Spatial Vision, Vol. 19, No. 6, pp. 485–504 (2006)

Also available online - www.brill.nl/sv

Dynamics of target selection in Multiple Object Tracking (MOT)

ZENON W. PYLYSHYN * and VIDAL ANNAN, JR.

Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, USA

Received 27 May 2005; accepted 27 March 2006

Abstract—In four experiments we address the question whether several visual objects can be selected voluntarily (exogenously) and then tracked in a Multiple Object Tracking paradigm and, if so, whether the selection involves a different process. 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.

Keywords: Tracking; attention; selection; attention capture; multiple object tracking.

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 and Papathomas, 1984; Shulman and

*To whom correspondence should be addressed at Rutgers University, Center for Cognitive Science, Psychology Building Annex, 152 Frelinghuysen Road, Piscataway, New Jersey 08854-8020, USA. E-mail: zenon@ruccs.rutgers.edu

Wilson, 1987), and on such features as color (Friedman-Hill and Wolfe, 1995; Green and Anderson, 1956), local shape (Egeth *et al.*, 1984), motion (McLeod *et al.*, 1991) (particularly looming, Franconeri and Simons, 2003), stereo-defined depth (Nakayama and Silverman, 1986) or other dynamic events (Franconeri *et al.*, 2004). Such selection is typically course-grained compared to the main types 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 and 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 automatically (and preattentively) or whether some voluntary and perhaps serial process may be involved. Our position has been that a 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 is automatic or voluntary and whether it is subject to any special conditions has not been tested outside a few tasks such as searching through subsets (Burkell and Pylyshyn, 1997) or subitizing (Trick and Pylyshyn, 1994).

Another closely related question concerns the maintenance or retention 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

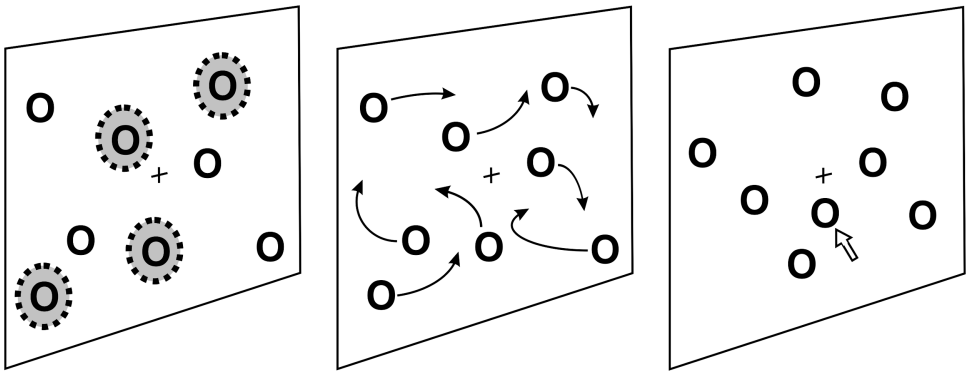

Figure 1. In Multiple Object Tracking a set of simple objects is briefly identified as the target set (e.g. by flashing the ones shown here with shadows around them) and then they move around unpredictably among identical nontargets for 5–10 s. The observer must identify the targets using a computer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

the Multiple Object Tracking (MOT) task developed by Pylyshyn and 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 Fig. 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 s). 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 refer to the FINST theory in discussing the hypotheses and findings in the final section.

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 — see Pylyshyn and 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. Performance in MOT depends on two processes, which may be sensitive to different properties; selection and tracking. With short trial durations leading to high tracking performance, we assume that the variability arises primarily from the selection stage; and that that can be strengthened to some extent by examining whether cue-presentation factors affect the performance.

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 when 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. 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), the basic prediction is that automatic capture is likely to operate in parallel across the visual field whereas voluntary assignment of indexes may require that targets be visited serially or at least that a different and effortful process is involved. 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 locally-distinct features (Bauer *et al.*, 1999; Johnson *et al.*, 2001). Such features have sometimes been referred to as 'popout' features since they typically lead to fast detection in visual search, but we are not committed to the equivalence of popouts and index capturing (moreover, as Wolfe, 1992, has pointed out, effortless detection may involve different mechanisms in search than in other multi-feature processes such as texture segregation). The present studies address this general prediction and also explore the parameters that determine whether items are selected automatically or voluntarily.

