THE ALLOCATION OF ATTENTION IN SCENE PERCEPTION

Inaugural-Dissertation

zur Erlangung des Doktorgrades der Philophie an der

Ludwig-Maximilians-Universität München

vorgelegt von

Melissa Lê-Hoa Võ

im Oktober 2008

TO MY PARENTS

Tag der mündlichen Prüfung:

19. Dezember 2008

Erstgutachter:

Prof. Dr. Werner Schneider, Department Psychologie, LMU

Zweitgutachter:

Prof. Dr. Heiner Deubel, Department Psychologie, LMU

CURRICULUM VITAE

PERSONAL DETAILS

Date of Birth	18. 03. 1981
Place of Birth	Munich, Germany
Nationality	U.S. American
Home Address	Dieselstraße 5, 80993 Munich, Germany
Father	Quy Tu Võ, Vietnamese
Mother	Judy Kay Võ, U.S. American

EDUCATION

10/2000 - 9/2003	Diploma Study program in Psychology at the Katholische Universität Eichstätt-Ingolstadt
2002	Vordiplom Psychologie (very good)
10/2003 - 4/2006	Diploma Study program in Psychology at the Freie Universität Berlin
2006	Diplom Psychologie (very good) "Effects of Emotional Valence on Implicit and Explicit Memory for Words: Can Pupil Dilation Give Clarification?"
11/2006 - 10/2008	Ph.D. student at the Ludwig-Maximilians-Universität Munich; (project P167 within the Cluster of Excellence “Cognition for Technical Systems - CoTeSys”)
2008	Dissertation Psychologie (summa cum laude) “The Allocation of Attention in Scene Perception”
since 11/2008	Post-Doctoral Fellow at Edinburgh University

CONTENTS

CHAPTER 1: INTRODUCTION

I. WHERE DO WE LOOK?

Our visual system is characterized by the restriction of high quality visual processing to a small region surrounding the center of gaze — the so-called fovea. Visual acuity rapidly decreases from the fovea to the low-resolution visual periphery. In order to process information from multiple locations of the visual field, we have to move our eyes about three times each second alternating between periods of information uptake, i.e., fixations, and rapid eye movements, i.e., saccades. Goal-directed eye movements and the deployment of visual attention are tightly linked (see Deubel, 2003; Deubel & Schneider, 1996; Paprotta, Deubel, & Schneider, 1999; Schneider & Deubel, 2002; for an overview see Deubel, O'Regan, & Radach, 2000). Thus, what we see and understand about the visual world is closely linked to where we direct our eyes. How active vision, i.e., the active control of human gaze, operates over complex real-world scenes has become an important issue in several core cognitive science disciplines such as cognitive psychology, visual neuroscience, or machine vision.

What is meant by complex, naturalistic scenes? Henderson and Hollingworth (1999b) defined *scene* as a semantically coherent human-scaled view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner. Just imagine what you see when standing in your own kitchen. The scene's background will consistent of the room itself with a floor, a ceiling, and walls. The room will probably be equipped with a kitchen counter and electric appliances like a fridge and a stove. Additionally, you will also find individual objects, spatially arranged according to the laws of physics and the constraints associated with the typical location of the various objects in a kitchen. Thus, depictions of real-world scenes differ from the type of stimulus material commonly used in psychological experiments, e.g., visual search displays containing rotated Ts amongst upright Ls, since their constituents are subject to known semantic and syntactic constraints (Biederman, Mezzanotte, & Rabinowitz, 1982). This implies that the active control of human gaze in naturalistic scenes draws not only on currently available visual input, but is strongly influenced by cognitive processes. These include stored short-term, episodic, or long-term information of previously encountered scenes as well as innate goals and expectations of the viewers (see Deubel, 1996; Henderson, 2007). The objective of this thesis was to investigate which cognitive factors influence the allocation of attention when actively inspecting complex, naturalistic scenes.

In the following, I will give a short overview of previous findings which form the theoretical framework of my thesis. The theoretical introduction will be followed by the description of five experiments I conducted in three separate studies dealing with a number of related issues currently being discussed in scene perception research. Finally, the thesis will close with a general conclusion.

II. BOTTOM-UP AND TOP-DOWN INFLUENCES ON VISUAL ATTENTION IN SCENE PERCEPTION

There is no doubt that when viewing a natural scene, attention and the human eye do not move around randomly. However, there is a dispute regarding the degree to which eye movements during scene perception are influenced by bottom-up image properties such as contrast or color on the one hand or by top-down factors such as the current task or scene knowledge on the other (for a review see Henderson, 2007). The first neuro-computational models of visual attention that dealt with natural scenes strongly relied on attention control by bottom-up image saliency (e.g., Itti & Koch, 2000; Itti, Koch, & Niebur, 1998; Parkhurst, Law, & Niebur, 2002). On the basis of combined information from different feature maps (e.g., color, intensity, and orientation), highly salient regions of an image can be located which are assumed likely to attract observers' attention. These models perform quite well when no specific task is driving the observer's exploration of an image (Underwood, Foulsham, van Loon, Humphreys, & Bloyce, 2006). However, in most real-world settings an observer's activity is influenced by a given task, for example, finding a target amongst a number of distractors. Underwood and colleagues (Underwood & Foulsham, 2006; Underwood et al., 2006) have tried to disentangle the specific contributions of bottom-up visual saliency and top-down task demands. In their experiments, participants inspected pictures of natural scenes in which two objects of interest were placed, one of which was characterized by high and the other by low visual saliency according to the Itti and Koch (2000) algorithm. The task was modified to determine whether visual saliency is invariably the dominant attractor of fixations, or whether task influences can provide a cognitive override that renders saliency secondary. When the participants were told to inspect the scene in preparation for a subsequent memory task, visually salient objects attracted early fixations in support of a saliency map model of scene inspection. However, when participants had to search for a specific target amongst distractors of higher visual saliency, the target would attract attention to a greater degree despite being less salient. These results support a version of the saliency map hypothesis in which task demands can cognitively override a purely bottom-up driven saliency map.

Recent computational models have taken the modulation of attention allocation by cognitive processes into account assuming the combined influence of both bottom-up and top-down information (see Navalpakkam & Itti, 2005; Torralba, Oliva, Castelhano, & Henderson, 2006). In those models, bottom-up processing based on low-level image features interacts with the top-down processing of scene gist and spatial layout thus making it possible to shift attention, and correspondingly the eyes, to locations which have a high probability of containing the search target. In their contextual guidance model, Torralba and colleagues (2006) propose that an image is analyzed in two parallel pathways: the local and the global pathway. Both pathways share the first stage during which the image is filtered by a set of multiscale-oriented filters. The local representation comprises each spatial location independently and is used to both compute local saliency peaks and perform object recognition. The global pathway, on the other hand, represents the entire image holistically by extracting global statistics from the image which makes it possible to activate knowledge and expectations regarding a specific scene — so-called scene priors. Task demands are supposed to influence only the global pathway by providing information on the expected location of the target as a function of scene priors. Thus, a key feature of the model is the interaction of local and global processing within the first glimpse in order to rapidly narrow down the search area to those parts of the scene that most probably contain the target.

III. INFLUENCE OF THE INITIAL SCENE REPRESENTATION ON TARGET SEARCH IN NATURALISTIC SCENES

In order to exert its influence on subsequent eye movements, the initial scene representation has to be stored in visual memory across several saccades. During search, more detailed information is then continuously acquired with each fixation adding to the evolving scene representation (e.g., Henderson & Castelhano, 2005; Hollingworth, 2005; Tatler, Gilchrist, & Rusted, 2003). To investigate whether the initial scene representation acquired from a flashed preview of a scene can be stored in such a way that it continuously exerts its influence on visual search in a real-world scene, Castelhano and Henderson (2007) used the "Flash-Preview Moving-Window Paradigm". This paradigm elegantly combines the brief tachistoscopic viewing method typically used in scene identification (or scene categorization) experiments with the moving window technique typically used to investigate eye movements under restricted viewing conditions.

In their study, participants were asked to search for target objects in scenes while their eye movements were recorded. Prior to presentation of the search scene, a scene preview was briefly presented for 250 ms. Then a word indicating the identity of the target was displayed, after which the search scene was presented. However, the search scene was only visible through a gaze-contingent moving window with a 2° diameter centered at fixation within a scene. This paradigm allows to selectively manipulate the information provided by the preview of the scene. At the same time it enables the investigation of subsequent eye movement behavior and controlling for the information uptake during the actual search by restricting the latter to foveal vision only. With the restricted view through the moving window during search, information uptake cannot be influenced by extrafoveal vision, which is needed for the viewer to extract a vector of global features and rapidly set up scene priors. Possible preview benefits due to global processing must therefore be attributed solely to the memory-based scene representation formed as a result of the processing of the briefly flashed preview.

In a number of experiments, Castelhano and Henderson (2007) were able to show not only that the initial scene representation can be used to predict highly probable target locations, but also that this initial representation continues to be available in an abstract manner, which allows it to survive multiple transsaccadic changes in the image falling on the retina. For example, an identical scene preview led to significant search benefits during subsequent target search as compared to a different or meaningless scene preview. Also, a scene preview still benefited subsequent target search when it was identical, but minimized in its size as compared to the search scene. However, a preview did not benefit search when it sustained the conceptual category of the following search scene while differing in its visual details. Thus, the initially crude scene representation seems to be stored in an abstract manner, but needs more specific information (e.g., the particular spatial layout of the scene) in order to benefit target search.

IV. INDIVIDUAL DIFFERENCES IN ATTENTION ALLOCATION DURING SCENE VIEWING

There is some evidence that individuals differ in the way they process visual input as a function of expertise — qualitatively different ways of processing information as a result of experience — or a more general processing efficiency unrelated to the specific visual input. For example, Underwood, Chapman, Brocklehurst, Underwood, and Crundall (2003) have shown that scan paths during driving differ as a function of expertise, i.e., experienced drivers monitored other road users more often than novice drivers, who showed little ability to switch the focus of their attention as potential hazards appeared. Further evidence of individual differences in processing visual input comes from a recent study by Brockmole, Hambrick, Windisch, and Henderson (in press), who found that expert chess players developed a contextual cueing effect during target search, which was four times greater than the one generated by novices. In a change detection experiment using alternating displays with a presentation rate of 500 ms, Werner and Thies (2000) could show that domain-specific expertise increased the ability to detect changes for flashed scenes implying that there are individual differences regarding rapid picture processing as a function of expertise as well. Apart from domain-specific expertise, studies on reading have shown that eye movement patterns differ between good and poor readers. Poor readers tend to fixate longer and make shorter saccades than good readers due to differences in general processing efficiency (e.g., Eden, Stein, Wood, & Wood, 1994; Hutzler & Wimmer, 2004). Thus, there seems to be evidence that individuals greatly differ in their ability to process visual information and to control attention allocation. The first study of this thesis was mainly concerned with the effects of the degree of initial scene processing rather than with the investigation of effects due to specific expertise. When the presentation time of a complex visual input is limited to a split second, individual processing efficiency may lead to differential benefits in using the flashed information for effective eye movement control.

V. THE THEORY OF VISUAL ATTENTION

An integrated theoretical and methodological approach, which permits the assessment of components of processing efficiency, is the Theory of Visual Attention (TVA; Bundesen, 1990, 1998). According to the TVA, visual recognition and attentional selection are based on making perceptual categorizations. Thus, an object *x* is selected — i.e., encoded into a capacity-limited visual short term memory (VSTM) — as soon as it is perceptually categorized. A further claim of TVA is that objects in the visual field are processed in parallel. Objects that are selected and therefore may be subsequently reported from a briefly exposed visual display, are those elements for which encoding is completed before the sensory representation of the stimulus array has decayed — provided that memory space is available in the store. In the framework of the TVA, the general efficiency of the visual processing system is reflected in the parameters visual perceptual processing speed *C* (number of visual elements processed per second) and visual short-term memory storage capacity *K* (number of elements maintained in parallel). These processes are formally described by a coherent, mathematical theory in terms of a set of (mathematically) independent quantitative parameters (for a detailed mathematical description, see Bundesen, 1990, 1998; Kyllingsbaek, 2006). Both parameters can be assessed using a whole-report task, in which participants are briefly presented with arrays of simple stimuli, e.g., letters, at varying exposure durations from which they have to identify (name) as many as possible. For some participants short presentation times may be sufficient to establish a conscious percept, for others the same presentation time might not suffice to allow a conscious report of presented scene details. Such individual differences in processing efficiency may also transfer to the perceptual processing of naturalistic scenes. In the first study of this thesis, eye movement data recorded during search for objects embedded in naturalistic scenes was combined with the assessment of perceptual efficiency parameters in order to further investigate the source of individual differences in the perception of naturalistic scenes.

VI. EFFECTS OF OBJECT-SCENE INCONSISTENCIES ON EYE MOVEMENT CONTROL

The second study of this thesis addressed another mechanism that has been shown to exert its effects on eye movement control during scene viewing, i.e., the processing of object-scene inconsistencies. There has been an ongoing debate concerning how quickly we can detect and process objects that do not fit the global gist of a scene, and whether initial eye movements can be modulated early on by the computation of such object-scene inconsistencies. Ever since the famous "octopus in farmyard" study by Loftus and Mackworth (1978) showed that scene inconsistencies can be detected early enough to affect initial eye movements, various research groups have either been able to replicate this finding (e.g., Becker, Pashler, & Lubin, 2007; Underwood & Foulsham, 2006; Underwood, Humphreys, & Cross, 2007; Underwood, Templeman, Lamming, & Foulsham, 2008) or have found evidence that argues against an early impact of scene inconsistencies on eye movement control (e.g., DeGraef, Christiaens, & d'Ydewalle, 1990; Gareze & Findlay, 2007; Henderson, Weeks, & Hollingworth, 1999; Rayner, Castelhano, & Yang, in press). The debate is based on the paradox that the gist of a scene can be perceived within a very short glance, while in the same amount of time only a few objects can be identified (e.g., Castelhano & Henderson, 2005; Henderson & Hollingworth, 1999a; 2003; Tatler, et al., 2003). The question, therefore, is whether object-scene inconsistencies can influence eye movement control prior to foveal processing of the inconsistent object.

Henderson and colleagues (1999) showed that, contrary to the results reported by Loftus and Mackworth (1978), there was no evidence for an effect of semantic inconsistency prior to the fixation of an inconsistent object. Participants viewed a set of line drawings of natural scenes modified from the ones used by DeGraef et al. (1990) in preparation for a memory task. The scenes included either a semantically consistent object, e.g., a cocktail glass in a kitchen, or a semantically inconsistent object, e.g., a microscope in the kitchen. Scenes were paired so that the inconsistent object in one scene would serve as the consistent object in another, i.e., the microscope would be the consistent object in a laboratory. The results showed that initial saccades were not controlled by semantic inconsistencies in the visual periphery, but upon fixation the semantic inconsistency of an object affected fixation densities and durations. Inconsistent objects were fixated longer and more often than their consistent counterparts, and viewers tended to return their gaze to inconsistent objects more often than to consistent objects. Even when participants were instructed to actively search for target objects which were either consistent or inconsistent with the scene context, there was no evidence for extrafoveal processing of semantic inconsistencies.

To date, most of the evidence bearing on the effect of object-scene inconsistencies has come from one type of manipulation: the semantic violation of a scene's gist. However, a different way to produce object-scene inconsistencies relates not to an object's semantic fit to the general scene gist, but to its position within the specific structure of scene elements, i.e., the scene syntax. In Study 2 of this thesis, two experiments were conducted which further investigated the effects of semantic as well as syntactic object-scene violations using 3D-rendered images of naturalistic scenes.

VII. EXPLICIT AND IMPLICIT CHANGE DETECTION IN NATURALISTIC SCENES

The third study of my thesis further elaborated on the effect of object-scene inconsistencies on eye movement control. However, instead of violating expectations on the general semantic and syntactic composition of a scene, the specific location of an object embedded in a scene was changed, thus contrasting what had been learnt across several presentations of the same scene. The objective of this study was, therefore, to investigate whether object location changes to episodically learnt scenes would be detected both explicitly — i.e., by means of explicitly indicating which object had changed its location— and implicitly — i.e., as mirrored in modulated eye movement behavior.

The following overview of some theoretical accounts is intended to emphasize the importance of stored object and scene representations for the detection of relational object changes. According to the object file theory (e.g., Kahneman & Treisman, 1984; Kahneman, Treisman, & Gibbs, 1992), the basic properties of an object are only linked once an object is attended to thus permitting the creation of a temporary representation of that object — an object file —, which maintains information about the object it represents. When an object or its location changes, the current object file will be compared to the previous object file with respect to the stored object features. A change in one or more object features can therefore only be detected if the respective object files are stored in memory so that a mismatch can be computed. Similarly, in visual memory theory, the detection of a change to an object in a scene is a function of the degree of attention deployed to the object during encoding, the retention of that representation in visual short- and long-term memory (VSTM/VLTM), and the generation of a new representation following the change which is necessary in order to compare it with the stored representation (see Henderson & Hollingworth, 2003; Schneider, 1995; 1999). In their model of visual processing, Rao and Ballard (1999) postulate that neural networks learn statistical regularities of the natural world, signaling deviations from such regularities to higher processing centers. A mismatch or "residual" signal due to differences between the stored and the currently processed scene representation leads to the deployment of attention to changed regions of scenes.

These theoretical implications have been supported by a number of experimental findings that provided evidence for memory-based as opposed to bottom-up triggered allocation of attention to changed objects in scenes without transient motion signals. While it has been shown that the sudden appearance of an object captures attention even when attention had originally been directed elsewhere (e.g., Jonides & Yantis, 1988; Theeuwes, 1994; Yantis, 1998), several studies by Brockmole and Henderson (2005a, 2005b, in press) found that in stationary scenes new objects also attracted gaze even when presented during a saccade, i.e., without transient motion signal. Thus, the current view of a scene had to be matched to an existing actively maintained memory representation that had been generated over the course of scene viewing (see Henderson & Castelhano, 2005).

Karacan and Hayhoe (2008) further showed that a higher degree of familiarity with an environment increased the time spent fixating regions in the scene where a change had occurred. This indicates that deviations from learnt scene schemas generate an increased allocation of attention. Thus, without a transient signal, changes to objects are preferentially processed as a result of the top-down, memory-based guidance of attention on the basis of stored scene representations and are not due to the bottom-up capture of attention. Moreover, it has been demonstrated that even when a change is not reported, fixation durations on a changed object are longer than on the same object when it has not changed (e.g., Hayhoe, Bensinger, & Ballard, 1998; Henderson & Hollingworth, 2003; Hollingworth, Williams, & Henderson, 2001; Ryan & Cohen, 2004). In the third study of my thesis, I investigated whether small location changes of objects in repeatedly presented scenes would be explicitly detected and whether eye movement behavior would implicitly mirror change detection.

VIII. OUTLINE OF THE EXPERIMENTS

This thesis was devoted to the investigation of how we deploy our attention and gaze during the viewing of naturalistic scenes. As outlined above, previous work has shown that human gaze control in scene viewing is not simply a function of current low-level saliency computation, but rather is influenced by higher-level cognitive factors such as task knowledge, prior experience, semantic scene analysis, or memory processes. In order to conduct experiments using naturalistic scenes as stimulus material, a set of 20 different 3D-rendered images of real-world scenes were created (see Appendix for sample scenes). These display a high degree of realism, while allowing for highly controlled manipulations of objects within a scene. All studies in this thesis used variations of the set of 20 rendered scenes while measuring eye movements thereon to shed more light on the allocation of attention during scene perception.

In Study 1, I investigated which information extracted from a short glimpse of a scene is most beneficial for subsequent target search. By using the "Flash-Preview Moving-Window" paradigm I was able to manipulate the amount of information available during the first glimpse of scene and to measure the effects of the initial scene representation on participants' eye movement behavior during search (Experiment 1). The additional assessment of processing efficiency parameters using the TVA approach allowed shedding more light on the source of individual differences in the efficiency of using shortly flashed scene information for target search (Experiment 2).

The aim of Study 2 was the examination of whether object-scene inconsistencies would attract attention prior to the fixation of inconsistent objects. While results regarding semantic inconsistencies have been mixed, the effect of syntactic violations on eye movement control during scene viewing has been largely neglected. I therefore manipulated scenes so that objects were either semantically inconsistent with the scene gist, syntactically inconsistent with the scene structure, or both. Participants were either instructed to view the scenes for later recognition (Experiment 3), or had to actively search for predefined target objects (Experiment 4). I analyzed eye movement behavior as an indicator for the allocation of attention in response to the two forms of object-scene inconsistencies.

Finally, Study 3 (Experiment 5) was designed to investigate whether deviations from episodically learnt scene representations would lead to increased attentional allocation to the changed object and whether effects of change would occur during earlier or later stages of scene viewing. Participants were asked to repeatedly inspect a randomized set of naturalistic scenes for later questioning. At the seventh presentation of each scene one object was shown at a new location. In this last study, attention was focused on whether participants were able to explicitly detect such location changes and whether eye movements would mirror both explicitly and implicitly detected changes.

CHAPTER 2: CONTEXTUAL GUIDANCE DURING SEARCH IN NATURALISTIC SCENES

What information can we extract from an initial glimpse of a scene and how do people differ in the way they process visual information? In Experiment 1, participants searched 3D-rendered images of naturalistic scenes for embedded target objects through a gaze-contingent window. A shortly flashed scene preview (identical, background, objects, or control) preceded each search scene. We found that search performance varied as a function of the participants' reported ability to distinguish between previews. Experiment 2 further investigated the source of individual differences using a whole-report task. Data were analyzed following the 'Theory of Visual Attention' approach, which allows the assessment of visual processing efficiency parameters. Results from both experiments indicate that during the first glimpse of a scene global processing of visual information predominates and that individual differences in initial scene processing and subsequent eye movement behavior are based on individual differences in visual perceptual processing speed.

Imagine visiting friends and helping out in their kitchen. When asked to fetch some plates you are probably able to find them at a location that seems plausible to you without having to search the entire room. This seems trivial, but involves a number of cognitive processes. For example, you have to recognize the scene you are about to act in as being part of a kitchen. In order to fulfill the given task you have to further activate your implicit knowledge about kitchens, i.e., their typical layout, their functionality, and typical locations of typical appliances. The combination of these long-term representations with the currently evolving representation of the specific kitchen as well as with task knowledge, will in most cases lead to search benefits due to the active exploration of only those parts of the kitchen that have a high probability of containing the plates (see Torralba, et al., 2006).

The study presented here investigated what information extracted from a first glimpse of a complex naturalistic scene can modulate the deployment of attention and eye movements during subsequent target search. We will see that the initial scene representation has significant influence on where we look next. However, the individual differences in perceiving shortly presented complex scenes is an issue that has seldom been discussed. People may differ in the way they benefit from an initial glimpse of a scene. Thus, we were also interested in the cognitive processes that differ between individuals during initial scene processing and whether these individual differences could then also affect attention allocation and eye movement control during subsequent search.

In order to investigate the influence of individual differences in rapid scene processing on subsequent search, we conducted two experiments. In Experiment 1, participants had to search for predefined target objects embedded in naturalistic scenes (see Castelhano & Henderson, 2007), while we varied the information provided in the flashed previews of the scenes. Each participant was then tested regarding the reportability of preview differences with a post-hoc questionnaire, which probed whether the participants had noticed differences between the previews and if so, which details in the previews differed. This allowed us to divide participants into two groups regarding their reported ability to differentiate between shortly flashed scene preview conditions: The 'conscious report group' consisted of participants who had processed the previews to such a degree that they were able to report preview differences, while participants in the 'no report group' were unable to report differences between previews. The efficiency of processing shortly flashed scenes may therefore not only influence the establishment of the initial scene representation, but may also determine the ability to consciously perceive and report differences between such scenes. For example, was a kitchen scene filled with a number of individual kitchen objects or was the same kitchen shown empty?

We hypothesized that the reported ability to discriminate between flashed scenes might not only indicate the degree of initial processing but could also explain differences in subsequent attention allocation and eye movement control, since these depend heavily on the initial scene representation. In Experiment 2, we further investigated the source of individual differences. We retested a subset of participants who had taken part in Experiment 1 using a whole-report task, which allowed assessing TVA parameters regarding the individual processing efficiency of each participant. We assumed that the 'conscious report group' would show higher processing efficiency than the 'no report group' with regard to the TVA parameters visual perceptual processing speed *C* and VSTM storage capacity *K*.

