ELSEVIER

COGNITION

Cognition 80 (2001) 159–177

www.elsevier.com/locate/cognit

What is a visual object? Evidence from target merging in multiple object tracking

Brian J. Scholl[a,*], Zenon W. Pylyshyn[b], Jacob Feldman[b]

[a]*Harvard University, Cambridge, MA, USA*
[b]*Rutgers University, New Brunswick, NJ, USA*

Received 16 October 1999; accepted 17 November 2000

Abstract

The notion that visual attention can operate over visual objects in addition to spatial locations has recently received much empirical support, but there has been relatively little empirical consideration of what can count as an 'object' in the first place. We have investigated this question in the context of the multiple object tracking paradigm, in which subjects must track a number of independently and unpredictably moving identical items in a field of identical distractors. What types of feature clusters can be tracked in this manner? In other words, what counts as an 'object' in this task? We investigated this question with a technique we call *target merging*: we alter tracking displays so that distinct target and distractor locations appear perceptually to be parts of the same object by merging pairs of items (one target with one distractor) in various ways – for example, by connecting item locations with a simple line segment, by drawing the convex hull of the two items, and so forth. The data show that target merging makes the tracking task far more difficult to varying degrees depending on exactly how the items are merged. The effect is perceptually salient, involving in some conditions a total destruction of subjects' capacity to track multiple items. These studies provide strong evidence for the object-based nature of tracking, confirming that in some contexts attention must be allocated to objects rather than arbitrary collections of features. In addition, the results begin to reveal the types of spatially organized scene components that can be independently attended as a function of properties such as connectedness, part structure, and other types of perceptual grouping.

Keywords: Visual object; Target merging; Multiple object tracking

* Corresponding author. Present address: Department of Psychology, Yale University, P.O. Box 208205, New Haven, CT 06520-8205, USA. Fax: +1-203-432-7172.
E-mail address: brian.scholl@yale.edu (B.J. Scholl).

0010-0277/01/$ - see front matter
PII: S0010-0277(00)00157-8

1. Introduction

Attention imposes a limit on our capacity to process visual information, but there has been much debate recently concerning the correct *units* for characterizing this limitation. It was traditionally argued or assumed that attention restricts various types of visual processing to certain spatial areas of the visual field – for example, in 'spotlight' and 'zoom lens' models of visual attention (e.g. Eriksen & St. James, 1986; Posner, Snyder, & Davidson, 1980). It has recently been demonstrated, however, that there must also be an object-based component to visual attention, in which discrete objects are directly attended, and in which attentional limitations are characterized in terms of the number of objects which can be simultaneously selected. The excitement which this shift has generated is apparent in the recent proliferation of empirical demonstrations of object-based attention using many different paradigms in both normal and impaired observers (see Scholl, 2001, for a review).

The notion that visual attention can select visual objects has thus been well confirmed, and has engendered many new theories of visual attention. The surprising lacuna in all of this research, however, is that we do not know what sorts of stimuli can count as visual objects in the first place. (By 'object' here, we simply mean an independently attendable feature cluster; for discussion of the distinction between 'objects', 'groups', 'parts', etc., see Scholl (2001).) The research program of object-based attention has until now largely concentrated on evidence *that* attention can be allocated to discrete objects, rather than on what the objects of attention can be. (For recent exceptions, see Avraham (1999) and Watson and Kramer (1999).) The importance of this issue is clear: among the most crucial tasks in the study of any cognitive or perceptual process is to determine the nature of the fundamental units over which that process operates. It has become abundantly clear that visual attention can operate over objects, but we do not as yet know what qualifies as an object except in certain simple cases.[1]

[1] Of course, there is a wealth of evidence concerning the factors that mediate perceptual grouping, starting with the seminal demonstrations of the Gestalt psychologists. These investigations are distinct from our question, though, for two reasons. First, as will be clear below, several of the manipulations used in our experiments do not involve standard grouping principles (e.g. to take a simple case, a line is typically thought of as a single unit, and not as a 'group' of points). Second, it is not a foregone conclusion that the units of attention will obey standard principles of perceptual grouping. It may be, for example, that attention will automatically spread only within a subset of those perceptual groups that we can intentionally perceive. Another way to put this is that attention may automatically spread only within groups defined primarily by 'bottom-up' factors, but 'top-down' factors may additionally form groups which are perceived as such, but which don't readily constrain the automatic spread of attention. For a more complete discussion of the relation between perceptual grouping and objecthood, see Feldman (1999). For a different perspective on the relation between grouping and attention, see Driver, Davis, Russell, Turatto, and Freeman (2001).

The experiments reported here address this question in the context of the multiple object tracking (MOT) paradigm, in which subjects must track a number of independently and unpredictably moving identical items in a field of identical distractors. What types of spatially organized components can be independently tracked in this manner? In other words, what counts as an 'object' in this task?

1.1. Multiple object tracking (MOT)

Most studies of object-based visual attention have employed static stimuli, for example in the large literature spawned by seminal studies of spatial cueing (e.g. Egly, Driver, & Rafal, 1994) and divided attention (e.g. Duncan, 1984). Others have used dynamic displays, but observers have typically had to attend to only a single moving item (e.g. Kahneman, Treisman, & Gibbs, 1992; Tipper, Brehaut, & Driver, 1990; Tipper, Driver, & Weaver, 1991). We suspect, however, that an entirely different set of constraints on objecthood may come into play when observers must simultaneously attend to multiple feature clusters which are allowed to move about the visual field while attentional selection is being assessed. One such context is the MOT task used here.