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 a G3 Apple computer. Targets and non-targets were identical except in the initial target-designation phase of each trial. Objects (targets and non-targets) consisted of white rings on a black background, subtending a visual angle of 2.7 degrees. The thickness 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. This trial duration was chosen so that (a) it was short enough that most of the variance in the tracking performance could be attributed to the selection process rather than errors in the tracking itself and (b) in the course of a trial the objects would have moved enough so they could not be selected based on their initial location alone. 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 0.10. 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 and Mingolla, 2002) that T-junction depth cues tend to minimize tracking errors when objects overlap.

Procedure and 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 s. 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-h 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 s, the digits disappeared and all the circles moved as described above for 5 s. 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 Fig. 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 s, 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 ($t = 4.26, df = 14, p < 0.001$). 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%. (For comparison with the 'effective number of targets tracked' or m-score, that will be introduced in analyzing Experiment 4, the m score in the number condition was 3.44 while in the baseline condition it was 3.74.)

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 and 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 s, this process involvers 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 *et al.*, 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 Flashed' 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 min 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 we 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 two selection conditions (Track Flashed and Track Nonflashed) and two cue durations (360 *vs* 1080 ms). All effects were significant. Overall tracking in the conditions in which targets were the flashed items, averaged over long and short cues, (93.1%) was significantly better than in the conditions where targets were the non-flashed items (89.9%); $F(1, 13) = 17.8$, $MS = 136.7$, $p < 0.001$. Overall tracking performance, averaged over track-flashed and track-nonflashed cue conditions, was significantly better when the cue duration was 1080 ms (93.3%) than when it was 360 ms (89.7%); $F(1, 13) = 9.1$, $MS = 187.8$, $p < 0.01$. In addition, as predicted, the interaction between these two effects was also significant, $F(1, 13) = 5.67$, $MS = 89.6$, $p < 0.03$. When the difference between short (360 ms) and long (1080 ms) cues was compared using a *post-hoc* t-test, the difference was not significant for the Track-Flashed condition (means were, respectively, 92.5% and 93.6%, resulting in $t = 1.0$, $df = 13$, $p > 0.34$) whereas it was significant for the Track-Nonflashed condition (means were, respectively, 86.8% and 93.0% resulting in $t = 3.1$, $df = 13$, $p < 0.01$). When conditions were compared at different cue durations, conditions were significantly different at the short duration ($t = 2.57$, $df = 14$, $p < 0.02$) whereas they were not significant at the long duration ($t = 0.94$, $df = 14$, $p > 0.36$), so additional time makes no difference for the automatic selection case. These results are shown in Fig. 2 (the 'effective number of items tracked' or m score is introduced in Experiment 4).

Discussion

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

There are, however, alternative possible explanations for these results. One is that subjects actually tracked the flashed objects and then, at the end of each

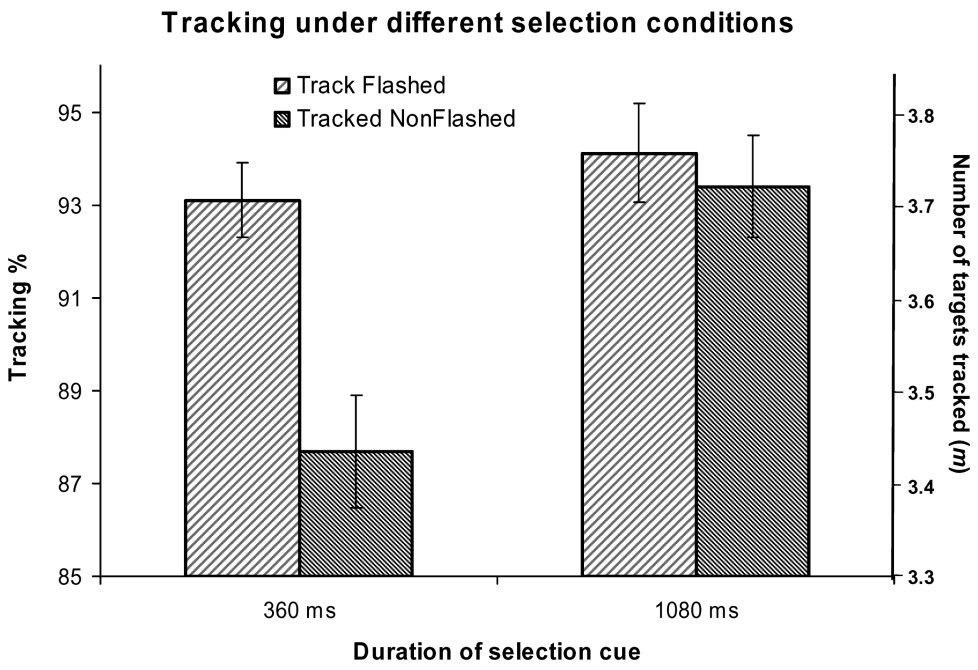

Figure 2. Comparison of exogenous, automatic (track flashed) and endogenous, voluntary (track non-flashed) target selection at different cue durations (error bars show standard errors).