EXPERIMENT 1

Experiment 1 investigated the specific contributions of both local and global processing of scene properties during the initial glimpse of a naturalistic scene to the control of subsequent eye movements during visual search. We used the Flash-Preview Moving-Window Paradigm introduced by Castelhano and Henderson (2007) to replicate findings of search benefits following identical scene previews. In addition to an identical preview, we modulated the information available during a short scene preview which either allowed for global processing of the scene or not. We presented three different preview variations of the same search scene plus a mask as control condition (see Figure 1.1). These three preview conditions varied in the information available for the participants when flashed before the search scene was shown: while the Background preview — e.g., an empty kitchen — contained spatial layout information and allowed for scene categorization, the Objects preview — e.g., a display of typical objects found in a kitchen — lacked spatial layout and could only convey the scene's category indirectly by the need to first identify most of the objects and then form a category from them. Combining Background and Objects condition results in the Identical preview — i.e., a fully equipped kitchen. Thus, the Identical and the Background condition allow for global processing, whereas the Objects preview would — in terms of the contextual guidance model — mainly be processed along the local pathway. We therefore hypothesized that both Identical and Background previews of the search scene would lead to search benefits, while previewing only the objects of the search scene would not benefit subsequent search.

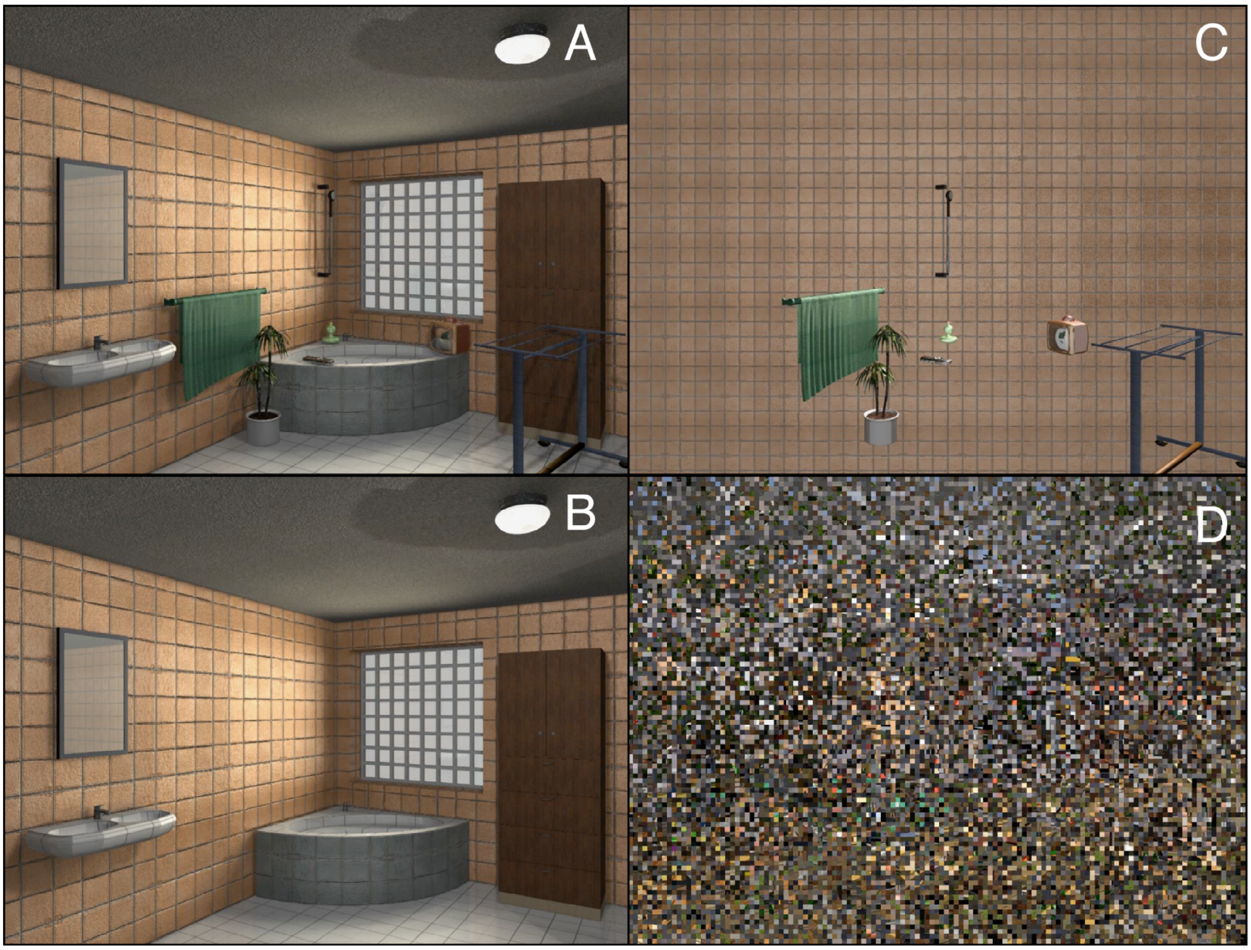

Figure 1.1: Sample scene previews of Experiment 1 with three different previews of the same bathroom scene (A: IDENTICAL, B: BACKGROUND, C: OBJECTS,) and D: the meaningless CONTROL preview also used as a mask.

Additionally, we were interested in whether the two groups of participants with different abilities to discriminate between preview conditions would also show differences in eye movement behavior during target search. If the 'conscious report group' is faster at processing visual information and can therefore process shortly presented scene information to a greater degree, this group should show superior target search performance as compared to the 'no report group'. Further, we hypothesized that the differences between the 'conscious report group' and the 'no report group' regarding preview processing would also lead to differential search benefits as a function of information provided in the flashed previews. For instance, the 'conscious report group' should benefit from previews containing a high degree of information (i.e., identical previews), while the 'no report group' should not show such a benefit due to the inability to completely process all the information only shortly provided in the previews.

METHOD

Participants

Forty students (26 female) from the Ludwig-Maximilians-Universität (LMU) Munich ranging in age between 19 and 31 (*M* = 22.87, *SD* = 2.72) participated in the study for course credit or for 8€/hour. All participants reported normal or corrected-to-normal vision and were unfamiliar with the stimulus material.

Stimulus Material

The search scenes consisted of 20 3D-rendered images of real-world scenes. The scenes were displayed on a 19-inch computer screen (resolution 1024 x 768 pixel, 100 Hz) subtending visual angles of 28.98 (horizontal) and 27.65 (vertical) at a viewing distance of

70 cm. The default background color was gray (RGB: 51, 51, 51). Each search scene was preceded by either an Identical, a Background, an Objects, or a Control preview (see Figure 1.1 for preview examples) none of which contained the search target. The Identical preview was a copy of the search scene except for the missing target object. The Background preview resembled the search scene in displaying the same background, but all distinct objects placed on background furnishings were deleted. The Objects preview consisted only of these distinct objects placed at exactly the same location as in the Identical preview but lacking its background. The Control was created from scrambled quadratic sections (8 x 8 pixel) taken from all search scenes and also served as a mask. Thus the Control was meaningless, but contained colors, orientations, and contours as is the case in unscrambled scenes. Each participant saw each search scene only once, and the four preview conditions for each scene were rotated across participants.

Apparatus

Eye movements were recorded with an EyeLink1000 tower system (SR Research, Canada), which tracks with a resolution of .01° visual angle at a sampling rate of 1000 Hz. The position of the right eye was tracked while viewing was binocular. Experimental sessions were carried out on an IBM compatible display computer running on OS Windows XP. Stimulus presentation and reaction recording was controlled by Experimental Builder (SR, Research, Canada). The eye tracker was hosted by another IBM compatible computer running on DOS, which recorded all eye movement data.

Procedure

The procedure of the study phase closely followed the procedure of the "Flash Preview Moving Window" paradigm used in the experiments of Castelhano and Henderson (2007). Experimental sessions were conducted in a moderately lit room (background luminance about 500 lx), in which the illumination was held constant. Each participant received written instructions before being seated in front of the presentation screen. Participants were informed that they would be presented with a series of scenes in which they had to search for a target as quickly as possible. They were also informed that short previews of the scene would precede the display of the search scene and that they should attend to these previews since they could provide additional information.

At the beginning of the experiment, the eye tracker was calibrated for each participant. The participants' viewing position was fixed with a chin and forehead rest, followed by a 9-point calibration and validation.

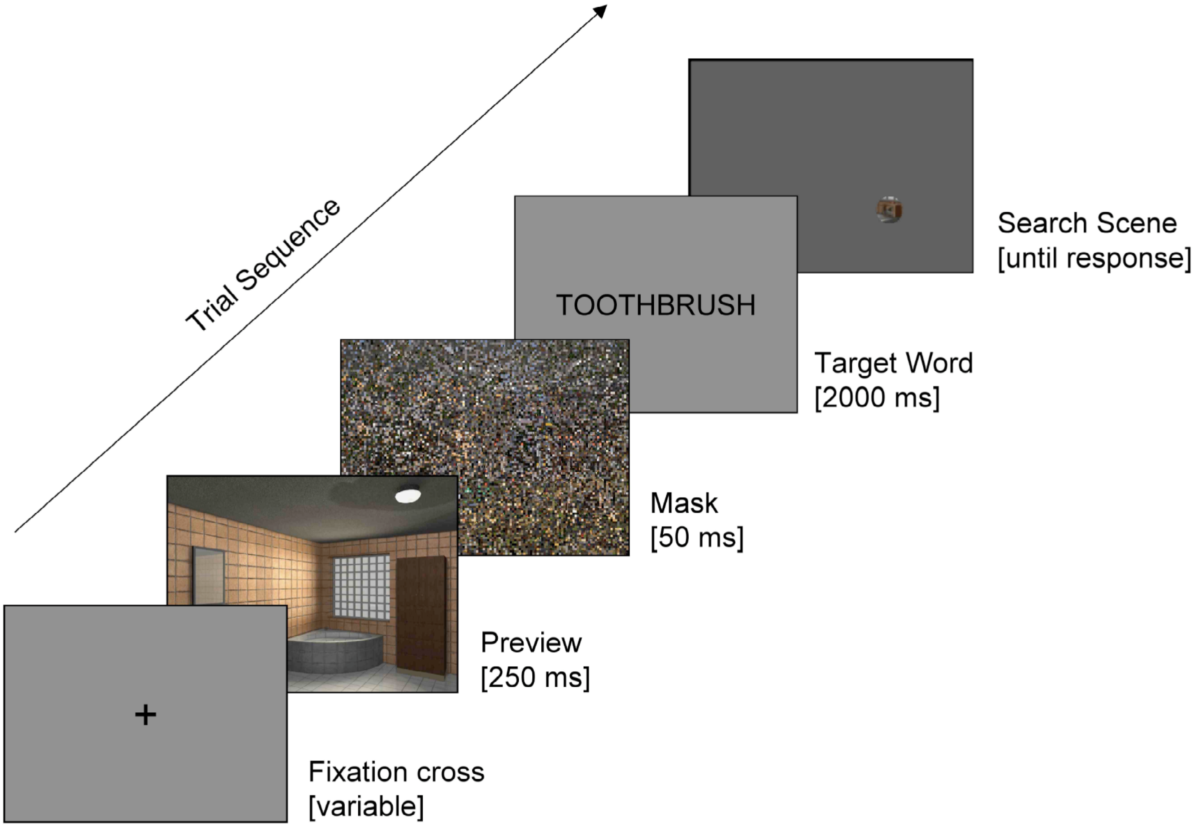

Figure 1.2.: Trial sequence of the "Flash-Preview Moving-Window" paradigm used in Experiment 1.

As can be seen in Figure 1.2, each trial sequence was preceded by a fixation check, i.e., in order to initiate the next trial, the participants had to fixate a cross centered on the screen for 200 ms. When the fixation check was deemed successful, the fixation cross was replaced by the presentation of the scene's preview for 250 ms. After the presentation of a mask for 50 ms, a black target word was displayed at the center of the gray screen for 2000 ms, which indicated the identity of the target object. Afterwards the search scene was shown through a 2° diameter circular window moving contingent on the participants' fixation location. The rest of the display screen was masked in gray. Thus, no peripheral vision was possible throughout the entire visual search. Participants had to search the scene for the target object and indicate the detection of the target object by holding fixation on the object and pressing a response button. The search scene was displayed for 15 s or until button press. Three practice trials at the beginning of the experiment allowed participants to get accustomed to the experimental set-up and the restricted vision during search due to the gaze contingent window. At the end of the study phase, participants were asked to fill out a post-hoc questionnaire to ascertain whether they were able to distinguish the previews that had been presented. The study phase lasted for about 20 minutes. Participants were not told that there would be a subsequent recognition memory test.

Data reduction and statistical analysis

Similar to Castelhano and Henderson (2007), a set of behavioral and eye movement data was analyzed. *Reaction Times (RTs)* were calculated from search scene onset until response button press. *Error Rate* was defined as the percentage of those trials in which participants failed to find and fixate the target object while simultaneously pressing the response button. *Latency to First Target Fixation* was measured from scene onset until the first fixation of the target object. *Number of Fixations to First Target Fixation* was measured as the sum of all fixations from search scene onset until the first fixation on the target object. Finally, the *Scan Path Ratio* was defined as the length of the scan pattern, i.e., the sum of all saccade amplitudes until the first fixation of the target object, divided by the shortest distance from the fixation cross to the centre of the target object.

For the analyses of both RT and eye movement data only correct responses were included, i.e., the participant pressed the response button while fixating the target object. Additionally, we excluded trials with a fixation number greater than 50. This was primarily caused by unstable calibration of the gaze contingent window [9.49 %].

Further, we had to exclude two participants who showed substantial instabilities in controlling the gaze-contingent window during search. The remaining 38 participants (25 female) ranged in age between 19 and 31 (*M* = 22.90, *SD* = 2.75). After completing Experiment 1, each participant was asked to fill out a questionnaire which included a question asking whether they were able to distinguish the different preview conditions. If a participant claimed to have noticed the preview differences he or she was asked to then describe the different conditions to the instructor in more detail. Only when the participants were able to differentiate between all three scene previews were they assigned to the 'conscious report group' (27 participants). All participants who only noticed a difference between the control preview and "other scenes" were assigned to the 'no report group' (11 participants). The 'conscious report group' (16 female) ranged in age between 19 and 31 (*M* = 22.73, *SD* = 2.79), while the 'no report group' (9 female) ranged in age between 20 and 28 (*M* = 23.27, *SD* = 2.87).

All data were submitted to an analysis of variance (ANOVA) with preview conditions (Identical, Background, Objects, Control) as within-subject factor and reportability ('conscious report group' vs. 'no report group') as between-subject factor. We confined post-hoc tests solely to theoretically driven comparisons of Identical vs. Control, Background vs. Control, Objects vs. Control, and Identical vs. Background preview conditions.

Since we expected to find similar patterns across all dependent variables, these planned contrasts were calculated for all dependent variables and for each participant group.

P values are reported with exact probabilities when a strong tendency is observed, but fails to reach statistical significance or when *p* values are presented in tables. All other *p* values that failed to reach significance are not reported. Further, for effects with multiple degrees of freedom, *p* values were Greenhouse-Geisser adjusted.

RESULTS

Reaction Time. RT data did not vary significantly across preview conditions, $F(3,37) = 2.16$. However, both the between-subject factor reportability as well as the interaction of both factors reached significance, $F(37) = 7.36$, $p < .01$ and $F(3,37) = 4.03$, $p < .05$, respectively (see Figure 1.3).

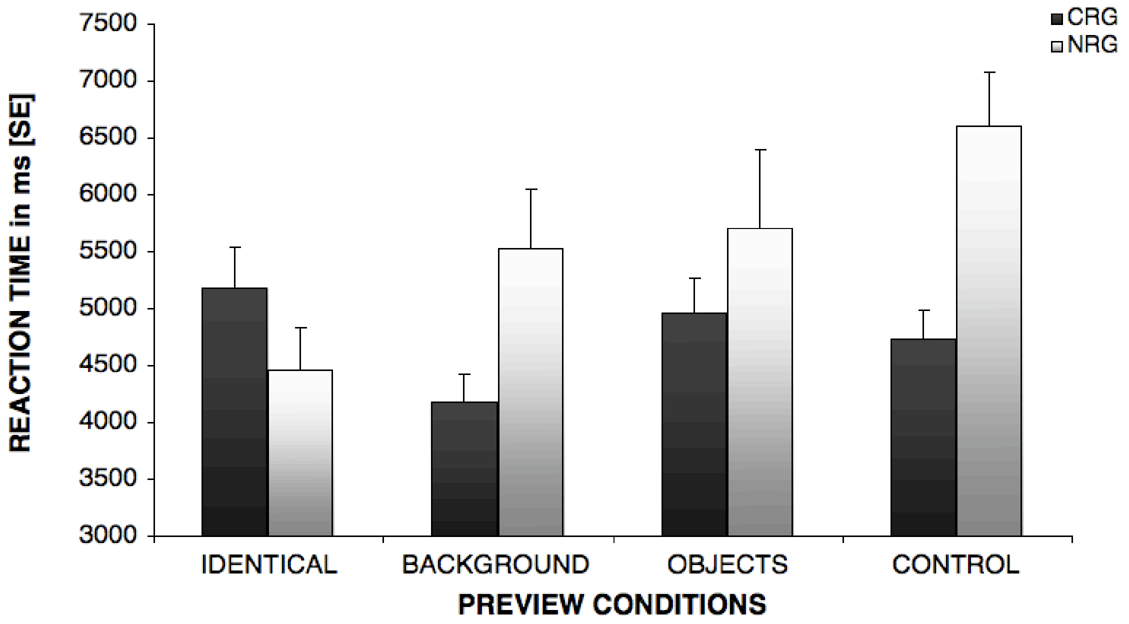

Figure 1.3: Mean reaction times [standard errors] for visual search in Experiment 1 across preview conditions (IDENTICAL, BACKGROUND, OBJECTS, CONTROL) split for participant groups ('conscious report group' = CRG, 'no report group' = NRG)

Overall, participants who were able to distinguish the three preview conditions showed faster RTs than participants who were not able to distinguish between previews (*M* = 4763, *SE* = 175 vs. *M* = 5574, *SE* = 182). Further, preview conditions showed differential preview benefits as a function of reportability (see Table 1.1). While the 'conscious report group' showed a preview benefit for Background vs. Identical, a tendency for Background vs. Control, and no statistical significance for Identical vs. Control and Object vs. Control, the 'no report group' showed a preview benefit for Identical vs. Control, a tendency for Background vs. Control, and no statistical significance for Objects vs. Control and Identical vs. Background.

Table 1.1: Summary of planned contrasts for Reaction Times and Latency to First Target Fixation across preview conditions (IDENTICAL = I, BACKGROUND = B, OBJECTS = O, CONTROL = C) during visual search in Experiment 1 split for participant groups ('conscious report group' = CRG, 'no report group' = NRG)

	Planned contrasts	T	d.f.	p
RT				
CRG	I vs. C	1.00	26	.16
	B vs. C	1.63	26	.05
	O vs. C	.58	26	.28
	I vs. B	2.72	26	.01
NRG	I vs. C	4.00	10	.00
	B vs. C	1.59	10	.07
	O vs. C	1.05	10	.16
	I vs. B	1.47	10	.09
Latency to First Target Fixation				
ALL	I vs. C	1.27	37	.11
	B vs. C	2.66	37	.01
	O vs. C	1.12	37	.13
	I vs. B	.65	37	.26
CRG	I vs. C	.51	26	.31
	B vs. C	2.00	26	.03
	O vs. C	.11	26	.46
	I vs. B	2.13	26	.02
NRG	I vs. C	3.85	10	.00
	B vs. C	1.76	10	.05
	O vs. C	1.63	10	.07
	I vs. B	1.34	10	.11

Error Rate. Error rates averaged at 21.32 % (Identical: 18.11 %, Background: 24.71 %, Objects: 14.66 %, Control: 27.78 %). There was a nearly significant main effect of preview, *F*(3,37) = 2.66, *p* = .05, but no effect of group, *F*(37) < 1, and no interaction, *F*(3,37) = 1.09. Participants produced more errors during target search when they had been presented with a Control preview than with an Identical or an Objects preview t(37) = 2.35, p < .05. All other contrasts failed to reach significance.

Latency to First Target Fixation. For latency data, all main effects as well as their interaction were significant. There was a main effect of preview, $F(3,37) = 3.14, p < .05$, a main effect of reportability, $F(37) = 11.15, p < .01$, and a significant interaction $F(3,37) = 4.24, p < .01$ (see Figure 1.4).

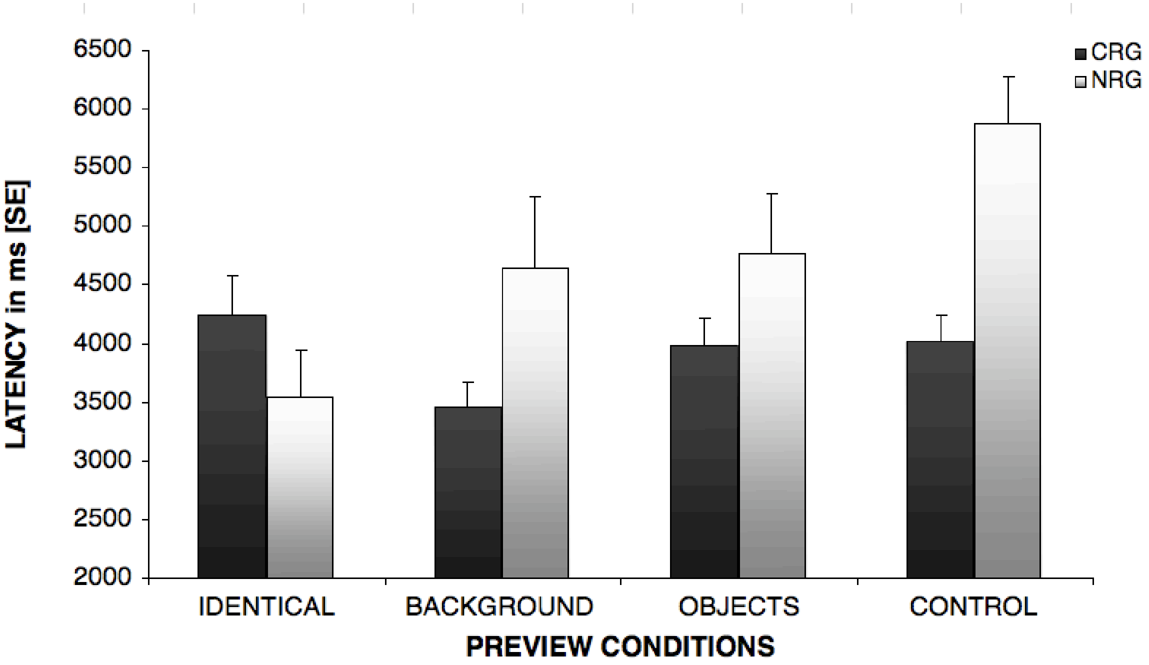

Figure 1.4: Mean latencies [standard errors] for visual search in Experiment 1 across preview conditions (IDENTICAL, BACKGROUND, OBJECTS, CONTROL) split for participant groups ('conscious report group' = CRG, 'no report group' = NRG)

As can be seen in Table 1.1, planned contrasts revealed a significant preview benefit for Background vs. Control across all participants, while all other planned contrasts failed to reach significance. Overall, participants who were able to distinguish the three preview conditions showed shorter latencies than participants who were not able to distinguish between previews (M = 3922, SE = 142 vs. M = 4708, SE = 111). The interaction of the factors preview and reportability can be characterized as follows: The 'conscious report group' showed strong search benefits after the presentation of the Background preview compared to Control as well as to the Identical preview. All other planned contrasts failed to reach significance. On the other hand, the 'no report group' showed a graded effect of preview conditions in that the effect was strongest for the Identical preview and showed a strong tendency to decrease for Objects and Background conditions.

Thus, the main effect of preview was characterized by a significant search benefit following a Background preview as compared to the presentation of the control. Preview benefit significantly varied as a function of groups.

Number of Fixations Until Target Fixation. There was a significant interaction between preview conditions and the reportability factor, $F(3,37) = 4.48$, $p < .01$, while the main effects of preview and reportability showed trends, $F(3,37) = 2.44$, $p = .07$ and $F(37) = 3.18$, $p = .08$, respectively (see Figure 1.5).

