Complex, Dynamic Scene Perception:

Effects of Attentional Set on Perceiving Single and Multiple Event Types

Thomas Sanocki & Noah Sulman

Department of Psychology,

University of South Florida

Word count, main text: 11, 895 (previous: 12,811)

Address:

E-mail: Sanocki@usf.edu

Thomas Sanocki

Psychology, PCD 4118

University of South Florida

Tampa, FL 33624

Abstract

Three experiments measured the efficiency of monitoring complex scenes composed of changing objects, or events. All events lasted about 4 sec but, in a given block of trials, could be of a single type (single-task) or multiple types (multi-task, with a total of 4 event types). Overall accuracy of detecting target events amidst distractors was higher for single event types relative to multiple types. Multiple event types were processed reasonably well when each event type was restricted to its own region, and much worse when even types were mixed in location. In most task conditions, observers reached an optimal level of performance (optimal attentional set). After one target was identified, performance for other targets dropped markedly and then recovered to optimal levels. However, set was not optimized when task locations were intermixed. The results support the idea that attentional set determines the efficiency of event perception in complex scenes. While single-event set was most efficient, there can be a reasonably efficient set for multiple event types.

How does the efficiency of perceiving objects in scenes vary with the number of observer goals, or attentional sets? The present experiments addressed this question with complex, dynamic scenes and objects — animated displays composed of many simple changing objects, or events. Set was manipulated by presenting either a single event type or multiple event types. Observers searched for target events amidst distractor events, and overall perceptual efficiency was indexed by the hit rate for target events. The attentional set hypothesis predicts that perceptual efficiency should be higher when observers can be set for a single event type than when observers must be set for multiple event types, requiring multiple visual sets. This research was motivated by several literatures, including those on top-down processing and attention, and scene perception.

Conceptual Background

Top-down Processing

The topics of everyday scene perception and top-down influences have been central to cognitive approaches to vision (e.g., Broadbent, 1977, Bruner, 1957, McClellend & Rumelhart, 1981, Neisser, 1967, Schyns, Goldstone, & Thibaut, 1998). Everyday scene perception is important ecologically because it is something people often do (see Neisser, 1976). Top-down effects are significant because they are cognitive mechanisms that can have strong influences on perceptual processing. However, the relative strength of top-down processing remains to be determined, because research has not adequately explored potentially relevant factors such as scene complexity, process preparation, and event time scale. Combining these considerations motivates an experimental paradigm that can be termed complex, dynamic scene perception.

Attentional Set Approach to Scene Perception

The core idea in this top-down approach is that everyday perception is an active process whose success depends on the state (set) of the perceiver (e.g., Most, Scholl, Clifford, & Simons, 2005, Postman & Bruner, 1949, Ekstrand & Wickens, 1954, Culbert, 1958). State is a deep concept involving multiple factors, including expertise (e.g., Goldstone, Braithwaite, & Byrge, in press), and motivation and beliefs (e.g., Balcetis & Dunning, 2006). When these factors are constant, the preparation of visual processing components can have a strong influence on perception (e.g., Postman & Bruner, 1949, Walther & Fei-Fei, 2007). We manipulated the number of event types observers must

attend to. When there is a single event type with one target stimulus, the observer can prepare for the stimulus in an optimal manner. This increases the likelihood that observers will become conscious of instances of the target, relative to unexpected or unattended stimuli (e.g., Mack, 2003, Most et al., 2005, Simons & Chabris, 2004). In this context, attentional set can be thought of as control settings describing target visual features or properties (e.g., Folk, Remington, & Wright, 1994, Leber, Kawahara, & Gabari, 2009, Most et al., 2005). Single-event conditions should require a single description, for the one event type, whereas multi-event conditions should require multiple descriptions, or one more general description. Perceptual efficiency should be reduced in multi-event conditions because of the more general descriptions used.

Attentional set is important in scene perception because of the nature of scenes and scene perception. We argue that scenes and scene perception defy narrow definition; at their most essential levels, they are characterized by their *ranges* — by the variation between scene types and between observer behaviors (cf. Henderson & Hollingworth, 2003, Greene & Oliva, 2009a, Torralba et al., 2009, Tsotsos, 2001). Scenes differ greatly on spatial factors such as scale and complexity, and on temporal factors as well. Observer activities may differ even more.

Given the wide range of scenes and observer activities, the mechanisms of scene perception are likely to be complex and powerful while also being adaptable. If so, change in mechanisms due to scene or task should be non-trivial. This is a fundamental implication of the attentional set hypothesis (e.g., Most et al., 2005, Neisser, 1976). And it corresponds to the well-studied cost of task-switching — reduced performance caused by the reconfiguration of processes when task changes (e.g., Monsell, 2003). Costs of task switching have been examined in detail within a large literature using mostly simple stimulus situations (e.g., Monsell, 2003, Van Loy, Liefooghe, & Vandierendonck, 2010). Current work on task switching is focusing on the specific processes that underlie these costs, and a number of hypotheses remain viable (e.g., Kiesel et al., 2010, Van Loy et al., 2010, Vandierendonck, Liefooghe, & Verbruggen, 2010).

Large effects of attentional set have also been found with complex stimuli, including real world scenes and events, such as the appearance of a gorilla (e.g., Most et al., 2005,

Simons & Chabris, 1979, Walther & Fei-Fei, 2007). Typically in these studies, there is a critical stimulus (e.g., gorilla) that can be perceived when observers expect it or have adopted a general set. However, in the main conditions, observers are induced to be set for another complex task; the gorilla is unexpected and requires a change of visual task. In these conditions, the gorilla is often not noticed nor consciously perceived (e.g., Macdonald & Lavie, 2008, Most et al., 2005, Simons & Chabris, 1979, White & Davies, 2008; see also Mack & Rock, 1998). In fact, unexpected stimuli do not cause attentional capture (e.g., Folk, Remington, & Wright, 1994, Lien, Ruthruff, & Johnston, 2010).

While lack of noticing is consistent with the attentional set hypothesis, the effects are somewhat limited in nature. Because the critical events are unexpected, they occur only once in most experiments.[1] Also, the responses are often indirect (e.g., answers to post-hoc questions), and can be influenced by expectation or memory. In the present experiments, we sought to extend attentional set effects to frequent, expected stimuli. The goal was to measure repeated missings of expected gorillas.

The Importance of Stimulus and Task Complexity

Stimulus complexity has major effects on perceptual processing,[2] and top-down processing becomes increasingly important as the stimulus situation gets more complex. For example, scene categorization becomes slower and more difficult when varying foreground objects are present (Walker, Stafford, & Davis, 2008). From a computational viewpoint, the challenge of computing interpretations from scenes in spite of their complexity is a significant, perhaps defining feature of general scene perception (Tsotsos, 1990, 2001).

Perceptual processing also varies in complexity. Much scene perception research has focussed on processes such as scene-categorization, which are *convergent* in nature — a stimulus image is processed to arrive at a single consensus label, which denotes (for

---

[1] Misperception has been studied for events with frequencies over one, but the events are peripherally presented (e.g., Macdonald & Lavie, 2008) or the measure is indirect (Folk et al., 1994).

[2] For example, effects of display size (e.g., Duncan & Humphreys, 1989, Schneider & Shiffrin, 1977), perceptual load and dilution (e.g., Lavie, Hirst, & Fockert, 2004; Tsal & Benoni, 2010), and crowding (e.g., Pelli et al., 2007).

example) the category, or the presence of an object in the scene. Such paradigms do not capture some major complexities of everyday scene perception. First, objects in live scenes change over time, creating events. Events can have many complexities, including time scale and history (see next section). Second, convergent paradigms require a single interpretation within a single task. In contrast, natural scenes vary in how many interpretations they support and how easy the interpretations are. Natural observers may engage in a variety of perceptual tasks within a period of time. These factors greatly compound the complexity of everyday scene perception, and are largely unexplored in the context of scene perception research.[3]

Costs of changing perceptual task may be more likely to appear in complex situations. Our scenes were composed of dynamic objects and involved divergent visual tasks. However, we avoided stimulus or response ambiguity because we did not want switching costs to be influenced by uncertainty. Uncertainty about stimuli can inflate response times to particular features, objects, and experimental situations (e.g., Sanocki & Oden, 1991).

Events and Their Time Scales

Most of the stimulus scenes in scene perception research consist of objects, surfaces, and their layout. However, natural scenes are live, and objects within them often change over time, producing events. Events compound the complexity of scene perception, by adding temporal dimensions of variation. There is a growing literature on event perception, and it involves a wide variety of event types, at a variety of temporal scales (e.g., Zacks & Tversky, 2001). The temporal structure of events plays a key role in the organization of mental life, and expectations regarding the temporal structure and scale of perceptual events are available in early in life (Ballargeon, 2004, Hespos & Baillargeon, 2001, Tversky, Zacks, & Hard, 2008). Tåemporal scale, like display complexity, may be a critical variable for determining the nature of perceptual processing.

---

[3] Other significant complexities of scene perception include the challenges of segmenting objects and events from background information, and the challenges of coordinating the different scene views obtained by an active observer across eye movements and changes of position.

In the present case, we decided to begin by examining a single time scale that, on intuitive grounds, seems to be near optimum for human scene perception -- the time scale of several seconds. As noted, many human events take place over seconds, and people like to watch events that last seconds. Our events were simple animated objects; research indicates that animated objects can be life-like and meaningful (e.g., Michotte, 1955/1991, Gao, Newman, & Scholl, 2009). The events were designed to be simple, predictable, and of a similar time scale, so that set effects would not depend on uncertainty about event structure or differences in times scales between events.

Surveillance as an Intense Scene Perception Process

One additional motivation concerns scene-surveillance. Security is critically important for communities, both in present times and throughout most human history. Modern surveillance methods present large amounts of information to observers, on multiple screens, or display cells. The widespread use of multiple displays suggests that this is an effective way of presenting scene information, although there are limits (e.g., Sulman, Sanocki, Goldgof, & Kasturi, 2012). In the present experiments, we presented the events on a grid on four display cells (Figure 1). Information about performance can be useful for the practical problem of providing security, as well as for perceptual theory.

Summary

Complex, dynamic scenes are especially interesting because their rich, many-dimensional structure seems to invite top-down influences. With this belief in mind, we developed the following experiments.

Introduction to the Experiments

The empirical goal was to measure overall perceptual efficiency as a function of the number of event types. In single-task conditions, a single event-type occurred in each of the 4 display cells, whereas there were 4 event types in the multi-tasking condition (Figure 1). Each trial involved a 60 sec stream during which 144 events occurred asynchronously. The event types required the same decision and response rules while varying markedly on visual factors. The events had different visual properties to minimize confusion between tasks. In the task switching literature, each event type would be termed *univalent*, because the event tokens had only one meaning. Two event types were

termed ventral events because their token instances remained stationary and the critical dimension was a ventral property (color and shape). The other two event types were termed dorsal events because their tokens moved and the critical dimension was location or type of motion.

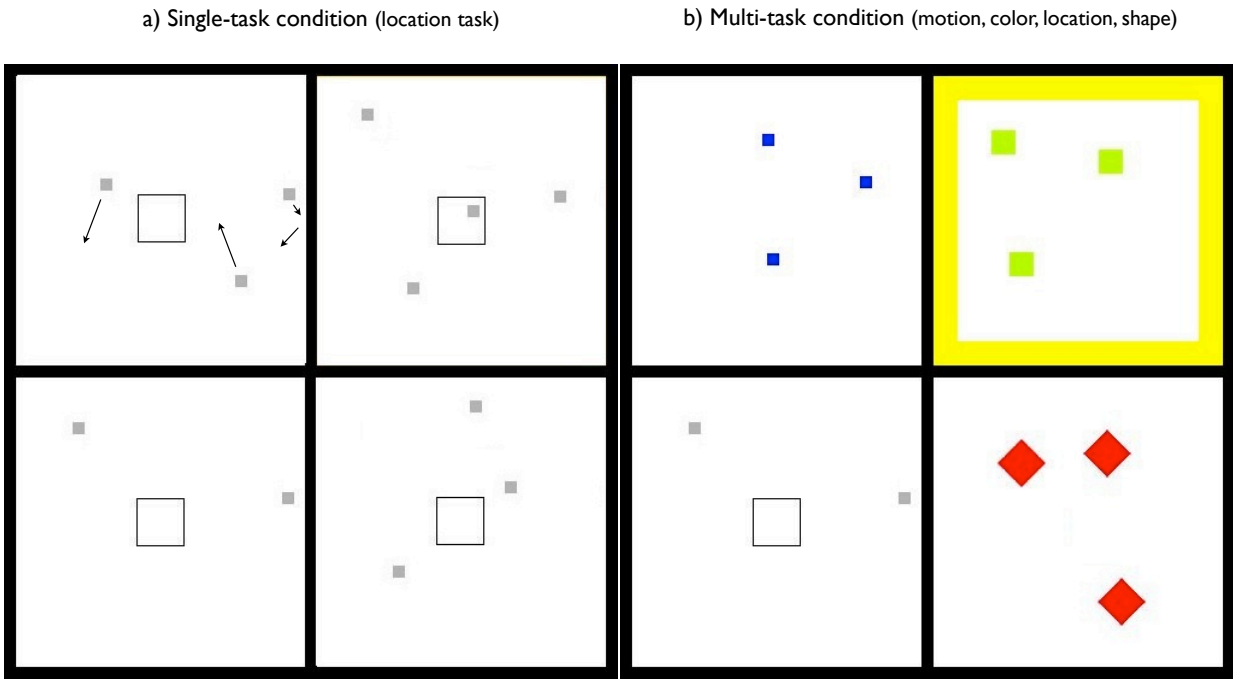

Figure 1. Main conditions Experiment 1. a) Single-task condition, with possible pathways shown for one cell. b) Multi-task condition, with tasks shown above (clockwise from upper left).