Observers in the first MOT experiment (Pylyshyn & Storm, 1988) were initially presented with a display containing a number of small identical crosses. After a moment, a subset of these crosses were blinked several times to indicate their status as targets. As soon as this blinking ceased, all of the identical crosses began moving independently and unpredictably about the display. At various times during this motion a small probe appeared, and observers had to indicate whether the probe had occurred on a target cross, a distractor cross, or neither. Note that this task can only be done by tracking the individual target items throughout the motion phase, since they all have identical features. Subjects were able to perform this task with over 85% accuracy when tracking up to five targets (but not more) in a field containing five other identical distractors, and computer simulations revealed that this performance could not be explained in terms of a single attentional spotlight which cyclically visited all of the targets in sequence (see Pylyshyn & Storm, 1988, for details). Other studies have demonstrated in other ways that attention is truly 'split' between the items in a MOT task, rather than being 'spread' between them (Intriligator, 1997; though cf. Yantis, 1992). Both Intriligator (1997, Experiment 2) and Sears and Pylyshyn (2000), for instance, demonstrated that various benefits which accrue to tracked items and locations (such as speeding response times to detect luminance increments) held only for the targets themselves, and not for the space between them or for other items which happened to be located within the polygon bounded by the targets. This selection is dynamic, and can survive occlusion, but not other similar disruptions in spatiotemporal continuity (Scholl and Pylyshyn, 1999). Recent neuroimaging experiments, which controlled for passive viewing, eye movements, and discrete attentional shifts, have localized the processes involved in this sort of task to parts of both parietal and frontal cortex

(Culham et al., 1998; Culham, Cavanagh, & Kanwisher, 2001). In this paper, we will regard this MOT as a paradigmatically *attention-demanding* task.[2]

1.2. What is an object that a person may track it?

To investigate the nature of visual objects in the context of MOT, we ask what types of configural properties a scene component must have in order to be tracked. Observers attempt to track arbitrary collections of features, for example as when a target (to be tracked) and a distractor (to be disregarded) are drawn as opposite endpoints of a single line segment. Here, in order to perform the tracking task, subjects must separately select one endpoint of the line segment, keeping track of which is the target and which is the distractor. (As in any tracking task, targets and distractors move independently.) But if, as we hypothesize, such undifferentiated ends are not parsed as objects, then this manipulation should substantially impair tracking performance. Subjects will be unable to separately track each of the two endpoints, but rather will only be able to track whole lines, and will be forced to guess which end is actually a target. This manipulation, which we call *target merging*, is the basis of all the experiments reported below.

We start with a baseline tracking task in which subjects are asked to track four targets from a total of eight independently moving items, a task most subjects can perform at or above 85% accuracy. In each trial, the eight items (drawn as identical boxes) are initially shown in a static display. After 1 s, four of these items are highlighted by small blinking probes which appear and disappear from the items to indicate that they are targets. Then all eight begin moving independently and unpredictably about the display. After 10 s of such motion, the items stop moving, and the subject must use the mouse to indicate which four of the eight items were the targets.

In all of the target merging conditions reported below, we used the same set of item trajectories and target selections as in the above baseline condition: the only difference is in how the targets and distractors are drawn. Instead of drawing each target and each distractor as a separate object (e.g. a dot), we randomly paired each target with a distractor and then *merged* the pair in some way so that they would be perceived as parts of the same object – for example, by drawing a line between them, drawing a convex hull around them, etc. (see details and schematic drawings of the conditions below). Because all items move independently (as in the baseline tracking task) the new 'combined' objects which result from this target merging are not rigid, but rather shrink as the constituent target and distractor approach each other and elongate as they recede from each other. Because one end of each pair is a target,

[2] MOT is clearly attentionally demanding and effortful, leading most researchers to talk of MOT as an attentional process (Culham et al., 1998, 2001; He, Cavanagh, & Intriligator, 1997; Intriligator, 1997; Scholl, 2001; Treisman, 1993; Viswanathan & Mingolla, in press; Yantis, 1992). However, Pylyshyn (1989, 1994, 2001) has proposed that this task may involve several stages, and that the mechanism responsible for tracking the continuing identity of individual objects could itself be preattentive. Pylyshyn has hypothesized such a mechanism, called a visual index or 'FINST', that individuates objects and keeps track of their identity in a data-driven manner, despite changes in their properties or locations. We will not discuss this hypothesis here, though it is discussed at length in Pylyshyn (2001).

while the other end is a distractor, in order to perform the task subjects need to select and track just part of each pair. Again, the actual trajectories to be tracked are exactly the same as in the baseline condition. Hence, this paradigm tests the hypothesis that attention can only be allocated to distinct objects, and not simply arbitrary collections of features. Moreover, by varying the exact manner in which target and distractor locations are merged, we can test exactly what configural cues – connectedness, part structure, and other aspects of perceptual organization – make a part of the scene count as a 'distinct object' (see also Watson & Kramer, 1999).

Though the design of our experiment is completely between-subjects, the conditions we used were of two main types, which we now discuss in turn. Given the inherently dynamic nature of these displays, readers may wish to view animations of these conditions, several of which are available for viewing or downloading with a web browser at http://pantheon.yale.edu/~bs265/bjs-demos.html.

1.3. Group #1: undifferentiated parts of objects

1.3.1. Boxes

In our baseline condition, each item was simply drawn as a small individual outlined square, similar to the items used in previous MOT experiments (see Fig. 1a). Because the items in this experiment employed completely independent trajectories, these items were allowed to intersect during their motion, and when this happened one of the squares would occlude the other, with T-junctions at the occluding borders to indicate a depth relation. Viswanathan and Mingolla (in press) demonstrated that this method results in tracking performance which is comparable to earlier studies (a fact which we confirmed in pilot studies).

1.3.2. Lines

In this condition, we simply connected each pair of target/distractor locations with a single line, so that the entire display consisted of four lines (see Fig. 1b). One end of each line was then highlighted during the target designation phase, so that subjects had to track one end of each of the four lines throughout the motion period. We expected impaired performance in this condition: if attention must be allocated to objects as wholes, then it should be very difficult to confine attention to undifferentiated ends of objects, even if such loci move through the very same trajectories that can be tracked quite well when using individual points or boxes. (Note that when there *are* easily differentiated parts or surfaces of objects, it seems likely that attention could be selectively applied to those parts or surfaces. For discussion, see the 'dumbbells' condition below, and see also Hochberg and Peterson (1987) and Peterson and Gibson (1991) for cases where attention does seem to select well defined intra-object surfaces.)