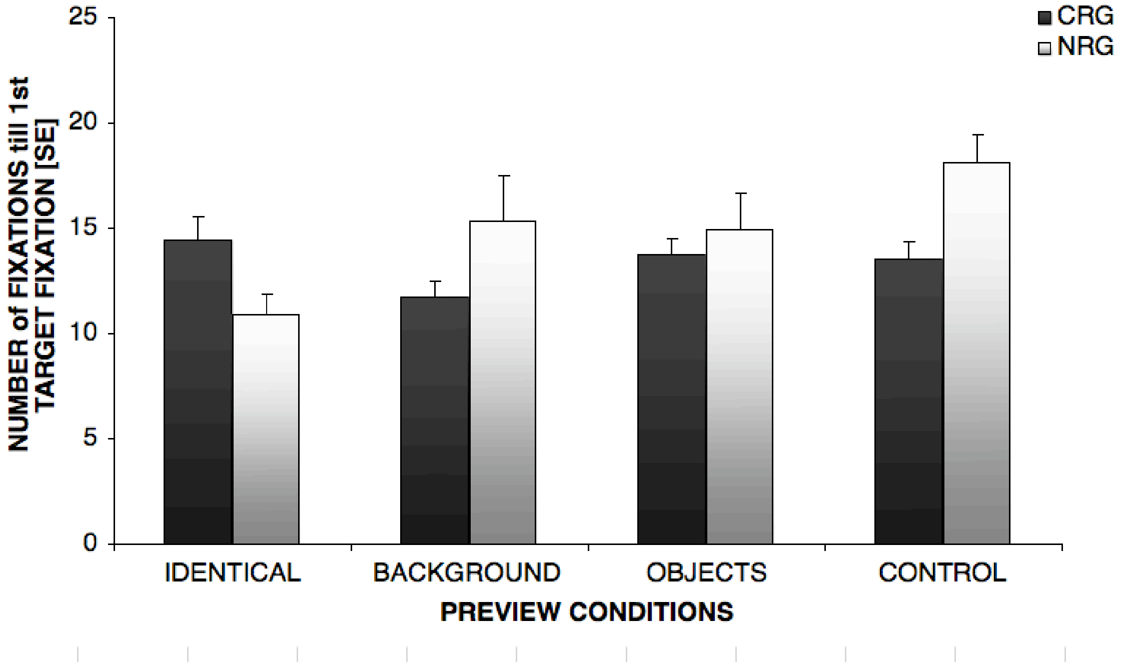

Figure 1.5: Mean number of fixations until first target fixation [standard errors] for visual search in Experiment 1 across preview conditions (IDENTICAL, BACKGROUND, OBJECTS, CONTROL) split for participant groups ('conscious report group' = CRG, 'no report group' = NRG)

As can be seen in Table 1.2, planned contrasts for the preview conditions showed a search benefit for Background vs. Control, while all other contrasts did not reach significance. For the grouping factor, the 'conscious report group' showed a trend to need fewer fixations until the first target fixation than 'no report group' (*M* = 13.39, *SE* = .45 vs. *M* = 14.82, *SE* = .58). The significant interaction between both factors for the number of fixations largely resembles the interaction observed for the Latency to First Target Fixation: Again, the 'conscious report group' showed strong search benefits after the presentation of the Background preview compared to Control as well as to the Identical preview. The other contrasts failed to reach significance. However, the 'no report group' showed a strong search benefit following an Identical preview and a tendency towards fewer fixations for Identical as compared to the Background condition. There were no significant differences between Background vs. Control and Objects vs. Control.

Table 1.2: Summary of planned contrasts for Number of Fixations until Target Fixation and Scan Path Ratio across preview conditions (IDENTICAL = I, BACKGROUND = B, OBJECTS = O, CONTROL = C) during visual search in Experiment 1 split for participant groups ('conscious report group' = CRG, 'no report group' = NRG)

	Planned contrasts	T	d.f.	p
Number of Fixations Until Target Fixation				
ALL	I vs. C	1.13	37	.13
	B vs. C	1.21	37	.02
	O vs. C	.74	37	.23
	I vs. B	.49	37	.31
CRG	I vs. C	.65	26	.26
	B vs. C	1.74	26	.05
	O vs. C	.11	26	.42
	I vs. B	2.13	26	.02
NRG	I vs. C	3.83	10	.00
	B vs. C	1.32	10	.11
	O vs. C	1.44	10	.08
	I vs. B	1.72	10	.06
Scan Path Ratio				
CRG	I vs. C	.97	26	.17
	B vs. C	.50	26	.31
	O vs. C	.45	26	.33
	I vs. B	1.52	26	.07
NRG	I vs. C	3.30	10	.00
	B vs. C	.1.38	10	.10
	O vs. C	.19	10	.43
	I vs. B	1.26	10	.12

Scan Path Ratio. The ANOVA for the scan path ratio did not show significant main effects of preview or reportability, *F*(3,37) = 1.51, and *F*(37) = 1.28, respectively. However, there was a strong tendency towards a significant interaction of both factors, *F*(3,37) = 2.62, *p* = .05 (see Figure 1.6).

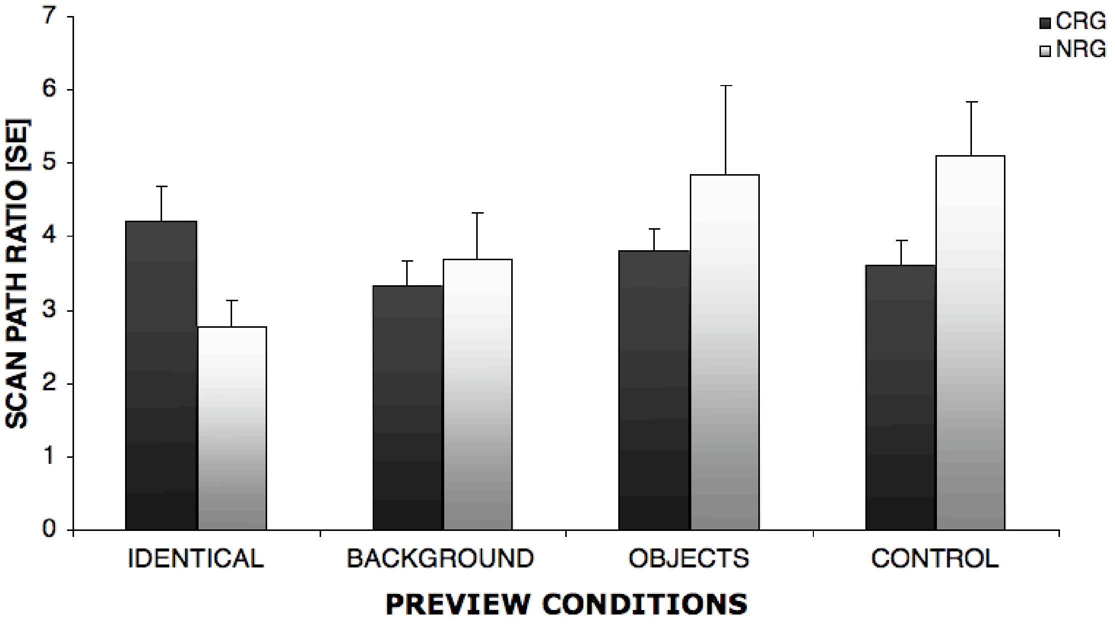

Figure 1.6: Mean scan path ratio [standard errors] for visual search in Experiment 1 across preview conditions (IDENTICAL, BACKGROUND, OBJECTS, CONTROL) split for participant groups ('conscious report group' = CRG, 'no report group' = NRG)

As can be seen in Table 1.2, no significant differences were found across preview conditions for the 'conscious report group'. There was a tendency towards a smaller scan path ratio for Identical as compared to Background preview. However, planned contrasts for the 'no report group' revealed a significantly decreased scan path ratio for Identical vs. Control preview.

Thus, all dependent variables except for error rates showed that the effects for preview conditions strongly interacted with the ability to differentiate between the three preview conditions. Participants who reported not being able to differentiate between previews showed greatest search benefit for Identical previews, while the group reporting the ability to differentiate between previews needed the least number of fixations to find the target in the Background preview condition. For this group of participants the Identical preview did not lead to search benefits compared to the Control condition.

DISCUSSION

Experiment 1 investigated the influence of both local and global processing of scene properties during the initial glimpse of a naturalistic scene on the control of subsequent eye movements during visual search. We were also interested in whether the ability to distinguish between flashed scenes would modulate search performance. The results reported above allow a number of interesting implications:

First, we found strong evidence for the influence of flashed scene previews on the guidance of eye movements. This indicates that participants were able to generate, store, and make use of an initial scene representation for subsequent target search. Since the visual field was restricted to a 2° diameter gaze-contingent window, the search benefits can be attributed to the information gathered from the initial glimpse of the scene. The initial scene representation must then be stored transsaccadically in order to play a functional role in eye movement control. This is in line with findings of Castelhano and Henderson (2007). They were able to show that the scene representation, which is used to effectively guide attention through search space, is robust to a 2 s delay between preview presentation and search scene display as well as to several intervening saccades. Thus, the results of Experiment 1 add to the growing evidence that there is a transsaccadic visual memory for initially generated scene representations which continuously exhibit their influence on eye movement control.

Second, compared to the Control condition, participants were faster and needed fewer fixations to find and fixate the target object when presented only with the scene's background. This shows that a scene representation generated from a preview, which mainly provides global scene information and therefore allows a rapid setup of scene priors, can already lead to an effective deployment of attention during subsequent search.

Third, the 'conscious report group' was faster overall (RT and Latency to First Target Fixation) and needed fewer fixations to find the target as compared to the 'no report group'. It seems that participants who had processed the previews to a greater degree were also able to control their eye movements more effectively during search. Additionally, the interaction found across all dependent variables showed that the two groups of participants, which differed in their reported ability to process the previews, also differed in their ability to benefit from scene previews. The main characteristic of the interaction, which was observable across RT, Latency to First Target Fixation, and fixation data was that while the 'conscious report group' showed the greatest search benefits after the presentation of Background previews, the 'no report group' profited most from Identical scene previews. This modulation of preview benefits as a function of group leads to the implication that the locus of the individual differences observed in search performance could lie in varying degrees of preview processing.

EXPERIMENT 2

Experiment 2 aimed at further investigating the locus of effects for the individual differences observed during visual search in Experiment 1. We therefore retested participants that had previously taken part in Experiment 1 using a TVA based whole-report task, which provided objective information on the general processing efficiency of each participant. This way we were able to examine whether the two participant groups differed in perceptual processing speed and/or VSTM storage capacity. It was specifically expected to find a higher degree of perceptual processing speed, i.e., higher *C* values, for the 'conscious report group' as compared to the 'no report group', would be found.

METHODS

Participants

Twenty-five students (18 female) from the LMU Munich ranging in age between 19 and 28 (M = 23.24, SD = 2.55) participated in the whole-report task for course credit or for 8€/hour. All participants reported normal or corrected-to-normal vision and were unfamiliar with the stimulus material. Further, all 25 participants had taken part in Experiment 1, eleven of whom had been previously assigned to the 'no report group' and 14 to the 'conscious report group'.

Stimulus Material

For the whole-report task, five equidistant red target letters (each 0.5° high x 0.4° wide) were presented in a vertical column, 2.5° of visual angle either to the left or to the right of a fixation cross, on a black screen. Stimuli for a given trial were randomly chosen from a pre-specified set of letters (ABEFHJKLMNPRSTWXYZ), with the same letter appearing only once per trial. In some trials letter displays were masked. Masks consisted of letter-sized squares (of 0.5°) filled with a '+' and an 'x'.

Apparatus

The TVA experiment was conducted in a dimly lit, soundproof room. Stimuli were presented on a 17" monitor (1024x768 pixel screen resolution, 70 Hz refresh rate). Participants viewed the monitor from a distance of 50 cm, controlled by the aid of a chin and forehead rest.

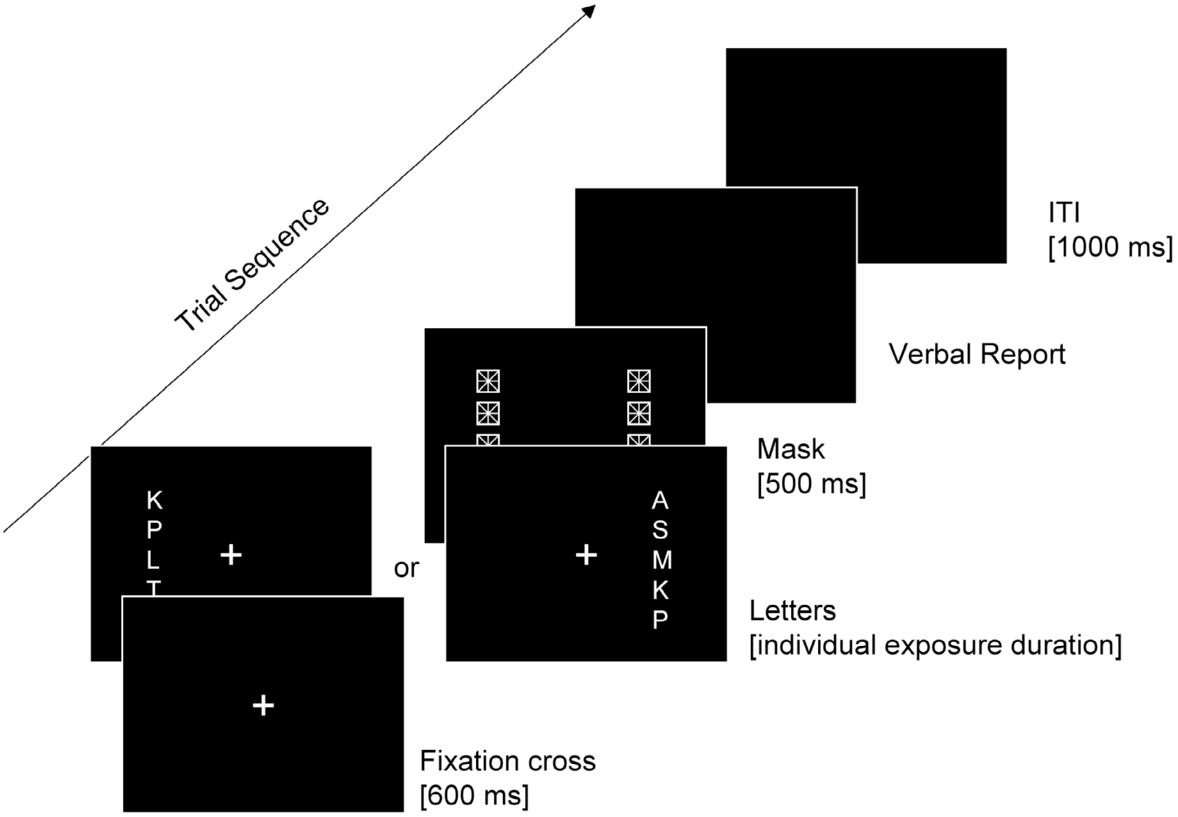

Figure 1.7: Trial sequence of the whole-report task used in Experiment 2.

Procedure

Figure 1.7 shows the trial sequence of the whole-report task. Participants were first instructed to fixate a white cross (0.3° x 0.3°) presented for 600 ms in the centre of the screen on a black background. Then five equidistant red target letters were presented in a vertical column either to the left or to the right of the fixation cross. The participants had to report as many letters as possible. The experiment comprised two phases: In phase 1 (pre-test), three exposure durations of the target letters were determined for phase 2 (main test), in which the data were collected. The pre-test comprised 24 masked trials with an exposure duration of 86 ms. It was assessed whether the subject could, on average, report one letter (20 %) per trial correctly. If this was achieved, exposure durations of 43 ms, 86 ms, and 157 ms were used in the main test. Otherwise, longer exposure durations of 86 ms, 157 ms, and 300 ms were used. Here, letter displays were presented either masked or unmasked. The masks were presented for 500 ms at each letter location. Due to 'iconic-memory' buffering, the effective exposure durations are usually prolonged by several hundred milliseconds in unmasked as compared to masked conditions (Sperling, 1960). Thus, by factorially combining the three exposure durations with the two masking conditions, six different 'effective' exposure durations were produced. These were expected to generate a broad range of performance, tracking the early and the late parts of the functions relating response accuracy to effective exposure duration. In several previous studies that used a similar paradigm (e.g., Finke, Bublak, Krummenacher, Kyllingsbaek, Müller, & Schneider, 2005), highly reliable estimates of the parameters *C* and *K* were obtained on the basis of 16 trials per target condition. In the present experiment, each participant completed 288 trials (2 hemi-fields x 2 masking conditions x 3 exposure durations x 16 trials per target condition). Before each phase, subjects were given written and verbal instructions.

RESULTS

The experimental results of the whole-report task are described by the TVA parameter estimates for ‘visual perceptual processing speed’ and ‘VSTM storage capacity’. These parameters were estimated using the standard procedure introduced by Duncan, Bundesen, Olson, Humphreys, Chavda, & Shibuya (1999) and used in several other recent studies (e.g., Bublak, Finke, Krummenacher, Preger, Kyllingsbaek, Müller, & Schneider, 2005; Finke, Bublak, Dose, Müller, & Schneider, 2006; Habekost & Rostrup, 2007; Hung, Driver, & Walsh, 2005). Detailed descriptions of the software used can be found in Kyllingsbaek (2006), detailed neural interpretations of the mathematically specified TVA concepts are described in Bundesen, Habekost, and Kyllingsbaek (2005). In short, the probability of identifying a given object x is modeled by an exponential growth function. The slope of this function indicates the total rate of information uptake in objects per second (perceptual processing speed, denoted by *C*), and its asymptote the maximum number of objects that can be represented at a time in VSTM (VSTM storage capacity, *K*).

Since we were interested in establishing whether the difference in reportability of preview differences was due to the participants' processing efficiency, we compared the 14 participants of the 'conscious report group' with 11 participants of the 'no report group' regarding both the perceptual processing speed C and the VSTM storage capacity K.

Perceptual Processing Speed C: C is defined as a measure of the perceptual processing speed in elements/second. *C* across all participants ranged from 6.22 to 33.46 (*M* =

17.06, *SD* = 7.10). Planned contrasts showed that participants from the 'conscious report group' (*M* = 19.79, *SE* = 1.74) are characterized by a higher perceptual processing speed *C* than participants from the 'no report group' (*M* = 13.60, *SE* = 1.96), *t*(1) = 2.36, *p* = .01.

VSTM Storage Capacity K: Parameter *K* reflects the number of items that can be simultaneously maintained in VSTM. *K* across all participants ranged from 2.38 to 4.00 (*M* = 3.31, *SD* = .57). There was no significant difference between groups regarding VSTM storage capacity *K* ('conscious report group': *M* = 3.43, *SE* = .15 vs. 'no report group': *M* = 3.17, *SE* = .17), *t*(1) = 1.11, *p* > .05.

Thus, the TVA parameters show that while participants of the 'conscious report group' did not differ from the 'no report group' in terms of VSTM storage capacity, they did show higher values in perceptual processing speed *C* as compared to the 'no report group'.

DISCUSSION

Experiment 2 was set out to further investigate the nature of individual differences observed in Experiment 1, where participants showed differences in their efficiency to search for target objects in naturalistic scenes. It had been hypothesized that the differences in the reportability of preview differences were due to varying degrees of information processing efficiency across participants leading to differential effects on eye movement control. Since post-hoc questionnaires can only provide subjective measures of information processing, we conducted a follow-up experiment on the basis of the TVA, i.e., a whole-report task using simple letters as stimulus material, which has shown to provide reliable estimates of individual processing efficiency parameters (e.g., Bublak et al., 2005; Finke et al., 2006; Habekost & Rostrup, 2007; Hung, et al., 2005). In this experiment, we observed a higher perceptual processing speed *C* for the 'conscious report group' than for the 'no report group', while both groups did not differ in their VSTM storage capacity *K*. These findings shed more light on the locus of the individual differences that emerge when being presented with only shortly visible scene previews: It seems that those participants who were able to distinguish between different scene previews were able to do so due to a higher degree of processing speed.

According to the TVA model, which is strongly related to the biased-competition conceptualization of visual attention (Desimone & Duncan, 1995), visual objects are processed in parallel and compete for selection (i.e., conscious representation). In TVA, selection of an object is synonymous with its encoding into limited-capacity VSTM, i.e., its 'conscious' representation within the information processing system. Objects that are selected, and hence may be reported from a briefly exposed visual display, are those elements for which the encoding is completed before the sensory representation of the stimulus array has decayed and before VSTM has filled up with other objects. Thus, when visual input is only available for a very limited amount of time, the number of items that can be encoded into VSTM greatly depends on the speed of processing visual information. Even though the whole-report task of Experiment 2 used much simpler stimulus material than the scenes presented during Experiment 1, it seems that the higher processing speed observed for processing letters enabled participants from the 'conscious report group' to better distinguish between the shortly flashed scene previews by extracting more detailed information than the 'no report group'. VSTM storage capacity, on the other hand, did not seem to play a decisive role in distinguishing scene previews and effectively controlling eye movements during subsequent search. As we will discuss in further detail, a higher processing speed while leading to increased performance in the TVA whole-report task, may not only benefit when searching for target objects in naturalistic scenes.

GENERAL DISCUSSION

One goal of the present set of experiments was to examine the contribution of both global and local processing to the initial scene representation which can be rapidly established from a first glimpse of a complex scene. Focus was placed on how this initially crude visual representation can control the deployment of attention and eye movements during subsequent target search, while more detailed object information is continuously added to the evolving scene representation. Additionally, the role that individual differences play in the generation of initial scene representations and how these can modulate eye movement behavior during target search was further investigated. The results of Study 1 show that the consideration of individual differences in information processing efficiency allows a more detailed understanding of the cognitive processes that underlie the processing of visual scene information and subsequent eye movement control.

DOMINANCE OF GLOBAL PROCESSING DURING THE FIRST GLIMPSE OF A SCENE

In Experiment 1, the information provided during flashed previews of the search scenes was varied in order to investigate the influence of both local and global processing on the control of subsequent eye movements during visual search. According to the contextual guidance model (Torralba et al., 2006), target detection is achieved by estimating the probability of the presence of the target object at different locations given the combined output of both local and global processing and moving the eyes to the location with the highest target probability. However, before attention is located to a particular part of a scene, scene context activates scene priors, which then allow the restriction of search space to those locations with the highest probability of containing the target. The observed search benefit following a preview containing only a scene's background implies that processing global features to compute spatial layout and set up scene priors combined with task knowledge seems sufficient to restrict search to highly probable locations in a scene. On the other hand, when the preview contained only individual objects, but lacked spatial layout and the possibility to quickly set up scene priors, there was no observable preview benefit, thus indicating that the local processing of individual objects in a scene is not as beneficial. In terms of the contextual guidance model, this suggests that isolated objects that are not embedded in a broader scene context do not allow for enough contextual guidance to effectively control eye movements, while processing along the global pathway does allow effective eye movement control without the need to additionally segregate and compute all displayed objects. This is in line with previous studies which have shown that the computation of a scene's gist can be done very rapidly (e.g., Oliva & Schyns, 1997; Oliva & Torralba, 2006; Potter, 1975; Thorpe, Fize, & Marlot, 1996), while only a few objects can be identified within a split second, thus preventing the establishment of a complete mental representation of a scene with all identities and visual details of objects within the first glimpse (e.g., Castelhano & Henderson, 2005; Henderson & Hollingworth, 2003; Tatler et al., 2003). For example, the gist of a scene can be inferred from its spatial layout, its global scene properties, or simply the spatial distribution of colors, major scales, and orientations (e.g., Greene & Oliva, 2006; McCotter, Gosselin, Sowden, & Schyns; Oliva &

Schyns, 2000; Schyns & Oliva, 1994). Accordingly, accuracy in scene recognition is not affected by the quantity of objects in a scene and can be achieved equally well when local object recognition is hampered by blur (Oliva & Schyns, 1997; Schyns & Oliva, 1994; for a review see Oliva, 2005).

Thus, while it is not possible to fully separate the effects of local and global scene processing and more work on this topic is needed, it can be argued from this data that global scene processing is a prerequisite for the rapid generation of an initial scene representation which makes it possible to effectively control subsequent eye movements while a more local object processing does not.

A GLIMPSE IS NOT A GLIMPSE

In Experiment 1, we found a main effect of the between-subject factor on reaction times and latency to first target fixation and a strong trend for the number of fixations in that the 'conscious report group' generally showed superior search performance as compared to the 'no report group'. What are the underlying cognitive processes that cause these observable differences? Verbal reports of participants may only provide subjective and indirect information and have to be treated with reserve. However, the reported inability to differentiate between the three preview conditions does imply a reduced degree of processing during a flashed preview as compared to the 'conscious report group'. When presented with a flashed preview of a scene, it is likely that the first wave of feed-forward processing in the visual brain is followed by a series of more complex processes required for generating conscious perception of the scene (Kirchner & Thorpe, 2006; for a review see Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006). Given the limited and brief presentation of a preview, for some subjects short presentation times may be sufficient to establish a conscious percept, for others the same presentation time may not be sufficient for a conscious report of presented scene details. The efficiency of processing shortly flashed visual scenes may therefore not only influence the establishment of the initial scene representation, but may also determine the ability to consciously perceive and report differences between such scenes, for example, whether a kitchen scene was filled with a number of individual kitchen objects or whether the same kitchen was shown empty.

The hypothesis that group differences in preview reportability were due to varying degrees of processing within the first glimpse of a scene was further supported by Experiment 2 which provided evidence that the 'conscious report group' is able to process shortly flashed visual information faster than the 'no report group'. Thus, it seems that the ability to efficiently process simple letters might also enable the extraction of more detailed information from scene previews, which can be subsequently used to efficiently control eye movements in the search for a predefined target object.