Event Lifetimes

Each object token started its lifetime as a distractor and then proceeded to change (increase and then decrease) along the single critical feature dimension, during a 4 sec lifetime. If the increase was large enough, the token became a target for a period of time. Figure 2 shows schematic event-trajectories — lifetimes for targets (solid lines) and distractors (dashed lines). Observers responded to targets with a keypress, while ignoring distractors. For example, in the dorsal location task, the tokens moved in linear pathways (usually oblique) up or down the display cell (Figure 1). The critical feature was location — a token became a target if it entered a central box. The target-distractor distinction was obvious when observers attended to the tokens.

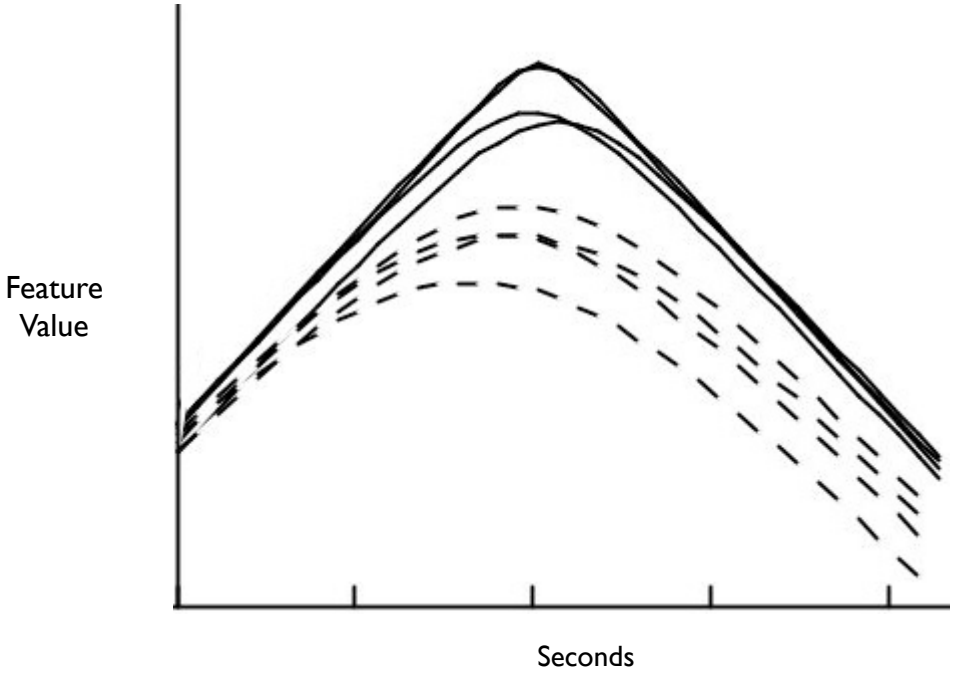

Figure 2. Schematic event trajectories for targets (solid lines) and distractors (dashed lines).

Although each event had a predictable trajectory, observers had to fixate near a token to process it optimally. This meant that observers had to move their eyes and attention from cell to cell throughout a trial.

Interference Between Targets Close in Time?

An additional empirical issue was how the detection of one event might influence detection of other events occurring close in time. The timing of targets was distributed randomly throughout trials, producing inter-target-intervals (ITI's) ranging from zero to beyond 10 sec. Temporally close targets may interfere with each other, perhaps in a manner roughly analogous to the attentional blink, which is obtained with very rapid stimulus presentations (Chun & Potter, 1995, Raymond, Shapiro, & Arnell, 1992). Our goal was to collect basic data on inter-target relations, which may be useful for understanding attentional set.

General Method

Observers participated in individual sessions that began with training on each of the four event types. Then single-task blocks were run, followed by multi-task blocks, and then a repeat of the single-task blocks, to permit assessment of learning across the test blocks.

Stimulus Displays

The stimulus for each trial was a 60 sec stream composed of 4 sec events, or event-tokens. The 144 tokens on a trial were divided among 4 adjacent display cells (Figure 1). Eight of the 144 tokens turned into targets. Observers were instructed to respond to targets only, by pressing a key. In Experiments 1 and 2, the key location corresponded to display cell location. There were four token types, one token type for each task. The display was 14 x 14 cm, with a visual angle of 29.5 deg at the viewing distance of (approximately) 54 cm. Figure 1 reproduces example displays in accurate relative scale. An iMac computer with standard keypad was used; the experiment was programmed in RealBasic 2008.

A token's lifetime began as it appeared and then lasted about 4 sec, when it disappeared. During most of a trial, there were 12 tokens in the display, at varying stages of their lifetime. When a given token reached the end of its lifetime, it disappeared and then a new token appeared in the same display cell, 1 sec later. Thus, consecutive tokens within a cell formed *token-streams*, and the stream spanned the 60 sec trial. Targets were constrained to appear only in lifetimes 2 - 11 of a token stream; the first and last lifetimes were buffer periods with no targets. Asynchrony was produced by delaying the start of token streams (the appearance of the first token) relative to each other.

Location task. In this task, grey square tokens traveled upward or downward at the same rate, in a linear path at varying angles (see Figure 1a). Proximity to a central critical box was the critical dimension; targets traveled through the box and distractors missed the box. The possible trajectories were predefined off-line; a token's trajectory was chosen from those produced by crossing 55 evenly spaced starting locations at the top or bottom of displays cell with 30 motion angles (between -45° and +45° relative to upright). When a token hit the wall, it was deflected with a reflection of its path. The 1650 possible trajectories were sorted into those that directed tokens through the central target region (target events) and those that did not (distractor events). Close cases, where tokens grazed the central region by .25° or less, were removed. This left 299 target paths and 1106 distractor trajectories. Tokens pathways were randomly selected from these (with no replacement within trials).

Motion task. Square blue tokens moved across the screen left to right, with varying degrees of vertical wobble (perturbation up or down; see Figure 1b, upper left). Targets had more extreme wobble, appearing “drunk”. The horizontal screen dimension was divided into eight sections (octants). For each token, at every new octant the vertical motion component reversed sign (up versus down) and the slope changed. The vertical slope increased with each octant through the 6th, and then decreased for the last two. The slopes were greater for targets than for distractors, producing the appearance of a stronger wobble, especially in octants 5 and 6.

Color task. Square color tokens appeared and remained at random locations; their color changed through color space across the lifetime, beginning as green-yellow and then changing toward pure yellow and then back to green-yellow. Targets changed more toward yellow, to match the yellow border of the color-task display cells (see Figure 2b). The color changes consisted of 40 gradual steps, beginning at a red-green-blue value of 127, 255, 0, on the RGB scale (0 = black, 255 = white). The red component was increased for the first 20 steps of a token lifetime and then decreased for the last 20 steps. Distractors and targets differed in the range of change in the red value; distractors changed in value between 3 and 6 at each step, whereas targets changed between 6 and 8.5; only targets reached the defining value of [255, 255, 0], which matched the border. Targets varied somewhat in how long they remained at the defining value.

Shape task. Red diamond-shaped tokens appeared and remained at random locations; their shape changed in concavity across the lifetime. Tokens began as fat diamonds (slightly convex sides) and changed to become star-like (very concave sides), with targets defined by increased concavity. Concavity decreased for the first 20 steps of a lifetime and increased for the last 20. The shape-pathways were chosen from 20 distractor and 20 target pathways generated before the experiment with varying sizes of shape-steps. As in the other tasks, only targets came close to and crossed the maximum value.

Main test procedure

There were four different single-task conditions, one for each task/event type. At the start of a trial, the 12 token-streams for that trial were assigned to display cells, with each display cell being randomly assigned a density of streams of 2, 3, 3, or 4 (for a total of 12). For new tokens, entry points and locations were random within ranges inside the cell. The eight target tokens on a trial were randomly assigned to token streams and lifetimes within streams (excluding the two buffer periods). This method distributed targets evenly throughout both space and time. Overlap between target lifetimes, which could produce competition between targets, ranged between 0 (same peak of critical feature function) to beyond 10 sec. The multi-task conditions used the same constraints as the single-task conditions except that the 12 token streams were divided among the 4 event types (2, 3, 3, or 4 per event, for a total of 12, randomly assigned at the beginning of each trial). For a given observer, each event type was assigned to a display cell that remained constant throughout the entire multi-tasking phase (grouped multi-tasking in Experiments 1 and 3).

Observers were instructed to press the spacebar to initiate a trial. They were then free to respond throughout the trial, by pressing the appropriate key (see below). A trial ended as the last token lifetime ended. There was no feedback during this main test phase. Observers learned about the task and target frequencies during the initial training period. During testing, responses were interpreted as *Hits* if they indicated the correct cell (Experiments 1 and 2) and if they occurred within a 4-sec window centered on the peak of the target's critical feature variation. We used this window to allow anticipatory responses, because event trajectories were predictable. In general, a target token's critical property first diverged from the distractors at about 1 sec into its lifetime (e.g., Figure 1a). The peak was reached between 2 and 3 sec into the lifetime. The 4-sec response window allowed anticipatory responses and continued out to about 3 sec after the first divergence of the critical property. If a second response occurred within the window but no new target had been presented, it was scored as a *False Alarm*.

Training

Sessions began with a training phase for each task. In training, observers monitored a single display cell for 60-sec trials, with 2 - 4 token streams (randomly determined) and 8 targets in the cell. Responses were scored as during the main test phase. However,

for each task, auditory feedback was provided during the first 4 trials (an incorrect or correct tone for each response). Then feedback was withheld for 2 more trials during which performance was measured. Performance was high for most observers (hit rate above 95%, false alarm (FA) rate < 2%). Observers who did not perform well on each task during this non-feedback phase were removed from analyses; the criterion for inclusion was a hit rate above 75% and false alarm rate below 10%.

Order of conditions

The order of the 4 tasks during training and the single-task conditions was determined randomly for each observer and maintained throughout the entire session. The test conditions followed the complete training phase. There was a single-task condition with each event type, followed by the multi-task condition, and then a repeat of the single-task conditions. This design (summarized in Table 1) allowed us to measure overall changes in performance during the session by comparing the initial and late single-task levels (a and c in Table 1), while also providing a balanced contrast of single- and multi-task performance (b versus a + c). In Experiments 1 and 2, the first single-task phase consisted of (a) for each event type, 4 practice trials followed by 4 test trials. The multi-task phase consisted of (b) 4 practice trials followed by 16 test trials. The second single-task phase consisted of (c) for each event type, one practice trial followed by 4 test trials.

Table 1. Session timeline in Experiments 1 and 2. Tasks are numbered by order 1 - 4; event types were randomly assigned to order for each observer.

TIME ——— > (~ 60 min total)			
Training (one cell)	Test Phase (four display cells)		
Learn-Test for Task: 1, 2, 3, then 4	(a) Single task, 4 cell: 1, 2, 3, then 4	(b) Multi-task, 4 cell: 1 - 4 one per cell	(c) Single task, 4 cell: 1, 2, 3, then 4

Experiment 1

The purpose of Experiment 1 was to measure the possible cost of switching between event types during a trial. We contrasted performance with single event types (e.g., Figure 2a) with multi-task performance, for 4 event types, with one event type in each cell (Figure 2b). In each condition, there were 8 target tokens and 136 distractor tokens. In

the multi-task conditions, there were 2 target-tokens and an average of 34 distractor-tokens for each task. If scene perception is most efficient when a single attentional set can be used throughout a trial, then performance should be highest in single-event (single task conditions). When the attentional set must be continuously changed for multiple events (multiple task condition), performance should be markedly reduced.