trial, indicated the complement set. We were cognizant of this possibility from the beginning and looked for evidence that observers had used this strategy in our post-experiment debriefing. None of the observers indicated that they had used this strategy — either in their open-ended remarks or in response to our direct question. Also, in piloting the experiment, we explicitly asked several observers to use this method. These observers reported that they had some difficulty in making the switch at the end of the trial and they made more errors when they did so. Another possible explanation for why more time was needed to select the non-flashed items is that after the automatic assignment of attention (or, in our terms, visual indexes), the subsequent disengagement and re-engagement of attention (in order to attend to the non-flashed circles) requires additional time (some reports estimated the time to disengage and re-engage at 300–600 ms, Duncan *et al.*, 1994; Mueller *et al.*, 1998). If that is the case in Experiment 2, then we would not be 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 increased 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. 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. The circles in this experiment were identical to those used in the previous experiments. When a line segment was superimposed over the circle, the line segment was the same thickness as the circular ring (0.11 degrees) and was about 25% longer than the circle diameter (the line segment subtended 3.3 degrees of visual angle while the circle itself subtended 2.7 degrees of visual angle).

Method

Subjects. Eighteen Rutgers University students participated in session lasting approximately 90 min 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. Conditions (1) *vs* (2) replicate 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. Conditions (3) and (4) tested the selection of objects using cues that, at least *prima facie*, require voluntary attention for identification — namely vertical *versus* horizontal line segments.

Although a horizontal line segment 'pops out' from a field of vertical line segments, *selecting all* circles with horizontal line segments from among those with vertical line segments does not work like a popout in this case, inasmuch as it takes longer to select items when there are more of them (see Experiment 4 and the Summary and general Discussion section). 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

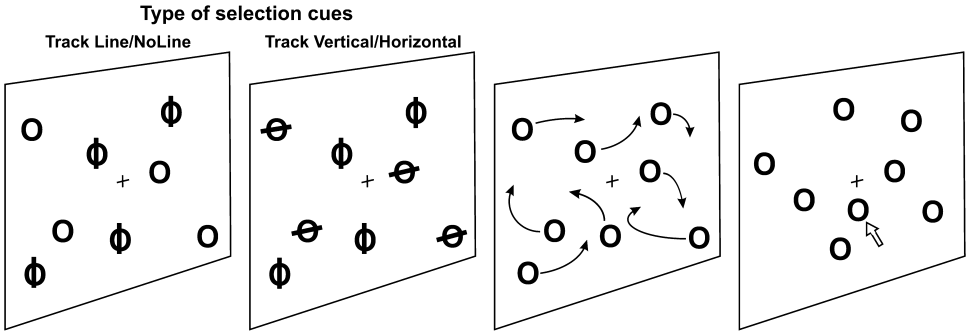

Figure 3. Displays used in Experiment 3, showing the four types of selection cues (only one type of cue appeared in each block).

were presented in separate blocks with 64 trials per block. The order of blocks was counterbalanced across subjects.

Results

Performance on all four conditions is shown in Fig. 4. A repeated-measures analysis of variance was performed on all four 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 $MS = 5791$, $F(3, 51) = 165.6$, $p < 0.001$, the effect of cue duration was significant with $MS = 1653$, $F(2, 34) = 131.9$, $p < 0.001$ and the interaction of these two was significant with $MS = 175$, $F(6, 102) = 11.3$, $p < 0.001$. Examining only the Track Line and Track No-line conditions (which are comparable to those of Experiment 2) showed that the main effect of condition was significant, with $MS = 232.8$, $F(1, 17) = 28.7$, $p < 0.001$, the effect of cue duration was significant, with $MS = 182.6$, $F(2.34) = 11.5$, $p < 0.001$, and the interaction was also significant with $MS = 41$, $F(2, 34) = 4.1$, $p < 0.02$, 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 $F(2, 34) = 1.43$, $p > 0.23$, 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, with the difference between the two conditions yielding $MS = 9862$, $F(1, 17) = 300$, $p < 0.001$, the effect of cue duration with $MS = 600$, $F(2, 34) = 101$, and $p < 0.001$ and the interaction with $MS = 342$, $F(2, 34) = 29.1$, $p < 0.001$. This finding replicates our earlier finding that a cue that requires voluntary allocation of attention produces poorer tracking performance and that this performance decrement decreases markedly with longer cue durations. (Figure 4 also shows performance using the m score or the 'effective number of objects tracked'. This score is shown solely for purposes of comparison with performance in Experiment 4 where the m score is introduced to allow comparison of performance across different number of targets, where the

'percent correctly identified' score is inappropriate — see discussion of Experiment 4 below.)