1.3.3. 'Rubber bands' (with occlusion)

One concern with the 'lines' condition is that such stimuli involve much less of an overall enclosed area than do the 'boxes'. In order to control for the possible confounding effect of size, we replicated the effect of the 'lines' by essentially stretch-

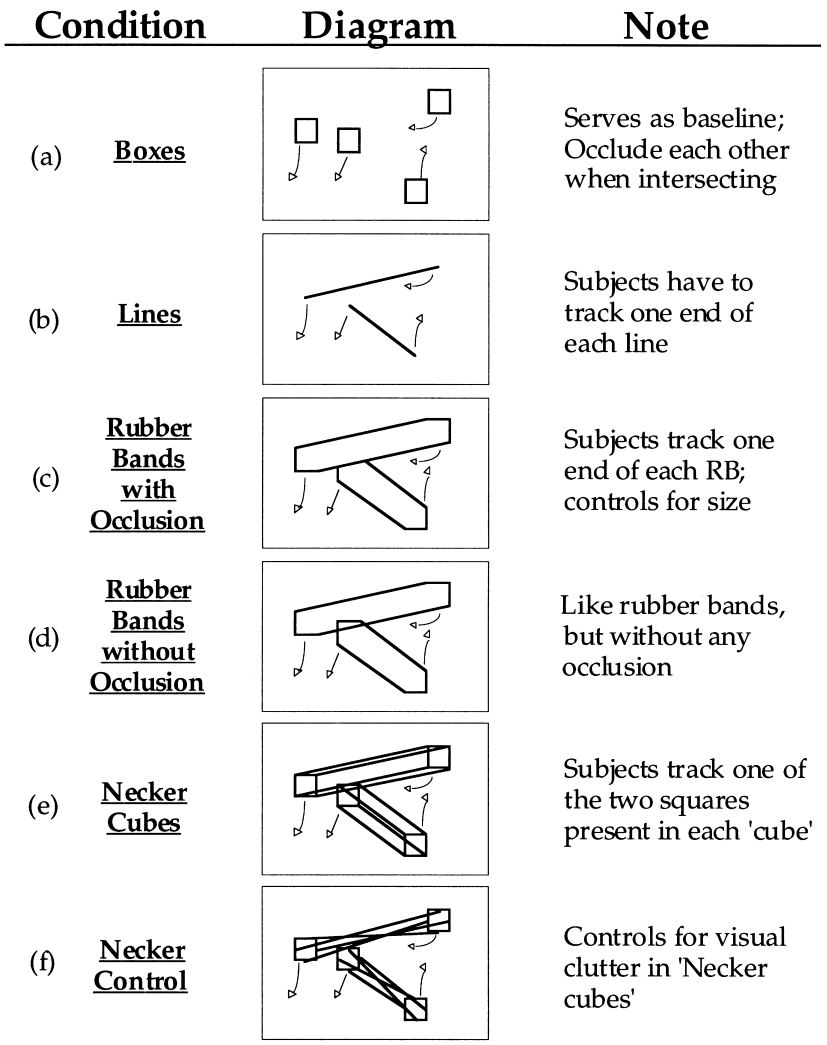

Fig. 1. A group of conditions designed to test whether undifferentiated parts or ends of objects can be independently selected by attention and tracked over time. See the text for discussion. The diagrams are not drawn to scale. Each diagram depicts only four items (two pairs), whereas the displays used in the experiments each involved eight items, four of which were targets and four of which were distractors. One 'end' of each item is a target; the other is a distractor. Dynamic animations of several of these conditions are available for viewing or downloading with a web browser at http://pantheon.yale.edu/~bs265/bjs-demos.html.

ing a single line 'rubber band' around each target/distractor pair of boxes, and then drawing only that rubber band (see Fig. 1c). This method again results in undifferentiated 'ends' of items which must be tracked, but such loci now occupy an equiva-

lent amount of space to the individual boxes. Since the items each move independently, however, it is possible for one such rubber band to occlude another.

1.3.4. 'Rubber bands' (without occlusion)

To control for the possible deleterious effects of such occlusion, we also tested this condition when the region bounded by a rubber band was not a 'solid' outlined shape, but just a set of lines, so that the full contours of each rubber band were always visible (see Fig. 1d).

1.3.5. Necker cubes

Another reason why boxes might be trackable but ends of rubber bands might not be is that the boxes provide a closed contour for selection and tracking (i.e. the square), whereas only two or three of a box's four sides are ever drawn in the rubber bands. To examine this, one could just draw both entire boxes in a target/distractor pair and then simply connect them with a line to form a 'dumbbell' (see below). The problem with this is that such a scheme might not necessarily involve a single object, but could rather involve two objects connected by a line! How can we draw both boxes in each pair, but still have those boxes subsumed into a configuration that is likely to be parsed as a single object? Our solution was to connect *each* vertex of a box to the corresponding vertex of its pair-mate, resulting in long thin 'Necker cubes' (see Fig. 1e). In this condition, each locus to be tracked still involves the enclosed contour of an entire square, as in the baseline 'boxes' condition, but these squares are subsumed into more global units.

1.3.6. Necker controls

The 'Necker cubes' condition also adds a substantial amount of visual clutter: each pair now involves four lines connecting the two squares. Any impairment of tracking for Necker cubes might thus simply reflect an intolerance for visual clutter. To control for this, we also employed a condition in which each pair of squares was connected by the same number of lines, but in a way which did support the perceptual interpretation of a single 'Necker cube': we instead attached the lines at the middles of each side of each square, and furthermore did not always connect equivalent sides (see Fig. 1f). We hypothesized that this condition would result in better performance than the Necker cubes, since, although it suffered from an equivalent degree of visual clutter, it might not be parsed as a single object. (In motion, this stimulus looked rather like two individual squares moving about independently, which happened to have some sticky substance stuck between them.)

1.4. Group #2: 'dumbbells'

The target merging conditions above were designed to examine whether undifferentiated parts or ends of objects can be independently selected and tracked over time. In three other conditions, we also explored how such 'attentional objecthood' is mediated by connectedness, part structure, and other configural properties. Connectedness and other aspects of perceptual grouping have been found to play