Method

Observers were instructed to respond only to targets, by pressing a key that corresponded spatially to the 4 cells (on the standard keypad; “4” for top left, “5” for top right, “1” for bottom left, “2” for bottom right). The response was a Hit only if the cell was correct as well as timing (see General Method). Fifteen college students (9 females) participated in exchange for course credit, from the University of South Florida. The data for an additional 2 subjects were not analyzed because they did not pass the training criteria.

Results

Hit rates provide a good measure of overall efficiency because false alarm rates were low in the test conditions (< 1% most conditions). Hit rates and sensitivity and bias values, are reported in Table 2. The main statistical analyses were planned comparisons between task conditions and factorial Analysis of Variance (ANOVA).

The hit rate was high at the end of training with one-cell displays and a single task (96.5%). When stimulus complexity was increased during the test phase to 4 cells with a single task, performance was reduced to an average of 78.4%, $F(1,14) = 7.80$, $p < .001$. The reduction suggests that stimulus complexity strained the observers’ resources in these single-task conditions.

When event types varied in the multi-tasking condition, performance was further decreased, to 64.3%. This 14.1% cost was highly reliable (*SE* of cost = 3.0%; $F(1,14) = 22.0$, $p < .001$; $\eta_p^2 = .61$). More concretely, subjects detected an average of 6.3 of 8 targets in the single task conditions, and 1.2 fewer targets in the multitasking condition. Thus, there was a cost for switching between multiple event types in this experiment.

Table 2. Sensitivity results in main conditions of each experiment.

	Hits (%)	FA's (%)	*d'*	*C*
Experiment 1 Single-event	78.4	0.8	3.53	0.87
Multi-event	64.3	0.9	2.97	1.09
Experiment 2 Single-event	79.3	0.8	3.50	0.81
Multi-event	45.2	2.0	2.17	1.21
Experiment 3 Single-event	85.5	0.3	4.07	0.81
Grouped Multi-Event	64.9	1.0	2.75	0.97
Distributed Multi-Event	53.1	1.0	2.48	1.17

Consistency across event type. Table 3 shows hit rates broken down by event type (task). There was the main effect of task condition reported above, but no main effect of event type (F[1,14] = 1.69, $p > .10$). The interaction of event type and task condition was reliable, $F(3,42) = 6.03$, $p < .01$; $\eta_p^2 = .30$. For 3 of the event types, performance was higher in the single-task conditions than in the multi-task condition. The exception is the location event, where there was a small advantage in multi-task conditions in this experiment. The lack of cost for the location event may reflect a confound in the location event — there was only one critical location in the multi-tasking condition (in the one cell for that event type) but four in the single-tasking conditions (one for each cell; see Figure 2b). Additional evidence on this issue is provided by the following experiments. Tentatively, while noting a confound for the location task, we conclude that the multi-tasking costs were generally consistent across tasks.

Table 3. Hit rates for each event type and task condition in Experiment 1.

	Color	Shape	Motion	Location
Single-event	0.78	0.85	0.79	0.73
Multi-event	0.63	0.65	0.54	0.75
Cost	0.15	0.19	0.25	-0.03

Effects of Inter-Target-Interval (ITI). Do target events that occur close in time interfere with each other? We divided up the data by ITI, after first separating out the data for initial (first) targets on each trial. For subsequent targets, the temporal interval after the previous target ranged from 0 (two target events with peak critical features values at the same time) to 10 sec and beyond. The data up to 10 sec were divided into one sec ITI windows and are shown in Figure 3, labeled by the longest ITI (e.g., the first interval includes ITI's of 0 to 1.0 sec). Shorter ITI's contained more trials — the 1 sec window contained 22%% of the trials, and frequency gradually decreased to (e.g.) 6% at 6 sec and 3% at 10 sec. The 10 ITI's shown captured 89% of the trials with non-initial targets.

If there was interference, we expected it to occur at the shorter ITI's. Performance may then return to asymptotic or baseline levels as ITI increased. There was in fact a decrement at the two shorter ITI's, as can be seen in Figure 3. In an ANOVA with task condition and ITI as factors, there was the main effect of task condition reported above, and a main effect of ITI ($F(9,135) = 3.01$, $p < .01$; $\eta_p^2 = .17$). The interaction of these variables was not reliable ($F[9,135]<1$). To measure the magnitude of decrements at short ITI's, we used as baseline the longer 8 ITI's (ITI's 3 - 10 sec). At the shortest ITI, the decrement was 15.1%, $t(15) = 5.36$, $p < .001$. At the 2 sec ITI, the decrement was 6.2%, $t(15) = 2.14$, $p = .049$. Thus, there was a decrement for temporally close targets. This decrement may be analogous to the attentional blink in research on temporal attention. However, the present decrement was much longer (over 2 sec) than in research with rapidly presented stimuli, where the effect typically ends after about 0.5 sec.

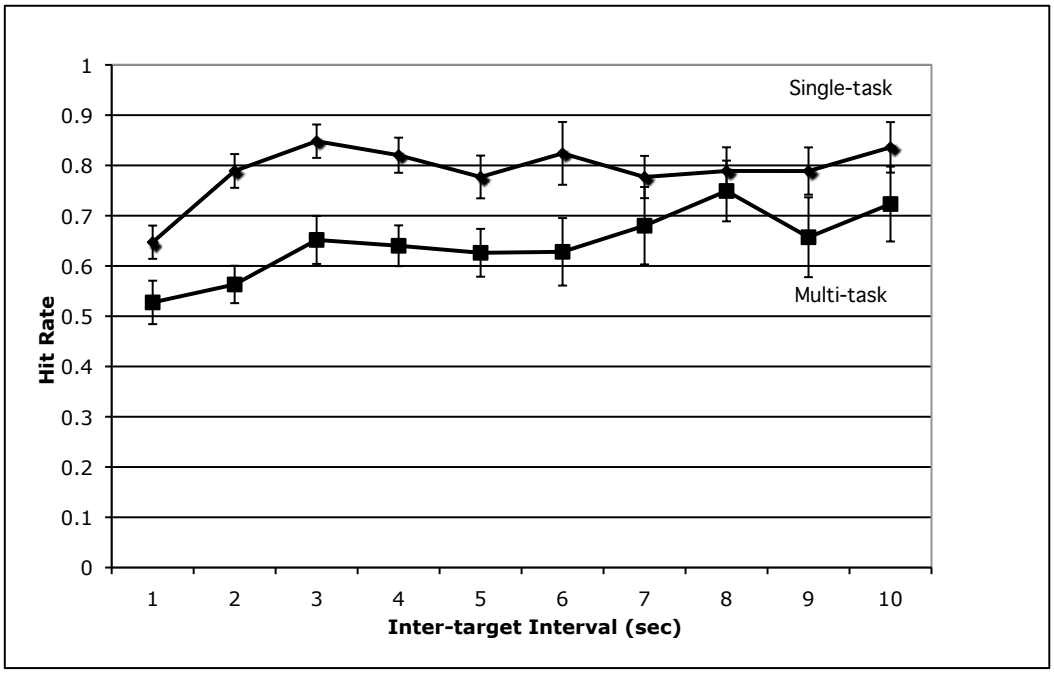

Figure 3. Proportion hits for each ITI window in Experiment 1, with standard errors.

Learning effects during the session. How much did performance change during the test session? The single-task blocks were run before and after the multi-task block (Table 1), to bear on this issue. Performance stayed constant across the early and late single-task blocks, averaging 78.6% and 78.3%, respectively (t<1). Thus, after observers learned the tasks during training, performance was generally constant across the experiment.

Discussion

The results indicate that there was a highly reliable cost of multi-events — observers were 14% less efficient when monitoring 4 types of events than when monitoring single events. The results are consistent with the attentional set hypothesis and the claim that switching between events would be costly. These costs appear to be perceptual rather than response-related because the decision and response rules were the same across task conditions.

The results also included large interference effects for targets occurring close together in time, in a manner roughly analogous to the attentional blink. However, the present

deficit lasted over 2 sec. Once the deficit is over, there appears to be an asymptotic period from 3 to 10 sec. This may represent an optimized set in that condition. We discuss these effects in more detail at the end of the experiments, because combining data across experiments provides a more systematic body of evidence.

Although the 14% multi-tasking cost was reliable and reasonably in size, we suggest that the magnitude is not as large as one might expect if a single attentional set was *necessary* for complex scene perception. In the multi-event conditions, many changes in set are required on each trial, because observers had to moves eye and attention throughout the display. Yet, the cost of perceiving multiple events was not large; it was not a missed gorilla (Simons & Chabris, 1994). One could say that bottom-up processing driven by the four types of event tokens was somewhat efficient.

Experiment 2

Experiment 1 produced the multi-tasking cost predicted by the attentional set hypothesis, but our tentative conclusion was that multi-event perception was fairly good in the present situation. Observers appear to adopt a reasonably efficient set for multi-event perception (MEP). The set may involve strategies for efficient time sharing between temporally overlapping perceptual processes. Such time sharing is likely in natural environments when events have time scales of seconds. The original attentional set hypothesis was based on the idea that there is a single set. However, the hypothesis may need to be modified to include the idea of an attentional set for MEP that is less efficient but still generally effective.

GIven the fairly efficient MEP observed in Experiment 1, we changed perspectives and considered the basis of MEP. What types of factors might support reasonably efficient MEP in complex scenes? One possibly general and important factor is spatial organization of task. Spatial organization means that events of the same type are grouped or located in the same region. This general factor appears to be essential in many domains, and to be embedded in our cultural thinking about scenes. When spaces are designed, tasks or purposes are assigned to specific regions (e.g., rooms in homes, regions in urban neighborhoods or factories or campsites). For example, urban design devotes separate lanes to vehicles and to pedestrians, and separate spaces for sitting and for

playing. Mixing functions across locations could produce perceptual chaos. In scene perception research, the learned organization of locations is a new and relevant topic (e.g., Torralba, Oliva, Castelhano, & Henderson, 2006). In the task switching literature, switching costs for simple tasks can be eliminated or reduced by the use of location as a task cue (Arbuthnott & Woodward, 2002; Mayr & Bryck, 2007).

Does reasonably efficient MEP depend on the spatial organization of the tasks? We reduced spatial organization in the multi-task condition of Experiment 2. In the new, *distributed* multi-tasking condition, the tokens for different tasks were distributed throughout the display. Figure 4 shows the old and new multi-task displays. Note that spatial organization can be viewed as a bottom-up factor that can influence attentional set. The hypothesis was that spatial organization would be important for MEP, and that multi-tasking costs would be greater in this experiment than in Experiment 1.

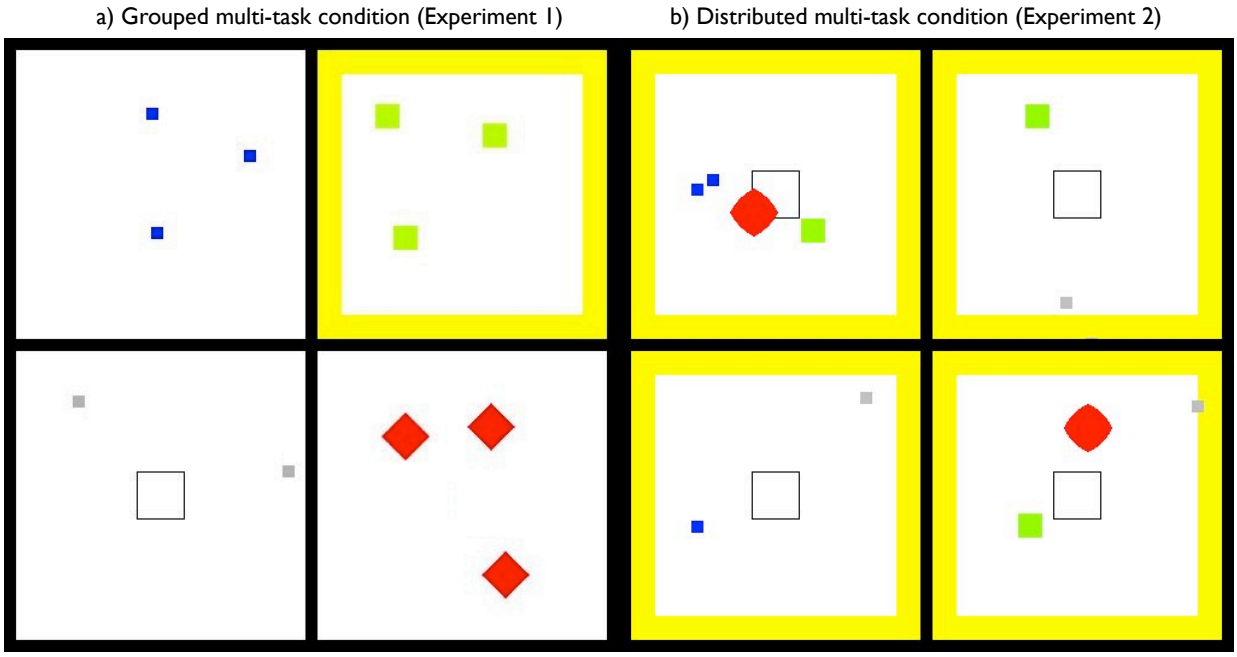

Figure 4. Muli-task conditions from Experiments 1 and 2.