INTERACTION OF GLOBAL AND LOCAL PATHWAYS AS A FUNCTION OF THE DEGREE OF PROCESSING

At first glance, it seems surprising that contrary to the sparser Background preview the Identical preview did not result in significant search benefits. Since the Identical and Background previews share the same global features, restriction of search space by a combination of setting up scene priors, spatial layout computation, and task knowledge should be possible for both previews alike. Also, Castelhano and Henderson (2007) observed clear search benefits when presenting identical as compared to the meaningless previews even when these were downscaled in size. However, the individual differences in preview processing may explain these seemingly contradictory findings.

Both experiments taken together provided clear evidence for a strong interaction between the degree of preview processing and the degree of information available in the different preview conditions. Participants who had reported being able to distinguish between the three flashed preview conditions (Experiment 1) and who showed a greater perceptual processing speed (Experiment 2) benefited most from the Background preview of the search scene, while participants who had reported not being able to distinguish between the three preview conditions and who were characterized by a lower processing speed searched the scenes most efficiently when presented with an Identical preview.

Interestingly, the 'conscious report group' — which represents about two-thirds of all participants — did not significantly benefit from an Identical preview as compared to the Control condition although it contained more information than the Background preview. It seems as if the additional objects in the Identical preview led to detrimental effects on the generation of an initial scene representation when the flashed scene was processed to a higher degree. In the contextual guidance model (Torralba et al., 2006), the setup of scene priors takes place solely on the global pathway, which parallels the processing of local objects, while the output of both pathways is later combined to interact in a scene-modulated saliency map which controls eye movement behavior. While the Background preview provides less but unequivocal information needed for the setup of scene priors due to its predominant processing on the global pathway, the Identical preview additionally provides local object information, which can be processed parallel to the global pathway before their outputs combine. We argue that due to the enhanced processing speed of the 'conscious report group' more objects can be segmented from the background and processed up to the level of identification, which in turn may activate additional priors regarding scenes or objects generated along the local pathway, i.e., plates and glasses on a dining room table could also elicit scene priors related to the context "kitchen". In line with these considerations, there is evidence that over the course of time, contingencies between objects are learned and the perception of one object can generate strong expectations about the probable presence and location of other objects (e.g., Chun & Jiang, 1999; Green & Hummel, 2006; for a review see Oliva & Torralba, 2007). Thus, the 'conscious report group' in contrast to the 'no report group' may activate competing scene priors, i.e., one generated along the global pathway and another generated along the local pathway. This competition needs to be resolved leading to detrimental effects on the effective control of subsequent search behavior. The 'no report group' on the other hand, cannot process local information to such an extent that individual objects presented in the Identical preview could elicit locally generated scene priors. In this case no detrimental competition amongst equivocal scene categories impedes effective eye movement control.

How strongly object and background information can interfere was demonstrated in a study by Joubert and colleagues (2007) in which they showed that the processing of scene context is fast enough to allow for early interactions between object and context processing. They used a go/no-go rapid visual categorization task in which participants had to distinguish as quickly as possible whether a scene that was only present for 26 ms was a 'man-made-environment' or a 'natural environment'. An interesting finding was that the presence of a salient object in a scene delayed processing of the background and induced an accuracy drop of up to 4.8%. When an object was also incongruent with the scene context, its detrimental effects on scene categorization were further increased. Similarly, Davenport and Potter (2004) had found evidence for an early interaction between scene background and objects in that inconsistent objects led to decreased performance in an object and background naming task. While not intended, some of the objects in the scenes presented in our study may not have led to the same setup of scene priors as the ones generated by a scene background. For participants with a high processing speed, this might result in detrimental background-object interactions when presented with an Identical preview containing both background and object information, which in turn could impede the effective restriction of search space thereafter.

Contrary to Castelhano and Henderson (2007), we only found search benefits following Identical previews for the participants of the 'no report group', which accounted only for a third of all participants. Thus, we did not observe an overall search benefit for Identical previews across all participants. A possible explanation for these contradictory findings could be the manipulation of previews used in our study as compared to the ones used in the Castelhano and Henderson study. While in the latter, identical, different, concept, or miniature previews were compared across several experiments, we contrasted different versions of the same scene preview, which varied only in the amount of information presented. The 'conscious report group' in particular may have been distracted by the presence of objects in the Identical preview, since they reported noticing the absence or presence of objects across previews. Another reason for the diverging results could be that we used 3D-rendered scenes, while Castelhano and Henderson used photographs of scenes. While the 3D-rendered scenes are very realistic, their scene composition may be more artificial than that of photographs. An artificially created scene generally tends to contain fewer and more isolated objects and might therefore be less cluttered than when simply using a photograph of a real living room. This may have caused the objects displayed in our scenes to be more salient than the ones used by Castelhano and Henderson thus increasing the possibility of detrimental effects especially in the 'conscious report group'.

CHAPTER 3: EFFECTS OF SEMANTIC & SYNTACTIC VIOLATIONS ON THE ALLOCATION OF ATTENTION DURING SCENE VIEWING

It has been shown that attention and eye movements during scene perception are preferentially allocated to semantically inconsistent objects as compared to their consistent controls. However, there has been a dispute over how early during scene viewing such inconsistencies are detected. In the study presented here, we introduced syntactic object-scene inconsistencies (i.e., floating objects) in addition to semantic inconsistencies to investigate the degree to which they attract attention during scene viewing. In Experiment 3 participants viewed scenes in preparation for a subsequent memory task, while in Experiment 4 participants were instructed to search for target objects. In neither experiment were we able to find evidence for extrafoveal detection of either type of inconsistency. However, upon fixation both semantically and syntactically inconsistent objects led to increased object processing as seen in elevated gaze durations and number of

fixations. Interestingly, the semantic inconsistency effect was diminished for floating objects, which suggests an interaction of semantic and syntactic scene processing. This study is the first to provide evidence for the influence of syntactic in addition to semantic object-scene inconsistencies on eye movement behavior during real-world scene viewing.

There is ample evidence that a short glimpse of a scene is enough to extract the global meaning — the so-called gist — of a scene (e.g., Oliva & Schyns, 2000; Oliva & Torralba, 2006; Potter, 1975; Thorpe et al., 1996). Extraction of gist leads to a set of expectations regarding the scene's composition, e.g., expectations of *which* objects a certain scene should contain or *where* within the scene such objects should be located. In the study presented here, we compared the effects of violating such expectations on the control of eye movements during scene viewing.

A key question was whether object-scene inconsistencies would attract early eye movements as, for example, Underwood and colleagues have reported (e.g., Underwood and Foulsham, 2006; Underwood et al., 2007; 2008), arguing for extrafoveal processing of scene inconsistencies, or whether scene inconsistencies would only exhibit additional processing once an inconsistent object has been fixated (e.g., Gareze & Findlay, 2007; Henderson et al., 1999). The discrepancy between these findings has been attributed to the differences in stimulus material used. While the scenes in the "octopus in farmyard" study (Biederman et al., 1982) were rather sparse, containing only a few objects displayed in a large field of empty space, the scenes used by Henderson and colleagues (1999) were derived from photographs and were therefore more cluttered. In sparser scenes, inconsistent objects might more readily "pop out" of the scene than when objects first have to be segregated from their background to allow for a greater degree of processing in the periphery of the visual field. One objective of this study was to put these earlier findings to the test by using highly controlled 3D-rendered images instead of photographs or line drawings.

More importantly, we were interested in directly comparing two different forms of object-scene inconsistency, namely semantic versus syntactic. While the effect of semantic inconsistency has been investigated and discussed controversially over the past decades, to date only few studies have dealt with the effect of syntactic object-scene inconsistencies. In the early 1980s, Biederman and colleagues (e.g., Biederman et al., 1982) investigated the effects of different object-scene inconsistencies including "probability" (objects tend to be found in one scene but not in others) and "support" (objects tend to rest on surfaces). These studies measured object detection performance using 150 ms scene presentations. One outcome was that both inconsistencies equally led to decreased object identification performance. Further, when an object was inconsistent in both support and probability, identification was even further decreased, arguing for the rapid detection of such object-scene inconsistencies. However, since object processing was measured by asking participants post-perceptually whether a certain object was absent or present, response bias and decision uncertainty may have produced the results (Hollingworth & Henderson, 1998; see also Henderson & Hollingworth, 1999b).

Eye movements provide a more unobtrusive, on-line measure of attention allocation and object processing. DeGraef et al. (1990) compared the effect of a variety of scene inconsistencies by measuring first fixation and gaze durations on objects embedded in line drawings of scenes in which participants were instructed to search for non-objects. Each scene contained two objects that were manipulated to create sets of inconsistencies. When the eye movement data were analyzed for objects that were fixated early versus late during viewing, there was no evidence that contextual information modulated object perception in the early stages of scene viewing. Only later did semantic as well as support violations lead to prolonged first fixation durations on these objects.

Taken together, both studies indicate that semantic as well as syntactic violations affect attention allocation during scene viewing. However, Biederman et al.'s (1982) findings of early effects were not based on eye movement data and might therefore have resulted from response biases, while DeGraef et al. (1990) never tested the effect of inconsistencies individually, but used pairs of different inconsistencies within one scene making it more difficult to interpret the effect of a single manipulation. Moreover, both studies used line drawings, which might have diminished the effect of syntactic violations due to a lack of depth perception in such reduced scenes (see Becker et al., 2007; Underwood et al., 2007).

The current study directly compared the effects of semantic and syntactic violations on eye movements during scene viewing. We used 3D-rendered images of real-world scenes instead of line drawings or photographs to create both semantic and syntactic inconsistencies of objects embedded in otherwise consistent scene contexts. Semantic violations of the scene context were created by replacing a semantically plausible object within a scene, e.g., a pot in the kitchen, with an implausible object, e.g., a printer. We operationalized syntactic inconsistencies by violating the local scene structure, i.e., having objects that normally rest on surfaces float. This resulted in four versions of each scene mirroring all possible combinations of semantic and syntactic manipulations (see Figure 2.1).

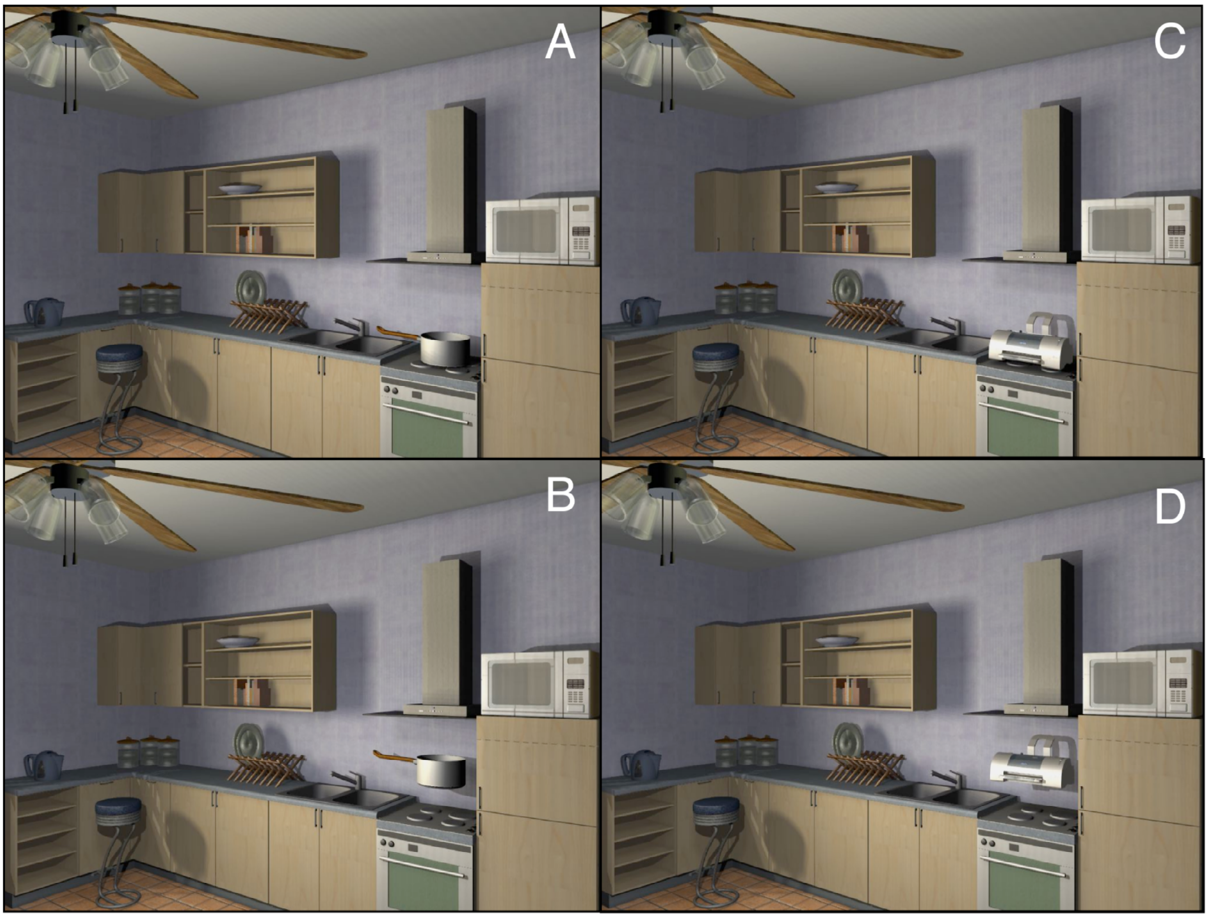

Figure 2.1: Sample of four versions of a kitchen scene containing A) a semantically consistent, non-floating object, B) a semantically consistent, floating object C) a semantically inconsistent, non-floating object, or D) a semantically inconsistent, floating object.

We hypothesized that if object identification is a prerequisite for the detection of both semantic and syntactic inconsistencies, no early effects on eye movements are expected and the eyes should not be drawn to the inconsistencies. However, once fixated, the violation of expectations regarding object-scene relations should lead to prolonged allocation of attention in order to resolve the detected anomaly. According to Itti and Baldi

(2005), the difference between prior and posterior expectations about the world constitutes "surprise" in a Bayesian framework, which subsequently leads to increased allocation of human attention and gaze to surprising events. If the degree of attention allocation to an inconsistent object represents a function of expectations or the probability of encountering such inconsistencies, floating objects should lead to more and longer fixations than semantically inconsistent objects due to the fact that we are more often exposed to semantically inconsistent than floating objects. You might, for example, have come across a cocktail glass in the lab, but have probably not encountered a floating microscope. In order to investigate these questions we recorded eye movements while participants either viewed a scene for later recognition (Experiment 3) or while searching for pre-specified target objects (Experiment 4).

EXPERIMENT 3

METHOD

Participants

Twenty-four students (21 female) from the University of Edinburgh ranging in age between 18 and 24 (*M* = 19.8, *SD* = 1.83) participated in Experiment 3 for course credit or for 6£/hour. All participants reported normal or corrected-to-normal vision and were unfamiliar with the stimulus material. Two participants had to be replaced due to unstable recording of the eye.

Stimulus Material

The stimulus material consisted of 20 3D-rendered images of real-world scenes. The scenes were displayed on a 21-inch computer screen (resolution 1024 x 768 pixel, 140 Hz) subtending visual angles of 25.66 (horizontal) and 19.23 (vertical) at a viewing distance of 90 cm. Each scene was manipulated so that it conformed to one of the four different experimental conditions: In the consistent-surface condition the object of interest was semantically consistent with the scene context and rested on a surface (e.g., a pot on a kitchen stove), whereas in the consistent-float condition the same object was displayed as hovering above the surface in mid-air. In the inconsistent-surface and inconsistent-float conditions, the semantically consistent object was replaced by an inconsistent object (e.g., a printer on a kitchen stove) resting on a surface or hovering in mid-air, respectively. Figure 2.1 displays a sample scene in its four versions.

Scenes were then paired so that the semantically inconsistent object of one scene was consistent in its paired scene (e.g., a printer on an office desk). Semantically consistent and inconsistent objects were matched for their size and were placed in the same position within each scene away from the center, where the initial fixation was to be made. Further, scenes were processed using the Itti and Koch (2000) MatLab Saliency Toolbox in order to determine the most salient regions according to low-level saliency calculations of brightness, color, contrast, and edge orientation. The rank order of saliency peaks — with rank 1 assigned to the most salient region of the scene — was used to ensure that consistent and inconsistent objects did not differ in their mean low-level saliency ($M = 8.45$, $SD = 3.09$ vs. $M = 8.9$, $SD = 2.38$, $p > .05$).

Apparatus

Eye movements were recorded with an EyeLink1000 tower system (SR Research, Canada) which tracks with a resolution of .01° visual angle at a sampling rate of 1000 Hz. The position of the right eye was tracked while viewing was binocular. Experimental sessions were carried out on an IBM compatible display computer running on OS Windows XP. Stimulus presentation and response recording were controlled by Experimental Builder (SR, Research, Canada).

Procedure

Each participant received written instructions before being seated in front of the presentation screen. Participants were informed that they would be presented with a series of scenes which they had to memorize for a later memory test.

At the beginning of the experiment, the eye tracker was calibrated for each participant using 9-point calibration and validation. The participants' viewing position was fixed with a chin and forehead rest. Each trial sequence was preceded by a fixation check, i.e., in order to initiate the next trial, the participants had to fixate a cross centered on the screen for 200 ms. When the fixation check was deemed successful, the fixation cross was replaced by the presentation of a scene for 15 seconds during which the participants inspected the scene freely in preparation for a memory task. After an inter-trial-interval of one second the next trial followed. Two practice trials at the beginning of the experiment allowed participants to become accustomed to the experimental set-up. The experiment lasted about 15 minutes. Subsequently, an offline memory test was administered without recording eye movements. Since we were only interested in the eye movement data during scene memorizing, the data from the memory test will not be reported here.

Eye movement data analysis

The interest area for each target object was defined as the rectangular box that was large enough to encompass the consistent and inconsistent target objects when located on a surface as well as when floating. Thus, the scoring regions were the same for all conditions to allow for better comparison. Fixation durations of less than 90 ms and more than 1000 ms were excluded as outliers. Raw data was subsequently filtered using SR Research Data Viewer and then submitted to an analysis of variance (ANOVA) with semantic consistency (consistent vs. inconsistent) and syntactic consistency (surface vs. float) as within-subject factors.

RESULTS

A set of measures was calculated in order to analyze viewers' eye movement patterns as a function of both the semantic and syntactic consistency manipulations. We have divided these measures into those that mirror extrafoveal processing of inconsistencies on the one hand and foveal processing of inconsistencies on the other.

Extrafoveal Processing of Scene Inconsistencies

The main aim of the current study was to investigate whether initial eye movements during scene viewing would already be modulated by the processing of peripheral scene inconsistencies. In order to investigate whether semantic as well as syntactic inconsistencies affect eye movements prior to their fixation, four measures were examined (see Table 2.1): Probability of Immediate Target Fixation, Latency to First Target Fixation, Number of Fixations to First Target Fixation, and Entering Saccade Amplitude.

Table 2.1: Summary of mean values [standard errors] for dependent variables in Experiment 3 reflecting extrafoveal processing as a function of semantic (consistent vs. inconsistent) and syntax (surface vs. float) manipulations. Dependent variables were Probability of Immediate Target Fixation, Latency to First Target Fixation, Number of Fixations to Target Fixation, and Entering Saccade Amplitude.

MEASURES	SEMANTIC		F	SYNTAX		F
	consistent	inconsistent		surface	float	
Probability of Immediate Target Fixation in %	11.70 [3.78]	7.04 [2.62]	3.59	10.63 [3.89]	7.88 [2.50]	2.66
Latency to First Target Fixation in ms	3808 [345]	4124 [334]	< 1	4034 [365]	3898 [313]	< 1
Number of Fixations till Target Fixation	10.40 [.98]	11.27 [.94]	< 1	11.05 [1.05]	10.61 [.86]	< 1
Number of Fixations to Target Fixation	10.40 [.98]	11.27 [.94]	< 1	11.05 [1.05]	10.61 [.86]	< 1
Entering Saccade Amplitude in degree visual angle	5.39 [.27]	5.67 [.38]	1.24	5.45 [.28]	5.61 [.37]	< 1

* p < .05, ** p < .01

Additionally, we analyzed the probability of immediate target fixation defined as the percentage of trials in which the initial saccade landed on the target object. There was neither an effect of semantic, $F(1,23) = 3.59, p > .05$, nor an effect of syntactic manipulation, $F(1,23) = 2.66, p > .05$, and no interaction, $F < 1$. In addition, we analyzed the cumulative probability of target fixation after the second saccade and also found no effects, all Fs < 1. The probabilities of target fixation as a function of ordinal fixation number can be seen in Figures 2.2 and 2.3. There is no indication of an early effect of either semantic or syntactic manipulations on initial eye movements.

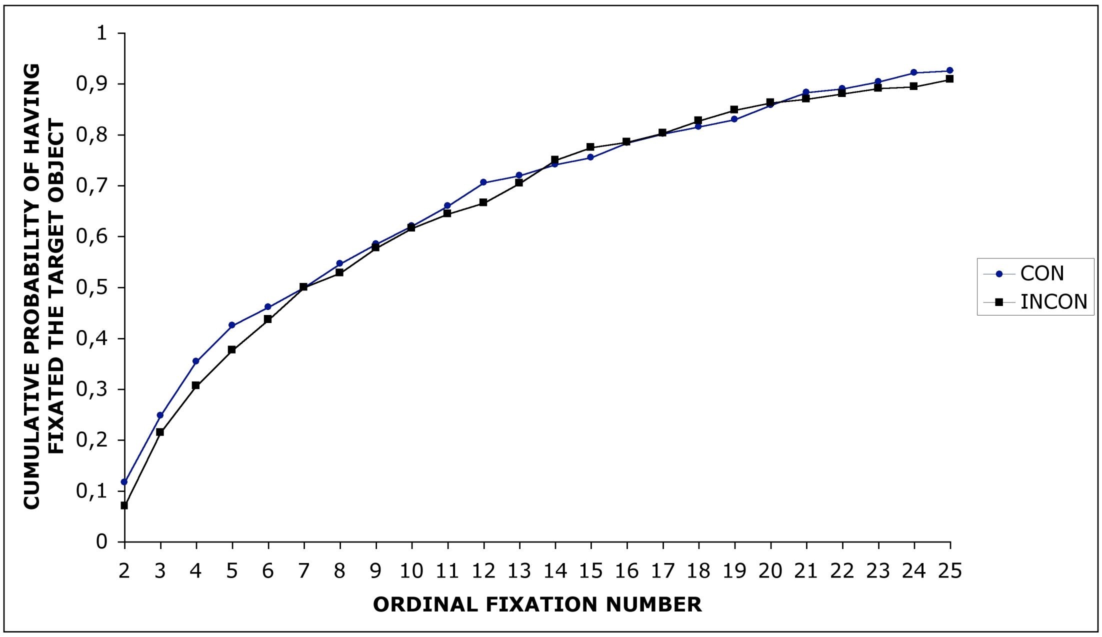

Figure 2.2: Cumulative probability of having fixated the target object as a function of the ordinal fixation number and semantic consistency (semantically consistent = CON, semantically inconsistent = INCON) in Experiment 3. Note that first fixations on the target were excluded because these were located on the initial fixation cross.

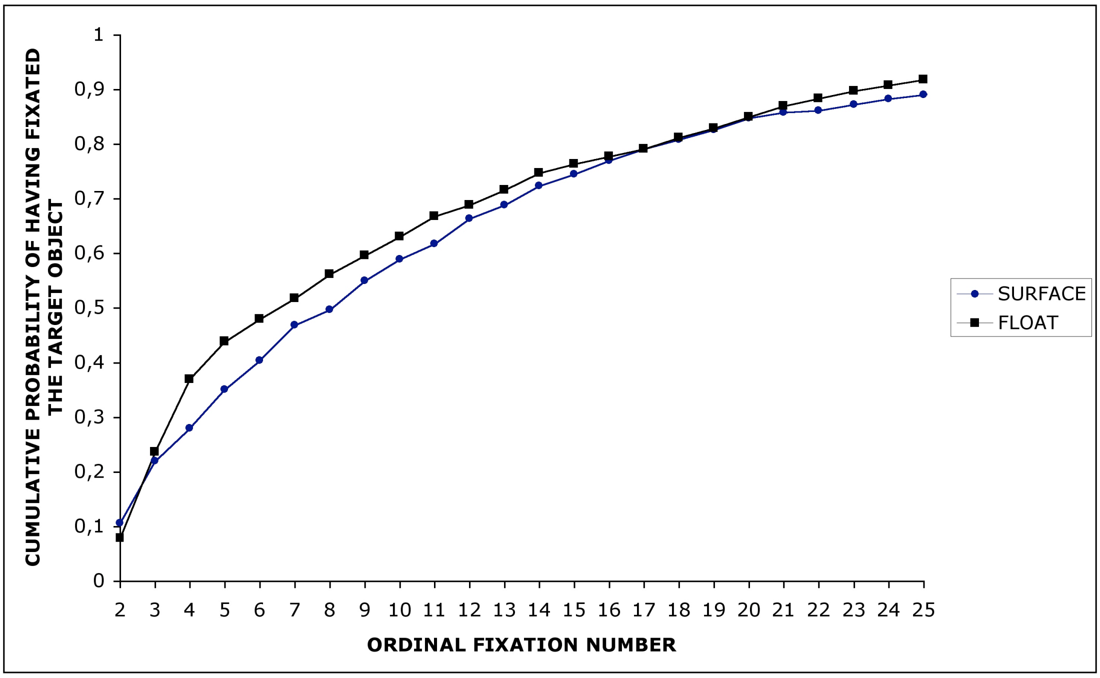

Figure 2.3: Cumulative probability of having fixated the target object as a function of the ordinal fixation number and syntactic manipulation (support, float) in Experiment 3.