In Experiment 2 we found that when cues were available for 1080 ms the difference between the automatic and the attentive selection disappeared. In the present experiment the difference does not disappear but remained significant for all three cue durations. However, it decreased greatly as cue duration increased, as shown by the significant interaction between the effect of condition (Track Line *vs* Track Horizontal/Vertical) and cue duration. When we compared the performance difference on the Horizontal/Vertical conditions between 300 ms and 600 ms with the performance difference between 600 ms and 900 ms this differences was found to be highly significant ($t = 2.96$, $df = 17$, $p < 0.01$), so the gap between baseline and the attentive cue conditions decreased significantly as cue duration increased. Moreover, as in Experiment 2, 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 $F(2, 34) = 2.9$, $MS = 26$, $p > 0.07$. By contrast, the added cue time markedly improved tracking in all the endogenously cued conditions (for the Track No-line condition the repeated measures ANOVA gave $F(2, 34) = 11.5$, $MS = 197$, $p < 0.001$, and for the combined track horizontal/track vertical $F(2, 34) = 104$, $MS = 976$, $p < 0.001$).

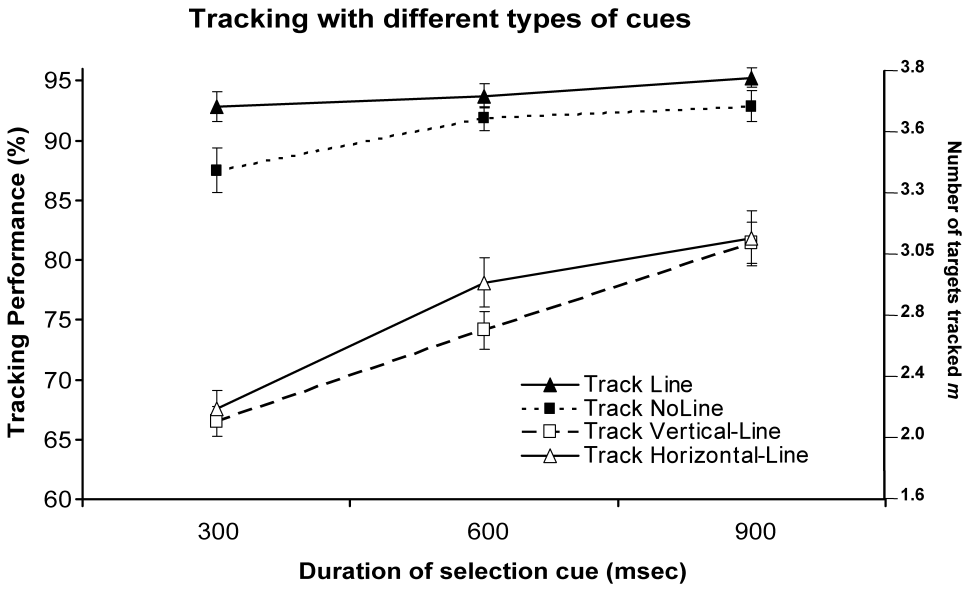

Figure 4. Tracking performance when the task was to track the circles indicated by the brief presence of a vertical line among circles without any line, by the absence of a vertical line among circles with vertical lines, by the presence of a vertical line among circles with horizontal lines or by the presence of a horizontal line among circles with vertical lines (for *m* score see description of Experiment 4). (Error bars are standard errors.)