a role in object-based attention in earlier studies. Some of the relevant evidence concerning perceptual grouping involves the role of occlusion. Several studies have shown that mechanisms of object-based attention treat partially occluded objects as wholes, as they are perceptually grouped (e.g. Behrmann, Zemel, & Mozer, 1998; Moore, Yantis, & Vaughan, 1998), and that dynamic object representations can survive even moments of complete occlusion (e.g. Scholl & Pylyshyn, 1999; Tipper et al., 1990; Yantis, 1995). In addition, some neuropsychological studies suggest a more direct role of grouping (e.g. Boutsen & Humphreys, 2000; Driver, Baylis, Goodrich, & Rafal, 1994; Ward, Goodrich, & Driver, 1994). Driver et al. (1994), for example, had neglect patients report whether a small triangle had a gap in its contour, where this triangle was surrounded by other triangles such that it was perceptually grouped into a right-leaning or a left-leaning global figure. When the critical triangle was grouped into the left-leaning global figure, the gap was on the right side of this overall group; when it was perceptually grouped into the right-leaning figure, in contrast, the gap was on the left of the overall group. This manipulation greatly affected whether the patients perceived the gap, even though the critical triangle was always drawn identically. With regard to connectedness in particular, several recent studies have demonstrated that connected regions are often represented as single objects (e.g. Kramer & Watson, 1996; Van Lier & Wagemans, 1998; Watson & Kramer, 1999). The literature on perceptual grouping itself, however, has been largely silent on what configural properties make a scene component count as a distinct whole 'object' (as opposed to a contour, surface, 'unit', or other figural component at some lower level of the hierarchy; see Feldman, 1999, for discussion).

To explore the roles which such issues play in the ability to track multiple items, we employed three other types of 'dumbbell' conditions (see Fig. 2).

1.4.1. Dumbbells

In the simplest of these conditions, we just combined the 'boxes' and 'lines', so that a line was drawn between each pair of boxes (with the boxes occluding the lines, so that the lines began at the boxes' contours; see Fig. 2a). It is difficult to formulate a specific prediction for this condition. On the one hand, unlike lines and rubber bands, these stimuli now possess salient curvature minima indicating the ends to be tracked. There is a large body of research stemming from Hoffman and Richards (1984) which has demonstrated that such discontinuities are precisely where perceptual *parts* of objects are parsed, and indeed the parts in this context may be parsed as visual objects of their own (at a different hierarchical level), which might not predict impaired performance. On the other hand, each pair of items in a dumbbell is still physically connected, and we might expect this to lead to impaired performance, as we do with lines and rubber bands. Several neuropsychological studies have demonstrated that merely connecting two items with a thin line will greatly affect the percepts of both neglect patients (e.g. Behrmann & Tipper, 1994; Tipper and Behrmann, 1996) and of patients suffering from Balint syndrome, who will typically perceive only one of two unconnected discs, but yet will perceive an entire dumbbell (e.g. Humphreys & Riddoch, 1993; Luria, 1959).

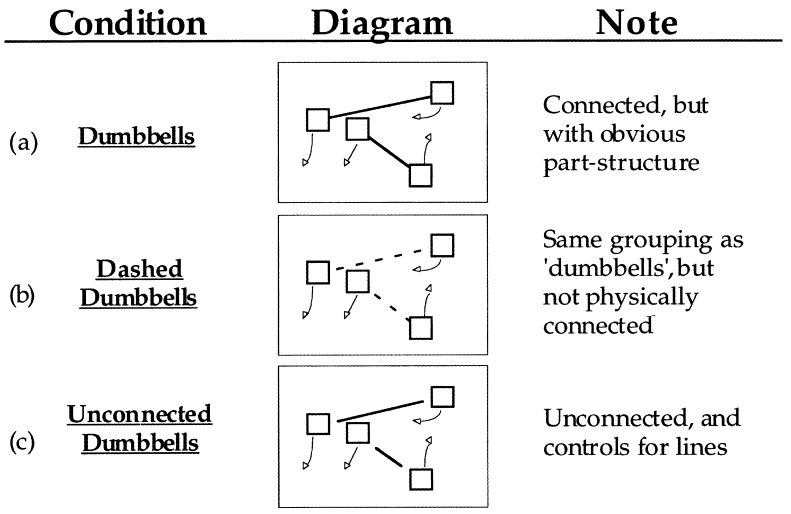

Fig. 2. A group of conditions (not to scale) designed to examine the roles of grouping, connectedness, and part structure. Each diagram depicts only four items (two pairs), whereas the displays used in the experiments each involved eight items, four of which were targets and four of which were distractors. One 'end' of each item is a target; the other is a distractor. Dynamic animations of several of these conditions are available for viewing or downloading with a web browser at http://pantheon.yale.edu/~bs265/bjs-demos.html.

1.4.2. Dashed dumbbells

In the basic dumbbells case, we might expect physical connectedness to impair tracking. In an effort to distinguish actual physical connectedness from more general perceptual grouping, we also tested a condition wherein the lines connecting the boxes were 'dashed' (see Fig. 2b), which disrupted the physical connectivity but not the grouping.

1.4.3. Unconnected dumbbells

In another condition, we used a solid connecting line which ended before it actually contacted a square, leaving a gap on either end the length of which was always 75% of the item size (see Fig. 2c). In addition to testing the role of connectedness, this condition also serves as another control for possible impairments arising simply from the existence of other lines in the display.

These conditions collectively embody an initial exploration of what types of feature clusters can be tracked in MOT. To ensure that any differences in performance are due to the stimulus manipulations and not to any other haphazard differences in trajectories, we used an entirely between-subjects design in which the same trajectories and target selections are used in all conditions – such that the *only* difference between the trials in each condition is the way in which the stimuli are drawn. In particular, each 'item' (or 'end' of an item pair) still moves on an independent trajectory, and on the *same* independent trajectory that it moves on in each of the other conditions.

2. Method

2.1. Participants

Eighty-one Rutgers University undergraduates, nine for each of the nine conditions, participated in one individual session either to fulfill an introductory psychology course requirement or to receive extra credit in another course. Each participant completed only one of the target merging conditions. Three subjects chose to terminate the experiment before completion, and were replaced. All subjects had normal or corrected-to-normal vision.

2.2. Apparatus

The tracking displays were presented on a monitor controlled by a Power Macintosh 6500 computer. Subjects were positioned with their heads in a chinrest 36.8 cm from the display monitor, the viewable extent of which subtended $45 \times 33.75°$. All displays were controlled by custom software written in the C programming language, using the VisionShell libraries of programming routines (Comtois, 1998).