Method

The method was similar to Experiment 1, with the exception of the distributing of task tokens in the multi-tasking displays. Whereas each token-stream was assigned to its task-cell before each trial in Experiment 1, each token-stream was randomly assigned

to any one of the four cells in Experiment 2, with the same constraints on cell-density as previously. For example, as in Figure 4b, one stream of shape-tokens (e.g., the diamond-like shape in the upper left) was assigned to that cell, and each token in that stream appeared in a random position of the cell, throughout the trial. Displays cells still contained the same number of token streams (2, 3, 3, or 4) as previously. Also, as can be seen in the Figure 4, task reference information (the critical square, and the color border) was provided in each display cell (as in single-task conditions for the two tasks). Fifteen college students (12 females) participated in exchange for course credit. The data for an additional 2 subjects were not analyzed because they did not pass the training criteria.

Results

The hit rate was high at the end of training with one-cell displays and a single task (96.0%). When stimulus complexity was increased during testing to 4 cells with a single task, performance was reduced to an average of 79.3%, $F(1,14) = 9.40$, $p < .001$. When observers had to monitor four events in the new multitasking displays, the hit rate was reduced to 45.2%. This is an absolute cost of 34.1% (*SE* of cost = 1.8%; $F(1,14) = 338.7$, $p < .001$, $\eta_p^2 = .96$). In more concrete terms, subjects went from identifying most targets on single-task trails (6.3 of 8) to less than half (3.6 of 8) on multi-task trials. This appears to be a fairly profound cost. Compared to Experiment 1, the multi-task cost was 20.0% greater in absolute size ($F(1,28) = 31.787$, $p < .001$, $\eta_p^2 = .53$). Thus, reducing spatial organization by distributing the tasks across the four cells made the perception of multiple events much more difficult. The results suggest that in complex scene perception, the challenges of switching between events are especially large when events are not spatially organized by type.

Consistency across event type. Table 4 shows hit rates broken down by event type (task). There was the main effect of multi-tasking reported above, and no main effect of event type ($F<1$). The interaction of multi-tasking and event did not approach reliability, $F[3,42] = 1.32$, $p > .20$. Single-task performance was higher than multi-task performance for all 4 event types. Note that in this experiment, there were four critical locations for the location event, in both the single-tasking and multi-tasking conditions (see Figure 3), removing the confound in Experiment 1. This appears to erase most of the interac-

tion with event type found that experiment. Thus, the multi-tasking cost was strong in the present experiment, across all event types. In fact, for each event type the costs were greater in magnitude than in Experiment 1.

Table 4. Hit rates for each event type and task condition in Experiment 2.

	Color	Shape	Motion	Location
Single-event	0.84	0.81	0.78	0.74
Multi-event	0.46	0.46	0.41	0.47
Cost	0.38	0.35	0.37	0.27

Effects of Inter-Target-Interval (ITI). The data were divided into 1 sec ITI windows as previously, capturing 89% of the post-initial target events, and are shown in Figure 5. In this experiment, the ANOVA produced the main effect of task condition reported above. There was no main effect of ITI, $F(9,126) < 1$, but there was an interaction of ITI and task condition, $F(9,126) = 2.34$, $p = .02$, $\eta_p^2 = .16$. In the single-task conditions, the main effect of ITI approached reliability, $F(9,126) = 1.93$, $p = .05$, $\eta_p^2 = .121$. The decrement relative to baseline was 14.1% at 1 sec, $t(14) = 6.20$, $p < .001$, and 11.8% at 2 sec, $t(14) = 2.80$, $p = .01$. Performance appears to asymptote after that, from 3 - 10 sec (or perhaps 5 - 10 sec).

In contrast, in the multi-tasking condition the effect of ITI was not reliable, $F(9,126) = 1.41$, $p > .10$. The ITI function is fairly noisy, and performance tended to be (unreliably) better at two shortest ITI's than at baseline (3.0% and 9.1% advantages, respectively; p's $> .10$). Thus, the ITI function for distributed grouping was different from that for single-tasking, and different from both task conditions in Experiment 1. The difference in the ITI functions for the two task conditions may relate to whether an attentional set can be re-established after a target is identified. The combined data will provide stronger evidence on this issue.

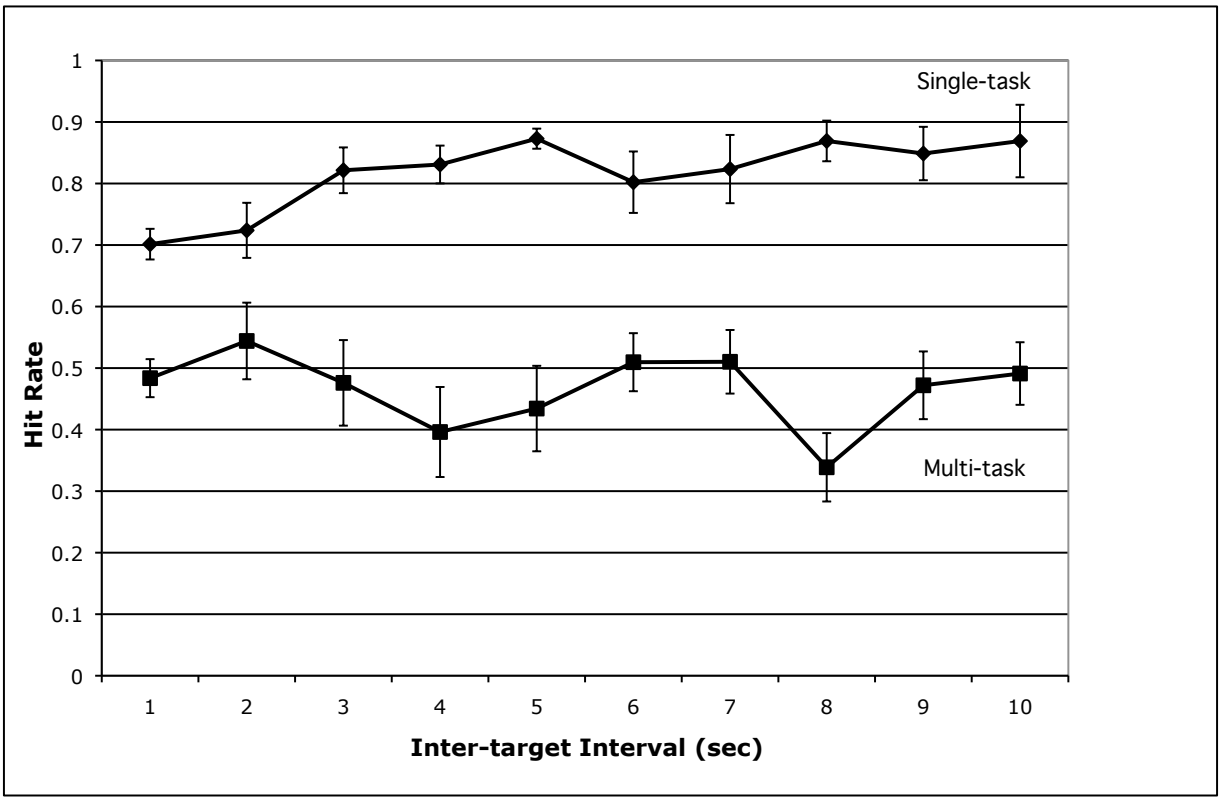

Figure 5. Proportion hits for each ITI window in Experiment 2, with standard errors.

Learning effects during the session. Performance was generally constant from the first (early) set of single-task blocks to the final set, averaging 80.5% and 78.0%, respectively, (t[14] = 1.33, $p > .20$). Thus, single-tasking performance was generally constant across the experiment.

Control Study Examining Perceptibility of Tokens

We would like to interpret the fairly profound costs for distributed multi-tasking as a cost of task-switching. However, an alternative explanation stems from possible differences in the perceptibility of tokens per se, in the distributed displays. Performance could be low in the new distributed condition because of the tokens are difficult to see, rather than because of task switching. Perhaps the mixture of different token types within a cell creates a busier display that reduces greatly perceptibility of the tokens themselves.

To test this explanation, we designed a control study to check the perceptibility of tokens. To separate token perceptibility from MEP, we used selective attention instructions

with the distributed multi-tasking displays. The design was the similar to Experiment 2, aside from the multi-task displays and instructions. During training with individual tasks, 14 of 16 new observers passed the criteria. The single-tasking condition was the same as previously (including instructions), and the average performance level was 73.3%. In the multi-tasking condition, the distributed displays (Figure 4B) were used, with two modifications. First, the task was a selective attention task; observers were instructed to detect targets of only one type at a time. There were 4 separate blocks, one for each event type. Second, to equate target frequencies with the main experiment, there were 8 targets of each task type instead of 2, and 6 fewer distractor tokens of each type. This made the overall task (detect 8 tokens) the same as previously. The 8 targets could appear in any of the 4 display cells, so observers had to monitor the entire display.

If the tokens themselves were difficult to perceive, because of the busy nature of the multi-tasking displays, then the tokens should also be difficult to detect in this selective attention condition. The overall hit rate in this selective attention condition was 88.3%, and performance was high with each task (color 92.8%, shape: 91.2%, motion: 82.6%, location: 86.7%). False alarm rates were less than 1%. Thus, the targets were not difficult to see in the distributed multi-tasking displays. In fact, performance was better than each of the single-task conditions in this experiment (which averaged 73.3%, as noted), and better than the single task conditions of the main experiment (Table 4). The high performance levels in this selective attention condition may occur because only the task-relevant tokens needed to be processed under the selective attention instruction. If observers can effectively selectively attend to one event type throughout the multi-task display, the effective number of distractor tokens is reduced to 28. Most importantly, the results imply that the tokens and the targets of each event type are perceptible.

This control study was not designed to provide a comprehensive assessment of token perceptibility across different conditions. Instead, it was designed to check whether the tokens were perceptible in the distributed multi-tasking conditions. The high level of performance indicates that they were. The large decrements when observers multi-tasked in the main experiment can not be explained by difficulties perceiving tokens.

Discussion

Experiment 2 produced large multi-tasking costs that were consistent with the idea that spatial organization of task is critical for reasonably efficient MEP. When tasks were no longer organized by cell in the multi-tasking conditions, performance was considerably lower than previously, falling to a hit rate of less than 46%. Further, there was no evidence of recovery at short ITI's in the distributed multi-task condition after a target had been identified. From an attentional set perspective, this suggests that an efficient set could not be re-instated after target detection.

Overall, Experiments 1 and 2 suggest that there are 3 levels of efficiency in the single-tasking and multi-tasking conditions. Optimal efficiency is obtained when there is a single task and observers can use a single attentional set. Efficiency is reduced when observers must change set continuously to handle 4 event types (Experiments 1 and 2). However, if the attentional set involves dedicated regions for each event type (Experiment 1), a reasonably efficient set for MEP can be maintained. When event types are distributed throughout space (Experiment 2), levels of perceptual efficiency are markedly lower, and a reasonably efficient set for MEP cannot be established. Interestingly, spatial organization may be a primarily bottom-up factor because it is an aspect of stimulus organization. This factor interacts with the observer's intention to successfully multi-task.

Experiment 3

We interpreted the large multi-tasking decrement in Experiment 2 in terms of the importance of spatial organization of the visual tasks for MEP. However, the response was also spatial in nature — target display cell was mapped onto a spatially corresponding key. Does the nature of the response contribute to the spatial organization effect? Could response complexity influence control processes in the present situation? To begin examining these questions, we used a simple, non-spatial response in Experiment 3 — observers pressed the spacebar whenever a target occurred, in any cell. Also, to more directly compare two multi-tasking conditions, we included both the distributed and the grouped multi-tasking displays.

