978). We needed to be able to select and keep track of token visual elements independently of (and prior to) having encoded any of their properties, in order to maintain the coherence of a representation being constructed incrementally over time (some of the motivation for this idea is reviewed in, Pylyshyn, 2000; Pylyshyn, 2001). Because the FINST mechanism was viewed as the first preconceptual contact between things in the world and cognitive representations, it was assumed that FINST indexes
were captured in a data-driven manner, rather than being assigned under top-down voluntary control. Notwithstanding this basic assumption about the bottom-up nature of index assignment, it was recognized that even within the framework of the FINST theory a certain capacity for voluntary influence on index assignment was possible. The theory (as described in, Pylyshyn, 2001, p147) suggested an indirect way in which voluntarily scanned attention could provide enabling conditions for the automatic assignment of indexes to occur at voluntarily chose loci (see also the assumptions of FINST theory listed in, Pylyshyn, 2003, p211). The way this could work is that focal attention would be scanned serially to the objects that an observer wished to index. While focusing on each such object with a narrow “beam” of attention, the object could provide locally-distinct cues that would allow a pop-out to occur, the way that singletons pop out in visual search (Theeuwes, 2004; Yantis, 1993). This is consistent with the findings that drawing attention to objects improves sensitivity and target detection at those objects even when there are several of them (Burkell & Pylyshyn, 1997; Solomon, 2004) and even when they are cued endogenously (Dosher, Liu, Blair, & Lu, 2004; Eckstein, Pham, & Shimozaki, 2004). In any case this line of reasoning is committed to the prediction that when endogenously guided selection takes place it does so by serially visiting the items that are selected. The findings of the present study are consistent with this view since they suggest that for cues other than “pop-out” properties, such as onsets or singletons, targets must be attended serially, a process that takes an additional increment of time for each target selected in this way.

These findings are not only relevant to the development of theories of object selection and tracking, such as the Visual Index Theory, but they are an important step towards understanding dynamic process of visual selection itself, a process that constitutes the first stage in the allocation of visual attention to the world, and therefore of visual information processing.
Appendix: Computing the Effective Number of Targets Tracked (m)

Let $N =$ total number of objects  
$n =$ number of targets (in our case 3, 4 or 5)  
$T =$ observer's score (number correctly identified at the end of a trial)  
$m =$ estimated number that the observer actually tracked

The observers' tracking score consists of the number actually tacked ($m$) plus the number correctly guessed ($\hat{E}$) or, $T = m + \hat{E}$

Assuming that observers are forced to make exactly $n$ responses (which is the case in all our experiments), then once they select the $m$ items actually tracked, they are forced to make $n-m$ guesses from among the remaining unselected $N-m$ items. Since by hypothesis $m$ correct targets have already been chosen, the proportion of remaining objects that are targets is $(n-m)/(N-m)$. Thus after selecting the $m$ tracked items the estimated guess score is:

$$\hat{E} = \text{total guesses made} \times \text{proportion of the to-be-guessed items that are actually targets}$$

$$= (n-m) \times \left(\frac{n-m}{N-m}\right)$$

Substituting,

$$T = m + \hat{E} = m + (n-m) \times \left(\frac{n-m}{N-m}\right)$$

Solving for $m$:

$$m = \left(\frac{n^2 - NT}{2n - N - T}\right)$$

In the first three experiments we had $N = 8$ and $n = 4$, and the performance measure shown in the graphs is expressed in terms of the percent $100(T/8)$. In Experiment 4, $n$ varies between 3 and 5 and $T$ varies between 8 and 10 (since there are always 5 nontargets).
References


Pylyshyn, Z. W., & Dennis, J. L. M. (in preparation). Can Multiple Object Tracking make use of individual object properties?