2.3. Stimuli

Each trial employed four target items and four distractor items, drawn as described below. Initial item positions were generated randomly, with the constraint that each had to be at least 5.62° from the edges of the display monitor and at least 4.22° from each other. A 10 s animation sequence was generated for every trial to produce unpredictable trajectories for each item as follows. Items were each assigned initial random horizontal and vertical velocity vectors which could vary by single integer steps between -3 and 3 (with '0' indicating a stationary position with regard to that dimension), and which determined how fast an item moved in the specified direction. There was a 10% chance after each frame of motion that this value would be updated by a single step (i.e. ± 1) in a randomly chosen direction (with -3 and 3 always serving as the most extreme values possible). Each item was updated independently, resulting in completely independent and unpredictable trajectories. The resulting set of trajectories for a trial, along with randomly selected target items, were stored offline as 335 static frames to be presented for 30 ms each for a total of 10 s of motion. In the resulting motion, items could move a maximum of 0.21°/frame. Since frames were displayed for 30 ms each, the resulting item velocities were in the range from 0 to 7.02°/s, with an average velocity across all items and trials of 2.37°/s.

2.4. Drawing conditions

The individual lines comprising each item were all one pixel (0.07°) wide, except as noted below, and were clearly visible. In the *boxes* condition, each item was drawn independently as a square subtending 2.81°, centered on the item position. The squares were all randomly assigned to different depth planes, such that they would occlude each other when their trajectories intersected. In the *lines* condition,

each target was randomly paired with a distractor to result in four target/distractor pairs, and each pair was drawn as a single line connecting the center of each item. In the *rubber bands with occlusion* condition, each target/distractor pair was drawn as the smallest convex polygon encompassing both of the areas drawn as squares in the 'boxes' condition. Each of the four resulting polygons was randomly assigned to a separate depth plane, such that they would occlude each other when they intersected. The *rubber bands without occlusion* condition was identical to the 'rubber bands with occlusion' condition, except that items never occluded each other: each polygon was always drawn simply as a collection of six line segments (or possibly four line segments, when the items were momentarily horizontally or vertically aligned). In the *Necker cube* condition, the squares were drawn as in the 'boxes' condition, and in addition each vertex of each square was connected to the same vertex of its pair-mate square with a single line. The resulting figures appeared to be extended three-dimensional boxes, with their depth relation appearing bistable. There was no occlusion involved. The *Necker control* condition was identical to the 'Necker cube' condition, except in the organization of the extra lines: instead of connecting the corresponding vertices, the four lines were drawn from the midpoints of the squares' sides, connecting the left side of one square to the top side of the other, and similarly top to bottom, bottom to left, and right to right (see Fig. 1e). The *dumbbell* condition was implemented simply by drawing both the lines and the boxes, with the boxes occluding the lines such that the lines connected at the borders of the squares (see Fig. 2a). The *dashed dumbbell* condition was identical to the 'dumbbell' condition, except that the line was dashed, with the lengths of both the line segments and gaps always 0.7° each, regardless of the total line length. Finally, the *unconnected dumbbell* condition was identical to the 'dumbbell' condition, except that the lines always terminated 2.11° from the squares' borders.

2.5. Procedure and design

At the beginning of each trial, the eight items were displayed, drawn as described above. After 1 s, the four target items were highlighted with small flashing probes (disappearing and reappearing for 165 ms each on each of five flashes). The 10 s of item motion then ensued. After 10 s, all of the items stopped moving, and the subject had to indicate the four target items using the mouse. The fourth mouse-click caused the display to disappear, and the subject initiated the next trial with a keypress. Eye movements were not monitored, and no special instructions were given concerning fixation, since different fixation conditions have been found not to affect performance on this task.[3]

Forty sets of trajectories (along with target selections) were generated and stored

[3] Pylyshyn and Storm (1988) monitored and ensured fixation by discarding trials on which subjects made eye movements, and obtained qualitatively identical results to other investigators who either employed no special constraints or instructions concerning fixation – for example, Intriligator (1997) and Yantis (1992) – or else instructed subjects to maintain fixation but did not monitor eye movements – for example, Scholl and Pylyshyn (1999).

off-line. Nine different subjects were run on these 40 trials in each of the nine different conditions (for total of 81 subjects in total). Subjects first completed ten practice trials for which data were not collected, and then completed the 40 experimental trials in a randomized order (different for each subject). The entire experimental session took about 20 min.

3. Results

Tracking accuracy was recorded on each trial. Because there were always four targets, percent correct P was always 0, 25, 50, 75, or 100%. Mean P and standard errors for each condition averaged across the nine subjects per condition are shown

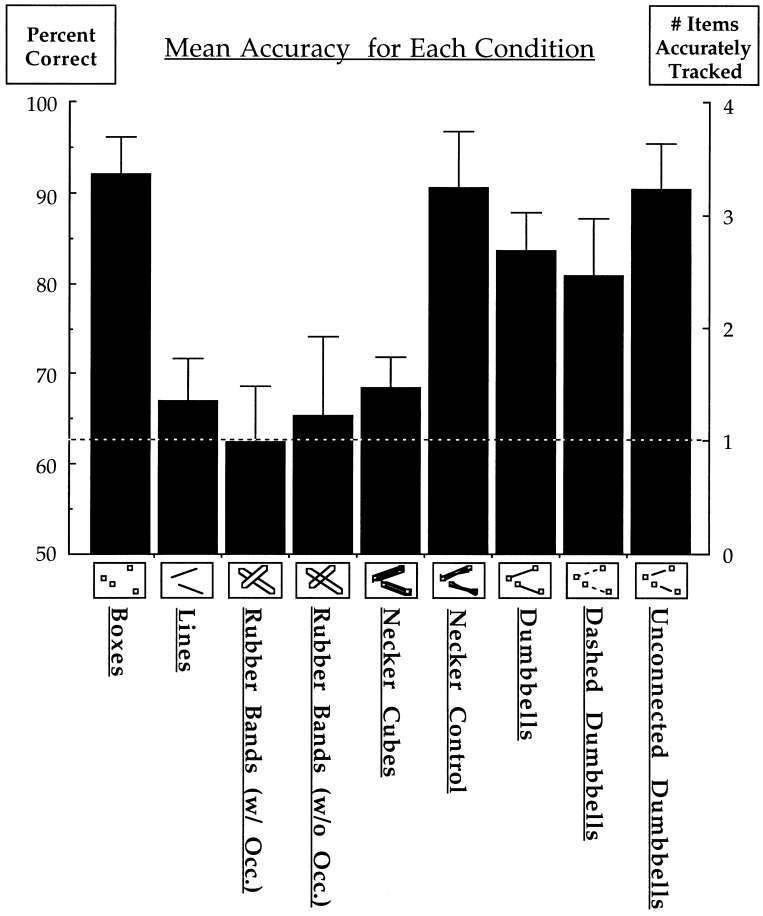

Fig. 3. Mean performance and standard errors in all conditions plotted in terms of both percent correct and the number of items successfully tracked (see Appendix A for details). Chance performance, as discussed in Section 3, is indicated by the dashed line.