Method

There were 3 main task conditions arranged in blocks: Single-tasking, grouped multi-tasking, and distributed multi-tasking. Each session began with training for each task with one display cell, followed by the four cell test conditions: single-tasking, and then the two multi-tasking conditions (in order counterbalanced across observers), and then the single-tasking condition again. Training involved, for each task, 3 trials with feedback and 3 trials without feedback. The first single-task condition involved 1 practice trial and 1 test trial for each event type. Each multi-task condition involved 4 practice trials followed by 12 test trials. The final single-task condition involved 2 test trials for each event type. The observers were instructed to press the spacebar once whenever they detected a target. A response was scored as a hit if it was the first response to occur within the temporal response window; location could not be scored. Data from twenty two observers (11 in each counterbalance group; 14 females in total) were analyzed; data for an additional 3 observers were not included because they failed to meet training criteria.

Results

The hit rate was high at the end of training with one-cell displays and a single task (96.6%). When stimulus complexity was increased to 4 cells but a single task, performance was reduced to an average of 85.5%, $F(1,21) = 34.26$, $p < .001$. This is a higher level of single-task performance than in the previous experiments, possibly because of the simplified response method.

When event types changed but in a spatially organized manner (grouped multi-tasking), performance was decreased 20.6% relative to the single-task conditions (to 64.3%; *SE* of cost = 1.8%; $F[1,21] = 127.49$, $p < .001$). Performance was reduced 11.7% more in the distributed multi-tasking condition relative to grouped multi-tasking, to a level of 53.2% (*SE* of cost = 1.8%; $F [1,21] = 41.41$, $p < .001$). Thus, even with a response rule that has no spatial component, there are multi-tasking costs, and the costs is especially high when the tasks are distributed throughout the display. In more concrete terms, the number of targets detected out of 8 ranged from 6.1 in single-task conditions, to 5.2 in grouped multi-task conditions, to 4.2 in distributed multi-tasking conditions.

Consistency across event type. Table 5 shows hit rates broken down by event type (task). The main effect of task condition was reliable, $F(2,42) = 160.09$, $p < .001$, $\eta_p^2 = 0.88$, as was the main effect of event type, $F(3,63) = 39.91$, $p < .001$, $\eta_p^2 = 0.65$. Also reliable was the interaction of task condition and event type, $F(6,126) = 10.48$, $p < .001$, $\eta_p^2 = 0.33$. Both the main effect of event type and the interaction can be attributed to the motion event, where performance was low in the two multi-tasking conditions. In this experiment and in the previous experiments, performance for the motion event tended to be lowest of the single-event conditions, and costs tended to be higher. The costs were especially high in the present experiment. However, there were substantial costs for all four event types.

Table 5. Hit rates for each task and condition in Experiment 3.

	Color	Shape	Motion	Location
Single-event	0.88	0.85	0.83	0.86
Multi-event Grouped	0.72	0.75	0.41	0.71
Multi-event Distributed	0.62	0.58	0.35	0.57
Costs (S-G)	0.15	0.10	0.42	0.15
Costs (G-D)	0.10	0.17	0.06	0.13

Effects of Inter-Target-Interval (ITI). The ITI data for 1 sec windows are shown in Figure 6 (capturing 91% of the post-initial target events). In this experiment, there was a main effect of ITI, $F(9,189) = 52.96$, $p < .001$; $\eta_p^2 = .72$, and statistically, it was consistent across task condition, ($F(18,378) = 1.28$, $p < .10$). The overall decrement relative to baseline was 26.1% at 1 sec, $t(21) = 11.07$, $p < .001$, and 11.4% at 2 sec, $t(21) = 4.44$, $p < .001$. Thus, there was a substantial decrement when targets were close together in time in this experiment. Further analyses were conducted with order-group as a factor (observers receiving grouped multi-tasking first versus those receiving distributed multi-tasking first). However, none of the interactions involving this variable approached reliability.

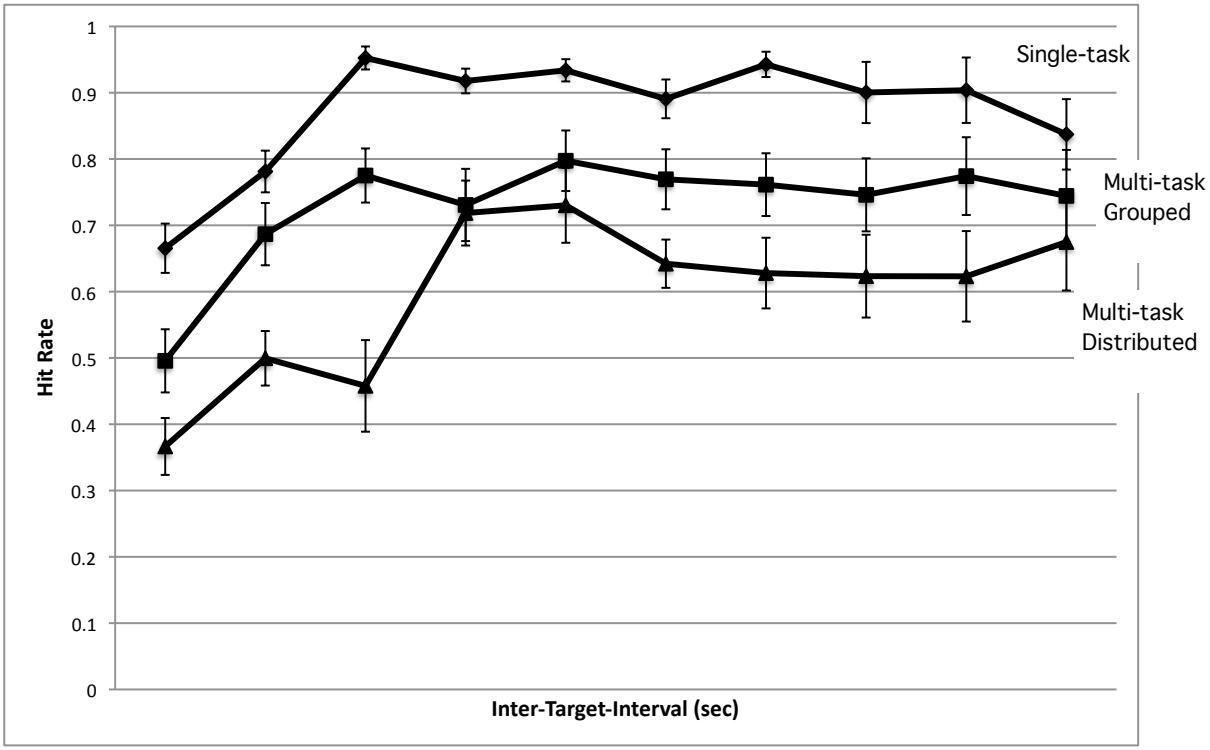

Figure 6. Proportion hits for each ITI window in Experiment 3, with standard errors.

Although the interaction of task and ITI was not reliable, there is some evidence of differences between ITI functions, and in particular, a somewhat slower recovery in the distributed multi-tasking condition — reduced performance at the 3 sec ITI. To provide a more powerful look at the ITI functions, we re-grouped and combined the ITI data across experiments, and present them in the next major section.

Learning effects during the session. Again, performance was similar from the first set of single-task blocks to the final set, averaging 80.5% and 78.0%, respectively, ($t < 1$). Thus, performance was generally constant across the experiment.

Discussion

Experiment 3 indicates that perceiving multiple events is very difficult when the events were intermixed spatially, even with a simple response rule. Multiple events are easier to perceive when they are spatially organized by task, regardless of the response or-

ganization. However, there was also a decrement for grouped multi-tasking in the experiment; optimal perceptual efficiency was obtained only in single-task conditions.

The change to a simple one-button response rule did not alter the overall pattern of results. However, note that overall performance levels increased relative to the previous experiments. The effects of response complexity can be examined further by comparing results between experiments. The combined results are interesting, and include interactions with response complexity, and conditions under which attentional set was optimized.

ITI and Response Complexity Across Experiments

To further examine the ITI functions and response complexity, we re-grouped the data from the three experiments into two mixed-design "experiments." Each re-grouped experiment has 2 levels of task condition, crossed with 2 levels of response complexity, and the ITI's (2 X 2 X 10). The first re-grouped experimental compares single-tasking to grouped multi-tasking and 4-button responses to 1-button responses, using data from Experiments 1 and 3. The full design and conditions are specified below[4]; together, the two re-grouped experiments use all of the data from the three experiments.

The figures were formatted to emphasize the systematic functions. Figure 7 shows the data for singe-task and grouped multi-task conditions. The two functions are similar in shape, and indicate that the main difference between single-task and grouped multi-task conditions is a constant (intercept) effect. For each task condition, there is a linear rise to an asymptotic performance level at 3 sec, followed by generally constant performance at 3 sec and beyond. A simple interpretation of the functions is as follows. As a target is detected, detection of temporally close targets in the 1 sec window suffers considerably. Then there is a recovery process causing a linear increase in performance and taking a total of 3 sec. Throughout longer ITI's, performance remains high, defining an asymptote. The constant difference between the two task conditions has been explained by differences in the efficiency of single- and multi- event perception. The rate of recovery appears similar in the two task conditions; the single- and multi- event slopes

[4] Task condition (single-task versus grouped multi-task) X response (4 buttons versus 1), with the following conditions: Single-task and grouped multi-task from Experiments 1 and 3.

(increase from 1 to 3 sec, per sec) were 12.5% and 10.7% in the single-task and multi-task conditions, respectively ($F<1$ in a fairly sensitive within-observer comparison; $SE_{difference} = 2.1\%$).

In further interpretations of the ITI functions, the asymptotic levels at 3 to 10 sec provide a good landmark— that is, they indicate the optimal level for the task condition, where set has been optimized. These levels appear as dashed lines in the figure. Relative to this landmark, the decrements in performance at short ITI's can be interpreted as a temporary loss of optimal efficiency, or perhaps loss of set. Further, the recovery slopes can be reframed, in terms of extents of recovery. In Figure 7, the extents of recovery were 20.2% and 20.4%, for singe-task and multi-task, respectively. These extents can be compared to other conditions.

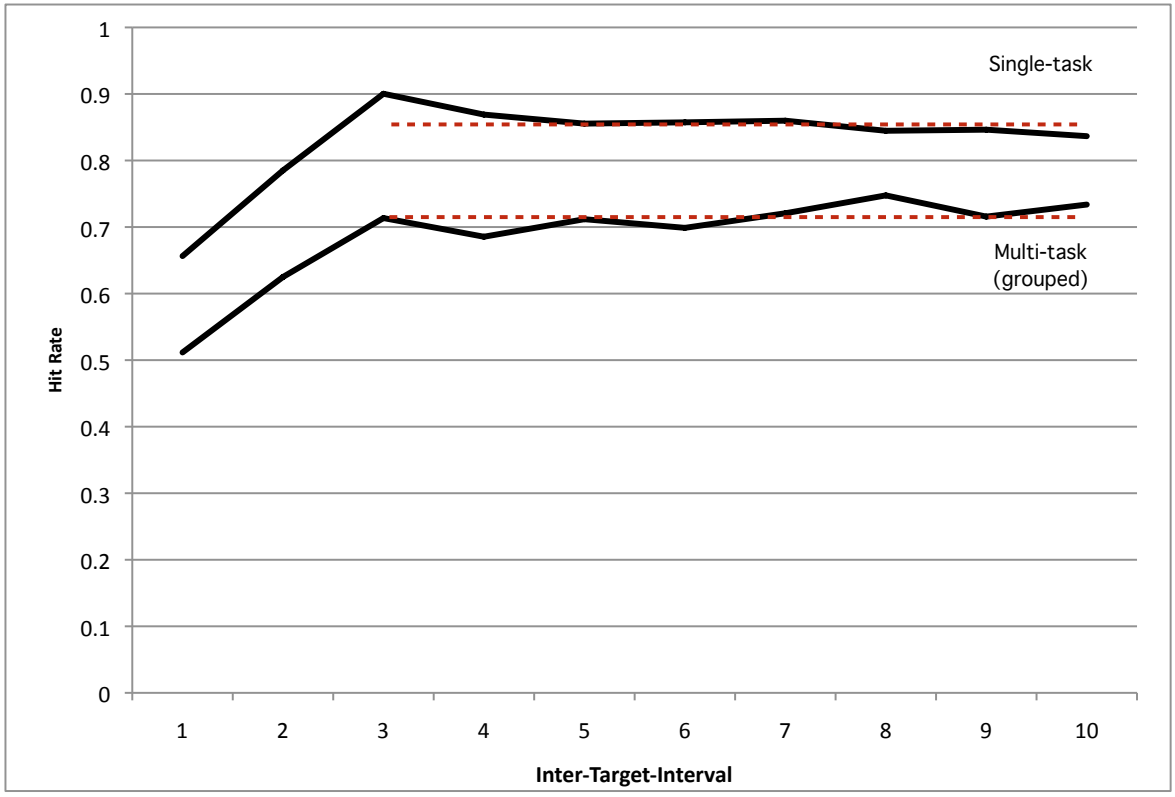

Figure 7. Means of single-task and grouped multi-task conditions of Experiments 1 and 3. (Error bars have been removed to emphasize the functions; variability is represented in the previous figures.)