Latency to First Target Fixation. Latency was measured from scene onset until the first fixation of the target object and averaged 3966 ms across all conditions. There was neither an effect of semantic, $F(1,23) = 1.15$, $p > .05$, nor an effect of syntactic consistency, $F < 1$, and no interaction, $F < 1$.

Number of Fixations to First Target Fixation. This measure is defined as the discrete number of fixations until the target object was first fixated. The values include both the initial fixation on the scene and the first fixation on the target object. On average, participants performed 10.83 fixations to the first fixation of the target object. There was neither an effect of semantic, nor an effect of syntactic consistency, and no interaction, all *F*s < 1.

Entering Saccade Amplitude. The amplitude of the saccade that first entered the target region was defined as in-coming saccade amplitude, and averaged 5.53 degrees visual angle. There was no effect of semantic consistency, $F(1,23) = 1.24$, $p > .05$, no effect of syntactic consistency, $F < 1$, and no interaction, $F < 1$. Taken together, none of these measures provided evidence that extrafoveal processing of either semantic or syntactic inconsistencies could draw the eyes to peripheral scene regions.

Foveal Processing of Scene Inconsistencies

In order to investigate whether the semantic or syntactic manipulations affected object processing once the object was fixated, we calculated five additional measures that mirror the degree of attention allocated to the target objects. The measures include total gaze duration and gaze count, first-pass gaze duration and gaze count, as well as first fixation duration (see Table 2.2).

Table 2.2: Summary of mean values [standard errors] of Experiment 3 regarding dependent variables on foveal processing as a function of semantic (consistent vs. inconsistent) and syntax (surface vs. float) including Total Gaze Duration and Gaze Count, First-Pass Gaze Duration and Gaze Count, and First Fixation Duration.

MEASURES	SEMANTIC		F	SYNTAX		F
	consistent	inconsistent		surface	float	
Total Gaze Duration	1633	1887	6.36*	1489	2030	42.43**
in ms	[87]	[103]		[83]	[108]	
Total Gaze Count	5.36	6.16	7.07**	5.00	6.52	29.91**
	[.32]	[.37]		[.29]	[.40]	
First-Pass Gaze	586	798	8.24*	577	806	12.26*
Duration in ms	[56]	[77]		[60]	[73]	
First-Pass Gaze Count	1,97	2.53	6.10*	1.91	2.59	10.87**
	[.18]	[.21]		[.16]	[.23]	
First Fixation Duration	280	293	< 1	268	305	10.99**
in ms	[14]	[14]		[12]	[17]	

* p < .05, ** p < .01

Total Gaze Duration. The total gaze duration was defined as the sum of all fixation durations on the target region from scene onset until scene offset. Across all conditions the mean total gaze duration was 1760 ms. There was a main effect of semantic consistency, $F(1,23) = 6.36$, $p < .05$, in that semantically inconsistent objects were fixated for a longer amount of time than objects that were consistent with the semantics of the scene. In addition, we observed a strong effect of the syntactic manipulation, $F(1,23) = 42.43, p < .01$, according to which floating objects were looked at longer than objects resting on surfaces. The interaction failed to reach significance, $F(1,23) = 1.94, p > .05$.

Total Gaze Count. Total gaze count was defined as the sum of all fixations located in the target region from scene onset until scene offset and averaged 5.76 fixations. Similar to the total gaze duration, we observed main effects for both the semantic, $F(1,23) = 7.07, p = .01$, and the syntactic manipulation, $F(1,23) = 29.91, p < .01$, while the interaction was not significant, $F < 1$. Semantically inconsistent as well as floating objects led to a greater number of fixations than semantically consistent objects or objects resting on a surface.

First-Pass Gaze Duration. In order to investigate the effect of inconsistency processing on the encoding of objects, we calculated the first-pass gaze duration, which was defined as the sum of all fixation durations from the first entry of the eyes to the target region until their first exit. It has been shown that the first-pass gaze duration increases when processing semantic inconsistencies (e.g., DeGraef et al., 1990; Henderson et al., 1999; Loftus & Mackworth, 1978). On average, participants spent 692 ms on the target before leaving the target region for the first time. As with the total gaze duration we found effects for both the semantic, $F(1,23) = 8.24, p < .01$, and the syntactic inconsistency, $F(1,23) = 12.26, p < .01$. In addition, there was a significant interaction of both factors, $F(1,23) = 8.64, p < .01$. As can be seen in Figure 2.4, the interaction was characterized by a strong effect of semantic inconsistency for objects resting on surfaces, $t(23) = 4.06, p < .01$, while this effect was eliminated for floating objects, $t(23) < 1$.

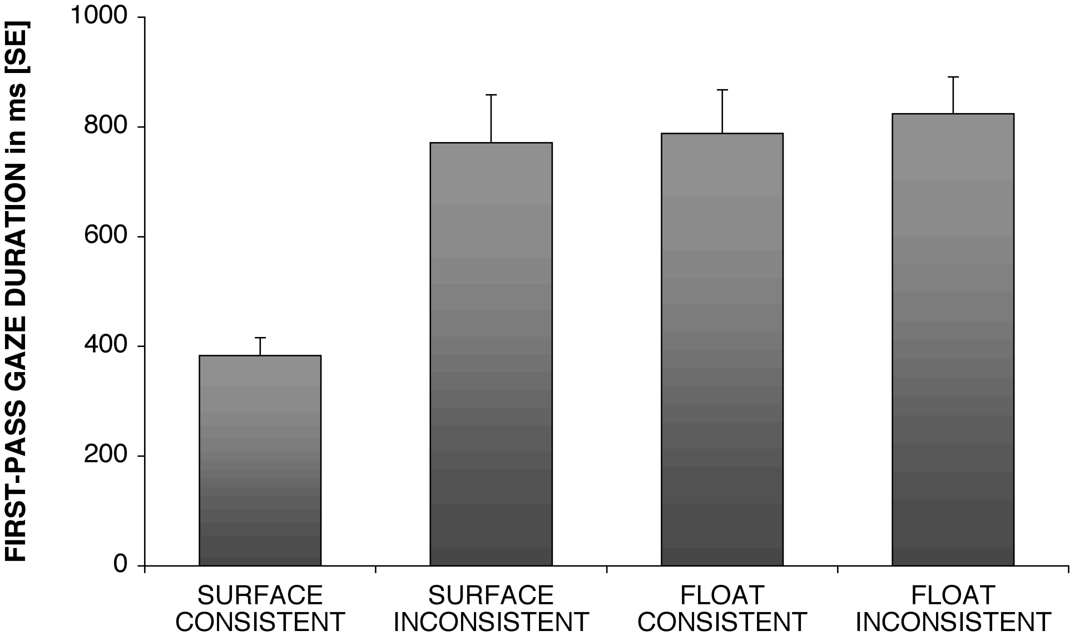

Figure 2.4: Mean First-Pass Gaze Durations [standard errors] for Experiment 3 as a function of semantic (consistent versus inconsistent) and syntax (consistent versus inconsistent) manipulations.

First-Pass Gaze Count. The first-pass gaze fixation count was defined as the number of fixations from the first entry of the eyes to the target region until their first exit. Similar to the first-pass gaze duration, we observed significant main effects for semantic inconsistency, $F(1,23) = 6.10$, $p < .05$, and syntactic inconsistency, $F(1,23) = 10.87$, $p < .01$, as well as a significant interaction, $F(1,23) = 5.69$, $p < .05$. Participants fixated semantically as well as syntactically inconsistent objects more often than consistent objects and objects resting on surfaces. While there was a significant effect of semantic inconsistency for objects on surfaces, $t(23) = 3.75$, $p < .01$, this effect disappeared for floating objects, $t(23) < 1$.

First Fixation Duration. As another indicator of initial object encoding we analyzed the first fixation duration defined as the duration of only the initial fixation made on the target object. First fixation durations are believed to provide a measure that more directly reflects object-identification time (e.g., Friedman, 1979; Henderson et al., 1989). The average first fixation duration amounted to 286 ms. While there was no significant main effect of semantic consistency, $F < 1$, and no significant interaction, $F(1,23) = 1.19$, $p > .05$, we found a significant main effect of the syntactic manipulation, $F(1,23) = 10.99$, $p < .01$, in that the first fixation duration was increased for floating objects compared to objects resting on surfaces. This suggests that syntactically inconsistent objects might require a greater degree of object processing during initial encoding.

In sum, the data show strong effects of both the semantic and syntactic consistency manipulation once the target object has been fixated. Further, it seems that the effect of semantic inconsistency is weakened when the target object is already syntactically inconsistent.

DISCUSSION

The aim of Experiment 3 was to investigate whether scene inconsistencies would attract early eye movements prior to the fixation of inconsistent objects when participants were asked to view a scene for a later memory test. There was no evidence that initial eye movements were drawn to objects that either violated expectations of scene semantics or syntax. Object-scene inconsistencies were neither fixated earlier during scene viewing nor from a greater distance than their consistent counterparts, arguing against an extrafoveal processing of scene inconsistencies. This largely replicates the findings by Henderson and colleagues (1999), who also found no early effects of semantic inconsistencies prior to their fixation using line drawings of naturalistic scenes. However, upon fixation, objects that did not fit the semantic context of the scene attracted a higher degree of attention than consistent objects, as seen in a greater number of fixations and therefore longer gaze durations. Again, this is in line with findings of increased object processing for inconsistent objects once fixated (e.g., DeGraef et al., 1990; Gareze & Findlay, 2007; Henderson et al., 1999; Hollingworth et al., 2001; Loftus & Mackworth, 1987; Underwood & Foulsham, 2006; Underwood et al., 2007, 2008).

Additionally, we found that objects that violated expectations of their syntactic properties, i.e., were floating, also resulted in increased processing compared to objects resting on surfaces. Especially when first encountering an object, the effects of the semantic and syntactic manipulations interacted in such a way that while non-floating objects showed clear modulation according to their semantic fit to the scene context, this semantic inconsistency effect was eliminated when the object was floating. A semantically consistent, but floating object held gaze to the same degree as an object that was resting on a surface, but incongruent with the scene semantics. A double inconsistency, i.e., a semantically inconsistent and floating object, did not yield more processing time than each individual inconsistency. Thus it seems that once an object violated syntactic regularities, it no longer mattered whether it fit the overall gist of the scene or not. The syntactic manipulation also affected the first fixation duration on an object, which was prolonged for floating objects, whereas the semantic manipulation did not affect initial encoding time. This implies that initial encoding of syntactically inconsistent objects required more time than the encoding of syntactically consistent objects.

A possible explanation for the strong impact of the syntactic manipulation is the probability of encountering such an inconsistency in everyday life. While we do come across misplaced objects from time to time, we rarely encounter floating objects. The stronger and more restricted the scene priors are, the greater the effect of their violations seems to be. Contrary to this view, Biederman and colleagues (1982) did not find stronger effects for the support compared to the semantic manipulation in their tachistoscopic object detection paradigm. Besides the lack of control for response biases, another reason for the lack of stronger effects of additional processing for floating objects could have been the task itself. While participants in the Biederman et al. study had to decide whether a certain object was absent or present, participants in our study were asked to memorize objects for later recognition. Floating objects tend to enable easier figure-ground-segmentation due to their position in the scene, e.g., in the Biederman et al. study the couch floating in the sky.

This is particularly true in line drawings as used in the Biederman et al. study and could have increased performance in the object detection task counteracting the detrimental effect of having to resolve the syntactic violation. As a result, semantic and syntactic inconsistencies yielded similar detection performance.

In sum, the data of Experiment 3 did not lend support to the claim that extrafoveal processing of object inconsistencies in scenes can guide eye movements to these inconsistent objects. Rather, our data clearly speak for a limited region around the fovea in which semantic as well as syntactic inconsistencies can be processed to such a degree that attention allocation and eye movements are modulated.

EXPERIMENT 4

The data of Experiment 3 seem to imply that object-scene inconsistencies cannot be processed when they are outside of foveal viewing, but they exhibit strong effects on attention allocation and eye movement control once fixated. According to Henderson and colleagues (1999), an alternative interpretation of the data from Experiment 3 could be that the scene inconsistencies could not exhibit an early effect on initial eye movements because participants were not motivated to fixate inconsistent objects quickly due to the unspeeded nature of the memorization task. In contrast to the 15 s viewing time in our study, participants in the Loftus and Mackworth (1978) study only had 4 s to inspect a scene, which could have increased the need for extrafoveal processing.

In order to address this possibility, we conducted a second experiment using the same experimental design. Instead of allowing participants to view each scene for 15 seconds, we asked them to search for pre-specified target objects as quickly as possible. The additional motivation to quickly find the target object should increase the effect of object-scene inconsistency on the attraction of eye movements, in that semantically and syntactically inconsistent objects should be fixated earlier and with a greater incoming saccade amplitude than their consistent counterparts.

METHOD

Participants

Twenty-four students (16 female) from the University of Edinburgh ranging in age between 19 and 26 (*M* = 21.8, *SD* = 2.5) participated in Experiment 4 for course credit or for 6£/hour. All participants reported normal or corrected-to-normal vision and none had taken part in Experiment 3. One participant had to be replaced due to misunderstandings of target words.

Stimulus Material

The search scenes were identical to the scenes used in Experiment 3. All target objects of Experiment 3 served as search targets in Experiment 4, which were specified by target words preceding each search scene. The 20 target words were displayed in uppercase black Arial typeset centered on a gray background (RGB: 51, 51, 51). Target words were chosen to be comprehensible and unambiguous in indicating the target object.

Apparatus

The apparatus was identical to the one used in Experiment 3. A joypad was added to collect reaction time data.

Procedure

As in Experiment 3, each participant received written instructions before being seated in front of the presentation screen. Participants were informed that they would be presented with a series of scenes, each of which contained a pre-specified target object that they had to find as quickly as possible. Once found, they were to press a button on a joypad.

At the beginning of the experiment the eye tracker was calibrated for each participant. Each trial sequence was preceded by a fixation check. When the fixation check was deemed successful, the fixation cross was replaced by the presentation of a word (2000 ms) indicating the identity of the target object. An additional fixation cross followed (500 ms) to make sure that after reading the target word the eyes were repositioned at the centre of the screen when the search scene appeared. Participants were instructed to search the scene for the target object as quickly as possible and to indicate the detection of the target object by holding fixation on the object and pressing a joypad button. The search scene was displayed for 15 s or until button press. After an inter-trial-interval of one second the next trial followed. Two practice trials were administered at the beginning of the experiment which lasted a total of about 10 minutes. Again, an offline memory test without recording eye movements followed. The data of the memory test will not be reported here.

RESULTS

In the following analyses, trials were excluded that did not result in successful target search or were subject to unstable tracking of the eye (3.94%). As in Experiment 3, raw data was preprocessed by the SR Research Data Viewer and then submitted to an analysis of variance (ANOVA) with semantic consistency (consistent vs. inconsistent) and syntactic consistency (surface vs. float) as within-subject factors.

Eye movement data recorded after fixation of the target object were sparse and truncated, because the scene disappeared once participants had pushed the button to indicate that they had found the target object. Due to this artificial termination of fixations by the button press we only report extrafoveal processing measures for Experiment 4.

Extrafoveal Processing of Scene Inconsistencies

In addition to the dependent variables reported in Experiment 3, reaction times are reported since participants had to press a button as soon as the target object had been found (see Table 2.3).

Table 2.3: Summary of mean values [standard errors] of Experiment 4 regarding dependent variables on extrafoveal processing as a function of semantic (consistent vs. inconsistent) and syntax (surface vs. float) including Probability of Immediate Target Fixation, Reaction Time, Latency to First Target Fixation, Number of Fixations to Target Fixation, and Entering Saccade Amplitude.

MEASURES	SEMANTIC		F	SYNTAX		F
	consistent	inconsistent		surface	float	
Probability of Immediate Target Fixation in %	17.02 [3.83]	19.79 [4.19]	< 1	16.77 [3.98]	20.04 [4,03[	< 1
Reaction Time in ms	1964 [146]	2067 [121]	2.71	1936 [122]	2094 [144]	< 1
Latency to First Target Fixation in ms	1237 [104]	1327 [104]	1.13	1248 [89]	1317 [119]	< 1
Number of Fixations till Target Fixation	3.83 [.21]	3.98 [.23]	< 1	3.97 [.23]	3.84 [.02]	< 1
Number of Fixations to Target Fixation	3.83 [.21]	3.98 [.23]	< 1	3.97 [.23]	3.84 [.02]	< 1
Entering Saccade Amplitude in degree visual angle	6.02 [.37]	6.45 [.39]	1.34	6.45 [.43]	6.02 [.34]	1.38

* p < .05, ** p < .01

Probability of Immediate Target Fixation. As in Experiment 3, there was no effect of either the semantic or syntactic manipulation on the probability that the first saccade would be directed at the target object, nor an interaction, Fs < 1. The same was true for the cumulative probability of the second saccade landing on the target object, Fs < 1. The probabilities of target fixations as a function of ordinal fixation number can be seen in Figures 2.5 and 2.6.

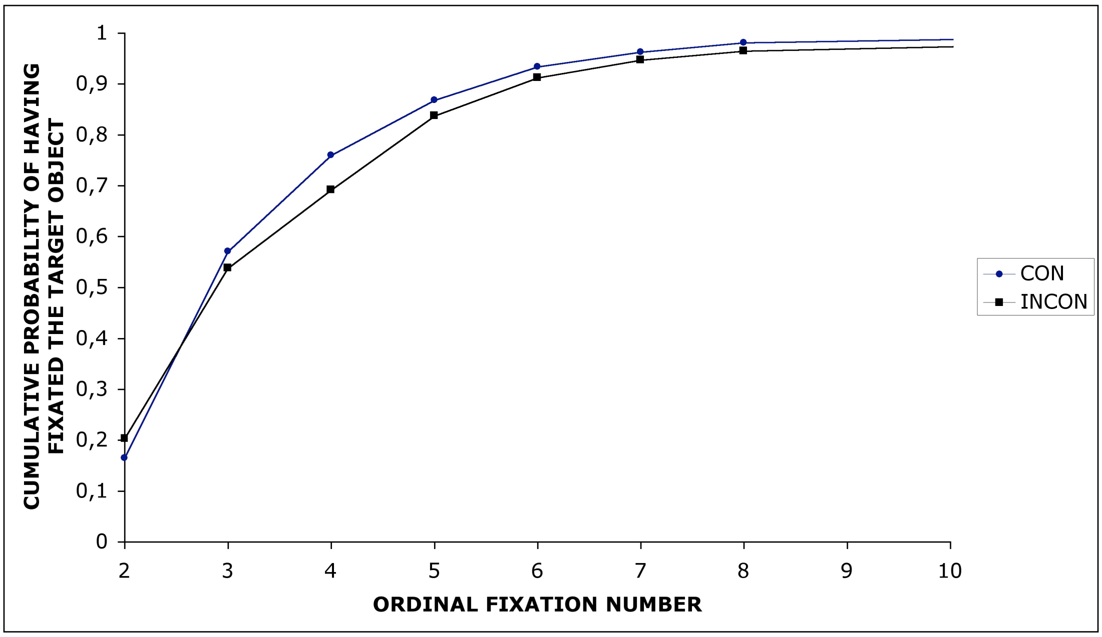

Figure 2.5: Cumulative probability of having fixated the target object as a function of the ordinal number and semantic consistency (semantically consistent = CON, semantically inconsistent = INCON) in Experiment 4.

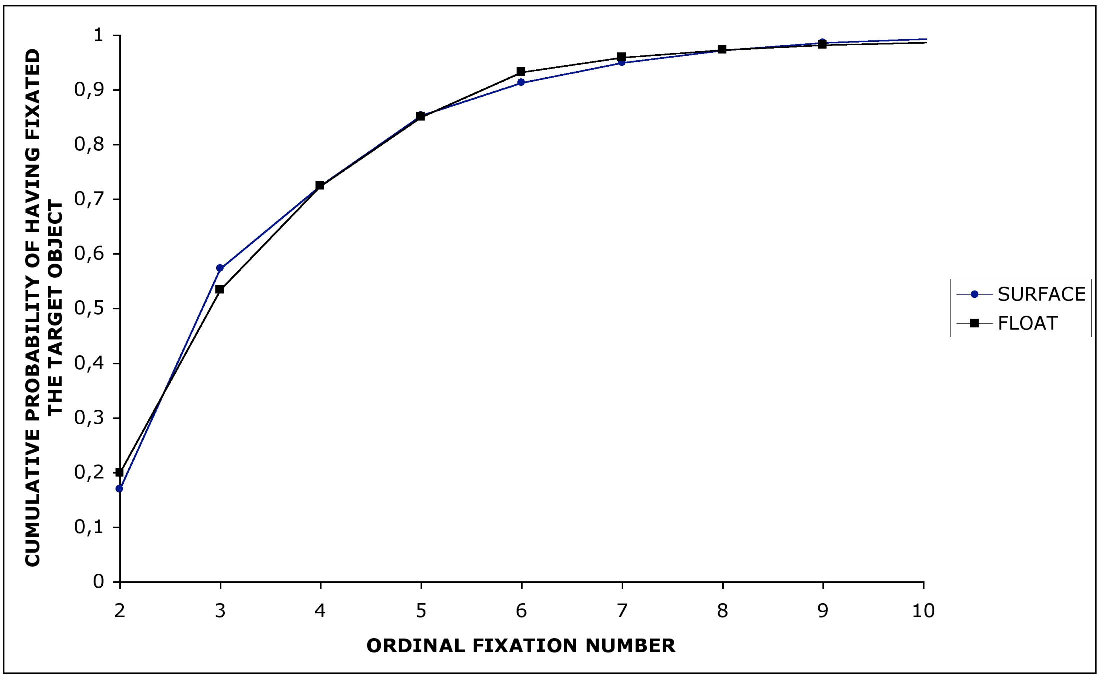

Figure 2.6: Cumulative probability of having fixated the target object as a function of the ordinal number and syntax manipulation (support, float) in Experiment 4.

Reaction Time. RT was defined as the time elapsed from scene onset until button press and averaged 2015 ms across all conditions. There was neither an effect of semantic, $F < 1$, nor an effect of syntactic consistency, $F(1,23) = 2.71, p > .05$, and no interaction, $F < 1$.

Latency to First Target Fixation. The time to the first fixation of a target object was much shorter than in Experiment 3 amounting to an average latency of 1282 ms. However as in Experiment 1, there were no main effects for either the semantic or syntactic manipulation, $F(1,23) = 1.11, p > .05$ and $F < 1$, respectively, and no interaction, $F(1,23) = 1.37, p > .05$.

Number of Fixations to Target Fixation. The average number of fixations needed to find the target object amounted to 3.90 fixations. Neither the main effects nor the interaction reached significance, all *F*s < 1.

Entering Saccade Amplitude. The amplitude of the saccade entering the target region for the first time averaged at 6.25° visual angle across all conditions. Again, there was no effect of semantic inconsistency, $F(1,23) = 1.34$, $p > .05$, no effect of syntactic inconsistency, $F(1,23) = 1.38$, $p > .05$, and also no interaction, $F(1,23) = 1.56$, $p > .05$.

These data suggest that despite the instruction to search for target objects as quickly as possible, eye movements were not modulated by extrafoveal processing of scene inconsistencies.

DISCUSSION

The purpose of Experiment 4 was to test whether the lack of extrafoveal effects of scene inconsistencies in Experiment 3 was due to the task, which did not motivate participants to actively move their eyes quickly to objects displayed in the scene. In order to check whether task instruction really had an effect on eye movement behavior when inspecting the scenes, we compared mean gaze durations and saccade amplitudes for Experiment 3 versus Experiment 4. We found that the average time the eyes spent fixating on the scene — excluding the time spent fixating target objects — differed significantly between Experiments 3 and 4, $t(1) = 10.39$, $p < .01$, with longer mean fixation durations for eye movements in the memorization task (*M* = 322 ms) than in the search task (*M* = 268 ms). Mean saccade amplitude — excluding those saccades that originated from or entered the target object — was shorter during memorization (*M* = 3.40° visual angle) than during search (*M* = 4.36° visual angle), $t(1) = 12.88$, $p < .01$. Thus, the task significantly affected eye movement behavior when viewing the scenes.