Discussion

The results of Experiment 3 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 onset cues. The difference between the baseline condition (Track Line) and the conditions involving endogenous cues, such as unmarked circles or horizontal lines among vertical lines or vertical lines among horizontal lines, showed that selection takes longer but that the difference decreases markedly when more time is available. This evidence is consistent with the assumption that voluntary selection of objects in MOT requires additional time 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 changed when different numbers of targets are 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 non-targets 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 (i.e. it was the same cue as in the Track No-line condition of Experiment 3). 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 (and in previous published studies), we used an estimate of the number of targets correctly tracked that took into account inflation due 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). (Note: The m score was not used in previous MOT studies because the performance, as measured by the proportion of targets correctly tracked, was sufficient to compare tracking under different conditions. However the proportion correct does not take into account the different contribution of guessing when

different numbers of targets or nontargets are used. For that reason we developed the *m* score (first reported in Annan and Pylyshyn, 2002) for comparing performance on conditions with different numbers of targets, as in the present Experiment 4. This score was developed prior to publication of the elegant analysis of (Hulleman, 2005), who showed that estimating how many objects were actually tracked is in general a complex problem whose solution must rest on assumptions about guessing strategies and on whether any non-targets are tracked. The method we use here (derived in the Appendix) is indeed based on simplifying assumptions, including the assumption that no non-targets are tracked and that the increase in probability of a correct guess with larger numbers of targets is solely a function of the increased chance of choosing a target from the remaining items, and not from any differences in guessing strategies. Further experiments are needed to determine whether these simplifying assumptions can be sustained.)

Subjects. A total of 18 Rutgers University students participated in sessions lasting approximately 90 min 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 Fig. 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 ($MS = 9.6$, $F(2, 26) = 84.8$, $p < 0.001$), as was the effect of cue

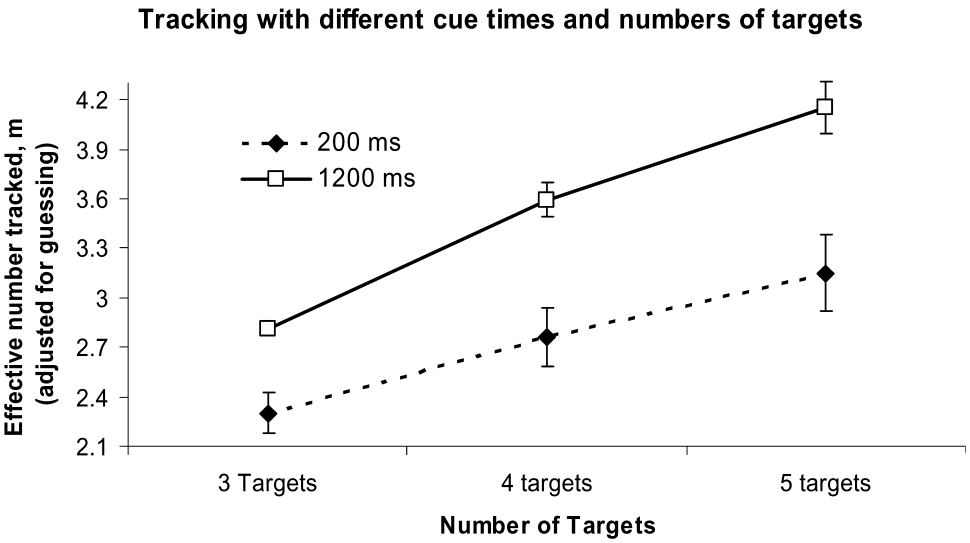

Figure 5. Differential effect of cue duration on endogenous selection of different numbers of targets (in terms of the *m* score — see Appendix). (Error bars are standard errors.)

duration ($MS = 13.4$, $F(1, 13) = 57.7$, $p < 0.001$), as well as the interaction of these two variables ($MS = 0.22$, $F(2, 26) = 5.1$, $p < 0.01$).

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 4. More relevant to our hypothesis, however, the data also show that the larger number of targets benefit more from increased cue duration (at least for the cues used here, where targets are identified as those that did not have a vertical line flashed on them briefly). 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. This is consistent with our hypothesis that the selection of multiple targets defined by features that do not capture attention in an automatic exogenous manner requires that the targets be visited serially.

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. As expected, performance in tracking by numeral designation was poorer than tracking where targets are designated by flashing, yet tracking performance was still very high (87.8% compared with the baseline performance of 93.8%). We interpreted this as suggesting that individual targets had to be visited serially, in order to read the digits. Experiment 2 showed that tracking performance for a cue that attracted attention automatically (e.g. a flash) was better than for a cue that required voluntary selection (i.e. selecting unflashed items) but that the difference disappeared if the cue was available for longer (i.e. 1080 ms as opposed to 360 ms). This was interpreted as suggesting that voluntary selection of the sort used in this experiment (i.e. the complement set of those that were flashed) can be successfully accomplished if more time is provided (i.e. 1080 ms) so that these items could be voluntarily located (presumably in a serial fashion). The fact that tracking performance with longer-lasting voluntary cues is the same as for the automatic attention-eliciting cues also suggests that once selected, items can be tracked independent of how the selection had been accomplished. This is consistent with the assumption inherent in these experiments that the different cuing methods affect selection rather than persistence of the selection through the tracking trial. A more convincing test would be to show that cue-presentation factors affect tracking performance independently of (and in addition to) clearly tracking-specific factors such as trial duration.