The similarity of the function shapes for single-tasking and grouped multi-tasking, and in particular the similar extent of recovery, is potentially interesting because these conditions differ in the requirements for attentional set. After target detection in single-task conditions, observers need only prepare for (or re-prepare for) one event type, and the next target will be of that same event type. In contrast, in the grouped multi-task condition, observers must eventually prepare for four events types, and the next target is most often a change in event type. The similar recoveries in these conditions is significant and considered further in the General Discussion.

Response complexity modifies the ITI effects. There was a substantial difference between the more complex (4-button Experiment 1) and simpler conditions (1-button Experiment 3), as can be seen in Figure 8. These data are separated by response complexity and combined across the 2 task conditions. The difference is the slopes and the extents of the recovery process — a faster and greater recovery with the simpler 1-button response of Experiment 3 (14.2% slope, 25.6% extent) than with the 4-button response of Experiment 1 (8.1% per sec slope, 15.1% recovery; $p < .01$). There is less (and slower) recovery when observers must re-instate the complex screen-to-response mapping required with the spatially organized, 4-button response rule. Overall perceptual efficiency is higher with the simpler response, perhaps because the easier response rule allows a more effective perceptual set to be maintained.

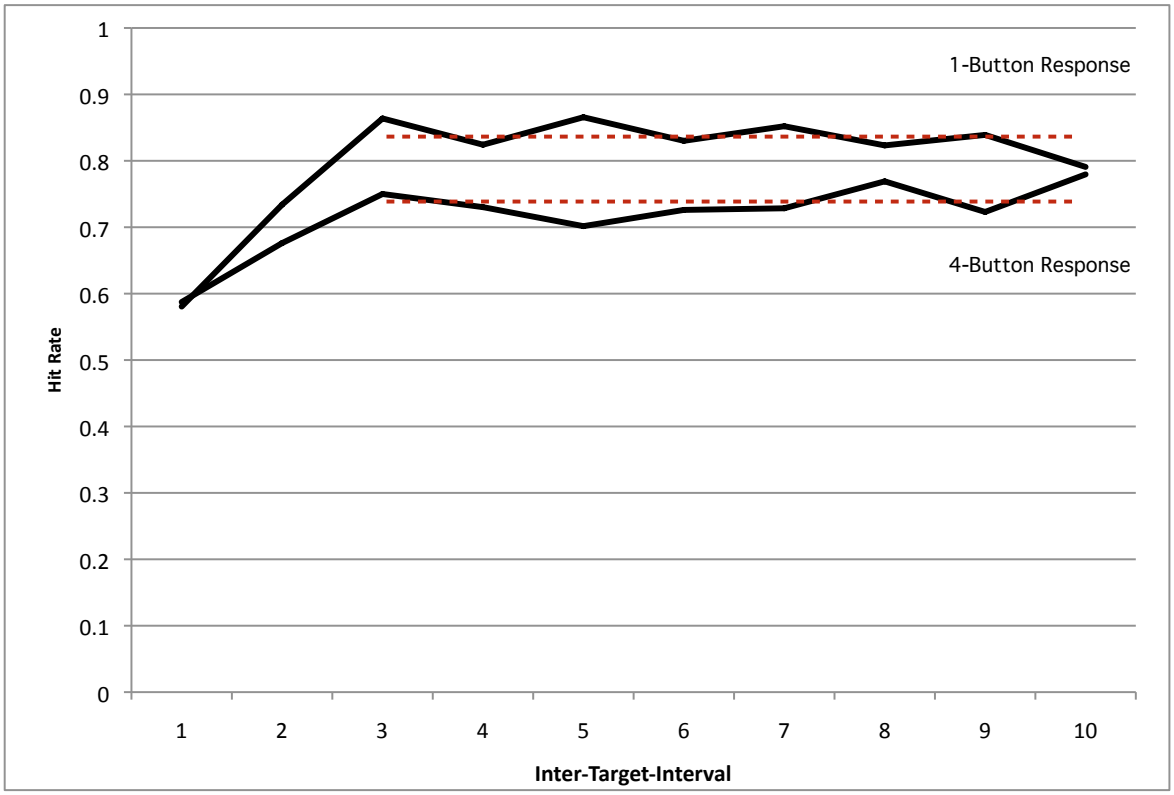

Figure 8. Mean of high response complexity (Experiment 1) and low complexity (Experiment 3) conditions; single-task and grouped multi-task data only.

The overall effects of task condition and response complexity shown above capture most of the variation in ITI data for these conditions. The significant task effects were reported with the original experiments. Further effects were evaluated in a mixed design ANOVA. The effect of response complexity was reliable, $F(1, 35) = 6.23$, $p = .02$. On the other hand, none of the overall interactions involving ITI approached significance.[5]

[5] We also note the small slope differences apparent in the asymptotic periods (3 - 10 sec) in both figures. Performance in simpler conditions (single-task and 1-button response) tends to decrease slightly across the target-less periods, whereas it tends to increase slightly in more complex conditions (multi-task and 4-button response). Perhaps these effects reflect slight changes in alertness or set efficiency over time. Similar small effects are seen in the next set of data.

We now turn to the second re-grouped experiment, which compares single-task and distributed multi-task conditions (Experiments 2 and 3).[6] The data for the two task conditions are shown in Figure 9. The single-task function is similar to previously, showing a 3-sec recovery to asymptote followed by a generally constant asymptotic period. The extent of recovery is 19.3%. In contrast, performance is much lower overall for the distributed multi-task condition, and recovery is less definitive. Recovery may take as long as 5 sec, and the amount recovered is smaller (12.0% extent). The slope for the first 3 ITI's was higher for the single-task condition (11.0% per sec) than for the distributed multi-task condition (2.6% per sec, $p = .01$).

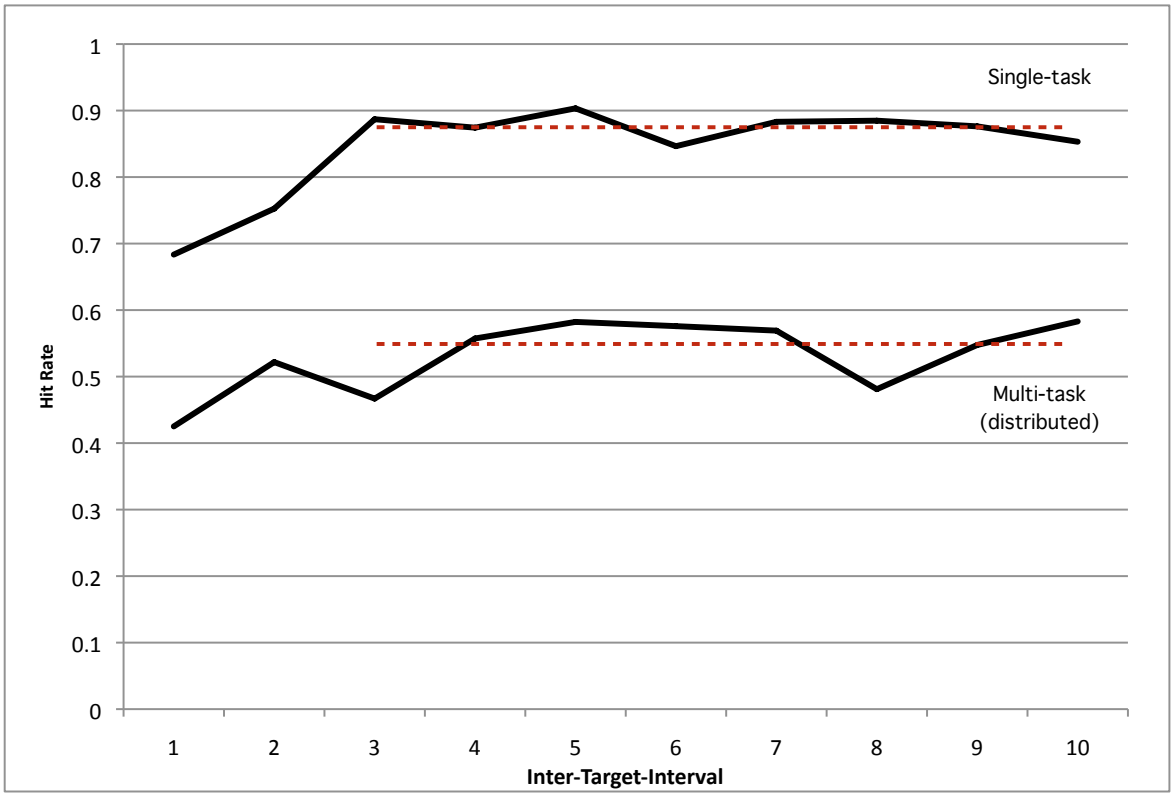

Figure 9. Means of single-task and distributed multi-task conditions of Experiments 2 and 3.

This same data-set can be analyzed for response complexity effects. As can be seen in Figure 10, there are marked differences in recovery slopes and extents. With 1 button responses, recovery is robust for the first 3 sec (9.5% slope, 25.8% extent), and appears to continue to 5 sec. With 4 buttons, however, little if any recovery occurs (slope =

[6] The design is task condition (single-task versus distributed multi-task) X response (4 buttons versus 1), with the following conditions: Single-task and distributed multi-task from Experiments 2 and 3.

2.8%, extent = 5.5%). The differences in recovery are marked in the 4-button distributed multi-tasking condition, as will now be seen.

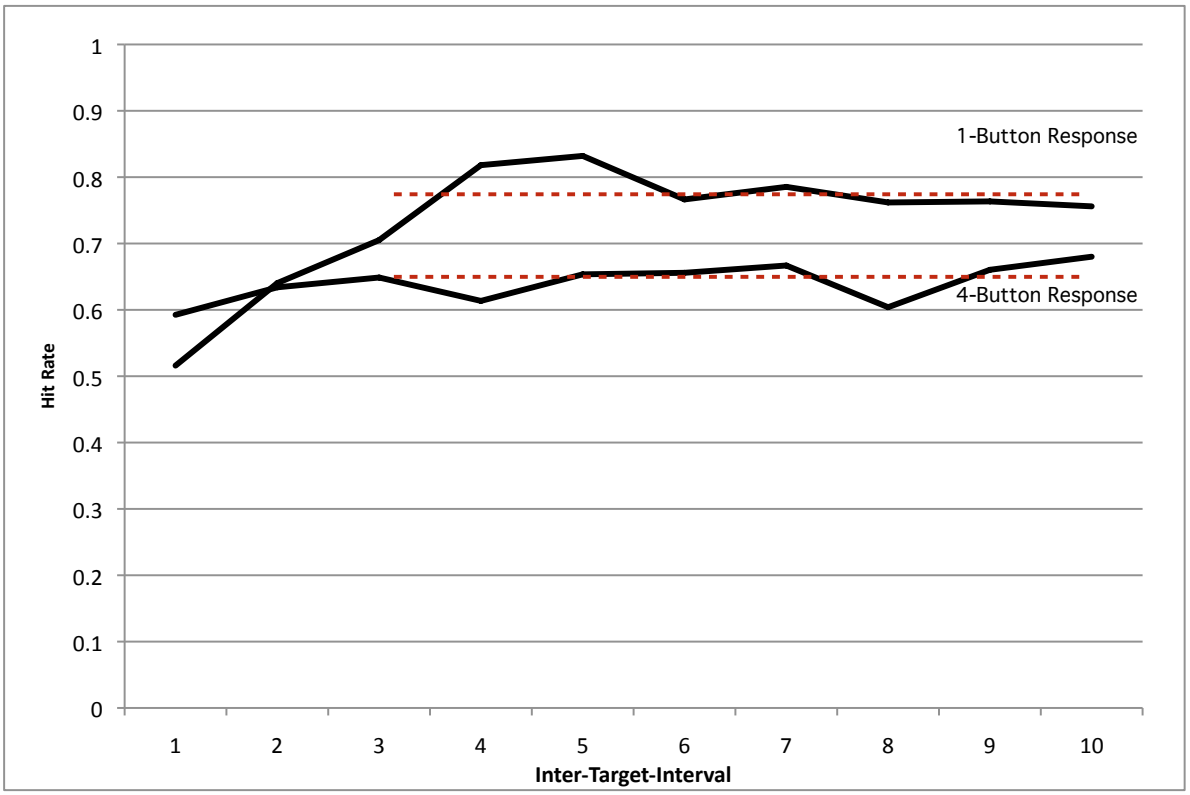

Figure 10. Means of high response complexity (Experiment 2) and low complexity (Experiment 3) conditions; single-task and distributed multi-task data only.