in Fig. 3. It is also possible to translate P linearly into a measure m giving the *effective number of items tracked* by each subject in each condition – i.e. the number of items that had to be independently tracked in order to give rise to the observed percent correct P. (A derivation and justification of m is given in Appendix A.) One advantage of the alternate performance measure m is that it allows us to mark the performance level consistent with the subject's having only tracked one target throughout a trial. A value of $m = 1$ means that his or her capacity to divide attention among multiple objects has been effectively obliterated, which in the current case is equivalent to $P = 62.5\%$ (though note that with other numbers of targets $m = 1$ will correspond to different values of P). In Fig. 3, percent correct P is given on the left scale and m is given on the right scale, and both chance ($P = 50\%$) and single-item tracking performance ($m = 1$) levels are marked.

An analysis of variance on these accuracy data revealed a significant effect of the drawing condition ($F(8, 72) = 41.12$, $P < 0.05$). Additional planned comparisons indicated that:

(a) performance with 'boxes' was significantly better than performance with 'lines' ($t(16) = 12.11$, $P < 0.01$), 'Necker cubes' ($t(16) = 13.44$, $P < 0.01$), and 'rubber bands' both with occlusion ($t(16) = 12.04$, $P < 0.01$) and without occlusion ($t(16) = 8.29$, $P < 0.01$);

(b) performance on 'rubber bands' did not differ based on whether the rubber bands occluded each other or not ($t(16) = 0.82$, $P > 0.3$);

(c) performance in the 'Necker control' was significantly better than performance with 'Necker cubes' ($t(16) = 9.35$, $P < 0.01$), but did not differ from performance with 'boxes' ($t(16) = 0.59$, $P > 0.3$);

(d) performance with 'dumbbells' was significantly *worse* than with 'boxes' ($t(16) = 4.24$, $P < 0.01$), and also significantly *better* than with both 'lines' ($t(16) = 7.85$, $P < 0.01$), and with 'rubber bands with occlusion' ($t(16) = 8.46$, $P < 0.01$);

(e) performance with 'dumbbells' was significantly worse than performance with 'unconnected dumbbells' ($t(16) = 3.07$, $P < 0.01$), but did not differ from performance with 'dashed dumbbells' ($t(16) = 1.07$, $P > 0.2$); and finally

(f) performance with 'unconnected dumbbells' did not differ from performance with 'boxes' ($t(16) = 0.74$, $P > 0.3$).

4. Discussion

These results from target merging displays in MOT are, in the first instance, strong evidence for the object-based nature of tracking, since the different stimulus conditions engendered very different levels of performance, despite the fact that the particular targets that subjects were being asked to track were identical in every other way across the conditions, and in particular moved through exactly the same trajectories. (The 'multiple *object* tracking' task, in other words, does appear to be aptly named.) Performance was greatly impaired when subjects had to track ends of lines, rubber bands, or Necker cubes; apparently, such undifferentiated *ends* of such

stimuli are not treated as objects, since observers were unable to independently individuate, select, and track them. In these target merging conditions, each pair in its entirety seems to constitute a single visual object, but this fact only impairs performance, since each pair consists of both a target and a distractor. This impaired performance is especially striking given that there are other reasons to expect that subjects should actually do better in such conditions. For instance, subjects know that each pair consists of a target and distractor, and therefore they know that any item about which they are confident can be used to fix the correct target status of another item: 'I know that this one way down here wasn't one of the targets, so that means that I can just follow the line up … to this one, which must be a target.' This interpretation is supported by the fact that performance in the 'Necker control' condition (see Fig. 1f) did *not* differ from the baseline, even though it differed only minimally in its physical constitution from the 'Necker cube' condition. This result may be related to the fact that 'Necker control' pairs tended to look like two separate items with some gooey substance stuck between them. (This odd percept may also be related to the intriguing result that 'Necker control' performance was actually better than with 'dumbbells', despite the additional clutter.)[4]

Another advantage of target merging is methodological. Other studies of object-based attentional effects have relied on reliable but small differences in accuracy or in reaction times across conditions (see Duncan, 1984; Egly et al., 1994; and the many follow-up studies generated by each), with differences between conditions seldom perceptually apparent to observers. By contrast, our method results in large differences in accuracy and a phenomenologically salient effect (like other phenomenological demonstrations of perceptual grouping, but unlike many measures of object-based attention). These results confirm not only that attention *can* be object-based, but that in some cases it *must* be object-based (see Driver & Baylis, 1998; He & Nakayama, 1995). Attention involuntarily spreads to the entire rubber band, for instance, even though subjects are attempting to track only the end of the rubber band.

Note that this explanation in terms of objecthood is distinct from concerns about segmentability in general. One might argue that perhaps observers are worse on 'Necker cubes' than on 'Necker controls', for instance, not because of how such displays are parsed in terms of objects, but simply because the ends of the pairs that subjects must track are more easily segmented from the 'Necker control' display: in the 'Necker cube' display, in contrast, there are many other rectangles and parallel lines that might make the ends harder to segment.[5] If so, then the deleterious effects of the Necker cubes should occur regardless of their connection to the targets – i.e.

[4] This 'Necker cube' versus 'Necker control' comparison also calls into question an explanation in terms of perceived depth: since both of these conditions appear to some degree to involve rotation in depth, that fact cannot be responsible for the impaired performance found here in some conditions. In addition, other studies have shown that multiple objects can be tracked in depth quite easily, as would be expected if this type of attention were to be useful in the real world (Blaser, Pylyshyn, & Domini, 1999; Viswanathan & Mingolla, in press).