There are several alternative explanations for the finding that more time is needed for non-flash cues to be effective. Before attention can be directed to the non-flashed items the initial automatic attention capture has to be overcome. Thus it may be that it takes longer to select the non-flashed items not because the items have to be visited serially, but because it takes additional time to disengage attention from the flashed objects and re-engage it on the non-flashed objects. Experiment 3 used a number of different cue types to show that the pattern of diminishing difference between flash-selected items and nonflash-selected items still occurs when selection does not involve disengaging attention from one set of items and re-engaging it to the another set. This experiment showed that selecting circles with vertical bars from among circles with horizontal bars requires more time than selecting circles with vertical bars from among circles with no bars and that this time difference significantly diminishes with increasing cue duration. We interpret this result as suggesting that voluntary selection requires additional time independent of any time required to disengage attention. This interpretation rests on the assumption that distinguishing circles with vertical lines from circles with horizontal lines requires voluntary attention — that it is effortful and probably serial. This assumption was motivated by the informal observation that locating four vertical line segment among the horizontal ones was effortful and slow. While search experiments have shown that detecting a vertical line segment among horizontal segments (and *vice versa*) is fast and independent of the number of distractors (i.e. it is a 'popout' search), the task of finding a single item with a unique property is quite different from that of locating all items of that type; for a particular set of features the first may be parallel and automatic whereas the second may be effortful and slow (as is the case in the texture segmentation examples discussed by Wolfe, 1992).

Another concern with the particular choice of stimuli used in Experiment 3 is that more time may be needed for selecting targets if their properties are pairwise less easily discriminated from those of non-targets. For example, a circle with a vertical line may be more easily discriminated from a circle without any line than from a circle with a horizontal line. Thus one might expect selection of items based on line *versus* no-line cues to be faster than selection based on vertical *versus* horizontal line segments. While this alternative explanation of the data found in Experiment 3 is plausible, it is also consistent with the present proposal since discriminating similar shape-features may require focal attention. Focal attention is known to enhance spatial resolution (Yeshurun and Carrasco, 1998), so more similar shape-feature pairs may require that focal attention be focused on the individual features. Indeed, Cave and Zimmerman (1997) reported evidence that in search, focal attention is allocated so as to distinguish similar items, and is allocated in proportion to the degree of similarity of items' shape properties. Thus although the increased time required to select vertical line cues from horizontal line cues (as opposed to circles with or without line segments) may be associated with their poorer discriminability, the underlying process may well be mediated by attention scanning.

Finally, Experiment 4 showed that the difference between tracking performance with a short (200 ms) cue and a long (1200 ms) cue increases as the number of targets increases. This is consistent with the hypothesis that the difference is attributable to serially attending to the individual targets since the more targets there are the longer one would expect a serial process to take and therefore the more the selection process would benefit from a longer cue duration. Once again, the conclusion that a serial process may be involved is based on the plausible assumption that more time indicates more operations. Yet, as in all discussions of parallel *versus* serial processing, other assumptions can replace the assumption that more time arises from more operations, such as the assumption of shared resources and 'competitive interaction' between parallel streams that results in slower reactions when there are more items. Arguments have been made for both points of view in search experiments (see, for example, Bricolo *et al.*, 2002; Mordkoff and Egeth, 1993; Townsend and Wenger, 2004; Woodman and Luck, 2003) and we suspect that the argument will only be settled in the context of a broader theory that accounts for a greater range of phenomena (as Newell, 1973, pointed out, individual experiments by themselves cannot settle questions such as whether a process is serial or parallel). So far, the evidence we have presented only shows that under certain conditions of cue presentation additional processing time is required.