Task condition and response complexity interact in this set of data. In the ANOVA, there were effects of task (reported in the original experiments) and response complexity, $F(1,35) = 11.33$, $p = 0.002$. In addition, there was an interaction of task, response complexity, and ITI, $F(9, 306) = 2.62$, $p < .001$, $\eta_p^2 = 0.07$. Therefore, we also present the 4 ITI functions defined by task condition and response. Because the individual functions are somewhat noisy (and were presented, in the slightly different groupings of Experiments 2 and 3), we grouped adjacent ITI windows to provide the more systematic functions shown in Figure 11.

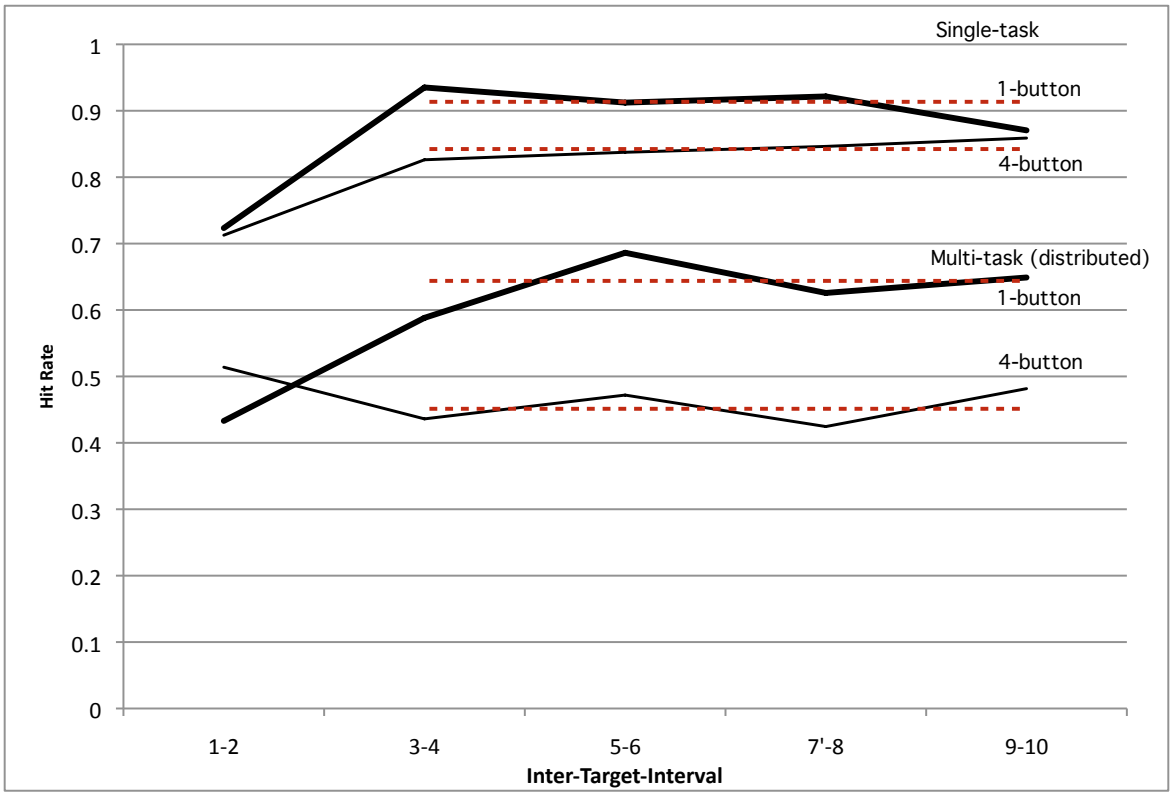

Figure 11. ITI functions for each condition of the re-grouped 2 X 2 (single-task [top two functions] and distributed multi-task data [bottom two] only). Adjacent ITI windows have been combined to reduce error.

The single-task functions reiterate that there is an initial recovery phase with a larger extent for the simpler 1-button responses than for 4-buttons. The distributed multi-tasking functions provide new information, however. First, in the 1-button condition there is a slower but sizable recovery. The extent of the recovery by 3 sec is 27.1% (calculated as above, from 1 sec windows). In this condition, a set for MEP appears to be constructed, but more slowly than in other conditions. The recovery appears to continue out to 6 sec. However, the asymptotic level is still well below single-task levels, and below the grouped multi-task conditions (comparable grouped functions are in Figures 6 and 7). In contrast, in the 4-button distributed condition there is no evidence of recovery; the recovered extent is actually negative! The combined demands of complex responses and distributed multi-task displays may overwhelm control processes, with the result that a MEP set cannot be set in this condition. Consequently, performance remains low throughout the ITI's. We consider interpretations of the ITI functions further below.

General Discussion

The experiments tested the idea that attentional set is critical in complex scene perception. We manipulated the number of event types and measured overall perceptual efficiency. Consistent with the attentional set hypothesis, perceptual efficiency was optimal when observers were set for a single event type — hit rates in single-task conditions averaged 81% across experiments, and asymptotic performance at moderate and longer ITI's reached 91% (Figure 11). When observers had to change set continuously to handle four event types in organized locations, performance was reduced by a substantial amount— overall hit rates fell an average of 17% relative to single-task, and asymptotic levels reached only 76% (Figure 6). This level of perceptual efficiency is reasonably high, however, and one could say that observers were set to handle four event types fairly well — a set for Multiple Event Perception (MEP). In fact, an attentional set for multiple grouped events was re-instated as quickly as in single-event conditions (e.g., Figure 7). Only when the spatial organization of the tasks was disrupted by distributing event types did performance fall drastically, bringing hit rates down to 48% in distributed multi-tasking conditions (Experiments 2 and 3). This is a miss rate above 50% and a repeated form of inattentional blindness, in which observers fail to detect targets present in their field of view (cf. Mack & Rock, 1998, Simons & Chablis, 1999). Furthermore, the ITI data indicate that there are difficulties establishing a set for MEP in distributed conditions. Set was reinstated slowly but somewhat effectively with simple responses, reaching an asymptote of 64% in Experiment 3 (Figure 11). More significantly, MEP could not be optimized with the more complex stimulus-to-response mapping; there was no rise to asymptote, with performance during that period averaging only 45% (Experiment 2, and Figure 11).

Overall, the range in optimal performance levels indicates that attentional and response set are important factors in scene perception. The results suggest that everyday scene perception can be severely limited in its efficiency, even when all target events are expected and predictable. Although we cannot generalize accurately to all types of complex scenes, the results suggest that efficiency limitations are pervasive in the perception of complex dynamic scenes.

Attentional Set Approach to Scene Perception

The core idea of the attentional set approach is that perception is an active process that depends on the state of the observer. Although early visual processing may be fairly automatic, the later stages of visual and cognitive processing are argued to depend on attentional set. The present results underscore the importance for overall perceptual efficiency of visual processing components related to specific target events. The experiments also revealed the importance of executive processes related to the response; response complexity determined whether recovery would occur after target detection and the optimal level of performance reached.

The simplest set hypothesis is that performance is optimal with a single attentional set, and this was confirmed in the present experiments. However, we argued that efficiency was reasonably high when there were multiple types of perceptual events, and we suggested the idea of MEP set. MEP set is likely to be important in a complex world, and it has clear boundary conditions, including grouping of event types by location, and complexity of response.

More detailed hypotheses about attentional set in the present situation are developed below. In general, attentional set is thought to influence the ease with which perceptual processes bring target events into consciousness (e.g., Mack, 2003, Most et al., 2005, Simons & Chabris, 2004). However, observers are unlikely to be aware of set itself. Observers are likely to be aware of goals, with control settings being implemented as the observer intentionally pursues goals, such as watching for a particular type of object or event. Control settings may also be activated by stimuli that are associated with certain actions or goals. Thus, perception is active in the following sense: Entry into awareness (sensitivity of perception), is influenced by control settings related to the observer's goals and recent experiences.

The critical factor of spatial organization deserves further comment. In many ways, it is a stimulus or bottom-up factor rather than top-down factor, because it concerns the organization of the environment. The effects of this factor emphasize the interactive nature of top-down and bottom-up processing (e.g., McClelland & Rumelhart, 1981, Most

et al., 2005, Neisser, 1976, Palmer, 1975); stimulus organization helped observers more effectively meet their goal of successful multi-tasking.

Measuring Scene Perception

The main measure of perceptual efficiency used here was global, summarizing overall accuracy across the 60 sec trials. A global measure is appropriate because everyday scene perception is a resultant outcome of a composition of more elementary processes. We argue that overall efficiency for natural composites of processes is a criterion that models of everyday scene perception should address. However, it is also important to decompose performance into component subprocesses, in order to understand the underlying processes.

Explanations in Terms of Underlying Processes

We present a brief eye-movement explanation and a more detailed attentional explanation. Both explanations are designed to address the three main types of results obtained here. The first type is differences in optimal performance levels, which were produced by task condition (Figures 7 and 11) and by response complexity (Figures 8 and 11). The second type of result is the large decrements at short ITI's, and the third result is the differing recovery rates after that. We begin with the short ITI decrements. A basic explanation is that the decrements occur because resources (foveal fixation in the eye movement account) are focussed on the initial target; consequently, subsequent targets in the rest of the display do not receive adequate resources. In most conditions, performance returns to asymptotic levels within 3 sec, because the allocation of resources to the rest of the display is restored.

Eye Movement Model

In this explanation the restoration of resources involves re-establishing a scanning pattern that is optimal for each event type. In single-event conditions there is one scanning pattern that is restored in 3 sec; once restored this pattern produces high optimal levels of performance. In grouped multi-event conditions, the scanning pattern is restored for the next display cell and task in 3 sec. The pattern must change again for each display cell (event type), and this could cause the moderate level of asymptotic performance that was observed. In distributed multi-event conditions, the event types are not organ-

ized by cell, increasing the difficulty of matching scanning pattern to event types. With a simple response (Figure 11), it is possible that spontaneous groupings of event types are created over the slow, 6 sec recovery period. Complex responses, on the other hand, may drain executive functions and thus prevent development of spontaneous groupings, so that there is no optimization of scanning.

An Attentional Set Model

The attentional explanation follows similar ideas but we add details from models of visual search, because the present task can be viewed as an instance of visual search in general. In models of visual search (e.g., Wolfe, Cave, & Franzel, 1989, Zelinsky, 2008), the search process can be divided into several subprocesses. The first set of processes initiate the search, as event tokens appear at the beginning of a trial. The tokens should cause descriptions of target-relevant features (*target templates*) to be activated. However, since targets did not occur during the initial portion of a trial, search initiation would not influence performance in the present experiments. Use of the target templates defines the second phase of processing. The templates would include information about how a target changes over time. And, to allow observers to detect targets at various eccentricities, the templates should exist at multiple levels of scale (Zelinsky, 2008). Examples of possible template properties are given below.[7] The template is applied to the stimulus-world in a scanning process in which the template is compared to features extracted from the display. When a display item matches the template to a reasonable degree, further processing is triggered, including a shift of additional attention to the item's location. If the match continues to be high, the item will be admitted to consciousness as a likely target. Processing would then continue until a decision about the item is made (target or distractor). At this point, resources have been focussed on the target location, with the consequence that other targets occurring soon are more likely to be missed (the decrement at short ISI's). In most cases, the scanning process resumes and becomes optimized within seconds. A critical issue is how templates for differing event types are used in this scanning process.

---

[7] Possible features for each task: Location task, proximity to goal box; motion task, abruptness of movement (speed/extent of movement); color task, increase of yellowness; shape task, increase in convexity.

The success of the scanning process appears to be well indexed by the asymptotic levels for the respective conditions. These levels indicate that even in single-task conditions, observers miss some targets. In single-task conditions, misses could occur because the complete set of tokens cannot be scanned quickly enough, with the result that some target events occur without being matched against the target template. The miss rate increases when multiple target templates must be used in multi-task conditions, and when a complex response rule must be maintained. The need to use multiple templates is a primary cause of multi-task costs. Consider first the grouped multi-task conditions.

Perhaps the simplest assumption about scanning is that one template is used at a time, and the templates are switched as the observer begins to process a new display cell (with a new category of event types). For explaining the grouped multi-tasking cost, the critical assumption is that during the template switch, no matching for targets would take place. These interruptions in scanning could increase the miss rate. The rapid recovery at early ITI's for grouped multi-task conditions could be explained by assuming that observers need to load only one target template, the one appropriate for the next display cell to be scanned. The selection of which template to use is made rapidly, perhaps driven by stimulus information. The ability to quickly switch templates with new display cells may be the basis of the fairly efficient attentional set for MEP in grouped conditions.