With the task to search for target objects in Experiment 4, participants might have tried to process more information from the periphery of their visual field, which could have increased the detection of inconsistencies outside the focus of a current fixation. However, despite the task to actively search for target objects, there was again no evidence that scene inconsistencies attracted eye movements. Neither semantically nor syntactically inconsistent target objects were found faster than their consistent controls. Also, the amplitude of the saccade entering the target region did not vary as a function of the semantic or syntactic manipulation, arguing against the view that scene inconsistencies attract attention prior to their fixation. This is in line with findings by Henderson and colleagues (1999), who also found no evidence for extrafoveal processing of semantic inconsistencies despite engaging participants in an active search task.

GENERAL DISCUSSION

The departure point of the present study was to shed new light on the discussion regarding the processing of object-scene inconsistencies during scene viewing. Therefore, I created 3D-rendered images of naturalistic scenes with a high degree of realism, while allowing for highly controlled manipulations of objects within a scene. A key question was whether inconsistent objects in the periphery of the visual field would be able to control initial eye movements prior to their fixation. In addition, we were interested in the direct comparison of two different scene inconsistencies: violations of the scene semantics on the one hand, and violations of scene syntax on the other. In the following we will discuss these issues in greater detail.

FOVEAL VERSUS EXTRAFOVEAL PROCESSING OF SCENE INCONSISTENCIES

Is there "semantic pop-out" in scene perception? That is, can an object-scene inconsistency be detected before the object has been foveally processed and identified? According to Loftus and Mackworth (1978), an object that does not fit the semantics of a scene exhibits control over eye movements very early during scene viewing, affecting initial eye movements prior to fixation of the inconsistent object. They not only found that inconsistent objects were fixated longer, but also earlier (in fact, immediately) when inspecting a scene for later recognition. Also, saccades entering the region of an inconsistent object were longer in amplitude than when entering the region of a consistent object, thus arguing in favor of an extrafoveal processing of inconsistencies in the visual periphery.

However, this interpretation has not gone unchallenged and there have been a number of studies showing that initial eye movements are not influenced by the processing of scene inconsistencies prior to their fixation (e.g., DeGraef et al., 1990; Gareze & Findlay, 2007; Henderson et al., 1999). Recently, Underwood and colleagues reinstated the claim that full identification of objects is not necessary in order to process object-scene inconsistencies in the visual periphery leading to attraction of eye movements prior to the objects' foveal processing (e.g., Underwood & Foulsham, 2006; Underwood et al., 2007; 2008). They used photographs instead of line drawings, manipulating both visual bottom-up saliency and the consistency of objects embedded in scenes, and found that semantically inconsistent objects were fixated earlier than their consistent counterparts.

In contrast to the early findings of Loftus and Mackworth (1978) as well as recent findings by Underwood and colleagues (e.g., Underwood & Foulsham, 2006; Underwood et al., 2007; 2008), we have not found early effects of scene inconsistencies, either when participants viewed a scene for later recognition (Experiment 3), or when they were instructed to actively search for target objects (Experiment 4). Thus, we argue that in complex naturalistic scenes, foveal processing of object-scene inconsistencies is necessary in order to influence the allocation of attention and exhibit control over initial eye movements.

How can these conflicting results be explained? The source of differences may lie in a combination of differences in presentation times, tasks, and stimulus material used across studies. Loftus and Mackworth (1978) as well as Underwood and colleagues (2007) used shorter presentation times (4 s and 5 s, respectively) in the memorization task than

Henderson and colleagues (1999) or we did (both 15 s). Shorter presentation times might have motivated participants to move their eyes around more quickly, thus widening the scope of attention allocation in order to process more information. However, the lack of evidence for extrafoveal processing of scene inconsistencies when participants were instructed to search for target objects in Experiment 4 (see also Henderson et al., 1999) argues against the idea that the different results across studies are solely due to the use of different presentation times or tasks.

The differences in stimulus material may have been a greater source of variance across studies. As has been discussed earlier, the scenes Loftus and Mackworth (1978) used were rather sparse and may have increased the impact of extrafoveal processing since there were only a few easily identifiable objects present. Also, inconsistent objects might have been more visually conspicuous attracting eye movements by means of low-level visual salience, since this was not controlled for. In contrast, DeGraef et al. (1990), Henderson et al. (1999), and Gareze and Findlay (2007) used more complex line drawings, which could have decreased the effect of extrafoveal processing. Underwood and colleagues (2007), on the other hand, used color photographs that were edited post-hoc, which may have introduced artificial low-level conspicuousness without the Itti and Koch (2000) algorithm detecting it, but with an effect on human observers. Since only the inconsistent objects were subject to post-hoc editing, these might have visually popped out as compared to consistent objects, which were originally part of the scene. Also, some of the inconsistent objects in the Underwood et al. study seem visually odd, due to inappropriate shadows and other artifacts from the pasting process.

In our study, we used 3D-rendered scenes which allowed for manipulations of objects within the scene without the need to edit the stimulus material post-hoc. Additionally, the scenes displayed a high degree of photorealism regarding colors, texture, and illumination of the scene and the embedded objects. Thus, consistent as well as inconsistent objects alike blended into the scenes.

However, comparing the mean saliency rank values of objects in our study (mean rank about 8.5) — calculated using the Itti and Koch (2001) algorithms as the rank of the scene region that contained the target object in relation to the rest of the scene — with the mean rank values of objects in scenes Underwood et al. (2007) used (mean rank about 3), reveals that the objects of interest in our scenes ranged relatively low in visual salience compared to other scene regions, while objects in the study by Underwood and colleagues ranged higher in visual salience within the scene context. According to Underwood and Foulsham (2006), objects of low salience values should especially exhibit effects of semantic inconsistency prior and upon fixation of the object. Thus, our stimulus material should have been more apt to produce effects of extrafoveal processing than that used by Underwood and Foulsham. However, this was not the case. The reason may lie in the definition of high and low salience. While Underwood and Foulsham defined high and low salience objects by comparing saliency rank values between two objects of a scene, it might be more important to relate the visual salience of an object to the entire scene in which it is embedded. In our case, a mean rank value of about eight implies that seven other regions in the scene were visually more conspicuous, while in the study by Underwood et al. (2007) on average only two other regions were more conspicuous. Thus, at least during free scene viewing, the effect of scene inconsistencies might depend on the relative visual salience of the inconsistent object. Specifically, there might be a greater impact of extrafoveal scene inconsistencies when there are not as many higher salient regions in the scene attracting gaze on the basis of low-level features. Follow-up studies explicitly manipulating the saliency ranks of inconsistent objects in relation to other parts of the scene may be able to shed more light on this possible source of variance across studies.

GRAVITY MATTERS: DIFFERENTIAL PROCESSING OF SEMANTIC AND SYNTACTIC OBJECT-SCENE VIOLATIONS

Most of the evidence on the allocation of gaze to inconsistent objects in naturalistic scenes has come from semantic violations of the scene context. However, another kind of inconsistency regards the local structural setting, i.e. the syntactic context, in which an object is placed within a scene. Certain constituents of a scene require certain syntactic structures. In this study, we directly compared the effects of semantic and syntactic violations on eye movement control.

While semantic inconsistencies referred to objects that did not fit the semantic context of the scene, syntactic inconsistencies were created by making objects float. Similar to our findings regarding semantic inconsistencies, we have produced the first evidence that syntactic inconsistencies do not attract gaze prior to the fixation of the inconsistent object. Thus, neither semantic, nor syntactic anomalies seem to be sufficiently processed outside of foveal viewing to control eye movements. Rather the detection of local syntactic structures seems to require fixation of the critical object. This would be in line with findings by

Tatler, Gilchrist, and Land (2005) according to which direct fixation of an object is required to extract meaningful position information, since only then can the position information of the stored representation be compared to position information of the one currently processed. This is also consistent with the results from transsaccadic change detection experiments, where it has been shown that changes to the structural relationship between an object and its scene (e.g., rotation of an object in a scene) is typically only detected when the object has been fixated before and after the change (Hollingworth & Henderson, 2002).

While we were not able to find early effects of scene inconsistencies, both semantic and syntactic scene violations led to increased gaze to inconsistent objects once fixated. Previous studies have reported increased fixation densities and durations for semantically inconsistent objects, implying prolonged allocation of attention necessary to resolve the object-scene inconsistency (e.g., Gareze & Findlay, 2007; DeGraef et al., 1990; Henderson et al., 1999; Hollingworth et al., 2001; Loftus & Mackworth, 1987; Underwood & Foulsham, 2006; Underwood et al., 2007; 2008).

Extending previous work, we found an increased degree of attention allocation to objects that were syntactically inconsistent with the scene context: floating objects were fixated longer and more often than objects resting on surfaces. Interestingly, we found an interaction of both types of inconsistencies during the first inspection of an object: While non-floating objects showed longer gaze durations when they were semantically inconsistent, this inconsistency effect was eliminated when objects were floating. Thus, when objects violated the scene's syntactic structure, their semantic fit to the rest of the scene was rendered secondary. We propose that the stronger effect of the syntactic violation is due to the lower probability of encountering such an object-scene inconsistency in everyday life. Coming across a floating cocktail glass in a kitchen will be more disturbing than finding a microscope on the kitchen counter. This disturbance can also be regarded as an extreme degree of surprise and therefore interpreted within the framework of surprise theories. Itti and Baldi (2005), for example, formulated surprise in a Bayesian framework as the difference between prior expectations of an observer about the world and new incoming data. Surprise then quantifies as the difference between the prior and posterior beliefs. The stronger the mismatch, the stronger the computed surprise, which will — when the mismatch is strong enough — lead to increased deployment of attention and human gaze to the surprising event. According to this framework, the mismatch between prior expectations regarding an object and its current representation is more extreme for syntactic compared to semantic violations. However, the surprise model in its current form computes surprise from low-level stimulus properties. Our data argue for increased attention allocation to surprising events based on prior experiences and higher-level cognitive processes of an observer. These should be included in models accounting for human gaze control during the viewing naturalistic scenes.

Future work could also investigate to what degree well-known effects of semantic and syntactic violations in sentence processing can be generalized to scene processing. For example, studies on reading have identified different event-related potentials (ERPs) marking either semantic or syntactic processing: While a semantic mismatch can be seen in the N400 component, syntactic violations have been observed in an early left anterior negativity (ELAN) at around 100 - 200 ms followed by positivity around 600 ms after presentation of the violation (P600) (e.g., Palolahti, Leino, Jokela, Kopra, & Paavilainen, 2005; for a review see Friederici & Weissenborn, 2007). The latter component is thought to additionally reflect the integration of both semantic and syntactic information during later stages of sentence processing. Similar to reading, the processing of naturalistic scenes might also involve an obligatory assignment of semantic and syntactic roles to individual objects within a scene. We need to further investigate the relationship of scene semantics and syntax and their effects on eye movement control in order to move ahead in our understanding of the processes involved during the perception of scenes.

CHAPTER 4: EXPLICIT AND IMPLICIT DETECTION OF OBJECT LOCATION CHANGES

In this study, participants had to repeatedly inspect a randomized set of naturalistic scenes for later questioning. At the seventh presentation of each scene one object was shown at a new location. Instead of a memory task, an unannounced change detection task was conducted at the end of the experiment. We were interested in whether deviations from episodically learnt scene representations would lead to increased attentional allocation to the changed object and whether effects of change would occur during earlier or later stages of scene viewing. Results showed that only upon fixation, eye movement control was largely affected in that the moved object was fixated to a higher degree. This was the case for both reported and not reported changes. Interestingly, we additionally found indications of increased attentional allocation to the previous, now vacant object location, when changes were subsequently reported. We surmise that the reportability of position change might depend on the binding of an object to its particular location permitting direct comparison of the previous object location stored within the established scene representation with the new object location.

When viewing naturalistic scenes, we tend to be quite insensitive to changes that occur while we move our eyes or when masked by transients such as a blank screen or a blink (O'Regan, 1992; O’Regan, Deubel, Clark, & Rensink, 2000; Rensink, 2000; for a review see Simons, 2000). Such "change blindness" has been taken as evidence for the inability of the visual system to store detailed scene representations across space and time. However, there are a variety of theoretical positions regarding the nature of scene representations ranging from theories suggesting that no detailed visual representations accumulate as attention is oriented from one view of a scene to another (e.g., Becker & Pashler, 2002; Horowitz & Wolfe, 1998; O'Regan, 1992; O'Regan, Rensink, & Clark, 1999; Rensink, 2000, 2002), to theories proposing that very detailed visual scene representations indeed can be stored (e.g., Hollingworth & Henderson, 2002; Melcher, 2006; for a review see Hollingworth, 2006). Recent studies suggest that a possible reason for the diverging views on scene representations could lie in the finding that different visual features show different rates of memory accumulation and decay (Melcher & Morrone, 2003; Tatler et al., 2003). For instance, Tatler and colleagues (2005) tested immediate recall of multiple types of information from naturalistic scenes. They were able to show that position information for a critical object accumulated with performance increasing with increasing number of fixations on the object, while this was not the case for identity or color information.

The final study of my thesis set out to investigate whether scene representations — established from repeated scene presentations and thus stored in episodic memory — are detailed enough to allow for the detection of spatial location changes of a critical object from one presentation of a scene to another. Assuming this to be the case, another interesting question would be exactly when during scene viewing the detection of change would impact eye movement control. Finally, we wanted to extend prior work on change detection by investigating whether eye movement behavior during scene viewing would qualitatively differ depending on whether or not a change was subsequently reported.

We repeatedly presented participants with a set of scenes in randomized order with the task to view all scenes for later questioning. On the seventh presentation of each scene, we introduced unannounced location changes to a critical object in the scene, which was then returned to its original location for all subsequent presentations of the scene. Participants were neither told beforehand that changes would occur, nor that they would be asked to indicate which objects within the presented scenes had changed their location at the end of the experiment. Since the changes occurred without transient motion signal and due to the randomized presentation of different scenes, we increased the dependency of change detection on episodic memory processes due to the need to store multiple scene representations throughout the experiment. Thus, we were able to test whether unannounced deviations from a set of learnt scene representations would lead to increased gaze and more refixations of the changed objects as well as explicit change detection.

Given the findings by Tatler and colleagues (2005), who found that direct fixation of an object is required to extract meaningful position information, we expected that deviations from stored scene representations would exhibit their effect only after fixation of the changed object. Thus, there should not be an effect of change on the time taken until the first fixation of the changed object, since only upon fixation can the position information of the stored object file be compared to position information of the new object file.

Moreover, it has been demonstrated that even when a change is not reported, fixation durations on a changed object are longer than on the same object when it has not changed (e.g., Hayhoe et al., 1998; Henderson & Hollingworth, 2003; Hollingworth et al., 2001; Ryan & Cohen, 2004). The present study extends previous work by examining not only whether eye movements regarding changed objects differed as a function of explicit change detection, but also whether the original object position, which was unoccupied during change, would be differentially attended depending on the reportability of change. We hypothesized that increased exploration, i.e., increased gaze durations and refixations, of the previous object location would benefit explicit change detection.

EXPERIMENT 5

METHOD

Participants

Twelve students (8 female) from the LMU Munich ranging in age between 20 and 27 (*M* = 23.83, *SD* = 1.90) participated in the study for course credit or for 8€/hour. All participants reported normal or corrected-to-normal vision and were unfamiliar with the stimulus material.

Stimulus Material

Twenty 3D-rendered images of real-world scenes were displayed on a 19-inch computer screen (resolution 1024 x 768 pixel, 100 Hz) subtending visual angles of 28.98 (horizontal) and 27.65 (vertical) at a viewing distance of 70 cm. As can be seen in Figure 3.1, each scene came in two versions: In the standard version, the critical object was presented in its original position within the scene, while in the changed version, the critical object was moved horizontally relative to another object with a distance of about three degree visual angle. The assignment of either position to the standard or the changed condition was counterbalanced.

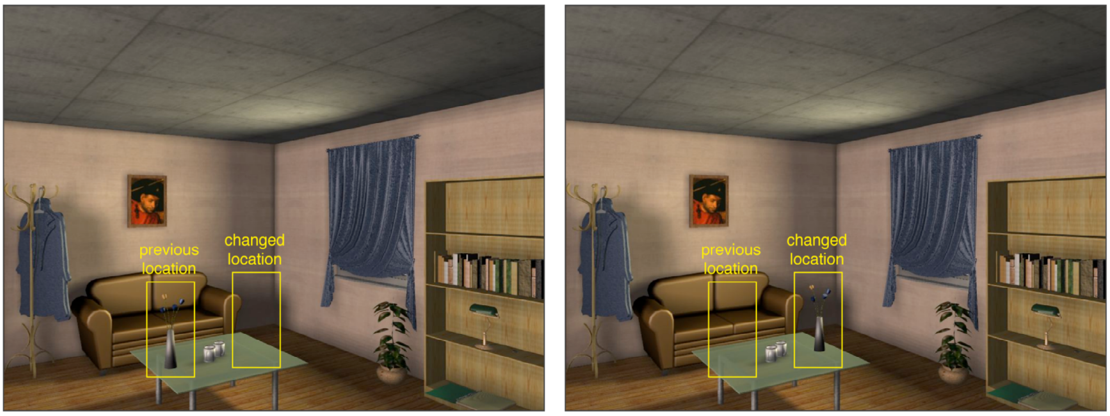

Figure 3.1: For the presentations one through six, the left picture showed the critical object (vase with flowers) in its original position. Upon the seventh presentation of the scene, the right picture was shown with the moved critical object at the changed location. Presentations eight through ten again showed the left picture with the critical object back in its original position. Note that at the seventh presentation the previous location of the object was vacant.

Apparatus

Eye movements were recorded with an EyeLink1000 tower system (SR Research, Canada), which tracks with a resolution of .01° visual angle at a sampling rate of 1000 Hz. The position of one eye was tracked while viewing was binocular. Experimental sessions were carried out on an IBM compatible display computer running on Windows XP. Stimulus presentation and reaction recording was controlled by Experimental Builder (SR, Research, Canada).

Procedure

After an initial 9-point calibration and validation, participants were informed that they would be presented with a set of repeating scenes, which they should view in preparation for questions at the end of the experiment. Each trial sequence was preceded by a drift correction. When the fixation check was deemed successful, the fixation cross was replaced by the presentation of a scene for seven seconds. Throughout the experiment, each scene was presented ten times in randomized order. Only on the seventh presentation was the critical object placed in a different location. The critical object was returned to its original position for all remaining presentations. At the end of the experiment, an unannounced change detection task was conducted in which all scenes in their standard version were presented to the participants again in randomized order with the instruction to indicate per mouse click, which object had changed position during the experiment.

Data analyses

For each scene a critical interest area was defined as the rectangular box that was large enough to encompass the critical object. This allowed us to analyze the following dependent variables: Latency was measured from scene onset until the first fixation of the critical object. Total Gaze Duration was defined as the sum of all fixation durations located in the critical region from scene onset until scene offset, while the Number of Refixations was defined as the number of times the critical interest area was entered and left.

For the analysis of eye movement data we excluded trials with first fixation durations on the critical interest areas outside the range of 2 standard deviations from the participant's mean fixation duration on these areas [3.71 %]. We also had to exclude two participants due to technical difficulties in assessing their eye movement data. The remaining raw data were subsequently preprocessed and then submitted to a repeated measures analysis of variance (ANOVA) with object change (pre-change — sixth presentation vs. post-change — seventh presentation) and change reportability (reported vs. unreported) as factors (for mean values see Table 3.1). In adddition to eye movement behavior regarding changed objects, we were interested in whether eye movements for the previous, but subsequently unoccupied object position differed as a function of reportability. Thus, we calculated a planned contrast for the previous object location in the changed scene for reported versus not reported changes (for mean values see Table 3.2).

RESULTS

Across all participants, about half of the position changes were correctly reported (*M* = 56 %, *SD* = 18.68). Note that participants had to select and indicate the object that had changed from several other objects within a scene in order to correctly report change greatly reducing the chance level.

Table 3.1: Summary of mean values [standard errors] regarding Latency, Gaze Duration, and Number of Refixations for the moved object as a function of object change (pre-change vs. post-change) and reportability (reported vs. unreported change).

MEASURES		REPORTED CHANGE		UNREPORTED CHANGE
	PRE-CHANGE	POST-CHANGE	PRE-CHANGE	POST-CHANGE
Latency in ms	2088 [227]	1982 [186]	2800 [850]	2639 [359]
Gaze Duration in ms	505 [43]	997 [110]	297 [46]	417 [54]
Number of Refixations	1.37 [.08]	2.30 [.20]	.94 [.06]	1.28 [.17]

Table 3.2: Summary of mean values [standard errors] regarding Latency, Gaze Duration, and Number of Refixations for the unoccupied object location on the seventh presentation as a function of reportability (reported vs. unreported change).

MEASURES	REPORTED CHANGE	UNREPORTED CHANGE
Latency in ms	3289 [343]	3294 [609]
Gaze Duration in ms	175 [37]	59 [17]
Number of Refixations	.52 [.10]	.26 [.09]

Latency. There was neither an effect of change, $F < 1$, nor of reportability, $F(1,9) = 3.71$, $p > .05$, and no interaction between change and reportability, $F < 1$.

Participants also did not fixate the previous, unoccupied object location earlier when the change was reported (report: 3289 [343] ms vs. no report: 3294 [609] ms), $t(9) < 1$.

Gaze Duration. Both change and reportability showed main effects, $F(1,9) = 21.29$, $p < .01$ and $F(1,9) = 26.45$, $p < .01$, respectively. On average, objects were looked at for an additional 306 ms after their movement compared to that prior to change, while reported changes were generally looked at 394 ms longer than unreported changes. There was also a significant interaction, $F(1,9) = 11.52$, $p < .01$, which was caused by the increased effect of change for reported, $t(9) = 4.72$, $p < .01$, as compared to unreported changes, $t(9) = 1.91$, $p < .05$: when the detection was reported, gaze duration was about two times longer than prior to change, while it was only about 1.4 times longer for unreported changes.

Additionally, the previous, unoccupied object position was fixated about three times longer when the change was reported (M = 175 ms, SE = 37) than when it was not (M = 59 ms, SE = 17), $t(9) = 2.81$, $p = .01$.

Number of Refixations. Changed objects were refixated about .63 times more often than objects in their original position, $F(1,9) = 18.92$, $p < .01$. Also, reported changes were refixated about .73 times more often than unreported changes, $F(1,9) = 15.77$, $p < .01$. Moreover, there was a significant interaction between both factors, $F(1,9) = 9.30$, $p < .05$, in that the effect of change was larger for reported, $t(9) = 5.02$, $p < .01$, as compared to unreported changes, $t(9) = 2.05$, $p < .05$. When changes were reported, participants refixated the moved object 1.7 times more often than when it had not changed position, while a moved object was refixated 1.4 times when the changes were not reported.

Also, the previous, unoccupied object position was refixated about twice as often when the change was reported (M = .52, SE = .10) than when it was not (M = .26, SE = .09), $t(9) = 2.01$, $p < .05$.

DISCUSSION

In the present study, we were able to show that eye movement behavior during scene viewing was strongly modulated by location changes as well as change reportability, in that gaze duration and refixations increased upon fixation of the changed object. For reported changes, the effect of change on eye movements was stronger and was accompanied by increased deployment of gaze towards the previous, unoccupied object position as compared to unreported changes. This implies that across repeated presentations of each scene, sufficiently detailed episodic scene representations were established allowing for both implicit and explicit change detection. Also a great number of object files — clearly exceeding the capacity of VSTM (see Kahneman & Treisman, 1984; Luck & Vogel, 1997) — had to be integrated into the scene representations and stored in VLTM for continuous comparison with new object files. It seems that over the course of scene viewing, position information for stored object files was able to accumulate (see Tatler et al., 2005) gradually refining the initially crude scene representations. When an object changed its position, the position information of the stored object file did not match the position information of the new object file creating a residual signal and prolonging the deployment of attention to the changed object (see Rao & Ballard, 1999). Previous findings of prolonged attentional allocation to changed objects in familiar as compared to unfamiliar environments (e.g., Brockmole & Henderson, 2005b; Karacan & Hayhoe, 2008) support the view that attentional allocation in scenes can be modulated by deviations of new scene representations from the ones stored and strengthened over the course of scene viewing.