[5] Thanks to an anonymous reviewer for suggesting this interpretation, and the resulting control experiment.

the impairment should simply be a function of their presence in the display as a whole. To test this, we ran an additional control experiment wherein nine new observers (the same number as in the main experiment) tracked four boxes in a field of eight boxes. In the background of the displays, however, were four additional boxes, connected into either 'Necker cube' or 'Necker control' pairs. Observers viewed 30 trials of each type in separate blocks, the order of which was counterbalanced. Overall, performance was impaired in these conditions (85% combined compared to 92% for the 'boxes' alone in the main experiment; $t(16) = 2.82$, $P < 0.05$), which demonstrates that the added background items were salient enough to impair performance, probably due to increased crowding in the display as a whole. The extent of this impairment did not differ depending on whether 'Necker cubes' or 'Necker controls' were in the background, however: subjects were 86% accurate for displays with background 'Necker cubes', and 85% accurate for displays with background 'Necker controls' ($t(8) = 0.71$, $P > 0.4$). The fact that these two stimuli yielded very different levels of performance in the main experiment, but did not do so here (and were in fact in the wrong direction), suggests that our object-based results are distinct from general concerns about segmentation.[6]

In addition, this initial exploratory study has begun to examine some of the factors which mediate the degree to which various feature clusters can 'count' as objects for purposes of MOT. Connectedness, for example, appears to play a role, since performance was impaired with 'dumbbells' but not with 'unconnected dumbbells'. On the other hand, this result might reflect perceptual grouping rather than physical connectedness, since there was no difference between performance with 'dumbbells' and 'dashed dumbbells', nor between 'rubber bands' and 'occluded rubber bands'. Finally, however, the deleterious effects of perceptual grouping and connectedness appear to be attenuated by the presence of easily parsable object 'parts', since performance with 'dumbbells' was still significantly better than with 'rubber bands'. In other studies, we are now investigating more precisely the roles of the curvature minima which signal the existence of these object parts (see also Driver & Baylis, 1995; Hoffman & Singh, 1997; Watson & Kramer, 1999).

The conditions reported here are only an initial investigation of the types of properties which can mediate 'attentional objecthood', but this paradigm has proven

[6] Yantis (1992) suggested that MOT can be enhanced by imagining the targets as being grouped into a single virtual polygon (VP), and then tracking deformations of this polygon. He demonstrated that such grouping does indeed play a role in MOT by showing that performance was facilitated simply by informing subjects of this strategy, or by constraining the items' trajectories such that the polygon could never collapse upon itself. Perhaps, then, performance is impaired simply because such manipulations disrupt the formation of the VP. While we agree that performance in the basic MOT task can be improved by using this strategy (or indeed, by any grouping strategy, e.g. pairing items into virtual line segments), the improvement seems likely to be due to an improved error-recovery process when one item is lost: when items are being perceptually tracked as virtual groups, one can make an educated guess as to where a lost item 'should' be, given the overall contour of the virtual shape (see Sears & Pylyshyn, 2000). In addition, Scholl and Pylyshyn (1999) have shown that information which is local to each item (or 'vertex' in the VP strategy) does greatly impact tracking performance. In any case, the VP strategy cannot easily explain the particulars of our results: for example, the 'Necker cube' and 'Necker control' conditions should disrupt performance to an equal degree on the VP story, but they do not.

a useful way to explore these issues. First, MOT provides a way to check the generalizability of other object-based attention results, the majority of which have been collected using only a few experimental paradigms involving static displays (or dynamic displays in which there is still a single locus of attention). Second, there are also several intrinsic advantages to using MOT to study object-based attention: the results obtained in this paradigm consist of large perceptually salient differences in accuracy, unlike most earlier results, and it also seems likely that this paradigm will have more power to reveal subtle differences between different conditions, since the 'objects' in this paradigm must not only be parsed from the display, but must be maintained throughout the tracking period. The experiments reported here, and others using this paradigm, will help us to provide a missing link in the study of object-based attention, by revealing what types of spatially organized visual components can be independently attended.

Acknowledgements

For helpful conversation and/or comments on earlier drafts, we thank Erik Blaser, Susan Carey, Patrick Cavanagh, Jon Driver, Steve Franconeri, Peter Gerhardstein, Glyn Humphreys, James Intriligator, Allan Kugel, Alan Leslie, Ken Nakayama, Mary Peterson, Dan Simons, Joe Tanniru, Anne Treisman, and Steve Yantis. We also thank Steve Franconeri, Damien Henderson, and Joe Tanniru for assistance with data collection. Some of these results were presented at the 2000 meeting of the Association for Research in Vision and Ophthalmology. B.J.S. was supported by NIH F32-MH12483-01, Z.W.P. was supported by NIH 1R01-MH60924, and J.F. was supported by NSF SBR-9875175.

Appendix A. Derivation of *m*, the effective number of items tracked

For simplicity, assume a display with n targets and n distractors ($2n$ objects total). A variation of the following with a target proportion other than one-half can easily be derived.

Assume the following idealized strategy: track m objects, and guess randomly on the others. We assume that the observer knows that half the items are targets and thus for each unknown item guesses 'target' with a probability of 0.5.

Assuming this strategy, the observer will correctly track m of the n targets, and guess correctly on half of the remaining $n - m$ targets. This yields a proportion correct P of

$$P = \frac{m}{n} + \frac{1}{2}\left(1 - \frac{m}{n}\right) = \frac{1}{2}\left(\frac{m}{n} + 1\right)$$

Solving for m, we have

$$m = n(2P - 1)$$

We interpret the performance score m as the 'effective number of items tracked'

(because it is the number of items which, when tracked correctly, gives rise to the given proportion performance P). This measure is advantageous because it allows percent correct scores from trials with different numbers of targets to be uniformly combined to estimate observers' tracking capacity.

Of course, m is based on an idealized conception of subjects' strategy, and must be interpreted with some caution. On a given trial a subject may track two items for a while, and perhaps lose track of one, thus in the end scoring $m = 1$ while having tracked more than one item for part of the trial. Note, however, that once lost a target cannot be 'picked up again' at better than a chance rate, so $m = 1$ performance does suggest that *at least* one object, but no more than one, was tracked from the beginning of the trial to the end. On the other hand, a capacity to track $m > n$ trials will still yield $P = 100\%$; hence, the estimate of tracking capacity derived from a given trial is necessarily capped at n.

References

Avrahami, J. (1999). Objects of attention, objects of perception. *Perception & Psychophysics*, *61*, 1604–1612.