While these findings are subject to several interpretations, it is of interest in the present context that they are consistent with the prediction of an independently motivated theory called Visual Index Theory (or FINST Theory). The FINST mechanism was hypothesized to fill a need for selecting and keeping track of token visual elements, in MOT as well as other phenomena, independently of encoding any of their properties (some of the motivations for these assumptions are reviewed in Pylyshyn, 2000, 2001). Because the FINST mechanism was viewed as the initial causal contact between objects in the world and cognitive representations, it was assumed that FINST indexes are *captured* in a data-driven manner, rather than being assigned under voluntary cognitive control. Yet the theory (as described in Pylyshyn, 2001, 2003, Chapter 5) also suggests a way in which voluntarily allocated focal attention might provide enabling conditions for index assignment to occur at voluntarily chosen loci. The way this would work is that focal attention would be scanned serially to the objects that an observer wished to index. While a narrow 'beam' of attention focused on each such object, locally-distinct cues would allow a feature pop-out to occur, the way that singletons pop out in visual search (Theeuwes, 2004; Yantis, 1993). This theory is consistent with the findings that precuing several such objects improves sensitivity and target detection at those objects (Burkell and Pylyshyn, 1997; Solomon, 2004) even when they are cued endogenously (Dosher *et al.*, 2004; Eckstein *et al.*, 2004). This line of reasoning is committed to the prediction that when voluntary selection takes place it does so because the items that are selected are visited serially with focal attention. The findings of the present study are consistent with the view that targets cued by features other than onsets or

singletons may be attended serially, a process that takes an additional increment of time for each target selected in this way.

These findings are not only important for 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.

Acknowledgements

We wish to acknowledge the support of NIH Grant Number 1R01-MH60924 to the first author and an NRSA post-doctoral aware to the second author.

We are grateful to Todd Horowitz for pointing out the problem of distinguishing selection from endurance and for suggesting this control experiment.

REFERENCES

Annan, V. and Pylyshyn, Z. W. (2002). Can indexes be voluntarily assigned in multiple object tracking? *Journal of Vision* **2**, 243a.

Atchley, P., Kramer, A. F. and Hillstrom, A. P. (2000). Contingent capture for onsets and offsets: attentional set for perceptual transients, *J. Exper. Psychol.: Human Perception and Performance* **26**, 594–606.

Bauer, B., Jolicur, P. and Cowan, W. B. (1999). Convex hull test of the linear separability hypothesis in visual search, *Vision Research* **39**, 2681–2695.

Bricolo, E., Gianesini, T., Fanini, A., Bundesen, C. and Chelazzi, L. (2002). Serial attention mechanisms in visual search: A direct behavioral demonstration, *J. Cognitive Neurosci.* **14**, 980–993.

Burkell, J. and Pylyshyn, Z. W. (1997). Searching through subsets: a test of the visual indexing hypothesis, *Spatial Vision* **11**, 225–258.

Cave, K. R. and Zimmerman, J. M. (1997). Flexibility in spatial attention before and after practice, *Psychol. Sci.* **8**, 399–403.

Comtois, R. (1999), VisionShell PPC Software Library.

Dosher, B. A., Liu, S.-H., Blair, N. and Lu, Z.-L. (2004). The spatial window of the perceptual template and endogenous attention, *Vision Research* **44**, 1257–1271.

Duncan, J., Ward, R. and Shapiro, K. L. (1994). Direct measurement of attentional dwell time in human vision, *Nature* **369**, 313–315.

Eckstein, M. P., Pham, B. T. and Shimozaki, S. S. (2004). The footprints of visual attention during search with 100% valid and 100% invalid cues, *Vision Research* **44**, 1193–1207.

Egeth, H. E., Virzi, R. A. and Garbart, H. (1984). Searching for conjunctively defined targets, *J. Exper. Psychol.* **10**, 32–39.

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

Franconeri, S. L. and Simons, D. J. (2003). Moving and looming stimuli capture attention, *Perception and Psychophysics* **65**, 999–1010.

Franconeri, S. L., Simons, D. J. and Junge, J. A. (2004). Searching for stimulus-driven shifts of attention, *Psychonomic Bulletin and Review* **11**, 876–881.

Friedman-Hill, S. and Wolfe, J. M. (1995). Second-order parallel processing: visual search for odd items in a subset, *J. Exper. Psychol.: Human Perception and Performance* **21**, 531–551.

Green, B. F. and Anderson, L. K. (1956). Color coding in a visual search task, *J. Exper. Psychol.* **51**, 19–24.

Hulleman, J. (2005). The mathematics of multiple object tracking: from proportions correct to number of objects tracked, *Vision Research* **45**, 2298–2309.