Further support for memory-based detection of changes without transient motion signals comes from the lack of finding evidence for an early capture of attention, which is commonly observed for object changes with transient motion signals (Jonides & Yantis, 1988; Theeuwes, 1994; Yantis, 1998). In our study, the moved object was not fixated earlier compared to the presentation of the scene in which the object was still in its original position arguing against attentional capture for the changed object. Only upon fixation did the moved object lead to prolonged gaze and an increased number of refixations. This is in line with findings by Tatler et al. (2005) according to which the detection of position changes depends on the fixation of the moved object. This is also consistent with the results of transsaccadic change detection experiments, which have shown that changes to the structural relationship between an object and its scene (e.g., rotation of an object in a scene) is typically only detected when the object has been fixated before and after the change (Hollingworth & Henderson, 2002). Fixation seems to be a prerequisite for associating position information with the object file, and follows the feature integration view that only directed attention permits the creation of a specific object file (Kahneman & Treisman, 1984). Thus, it seems that the processing of position information proceeds serially within the scene since positional mismatch calculation requires focused attention by means of fixation.

Moreover, we were particularly interested in whether eye movement behavior was modulated by the ability to explicitly report object changes. Our data allow drawing a number of interesting conclusions. First, we found that changed objects modulated attentional allocation regardless of explicit change detection: A changed object was fixated to a greater degree than when it had not changed its position. This finding implies that eye movements may be a more sensitive indicator for change detection than explicit report. While this has been observed for type, token, rotation, and color changes (e.g., Hayhoe et al., 1998; Henderson & Hollingworth, 2003), we report implicit change detection for spatial displacements (see also Karacan & Hayhoe, 2008). However, our findings differ from those of Ryan and Cohen (2004), who observed effects of change on eye movements regarding the moved object only when the change was explicitly reported. This may be due to the generally higher change-detection rates in their study despite the use of masks (at least 84%) as opposed to correctly reported change in our study (53%). While reported changes in Ryan and Cohen's study may have been based on relatively confident judgments, unreported changes may have mirrored a total lack of change detection both explicit and implicit.

Second, we found a main effect of reportability in that reported changes were characterized by longer gaze durations and increased refixations to the critical object both before and after the changed occurred. It seems that longer encoding of the object prior to its change increased the ability to explicitly report the subsequent change. This is in line with the visual memory theory (Henderson & Hollingworth, 2003) according to which the detection of a change to an object in a scene is a function of the degree of attention deployed to the object during encoding. Karacan and Hayhoe (2008) also reported a significant correlation between average fixation durations on the changing objects and explicit change detection scores providing further evidence for the gradual refinement of stored scene representations along increasing viewing time.

Third, the effect of object change on the deployment of attention towards the moved object — while existent for both reported and unreported changes — was strongest for reported changes. This finding further supports the notion that eye movements are able to reflect differences in explicit change detection.

Last but not least, our study went beyond previous findings in that we were not only able to show that explicit change detection involved a greater degree of fixating the changed objects as compared to unreported changes, but that this was also the case for the previously occupied object location. Thus, even though the previous object location was unoccupied at the time of viewing, i.e., there was no object to look at, participants fixated that particular region of the scene to a higher degree when they were subsequently able to report the position change. We surmise that explicit change detection may rely on the binding of an object to its stored location, i.e., the storage of the previous position tag. The successful reactivation of position information contained in the previously established object file during retrieval seems to allow for an extra validation check of the formerly occupied position, which may increase the probability of explicit change detection. When the reactivation of previous position information is possible, but not exact, a mismatch signal might nevertheless be generated, but does not allow for a precise validation check therefore impeding explicit change detection. Alternatively, the mismatch signal generated without prior validation check may not be strong enough to exceed a change detection threshold necessary for explicit report, but may still be able to affect eye movement control.

CHAPTER 5: GENERAL CONCLUSION

High quality vision is restricted to a small region at the center of gaze. As a result, we have to actively shift our eyes to different parts of a visual scene in order to continuously gather information from our surroundings. Understanding the mechanisms that underlie the selection of locations for the eyes to fixate is therefore fundamental to the understanding of visual behavior. Which of these mechanisms predominantly drive selection by the eye during the perception of natural scenes remains controversial. Current views range from accounts suggesting that where we look is determined mainly by the visual characteristics of the scene we are viewing to those suggesting that where we look is determined mainly by our internal goals and agendas. The main objective of the line of experiments described in my thesis was to provide further evidence for the predominant role of *cognitive* control on the allocation of attention in scene perception.

Study 1 provided evidence for the predominance of global processing in the generation of initial scene representations and in the effective control of attention and eye movements during visual search in naturalistic scenes. A short glimpse of a scene's background suffices to restrict search space when subsequently looking for a predefined target object. The underlying mechanism seems to be the rapid setup of scene priors, i.e., prior knowledge and expectations on the configuration of objects within certain scenes, which makes it possible to modulate a purely bottom-up saliency map resulting from local scene processing (see Torralba et al., 2006). Thus, when presented with naturalistic scenes we are able to draw on higher-level cognitive processes, which allow for a more effective target object search. Additionally, we found that people greatly differ in their ability to process flashed scene previews. We argue that varying degrees of processing either local or global scene information can lead to differential generations of initial scene representations, which could account for the individual differences observed in subsequent eye movement behavior. Future work will need to further investigate the relationship between varying degrees of information processing and degrees of effective attention and eye movement control, for example, by actively manipulating the degree of scene processing. While the approach to the investigation of individual differences in scene processing offered in Study 1 is by no means meant to be exhaustive, the results of the study make clear that individual differences can arise and should not be neglected when investigating the impact of the first glimpse of a scene.

Cognitive control of attention during scene perception was further investigated in Study 2, where we violated expectancies on either the semantic or the syntactic configuration of scenes. Objects in scenes are usually embedded in such a way that they adhere to certain physical laws and semantic constraints. While it has been shown that the violation of semantic object-scene expectations leads to prolonged gaze towards such inconsistent objects, the question of when during scene viewing this semantic inconsistency effect impacts on eye movement control has been controversially debated (e.g., Underwood & Foulsham, 2006; Underwood et al., 2007; 2008, arguing for extrafoveal processing of scene inconsistencies, or, e.g., Gareze & Findlay, 2007; Henderson et al., 1999, arguing against an effect prior to the object's fixation). By additionally introducing syntactic object-scene violations, i.e., having objects float, we were able to directly compare the effects of semantic and syntactic object-scene inconsistencies. The findings of Study 2 clearly speak against an early, extrafoveal influence of object-scene inconsistencies on initial eye movements during scene viewing. We used 3D-rendered scenes instead of line drawings or photographs and found that neither semantic nor support violations led to earlier fixations of inconsistent objects compared to their consistent counterparts. Upon fixation, both inconsistencies affected the deployment of attention and eye movements with inconsistent objects being fixated more often and longer than consistent objects. The direct comparison of semantic and syntactic violations showed that both inconsistencies interacted. We propose that the effect object-scene inconsistency on eye movements varies as a function of prior beliefs and expectations: the greater the beliefs and expectations, the greater the effect on eye movement control when these are violated. This further promotes the idea that we automatically assign certain expectations to objects within a scene regarding their semantic and syntactic integration, which can subsequently influence how we direct our gaze when viewing scenes in the real world.

The final study of this thesis tested another kind of object-scene inconsistency, namely the inconsistency stemming from the mismatch between a current scene representation and prior scene representations, which were previously generated and stored in episodic memory across multiple viewings of the same scene. We found that changing the location of objects from one view of a scene to another led to both implicit and explicit change detection. This provided support for the notion that we are able to establish and store episodic scene representations detailed enough to allow for the detection of non-transient position changes to objects embedded in naturalistic scenes. Regardless of the ability to explicitly report the change, moved objects led to increased gaze. However, similar to the findings of Study 2 eye movement control was only affected upon fixation of the changed object. Interestingly, we additionally found indications of prolonged memory-based allocation of attention to the previous object location when changes were subsequently reported. The successful reactivation of position information contained in a previously established object file seems to allow for an extra validation check of the previously occupied position, which seems to increase the probability of explicit change detection. These results imply that we are able to learn the structure of unknown naturalistic scenes over time and that attention and eye movements are modulated by deviations from scenes that we have learnt and stored across time and space.

Taken together, all studies of this thesis converge on the general conclusion that the control of where we allocate our attention and gaze during the viewing of naturalistic scenes is largely influenced by higher-level cognitive processes. The mere bottom-up processing of current visual input would not explain, for example, why we tend to look at a semantically abnormal, but otherwise visually inconspicuous object in a scene. Thus, the studies presented here went beyond previous work in several ways.

First, I created a set of highly controlled 3D-rendered scenes. These display a high degree of photorealism, while allowing for a variety of manipulations, for example, making objects float. In pictures that are manipulated post-hoc, the resulting shadows are often incorrect and thus confound the intended higher-level manipulation with editing artifacts (see Underwood et al., 2007). Additionally, the scenes were controlled for bottom-up visual salience using the Itti and Koch (2001) algorithm. This ruled out the possibility that the effects we found were based merely on visual conspicuousness. Especially the contradicting effects of semantic object-scene inconsistencies found across several studies in the literature seem to be at least partly based on differences in the stimulus material used. Without a higher degree of control for the scenes we use as stimulus material, it will be difficult to reach common ground.

A second contribution of this thesis to research on scene perception is the notion that there seem to be large individual differences in how people process shortly flashed visual displays and that neglecting them would disregard important information. Only few studies so far have looked at individual differences in scene perception (e.g., Brockmole et al., in press; Rayner, et al., in press; Underwood et al., 2003). None of them, however, investigated how people differ in the way they process the first glimpse of a scene. We could show that differences observed in eye movement behavior during scene viewing can be attributed to differences already evident in initial scene processing. This was possible by making use of the TVA approach introduced by Bundesen (1990) to objectively determine parameters of processing efficiency. While future studies will have to further test the applicability of the TVA approach to scene perception research, this thesis can be seen as a first step towards a more in-depth investigation of individual differences in scene perception.

Finally, in my thesis I have tried to introduce and apply concepts known from sentence processing to the investigation of scene perception. Specifically, I borrowed on the distinction between semantic and syntactic compositions of elements commonly discussed regarding the relationship between words within a sentence. Here, I propose that the relationship between objects embedded in a scene can also be regarded as being either semantic or syntactic in nature. We could show that violating expectations regarding these relationships leads to increased inspection of objects inconsistent with the scene. The next step in applying findings stemming from sentence processing research to the investigation of scene processing would be, for example, to look for electrophysiological markers known to reflect semantic versus syntactic processing, e.g., the N400 and the P600, respectively. The great advantage of using such electrophysiological markers would be the ability to investigate the time course of cognitive processes active during scene perception. This would enable us to gain more insight into how we perceive and construct the visual world surrounding us.

In sum, my thesis shows that our perception of the natural world is highly constructive in nature and that the allocation of attention in scene viewing is both the basis as well as the result of our perception of the world.

Deutsche Zusammenfassung

Die Allokation von Aufmerksamkeit bei der Szenenwahrnehmung

Unser visuelles System ist so aufgebaut, dass nur in einem eng umgrenzten Bereich, der Fovea, visuelle Information hochauflösend verarbeitet werden kann. Das Verarbeiten von Information, die auf diesen Bereich der Retina fällt, wird Foveales Sehen genannt. Aufgrund dieser Eigenart unseres Wahrnehmungsapparates bewegt der Mensch seine Augen circa drei bis vier Mal pro Sekunde, um Information an verschiedenen Orten des visuellen Feldes zu verarbeiten. Dies führt zu einer Abfolge von Fixationen — Phasen von ca. 200 bis 300 ms, in denen das Auge ruht, um Information aufzunehmen, und Sakkaden — schnellen Bewegungen des Auges, welche die Fovea auf den jeweils nächsten Ort ausrichten. Unter natürlichen Bedingungen geben somit Augenbewegungen Auskunft über die Verlagerung visueller Aufmerksamkeit in der Szene (offene Aufmerksamkeitsverschiebung).

Die Frage, welche Mechanismen die Bewegungen unserer Augen kontrollieren, ist daher essentiell bei der Erforschung von Aufmerksamkeitsprozessen in der visuellen Wahrnehmung. Die vorliegende Arbeit hat zum Ziel, solche kognitiven Kontrollmechanismen zu untersuchen, die bei der Wahrnehmung von natürlichen Szenen für die Steuerung der Blickbewegungen mitverantwortlich sind. Dabei ist es wichtig, sich die Besonderheiten von natürlichen Szenen bewusst zu machen. Hollingworth und Henderson (1999b) definieren Szene als "a semantically coherent human-scaled view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner". Der Aufbau einer natürlichen Szene ist also im Gegensatz zu einfacheren Displays, z.B. einer Buchstabenmatrix, wie sie häufig in der visuellen Aufmerksamkeitsforschung benutzt werden, bestimmten Restriktionen unterworfen. So sind manche Objekte in bestimmten Szenen mit höherer Wahrscheinlichkeit anzutreffen als andere. Zum Beispiel würde uns wohl ein Hydrant im Schlafzimmer als seltsam auffallen. Zudem haben wir bestimme Vorstellungen, wo in einer Szene bestimme Objekte am wahrscheinlichsten vorzufinden sind. Suchen wir zum Beispiel eine Tasse in einer Küche, werden wir diese kaum auf dem Fußboden vermuten.

Es erscheint daher plausibel, dass die Steuerung unserer Blickbewegungen nicht allein vom momentanen visuellen Input abhängt, sondern auch von höheren kognitiven Prozessen geleitet wird, wie zum Beispiel bestimmten Erwartungen und Zielen des Beobachters, die auf früheren Erfahrungen basieren. In einer Reihe von drei Studien konnte gezeigt werden, dass die Steuerung von Blickbewegungen beim Betrachten natürlicher Szenen in hohem Maße solcher kognitiven Kontrolle unterliegt.

Interindividuelle Unterschiede bei der Szenenwahrnehmung

In Studie 1 galt das Interesse der Verarbeitung von Information innerhalb des ersten Blickes auf eine natürliche Szene. In mehreren Arbeiten konnte bereits gezeigt werden, dass selbst bei Darbietungszeiten von wenigen 100 ms der Hauptinhalt, die sogenannte "gist", einer Szene erkannt werden kann (z.B., Oliva & Schyns, 2000; Oliva & Torralba, 2006; Potter, 1975; Thorpe, et al., 1996). So lassen sich Szenen innerhalb nur eines Blickes verschiedenen Szenenkategorien, z.B. Küche versus Schlafzimmer versus Bad, zuordnen. Im Gegensatz zu einer mehr bottom-up kontrollierten Steuerung von Blickbewegungen, wie sie das salienzbasierte Modell von Itti und Koch (2001) propagieren, gehen Torralba und Kollegen (2006) in ihrem "Contextual Guidance Model" von einer auf dem Szenenkontext basierenden Modulation von Aufmerksamkeit und Blickbewegungen aus. Laut dem Modell werden natürliche Szenen auf zwei parallelen Pfaden verarbeitet. Auf dem "Lokalen Pfad" wird ähnlich dem Itti und Koch Modell aufgrund von Helligkeitskontrasten und Merkmalsorientierungen eine Salienzkarte erstellt. Zusätzlich werden auf dem "Globalen Pfad" globale Merkmale der Szene zu einem Vektor zusammengefasst, der die spezifische Szene betreffende Erwartungen, sogenannte "scene priors", aktiviert und zusammen mit dem Wissen um die Aufgabe, z.B. einer Objektsuche, den Suchraum kontextmoduliert einschränkt. Das heißt, dass die rein auf bottom-up Verarbeitung basierende Salienzkarte des Lokalen Pfades durch die top-down Verarbeitung des Szenenkontextes auf dem Globalen Pfad so moduliert wird, dass die Augen zunächst zu den Teilen der Szene gesteuert werden, die im Hinblick auf den spezifischen Kontext mit höchster Wahrscheinlichkeit das zu suchende Objekt beinhalten.

Um zu untersuchen, welche Information für eine spätere Objektsuche am hilfreichsten ist, haben wir das "Flash-Preview Moving-Window" Paradigma herangezogen (siehe Castelhano & Henderson, 2007). Dabei wird den Versuchspersonen für 250 ms eine Szene dargeboten, die anschließend maskiert wird. Dann folgt ein Wort, dass das zu suchende Objekt indiziert. Im Anschluss wird erneut eine Szene präsentiert, die jedoch nur durch ein kleines, sich mit dem Auge mitbewegendes Sichtfenster von ca. 2° Sehwinkel hindurch inspiziert werden kann. Außerhalb des Sichtfensters erscheint die Szene nur grau maskiert. Mit Hilfe dieses Versuchsaufbaus lässt sich die Kontrolle der Blickbewegungen während der Suche auf die nur kurz präsentierten Previews zurückführen. Versuchspersonen dieser ersten Studie bekamen eine Reihe von verschiedentlich manipulierten Previews zu sehen. Diese unterschieden sich nur in ihrem Informationsgehalt. Ein "Identischer Preview" war identisch mit der später nur eingeschränkt sichtbaren Suchszene (mit der Ausnahme, dass das Suchobjekt im Preview fehlte). Der "Hintergrund Preview" zeigte nur den Szenenhintergrund an, d.h. individuelle Objekte wurden aus dem Preview entfernt. Bei dem "Objekte Preview" fehlte dagegen der Szenenhintergrund, während alle einzelnen Objekte unverändert in ihrer Position innerhalb der Szene verblieben. Dieser letzten Previewvariante fehlte daher die räumliche Tiefenkomponente des Raumes. Als Kontrollbedingung diente eine Maske, die aus durcheinandergewürfelten Teilen aller Szenen bestand. Aufgrund der Annahmen des "Contextual Guidance Models" erwarteten wir verbesserte Suchleistungen für Szenenpreviews, die global verarbeitet werden können, d.h. für sowohl den Identischen, als auch den Hintergrund Preview.

In Vorexperimenten hatte sich zudem bereits gezeigt, dass es bei der Verarbeitung der kurz dargebotenen Szenenpreviews durchaus große interindividuelle Unterschiede zwischen den Versuchspersonen gab. Deshalb sollten die Versuchspersonen am Ende von Experiment 1 in einem Fragebogen angeben, wie gut sie zwischen den verschiedenen Previewbedingungen unterscheiden konnten. Dabei stellte sich heraus, dass ca. ein Drittel der Versuchspersonen nicht zwischen den drei Szenenpreviews unterscheiden konnten, sondern nur angaben, einen Unterschied zwischen der Maske und "anderen Szenenpreviews" bemerkt zu haben. Die restlichen Versuchspersonen gaben hingegen an, die drei Previews — nämlich *Identischen*, *Hintergrund* und *Objekt Preview* — unterschieden zu haben. Aufgrund der Auswertung des Fragebogens wurden die Versuchspersonen dann in zwei Gruppen unterteilt. Die Analyse der Blickbewegungsdaten zeigte einen über alle Versuchspersonen hinweg deutlichen Suchvorteil nach der Präsentation eines Hintergrund Previews. Dies deutet darauf hin, dass der alleinige Hintergrund einer Szene bereits ausreicht, um kontextmodulierte Suchraumeinschränkungen zu gewährleisten. Der *Identische Preview* hingegen wurde von den beiden Versuchspersonengruppen äußerst unterschiedlich verarbeitet. Während die Versuchspersonen, die keinen Unterschied zwischen den Szenenpreviews feststellen konnten, am meisten vom Identischen Preview profitierten, schien dergleiche Preview bei der Versuchspersonengruppe mit besserer Previewdifferenzierung Interferenzen hervorzurufen und zu Suchleistungseinbußen zu führen. Wir nahmen an, dass die Effektivität von Augenbewegungen während der Suche durch die Verarbeitungsgeschwindigkeit des Previews bestimmt wurde. Aufgrund dieser Hypothese, erhoben wir mit Hilfe einer Testbatterie — basierend auf der "Theory of Visual Attention" (TVA) — im Experiment 2 von einigen Versuchspersonen zusätzliche Verarbeitungseffizienzparameter (Bundesen, 1990). In der Tat wiesen die Versuchspersonen mit besserer berichteter Differenzierung der Szenenpreviews eine signifikant höhere perzeptuelle Verarbeitungsgeschwindigkeit auf. Die unterschiedlichen Effekte des Identischen Previews zwischen den Versuchspersonengruppen können daher auf die Unterschiede in der Verarbeitungsgeschwindigkeit zurückgeführt werden. Bei höherer Verarbeitungsgeschwindigkeit reicht eine kurze Szenenpräsentation aus, um über den global-abstrakten Szeneninhalt hinaus einzelne Objekte über den Lokalen Pfad zu identifizieren. Die Identifizierung einer Anzahl von Objekten könnte wiederum zu einer Aktivierung alternativer Szenenkategorien führen, die mit Szenenkategorien auf dem Globalen Pfad interferieren und deshalb die Suchleitung mindern.

Studie 1 konnte also zeigen, dass es gravierende interindividuelle Unterschiede in der Szenenwahrnehmung gibt und dass diese Unterschiede zum Teil auf unterschiedlich hohen Verarbeitungsgeschwindigkeiten beruht. Besonders bei sehr kurzen Darbietungen von Szenen erscheint es wichtig, interindividuelle Unterschiede zu berücksichtigen. Während weitere Studien die Applizierbarkeit der TVA (Bundesen, 1990) auf die Untersuchung von Szenenwahrnehmung testen sollten, stellt diese Studie einen ersten Schritt in die gezielte Erforschung interindividueller Unterscheide bei der Verarbeitung von natürlichen Szenen dar.

Semantische und Syntaktische Verarbeitung von natürlichen Szenen

Studie 2 beschäftigte sich mit der Frage, wie semantische und syntaktische Inkonsistenzen in natürlichen Szenen verarbeitet werden und inwiefern eine solche Verarbeitung auf die Steuerung von Aufmerksamkeit und Blickbewegungen wirken. Als *semantisch* inkonsistent wird ein Objekt dann bezeichnet, wenn es inhaltlich nicht zur Szene passt, z.B. ein Feuerhydrant im Schlafzimmer. Als *syntaktisch* inkonsistent bezeichneten wir in dieser Studie Objekte dann, wenn sie nur die Struktur der Szene verletzten, aber inhaltlich konsistent sind, z.B. ein "schwebender" Mixer in einer Küche. Während die Effekte semantischer Inkonsistenz auf die Aufmerksamkeitssteuerung in der Literatur bisher kontrovers diskutiert wurden, wurde der syntaktischen Inkonsistenz in Szenen bisher kaum Beachtung geschenkt (vgl. Biederman et al., 1982; DeGraef et al., 1990). Studie 2 hatte zum Ziel die Effekte semantischer und syntaktischer Inkonsistenz direkt zu vergleichen. Dabei stellte sich die Frage, ob bereits erste Blickbewegungen in der Szene von derartigen Inkonsistenzen beeinflusst werden oder ob der Effekt der Inkonsistenz erst nach Fixation des inkonsistenten Objektes eintritt.

Bisherige Studien zum Effekt semantischer Inkonsistenz lieferten diesbezüglich unterschiedliche Befunde. Während einige Studien frühe Effekte semantischer Inkonsistenz auf die Blickbewegungssteuerung fanden (z.B., Loftus & Mackworth, 1978; Underwood & Foulsham, 2006; Underwood et al., 2007; 2008), konnten andere Studien nur nach Fixation des inkonsistenten Objektes eine Modulation der Blickbewegungssteuerung feststellen (DeGraef et al., 1990; Gareze & Findlay, 2007; Henderson et al., 1999). Einer der Gründe für derart divergierende Ergebnisse scheinen die verwendeten Szenen selbst zu sein, die sich in ihrer Komplexität und der Realitätsgetreue oft deutlich unterschieden. In den Experimenten der vorliegenden Arbeit wurden daher ausschließlich photorealistisch überarbeitete Abbildungen von Szenen verwendet, die mit spezieller Architektursoftware erstellt worden waren. Dies erlaubte größtmögliche Kontrolle des Stimulusmaterials und zugleich gezielte Manipulation einzelner Szenenbestandteile ohne dabei Gefahr zu laufen, durch die nachträgliche Bearbeitung von Bildmaterial Inkonsistenzen zu verursachen, die manchmal auf rein visuellen Artefakten der Editierung beruhen könnten (siehe Underwood et al., 2007).

In zwei Experimenten konnten wir zeigen, dass weder semantische noch syntaktische Inkonsistenzen eine frühe Modulation der Blickbewegungssteuerung verursachten. Erst nach Fixation des inkonsistenten Objektes wurden diese länger und öfter fixiert als szenenkonsistente Objekte. Interessanterweise zeigte sich ein starker Effekt für syntaktische Inkonsistenzen: Während syntaktisch konsistente Objekte einen Effekt semantischer Inkonsistenz aufwiesen, verschwand dieser für syntaktisch inkonsistente, also schwebende, Objekte. Die alleinige Tatsache, dass ein Objekt schwebte, schien also seine semantische Inkonsistenz irrelevant zu machen. Diese Interaktion kann darauf zurückgeführt werden, dass die Wahrscheinlichkeit, einer semantischen Inkonsistenz im Alltag zu begegnen, größer ist als bei syntaktischen Inkonsistenzen und daher die damit verbundene Verwunderung des Betrachters bei syntaktisch inkonsistenten, d.h. schwebenden Objekten, größer ist als bei semantisch inkonsistenten. Die Verletzung von apriori Erwartungen ist im ersten Fall also größer als im zweiten Fall, was eine stärkere Aufmerksamkeitsverlagerung auf ein schwebendes Objekt nach sich zieht. Weitere Studien könnten möglicherweise den graduellen Zusammenhang von Aufmerksamkeitsverlagerung und der Verletzung von apriori Erwartungen untersuchen.