Behrmann, M., & Tipper, S. (1994). Object-based visual attention: evidence from unilateral neglect. In C. Umilta, & M. Moscovitch (Eds.), *Attention and performance. Conscious and nonconscious processing and cognitive functioning* (Vol. 15. pp. 351–375). Cambridge, MA: MIT Press.

Behrmann, M., Zemel, R., & Mozer, M. (1998). Object-based attention and occlusion: evidence from normal participants and a computational model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 1011–1036.

Blaser, E., Pylyshyn, Z., & Domini, F. (1999). Measuring attention during 3D multi-element tracking (abstract). *Investigative Ophthalmology & Visual Science*, *40* (4), 552.

Boutsen, L., & Humphreys, G. (2000). Axis-based grouping reduces visual extinction. *Neuropsychologia*, *38*, 896–905.

Comtois, R. (1998). *VisionShell [Software libraries]*. Cambridge, MA: author.

Culham, J. C., Brandt, S., Cavanagh, P., Kanwisher, N. G., Dale, A. M., & Tootell, R. B. H. (1998). Cortical fMRI activation produced by attentive tracking of moving targets. *Journal of Neurophysiology*, *80*, 2657–2670.

Culham, J. C., Cavanagh, P., & Kanwisher, N. (2001). Attention response functions of the human brain measured with fMRI. Manuscript submitted for publication.

Driver, J., & Baylis, G. (1995). One-sided edge assignment in vision: II. Part decomposition, shape description, and attention to objects. *Current Directions in Psychological Science*, *4*, 201–206.

Driver, J., & Baylis, G. (1998). Attention and visual object segmentation. In R. Parasuraman (Ed.), *The attentive brain* (pp. 299–325). Cambridge, MA: MIT Press.

Driver, J., Baylis, G., Goodrich, S., & Rafal, R. (1994). Axis-based neglect of visual shapes. *Neuropsychologia*, *32*, 1353–1365.

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention, and phenomenal visual objects. *Cognition*, this issue, *80*, 61–95.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*, 501–517.

Egly, R., Driver, J., & Rafal, R. (1994). Shifting visual attention between objects and locations: evidence for normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, *123*, 161–177.

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

Feldman, J. (1999). The role of objects in perceptual grouping. *Acta Psychologica*, *102*, 137–163.

He, S., Cavanagh, P., & Intriligator, J. (1997). Attentional resolution. *Trends in Cognitive Sciences*, *1*, 115–121.

He, Z. J., & Nakayama, K. (1995). Visual attention to surfaces in 3-D space. *Proceedings of the National Academy of Sciences USA*, *92*, 11155–11159.

Hochberg, J., & Peterson, M. A. (1987). Piecemeal perception and cognitive components in object perception: perceptually coupled responses to moving objects. *Journal of Experimental Psychology: General*, *116*, 370–380.

Hoffman, D., & Richards, W. (1984). Parts of recognition. *Cognition*, *18*, 65–96.

Hoffman, D., & Singh, M. (1997). Salience of visual parts. *Cognition*, *69*, 29–78.

Humphreys, G. W., & Riddoch, M. J. (1993). Interactions between object and space systems revealed through neuropsychology. In D. Meyer & S. Kornblum, *Attention and performance* (Vol. XIV, pp. 183–218). Cambridge, MA: MIT Press.

Intriligator, J. M. (1997). *The spatial resolution of visual attention*. Unpublished doctoral dissertation, Harvard University, Cambridge, MA.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: object-specific integration of information. *Cognitive Psychology*, *24*, 174–219.

Kramer, A., & Watson, S. (1996). Object-based visual selection and the principle of uniform connectedness. In A. Kramer, M. Coles, & G. Logan (Eds.), *Converging operations in the study of visual selective attention* (pp. 395–414). Washington, DC: APA Press.

Luria, A. R. (1959). Disorders of 'simultaneous perception' in a case of bilateral occipito-parietal brain injury. *Brain*, *83*, 437–449.

Moore, C., Yantis, S., & Vaughan, B. (1998). Object-based visual selection: evidence from perceptual completion. *Psychological Science*, *9*, 104–110.

Peterson, M. A., & Gibson, B. S. (1991). Directing spatial attention within an object: altering the functional equivalence of shape descriptions. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 170–182.

Posner, M. I., Snyder, C. R. R., & Davidson, B. J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, *109*, 160–174.

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. (1994). Some primitive mechanisms of spatial attention. *Cognition*, *50*, 363–384.

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

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

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

Scholl, B. J., & Pylyshyn, Z. W. (1998). Tracking multiple items through occlusion: clues to visual objecthood. *Cognitive Psychology*, *38*, 259–290.

Sears, C. R., & Pylyshyn, Z. W. (2000). Multiple object tracking and attentional processing. *Canadian Journal of Experimental Psychology*, *54*, 1–14.

Tipper, S., & Behrmann, M. (1996). Object-centered not scene-based visual neglect. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 1261–1278.

Tipper, S., Brehaut, J., & Driver, J. (1990). Selection of moving and static objects for the control of spatially directed action. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 492–504.

Tipper, S. P., Driver, J., & Weaver, B. (1991). Object-centered inhibition of return of visual attention. *Quarterly Journal of Experimental Psychology*, *43A*, 289–298.

Treisman, A. (1993). The perception of features and objects. In A. Baddeley, & L. Weiskrantz (Eds.), *Attention: selection, awareness, and control* (pp. 5–35). Oxford: Clarendon Press.

Van Lier, R., & Wagemans, J. (1998). Effects of physical connectivity on the representational unity of multi-part configurations. *Cognition*, *69*, B1–B9.

Viswanathan, L., & Mingolla, E. (in press). Attention in depth: disparity and occlusion cues facilitate multi-element visual tracking. *Perception*.

Ward, R., Goodrich, S., & Driver, J. (1994). Grouping reduces visual extinction: neuropsychological evidence for weight-linkage in visual selection. *Visual Cognition*, *1*, 101–129.

Watson, S., & Kramer, A. (1999). Object-based visual selective attention and perceptual organization. *Perception & Psychophysics*, *61*, 31–49.

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

Yantis, S. (1995). Perceived continuity of occluded visual objects. *Psychological Science*, *6*, 182–186.