Johnson, J. D., Hutchison, K. A. and Neill, W. (2001). Attentional capture by irrelevant color singletons, *J. Exper. Psychol.: Human Perception and Performance* **27**, 841–847.

Julesz, B. and Papathomas, T. V. (1984). On spatial-frequency channels and attention, *Perception and Psychophysics* **36**, 398–399.

LaBerge, D. (1998). Attentional emphasis in visual orienting and resolving, in: *Visual Attention. Vancouver Studies in Cognitive Science*, R. D. Wright (Ed.), Vol. 8. Oxford University Press, New York, NY, USA, pp. 417–454.

McLeod, P., Driver, J., Dienes, Z. and Crisp, J. (1991). Filtering by movement in visual search, *J. Exper. Psychol.: Human Perception and Performance* **17**, 55–64.

Mordkoff, J. T. and Egeth, H. E. (1993). Response time and accuracy revisited: converging support for the interactive race model, *J. Exper. Psychol.: Human Perception and Performance* **19**, 981–991.

Mueller, M. M., Teder-Saelejaervi, W. A. and Hillyard, S. A. (1998). The time course of cortical facilitation during cued shifts of spatial attention, *Nature Neuroscience* **1**, 631–634.

Nakayama, K. and Silverman, G. H. (1986). Serial and parallel processing of visual feature conjunctions, *Nature* **320**, 264–265.

Newell, A. (1973). You can't play 20 Questions with nature and win: projective comments on the papers of this symposium, in: *Visual Information Processing*, W. Chase (Ed.). Academic Press, New York, USA, pp. 283–308.

Posner, M. I. (1980). Orienting of attention, *Quart. J. Exper. Psychol.* **32**, 3–25.

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

Pylyshyn, Z. W. (2000). Situating vision in the world, *Trends in Cognitive Sciences* **4**, 197–207.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision, *Cognition* **80**, 127–158.

Pylyshyn, Z. W. (2003). *Seeing and Visualizing: It's Not What You Think*. MIT Press/Bradford Books, Cambridge, MA, USA.

Pylyshyn, Z. W. and Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism, *Spatial Vision* **3**, 1–19.

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

Pylyshyn, Z. W., Burkell, J., Fisher, B., Sears, C., Schmidt, W. and Trick, L. (1994). Multiple parallel access in visual attention, *Canad. J. Exper. Psychol.* **48**, 260–283.

Saarinen, J. (1996). Localization and discrimination of 'pop-out' targets, *Vision Research* **36**, 313–316.

Scholl, B. J. (2001). Objects and attention: the state of the art, *Cognition* **80**, 1–46.

Shulman, G. L. and Wilson, J. (1987). Spatial frequency and selective attention to local and global information, *Perception* **16**, 89–101.

Solomon, J. A. (2004). The effect of spatial cues on visual sensitivity, *Vision Research* **44**, 1209–1216.

Theewues, J. (1994). Endogenous and exogenous control of visual selection, *Perception* **23**, 429–440.

Theeuwes, J. (2004). Top-down search strategies cannot override attentional capture, *Psychonomic Bulletin and Review* **11**, 65–70.

Townsend, J. T. and Wenger, M. J. (2004). The serial-parallel dilemma: a case study in a linkage of theory and method, *Psychonomic Bulletin and Review* **11**, 391–418.

Treisman, A. and Gelade, G. (1980). A feature integration theory of attention, *Cognitive Psychology* **12**, 97–136.

Trick, L. M. and Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited capacity preattentive stage in vision, *Psychol. Rev.* **101**, 80–102.

Viswanathan, L. and Mingolla, E. (2002). Dynamics of attention in depth: evidence from multi-element tracking, *Perception* **31**, 1415–1437.
Wolfe, J. M. (1992). 'Effortless' texture segmentation and 'parallel' visual search are not the same thing, *Vision Research* **32**, 757–763.
Woodman, G. F. and Luck, S. J. (2003). Serial deployment of attention during visual search, *J. Exper. Psychol.: Human Perception and Performance* **29**, 121–138.
Yantis, S. (1993). Stimulus-driven attentional capture, *Current Directions in Psychological Science* **2**, 156–161.
Yeshurun, Y. and Carrasco, M. (1998). Attention improves or impairs visual performance by enhancing spatial resolution, *Nature* **396**, 72–75.

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:

$\hat{E}$ = (total guesses made) × (proportion of the to-be-guessed items that are actually targets) = $(n - m) \times (\frac{n-m}{N-m})$.
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).