An alternative assumption involves more general control settings — either a more general template for multiple event types, or multiple templates that are matched against stimulus items in parallel manner. In this model, the decrement for grouped multi-tasking could be explained by the additional difficulty of using more general settings. However, this model does not explain the equally fast recovery, after detection of a target, in grouped multi-tasking and single-task conditions. It seems that one template should be re-set faster in single task conditions than more multiple templates or more general templates in multi-task conditions. Therefore, we favor the idea that a single template is used, and that it is switched rapidly in grouped multi-task conditions.

Returning to the single-template scanning model, consider now the distributed multi-task condition, in which events are not spatially organized. To handle each event type within a region of the display, the scanning process would have to continually change target templates. The more frequent template changes explains the lower level of performance in the distributed relative to the grouped conditions. The slow but sizable recovery for distributed multi-tasking when the response is simple (Figure 11) could occur because observers create spontaneous groupings of similar event types and are able to use these groups to swap target templates in a fairly efficient manner. Note that within a trial, each event-stream (sequence of tokens) was assigned to a particular cell. Observers may learn the distribution on each trial. As the groupings become organized over about 6 sec, the MEP set becomes optimized and distributed multi-tasking approaches the efficiency of grouped multi-tasking.

With the added complexities of 4-button responses, however, the challenge for the processes that organize template switching appears to be too great. Perhaps there is a single control process that coordinates template switching and response interpretation, and this process is overwhelmed. There could be competition between the stimulus organization and response mapping — between the spontaneous, stimulus-driven organization of task groups, and the need to map from display cells to keys.

A Bottom-Up Explanation?

An alternative to our top-down emphasis is the idea that bottom-up, stimulus-driven processing determines performance in the respective conditions. Could the event-switching costs be explained by bottom-up mechanisms? One possible approach is to posit priming processes that alter efficiency between conditions. For example, the consistency of event types in single-task conditions might result in stronger priming of target templates for that event type than in multi-task conditions. There are two problems for this account. First, priming does not seem strong enough of a mechanism to produce the large multi-event costs; even in multi-event conditions, each event type occurs often and therefore there is no reason for large priming differences. Second, priming does not explain the similar rates of ITI recovery in single-event and grouped multi-event conditions, since priming should be be stronger (causing a quicker return to optimal levels) in single-event conditions than in the multi-event conditions.

Is a Four-Cell Display a Scene?

The four-cell display of changing objects used are not the same as natural scenes. Is it reasonable to think of the present displays as representative of scene perception? The motivation for using display cells stemmed from our interest in the surveillance process, in which observers scan scenes looking for critical categories of events. In the surveillance profession, the use of multiple cells is a standard practice. Because security is important, the fact that surveillance operators use multiple cells suggests (anecdotally) that it is a reasonable and efficient way of representing information about events and scenes. Furthermore, psychological evidence suggests that multi-cell displays capture significant aspects of scene perception, including the perception of layout and meaning (e.g., Sanocki et al., 2006, Potter & Fox, 2009). Theories of scene perception assume that scene representations are derived from separate "snapshots" (e.g., Hochberg, 1978), and there is integration across cells (Sanocki et al., 2006). These observations support the assumption that the multiple display cell presentation is a useful approach.

The present displays lacked background, unlike natural scenes. Background layout introduces additional complexities to scene perception and could compound perceptual difficulties. Therefore, presence of background could increase the importance of attentional set.

Surveillance and the Design of Complex Scene Perception

The present experiments are representative of the surveillance process and illustrate that scene perception can be severely limited, even with known, predictable target events. The results indicate that several factors limit perceptual efficiency in scene perception and surveillance: The need to monitor multiple event categories, the need to make complex responses, the presence of multiple targets close in time, and the absence of task-specific locations. Further research on these phenomena should serve both psychological theory and human security.

References

Balcetis, E., & Dunning, D. (2006). See what you want to see: Motivational influences on visual perception. *Journal of Personality and Social Psychology, 91(4)*, 612-625. doi: 10.1037/0022-3514.91.4.612

Bruner, J. S. (1957). On perceptual readiness. *Psychological Review, 64*(2), 123-152. doi:10.1037/h0043805

Chun, M., & Potter, M. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 109-127.

Dreisbach, G., & Haider, H. (2009). How task representations guide attention: Further evidence for the shielding function of task sets. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(2)*, 477-486.

Dux, P. E. & Marois, R. (2009). The attentional blink: A review of data and theory. *Attention, Perception, & Psychophysics, 71(8)*, 1683-1700.

Folk, C., Leber, A., & Egeth, H. (2008). Top-down control settings and the attentional blink: Evidence for nonspatial contingent capture. *Visual Cognition*, *16*(5), 616-642.

Folk, C., Remington, R., & Wright, J. (1994). The structure of attentional control: Contingent attentional capture by apparent motion, abrupt onset, and color. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 317-329.

Gao, T., Newman, E., & Scholl, B.J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, in press.

Greene, M.R., & Oliva, A. (2009a). Recognition of Natural Scenes from Global Properties: Seeing the Forest Without Representing the Trees. Cognitive Psychology, 58(2), 137-179.

Greene, M.R., & Oliva, A. (2009b). The briefest of glances: the time course of natural scene understanding. *Psychological Science, 20 (4),* 464-472.

Goldstone, R. L., Braithwaite, D. W., & Byrge, L. A. (in press). Perceptual learning. In N. M. Seel (Ed.) Encyclopedia of the Sciences of Learning. Heidelberg, German: Springer Verlag GmbH.

Henderson, J. M., & Hollingworth, A. (2003). Eye movements, visual memory, and scene representation. In M. A. Peterson and G. Rhodes (Eds.), Analytic and holistic processes in the perception of faces, objects, and scenes (pp. 356-383). New York: Oxford University Press.

Kiesel, A., Steinhauser, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin, 136*(5), 849-874. doi:10.1037/a0019842.

Kirchner, H., & Thorpe, S. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762-1776.

Koivisto, M., & Revonsuo, A. (2007). How meaning shapes seeing. *Psychological Science, 18*, 845 - 849.

Lavie, N., Hirst, A., & Fockert, W., & Viding, E. (2004). Load theory of selective attention and cognitive control. Journal of Experimental Psychology General, 133, 3, 339-354.

Leber, A. B., Kawahara, J., & Gabari, Y. (2009). Long-term abstract learning of attentional set. *Journal of Experimental Psychology: Human Perception and Performance, 35*(5), 1385-1397. doi:10.1037/a0016470

Macdonald, J., & Lavie, N. (2008). Load induced blindness. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1078-1091.

Mack, A., Pappas, Z., Silverman, M., & Gay, R. (2002). What we see: Inattention and the capture of attention by meaning. *Consciousness and Cognition*, *11*(4), 488-506. doi:10.1016/S1053-8100(02)00028-4

Mack, A. & Rock, I. (1998). *Inattentional Blindness*. Cambridge, MA: MIT Press.

Malcolm G. L., Henderson J. M. (2009). The effects of target template specificity on visual search in real-world scenes: Evidence from eye movements. *Journal of Vision, 9*, (11):8, 1–13, http://journalofvision.org/9/11/8/, doi:10.1167/9.11.8.

Michotte, A. (1950/1991). The emotions regarded as functional connections. In M. Reymert (Ed.), *Feelings and emotions: The Mooseheart symposium* (pp. 114–125). New York: McGraw-Hill. [Reprinted in Thinès, G., Costall, A., & Butterworth, G. (Eds.). (1991). *Michotte's experimental phenomenology of perception* (pp. 103–116). Hillsdale, NJ: Erlbaum.]

Monsell, S. (2003). Task switching. *TRENDS in Cognitive Sciences, 7*(3), 134 - 140.

Moore & Weisman

Most, S. B., Scholl, B. J., Clifford, E. R., & Simons, D. J. (2005). What you see is what you set: Sustained inattentional blindness and the capture of awareness. *Psychological Review, 112*(1), 217-242.

Neisser, U. (1967). *Cognitive psychology*. East Norwalk, CT: Appleton-Century-Crofts.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. New York, NY: W H Freeman/Times Books/ Henry Holt & Co.

Paap, K. R., & Ogden, W. C. (1981). Letter encoding is an obligatory but not capacity-demanding operation. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 518-527.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*, 519–526.

Postman, L., & Bruner, J. S. (1949). Multiplicity of set as a determinant of perceptual behavior. *Journal of Experimental Psychology, 39*(3), 369-377. doi:10.1037/h0058224

Potter, M. C., & Fox, L. F. (2009). Detecting and remembering simultaneous pictures in rapid serial visual presentation. *Journal of Experimental Psychology: Human Perception and Performance, 35*(1), 28-38.

Pylyshyn, Z. (1999). Is vision continuous with cognition? The case for cognitive impenetrability of visual perception. *Behavior and Brain Sciences, 22,* 341-423.

Raymond, J., Shapiro, K., & Arnell, K. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink?. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 849-860.

Sanocki, T. (2003). Representation and perception of scenic layout. *Cognitive Psychology, 47*, 43-86.

Sanocki, T., & Oden, G. C. (1991). Adjustments on representations of familiar patterns: Change over time and relational features. *Perception & Psychophysics, 50*, 28-44.

Sanocki, T. & Epstein, W. (1997). Priming spatial layout of scenes. *Psychological Science*, 8, 374-378.

Schneider, W. & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84* (1), 1-66.

Schreij, D., & Olivers, C. (2009). Object representations maintain attentional control settings across space and time. *Cognition*.

Schyns P.G., Goldstone R.L. & Thibaut J.P. (1998) The development of features in object concepts. *Behavioral & Brain Sciences 21*(1) pp 1-17

Shapiro, K and Driver, J and Ward, R and Sorenson, RE (1997) Priming from the attentional blink: A failure to extract visual tokens but not visual types. *Psychological Science, 8*, 95 - 100.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception, 28*, 1059-1074.

Sulman, N., Sanocki, T., Goldgof, D., & Kasturi, R. (2012). *Display-based limitations in human video surveillance monitoring*. Paper under review.

Tversky, B., Zacks, J. M., Hard, B. M. (2008). The structure of experience. In T. Shipley and J. M. Zacks (Editors, p. 436-464), *Understanding events*. Oxford: Oxford University.

Torralbo, A., Chai, B., Caddigan, E., Walther, D., Beck, D., & Fei-Fei, L. (2009). Categorization of good and bad examples of natural scene categories [Abstract]. *Journal of Vision, 9*(8):940, 940a, http://journalofvision.org/9/8/940/,doi:10.1167/9.8.940.

Torralba, A., Oliva, A., Castelhano, M., & Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review, 113*(4), 766-786.

Tsal, Y., & Benoni, H. (2010). Diluting the burden of load: Perceptual load effects are simply dilution effects. *Journal of Experimental Psychology: Human Perception and Performance, 36*(6), 1645-1656. doi: 10.1037/a0018172

Tsotsos, J.K. (1990). Analyzing Vision at the Complexity Level. *Behavioral and Brain Sciences 13*, 423 - 445.

Tsotsos, J.K. (2001). Complexity, Vision and Attention. In L. Harris and M. Jenkin (Eds.), *Vision and Attention*. Springer-Verlag; New York.

Van Loy, B., Liefooghe, B., & Vandierendonck, A. (2010). Cognitive control in cued task switching with transition cues: Cue processing, task processing, and cue-

task transition congruency. *The Quarterly Journal of Experimental Psychology, 63*(10), 1916-1935. doi:10.1080/17470211003779160

Vandierendonck, A., Liefooghe, B., & Verbruggen, F. (2010). Task switching: Interplay of reconfiguration and interference control. *Psychological Bulletin, 136*(4), 601-626. doi:10.1037/a0019791

Walther, D. B., & Fei-Fei, L. (2007). Task-set switching with natural scenes: Measuring the cost of deploying top-down attention. *Journal of Vision, 7*(11):9, 1–12, http://journalofvision.org/7/11/9/, doi:10.1167/7.11.9.

White, R., & Davies, A. (2008). Attention set for number: Expectation and perceptual load in inattentional blindness. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(5), 1092-1107.

Wolfe, Cave, & Franzel (1989). Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance, 15(3),* 419-433.

Yu, A., Dayan, P., & Cohen, J. (2009). Dynamics of attentional selection under conflict: Toward a rational Bayesian account. *Journal of Experimental Psychology: Human Perception and Performance, 35(3)*, 700-717.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*, 3-21.

Zelinsky G. J. (2008). A theory of eye movements during target acquisition. *Psychological Review, 115*, 787–835.