Um festzustellen, ob die Art der Aufgabe möglicherweise nicht hinreichend zum raschen Aufsuchen der inkonsistenten Objekte motiviert haben könnte, ließen wir in einem zweiten Experiment Versuchspersonen nach vorgegebenen Objekten suchen, die entweder konsistent oder semantisch und/oder syntaktisch inkonsistent waren. Durch die Aufgabe aktiv nach Objekten innerhalb der Szenen zu suchen, sollte sich der Aufmerksamkeitsfokus erweitern und damit die Verarbeitung extrafovealer Inkonsistenzen erhöhen. Dennoch zeigte sich auch in Experiment 2 kein früher Effekt von Inkonsistenzen auf die Blickbewegungssteuerung.

Zusammengefasst lässt sich sagen, dass Studie 2 selbst mit Hilfe von hochkontrolliertem Szenenmaterial keinen Hinweis für die frühe Modulation der Steuerung von Aufmerksamkeit beziehungsweise Blickbewegungen liefern konnte. Wir schließen daraus, dass Effekte sowohl semantischer als auch syntatischer Inkonsistenzen erst nach Fixation des jeweilig inkonsistenten Objekts in Erscheinung treten, um sich dann in der Modulation von Blickbewegungen niederzuschlagen. In nachfolgenden Studien wollen wir zum Beispiel mit Hilfe von ereigniskorrelierten Potentialen (EKPs) die Abstufung von Inkonsistenzeffekten sowie deren zeitlichen Verlauf im Detail untersuchen. Dazu müsste zunächst überprüft werden, ob bereits aus der Satzverarbeitung bekannte EKPs wie die N400 und die P600, welche semantische beziehungsweise syntaktische Inkonsistenzen in der Satzverarbeitung signalisieren, auch bei der Verarbeitung von Inkonsistenzen in

Szenen auftreten. Wäre dies der Fall, spräche das zusätzlich für die Involviertheit höherer kognitiver Prozesse bei der Betrachtung von natürlichen Szenen.

Explizites und Implizites Erkennen von Objektverschiebungen in Natürlichen Szenen

Studie 3 beschäftigte sich mit dem Grad der Detailliertheit von Szenenrepräsentationen, die über wiederholte Betrachtungen derselben Szenen hinweg aufgebaut werden. Dazu präsentierten wir Versuchspersonen 20 verschiedene Szenen zehn Mal in randomisierter Reihenfolge (dieselbe Szene wurde jedoch nie zweimal hintereinander dargeboten). Aufgabe war es, alle Szenen mehrmals eingängig zu betrachten, da am Ende des Experiments Fragen zu den Szenen gestellt werden würden. Bei der siebten Präsentation einer Szene, wurde diese leicht verändert gezeigt, d.h. ein Objekt innerhalb der Szene wurde in einer neuen Position dargeboten. Die Versuchspersonen wurden über solche Veränderungen jedoch vorab nicht informiert. Erst am Ende des Experiments folgte ein Test, bei dem den Versuchspersonen alle 20 Szenen in ihrer unveränderten Version nochmals dargeboten wurden und sie per Mausklick angeben sollten, welche Objekte während des Experiments jeweils ihre Position verändert hatten. Da wir während des gesamten Experiments Blickbewegegungsdaten aufzeichneten, war es uns möglich den Zusammenhang zwischen explizitem Berichten von Objektverschiebungen und impliziter Modulation von Blickbewegungen zu untersuchen.

Die zugrunde liegende Frage war, ob gelernte und episodisch gespeicherte Szenenrepräsentationen detailliert genug sind, um Ortsverschiebungen einzelner Objekte explizit beziehungsweise implizit erkennen zu lassen.

Frühere Studien zur sogenannten "Change Blindness" haben den Eindruck erweckt, dass die menschliche Fähigkeit, Veränderungen in Szenen wahrzunehmen, äußerst gering ist. "Change Blindness" — oder zu Deutsch "Veränderungsblindheit" — bezeichnet ein Phänomen, bei dem es Versuchspersonen misslingt zum Teil große Veränderungen in einer Szene zu erkennen, wenn der Übergang zwischen beiden Versionen der Szene maskiert wird. Die Veränderung findet also in einer kurzen Unterbrechung der visuellen Wahrnehmung statt, d.h. wenn die erste Präsentation der Szene bereits erloschen, die veränderte Variante der Szene aber noch nicht erschienen ist. Solche Unterbrechungen können Sakkaden, Lidschläge, Blinzeln oder Maskierungen sein (siehe O'Regan, 1992; Rensink, 2000; Simons, 2000). Diese Studien legen den Schluss nahe, dass die Repräsentation einer Szene von einer Betrachtung zur nächsten sehr spärlich ist. Demgegenüber haben andere Studien gezeigt, dass Szenenrepräsentationen durchaus sehr detailliert sein und über längere Zeit abgespeichert werden können (Hollingworth, 2006; Hollingworth & Henderson, 2002; Melcher, 2006; Tatler et al., 2005). Studie 3 untersuchte inwiefern Objektverschiebungen in episodisch gespeicherten, natürlichen Szenen explizit erkannt werden können.

Die Ergebnisse dieser Studie liefern weitere Hinweise, die für durchaus detaillierte Szenenrepräsentationen im episodischen Gedächtnis sprechen. Zunächst konnten wir zeigen, dass mehr als die Hälfte der Objektverschiebungen explizit erkannt wurden. Dies ist beachtlich, wenn man bedenkt, dass bei der richtigen Indizierung von den verschobenen Objekten alle Objekte einer Szene potentiell zur Wahl standen. Die Zufallswahrscheinlichkeit lag also bei dieser Art des expliziten Berichts bei weit unter 50%. Auch in der Steuerung der Blickbewegung zeigte sich ein Effekt der Objektverschiebung: Verschobene Objekte wurden länger fixiert und im Laufe der Szenenbetrachtung öfter refixiert. Interessanterweise zeigte sich — wenn auch schwächer ausgeprägt — diese Modulation der Augenbewegungen selbst dann, wenn die Objektverschiebung nicht explizit erkannt wurde. Dies passt zu anderen Studien, die gezeigt haben, dass selbst beim Fehlen expliziter Berichtbarkeit von Objektveränderungen, Blickbewegungen dennoch entsprechend moduliert werden können (z.B., Hayhoe et al., 1998; Henderson & Hollingworth, 2003; Hollingworth et al., 2001; Karacan & Hayhoe, 2008; Ryan & Cohen, 2004). Blickbewegungen scheinen demnach ein sensitiverer Indikator für Veränderungen in Szenen zu sein als deren expliziter Bericht (vgl. Karacan & Hayhoe, 2008). Anders als bisherige Studien analysierten wir zusätzlich zu den Blickbewegungsdaten für verschobene Objekte auch Blickbewegungen bezüglich deren früherer Position im Raum, die zum Zeitpunkt der Verschiebung kein Objekt mehr beinhaltete. Dabei zeigte sich, dass die frühere Position des Objektes dann öfter und länger fixiert wurde, wenn die Objektverschiebung auch explizit erkannt wurde. Anscheinend wird die explizite Erkennung von Objektverschiebung dann besser gewährleistet, wenn zuvor das Objekt mit seiner genauen Position innerhalb der Szene abgespeichert wurde und dadurch eine Art Positionsüberprüfung durch Blickbewegungen zum früheren Standort des Objektes möglich ist.

Letztlich stellte sich die Frage, zu welchem Zeitpunkt Objektverschiebungen beim Betrachten der Szenen Aufmerksamkeit beziehungsweise Blickbewegungen modulieren. Ähnlich der Ergebnisse aus Studie 2, fand eine Modulation der Blickbewegungssteuerung erst statt, nachdem das jeweilige Objekt fixiert wurde. Auch Tatler und Kollegen (2005) fanden nur dann eine Verbesserung der Gedächtnisleistung für die Position von Objekten, wenn diese direkt fixiert wurden.

Studie 3 konnte also zeigen, dass die internen Repräsentationen von Szenen durchaus detailliert sein können und es ermöglichen selbst geringfügige Objektverschiebungen zu erkennen. Das Detektieren von Objektverschiebungen spiegelt sich nicht nur in deren explizitem Bericht, sondern auch in der Steuerung von Blickbewegungen wider. Wenn die im episodischen Kurzzeitgedächtnis gespeicherte Repräsentation einer Szene nicht mit der momentan generierten übereinstimmt, scheint ein sogenanntes "Mismatch-Signal" Aufmerksamkeit an die Stelle der Nichtübereinstimmung zu lenken, was wiederum die Augen dorthin lenkt (vgl. Rao & Ballard, 1999). Wenn zudem eine gewisse Schwelle überschritten wird, kann diese Nichtübereinstimmung auch explizit als Objektverschiebung berichtet werden. Weitere Studien könnten zum Beispiel die Detailliertheit von Szenenrepräsentationen in Abhängigkeit der Darbietungszeit oder der Anzahl der Präsentationen untersuchen. Eine offene Frage wäre hierbei, wie viele Darbietungen einer Szene nötig sind, um explizite beziehungsweise implizite Detektion von Objektveränderungen zu ermöglichen.

Schlussfolgerungen

Die in der vorliegenden Arbeit zusammengefassten Studien konnten zeigen, dass die Allokation von Aufmerksamkeit bei der Szenenwahrnehmung nicht allein vom visuellen Input gesteuert wird, sondern vielmehr zu hohem Grade auch kognitiver Kontrolle unterliegt. Diese umfasst sowohl frühere Erfahrungen und daraus resultierende Erwartungen, als auch spezifische kognitive Fähigkeiten und Ziele der Personen. Das schnelle Erkennen der Bedeutung einer natürlichen Szene — der sogenannten "gist" —lässt durch die Verbindung mit früheren Erfahrungen und gelernten Gesetzmäßigkeiten die Inferenz wahrscheinlicher Objektlokationen für die effektive Einschränkung des Suchraumes zu. Ferner führt die Verletzung von Erwartungen und Gesetzmäßigkeiten bezüglich des semantischen und syntaktischen Aufbaus einer Szene zu einer Modulation der Aufmerksamkeitssteuerung. Dabei scheint es eine wichtige Rolle zu spielen, wie stark diese Erwartungen sind. Erwartungen an den Aufbau einer Szene können aus lange zurückliegenden, immer wiederkehrenden Erfahrungen mit diversen Szenentypen im Allgemeinen resultieren, oder aber auch episodisch so aufgebaut werden, dass kleine Objektveränderungen innerhalb einer neu gelernten Szene Aufmerksamkeit auf sich ziehen können. All dies legt letztlich den Schluss nahe, dass die Wahrnehmung von Szenen konstruktiver Natur ist, wobei sich die Aufmerksamkeitsverlagerung innerhalb von natürlichen Szenen und die Konstruktion einer Gesamtrepräsentation gegenseitig bedingen.

References

Becker, M. W., & Pashler, H. (2002). Volatile visual representations: Failing to detect changes in recently processed information. *Psychonomic Bulletin & Review*, *9(4)*, 744-750.

Becker, M. W., Pashler, H., & Lubin, J. (2007). Object-intrinsic oddities draw early saccades. *Journal of Experimental Psychology: Human, Perception and Performance*, *35(1)*, 20-30.

Biederman, I., Mezzanote, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143-177.

Brockmole, J. R., & Henderson, J. M. (2005a). Object appearance, disappearance, and attention prioritization in real-world scenes. *Psychonomic Bulletin and Review*, *12*, 1061-1067.

Brockmole, J. R., & Henderson, J. M. (2005b). Prioritization of new objects in real-world scenes: Evidence from eye movements. *Journal of Experimental Psychology: Human Perception and Performance*, *31(5)*, 857-868.

Brockmole, J. R., & Henderson, J. M. (in press). Prioritizing new objects for eye fixation in real-world scenes: Effects of Object-Scene Consistency. *Visual Cognition*.

Brockmole, J. R., Hambrick, D. Z., Windisch, D. J., & Henderson, J. M. (in press). The role of meaning in contextual cueing: Evidence from chess expertise. *The Quarterly Journal of Experimental Psychology*.

Bublak, P., Finke, K., Krummenacher, J., Preger, R., Kyllingsbaek, S., Müller, H. J., & Schneider, W. X. (2005). Usability of a theory of visual attention (TVA) for parameter-based measurement of attention II: evidence from two patients with frontal or parietal damage. *Journal of the International Neuropsychological Society*, *11*, 843-854.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*, 523-547.

Bundesen, C. (1998). A computational theory of visual attention. *Philosophical Transactions of the Royal Society of London Series B - Biological Sciences*, *353*, 1271-1281. Bundesen, C., Habekost, T., & Kyllingsbaek, S. (2005). A neural theory of visual attention: bridging cognition and neurophysiology. *Psychological Review*, *112*, 291-328.

Bundesen, C., Habekost, T., & Kyllingsbaek, S. (2005). A neural theory of visual attention: bridging cognition and neurophysiology. *Psychological Review*, *112*, 291-328.

Castelhano, M., & Henderson, J. M. (2005). Incidental visual memory for objects in scenes. *Visual Cognition*, *12(6)*, 1017-1040.

Castelhano, M., & Henderson, J. M. (2007). Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *33(4)*, 753-763.

Chun, M. M., & Jiang, Y. (1999). Top-down attentional guidance based on implicit learning of visual covariation. *Psychological Science*, *10*, 360-365.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science*, *15(8)*, 559-564.

DeGraef, P., Christiaens, D., & d'Ydewalle, G. (1990). perceptual effects of scene context on object identification. *Psychological Research*, *52*, 317-329.

Dehaene, S., Changeux, J., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: a testable taxonomy. *Trends in Cognitive Sciences*, *10(5)*, 204-211.

Deubel, H. (1996). Visual processing and cognitive factors in the generation of saccadic eye movements. In W. Prinz and B. Bridgeman (Eds.), *Handbook of Perception and Action (Perception)*, London, New York: Academic Press, (pp. 143-189).

Deubel, H. (2003). Attention and awareness in goal-directed eye and hand movements. In: N. Elsner and H. Zimmermann (Eds.), *The Neurosciences from Basic Research to Therapy* (p. 338). Stuttgart: Thieme.

Deubel, H. & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36(12)*, 1827-1837.

Deubel, H., O'Regan, K., & Radach, R. (2000). Attention, information processing and eye movement control. In: Kennedy, A., Radach, R., Heller, D. & Pynte, J. (Eds). *Reading as a Perceptual Process* (pp. 355-376). Elsevier: Oxford.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193-222.

Duncan, J., Bundesen, C., Olson, A., Humphreys, G., Chavda, S., & Shibuya, H. (1999). Systematic analysis of deficits in visual attention. *Journal of Experimental Psychology: General*, *128*, 450-478.

Eden, G. E., Stein, J. E., Wood, H. M., & Wood, E. B. (1994). Difference in eye movements and reading problems in dyslexic and normal children. *Vision Research*, *34*, 1345-1358.

Finke, K., Bublak, P., Dose, M., Müller, H. J., & Schneider, W. X. (2006). Parameter-based assessment of spatial and non-spatial attentional deficits in Huntington's disease. *Brain*, *129*, 1137-1151.

Finke, K., Bublak, P., Krummenacher, J., Kyllingsbæk, S., Müller, H.J. & Schneider, W. X. (2005). Usability of a theory of visual attention (TVA) for parameter-based measurement of attention I: Evidence from normal subjects. *Journal of the International Neuropsychological Society*, *11*, 832-842.

Friederici, A. D., & Weissenborn, J. (2007). Mapping sentence form onto meaning: The syntax-semantic interface. *Brain Research*, *1146*, 50-58.

Friedman, A. (1979). Framing pictures: The role of knowledge in automatized encoding and memory for gist. *Journal of Experimental Psychology: General*, *108*, 316-355.

Gareze, L., & Findlay, J. M. (2007). Absence of scene context effects in object detection and eye gaze capture. In. R. P. G. van Gompel, M. H. Fischer, W. S. Murray and R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 618-637).

Greene, M. R., & Oliva, A. (2006). Natural Scene Categorization from Conjunctions of Ecological Global Properties. *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vancouver, July (pp. 291-296).

Green, C., & Hummel, J .E. (2006). Familiar interacting object pairs are perceptually grouped. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 1107-1119.

Habekost, T., & Rostrup, E. (2007). Visual attention capacity after right hemisphere lesions. *Neuropsychologia*, *45(7)*, 1474-1488.

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task constraints in visual working memory. *Vision Research*, *38*, 125-137.

Henderson, J. M. (2007). Regarding Scenes. *Current Directions in Psychological Science*, *16(4)*, 219-222.

Henderson, J. M., & Castelhano, M. S. (2005). Eye movements and visual memory for scenes. In G. Underwood (Ed.), *Cognitive processes in eye guidance* (pp. 213-235). New York: Oxford University Press.

Henderson, J. M., & Hollingworth, A. (1999a). The role of fixation position in detecting scene changes across saccades. *Psychological Science*, *5*, 438-443.

Henderson, J. M., & Hollingworth, A. (1999b). High-level scene perception. *Annual Review of Psychology*, *50*, 243-271.

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

Henderson, J. M., Weeks, P. A., & Hollingworth, A. (1999). The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 210-228.

Hollingworth, A. (2005). The relationship between online visual representation of a scene and long-term scene memory. *Journal of Experimental Psychologe: Learning, Memory, & Cognition, 31*(3), 396-411.

Hollingworth, A. (2006). Visual memory for natural scenes: Evidence from change detection and visual search. *Visual Cognition*, *14*, 781-807.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, *127*, 398-415.

Hollingworth, A., & Henderson, J. M. (2002). Accurate visual memory for previously attended objects in natural scenes. *Journal of Experimental Psychology: Human Perception and Performance, 28,* 113-136.

Hollingworth, A., Williams, C. C., & Henderson, J. M. (2001). To see and remember: Visually specific information is retained in memory from previously attended objects in natural scenes. *Psychonomic Bulletin & Review*, *8*, 761-768.

Horowitz, T. S., & Wolfe, J. M. (1998). Visual Search has no memory. *Nature*, *394*, 575-577.

Hung, J., Driver, J., & Walsh, V. (2005). Visual selection and posterior parietal cortex: effects of repetitive transcranial magnetic stimulation on partial report analyzed by Bundesen's theory of visual attention. *Journal of Neuroscience*, 25(42), 9602-9612.

Hutzler, F., & Wimmer, H. (2004). Eye Movements of Dyslexic Children when Reading in a Regular Orthography. *Brain and Language*, *89(1)*, 235-242.

Itti, L., & Baldi, P. (2005). Bayesian surprise attracts human attention. *Proceedings in Neural Information Processing Systems* (Vol. 19, pp 1-8). Cambridge, MA: MIT Press.

Itti, L., & Koch, C. (2000). A saliency-based mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*, 1489-1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1254–1259.Irwin, D. E. (1992). Memory for position and identity across eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*, 307-317.

Jonides, J., &Yantis, S. (1988). Uniqueness of abrupt visual onset in capturing attention. *Perception and Psychophysics*, *43*, 346-354.

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*, 3286-3297.

Kahneman, D., & Treisman, A. (1984). Changing views of attention and automaticity. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 29-61). New York: Academic Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*, 175-219.

Karacan, H., & Hayhoe, M. M. (2008). Is attention drawn to changes in familiar scenes? *Visual Cognition*, *16(2/3)*, 356-374.

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

Kyllingsbaek, S. (2006). Modeling visual attention. *Behavior Research Methods*, *38(1)*, 123-133.

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 565-572.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279-281.

McCotter, M., Gosselin, F., Sowden, P., & Schyns, P. G. (2006). The use of visual information in natural scenes. *Visual Cognition*, *12*, 938-953.

Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *Journal of Vision*, *6(1)*, 8-17.

Melcher, D., & Morrone, M. C. (2003). Spatiotopic temporal integration of visual motion across saccadic eye movements. *Nature Neuroscience*, *6*, 877-881.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research*, *45*, 205-231.

Oliva, A. (2005). Gist of the Scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 251-256). San Diego, CA: Elsevier.

Oliva, A., & Schyns, P. G. (1997). Course blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, *34*, 72-107.

Oliva, A., & Schyns, P. G. (2000). Diagnostic colors meduate scene recognition. *Cognitive Psychology*, *41*, 176-210.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: the role of global image features in recognition. *Progress in Brain Research*, *155*, 23-36.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11(12)*, 520-527.

O'Regan, J. K. (1992). Solving the' 'real'' mysteries of visual perception: The world as an outside memory. *Canadian Journal of Psychology*, *46(3)*, 461-488.

O'Regan, J. K., Rensink, R. A., & Clark, J. J. (1999). Change blindness as a result of "mudsplashes". *Nature*, *398*, 34.

O'Regan, K., Deubel, H., Clark, J. J., & Rensink, R. A. (2000). Picture changes during blinks: Looking without seeing and seeing without looking. *Visual Cognition*, *7*, 191-211.

Palolahti, M., Leino, S., Jokela, M., Kopra, K., & Paavilainen, P. (2005). Event-related potentials suggest early interaction between syntax and semantics during on-line sentence comprehension. *Neuroscience Letters*, *384*, 222-227.

Paprotta, I., Deubel, H., & Schneider, W. X. (1999). Object recognition and goal-directed eye or hand movements are coupled by visual attention (p. 241-248). In: W. Becker, H. Deubel & Th. Mergner (Ed.) *Current Oculomotor Research: Physiological and Psychological Aspects*. New York, London: Plenum.

Parkhurst, D., Law, K., & Niebur, E. (2002). Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, *42*, 107-123.

Potter, M. C. (1975). Meaning in visual scenes. *Science*, *187*, 965-966.

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience, 2, 79-87.

Rayner, K., Castelhano, M. S., & Yang, J. (in press). Viewing task influences eye movements during active scene perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition.*

Rensink, R. A. (2000). The dynamic representation of scenes. *Visual Cognition*, *7*, 17-42.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology*, *53(1)*, 245-277.

Ryan J. D., & Cohen N. J. (2004). The nature of change detection and on-line representations of scenes. *Journal of Experimental Psychology: Human Perception and Performance*, *30(5)*, 988-1015.

Schneider, W. X. (1995). VAM: A Neuro-Cognitive Model for Visual Attention Control of Segmentation, Object Recognition, and Space-based Motor Action, *Visual Cognition*, *2*, 331-375.

Schneider, W. X. (1999). Visual-spatial working memory, attention, and scene representation: A neuro-cognitive theory. *Psychological Research*, *62*, 220-236.

Schneider, W. X., & Deubel, H. (2002). Selection-for-perception and selection-for- spatialmotor-action are coupled by visual attention: a review of recent findings and new evidence from stimulus-driven saccade control. In W. Prinz & B. Hommel (Eds.), *Attention and Performance XIX: Common Mechanisms in Perception and Action*, 609–627. Oxford: Oxford University Press.

Schyns, P. G., & Oliva, A. (1994). From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition. *Psychological Science*, *5*, 195-200.

Simons, D. J. (2000). Current Approaches to Change Blindness. *Visual Cognition*, *7* (1/2/3), 1-15.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, *74*.

Tatler, B. W., Gilchrist, I. D., & Rusted, J. (2003). The time course of abstract visual representation. *Perception*, *32*, 579-592.

Tatler, B. W., Gilchrist, I. D., & Land, M. F. (2005). Visual memory for objects in naturalistic scenes: From fixations to object files. *The Quarterly Journal of Experimental Psychology*, *58A(5)*, 931-960.

Theeuwes, J. (1994). Stimulus-driven capture and attentional set: Selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 799-806.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in human visual system. *Nature*, *381*, 520-522.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (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.

Underwood, G, & Foulsham, T. (2006). Visual saliency and semantic incongruency influence eye movements when inspecting pictures. *Quarterly Journal of Experimental Psychology*, *59(11)*, 1931-1949.

Underwood, G., Humphreys, L., & Cross, E. (2007). Congruency, saliency, and gist in the inspection of objects in natural scenes. In. R. P. G. van Gompel, M. H. Fischer, W. S. Murray and R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 564-579).

Underwood, G., Templeman, E., Lamming, L., & Foulsham, T. (2008). Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. *Consciousness and Cognition*, *17*, 159-170.

Underwood, G., Chapman, P., Brocklehurst, N., Underwood, J., & Crundall, D. (2003). Visual attention while driving: Sequences of eye fixations made by experienced and novice drivers. *Ergonomics*, *46*, 629-646.

Underwood, G., Foulsham, T., van Loon, E., Humphreys, L., & Bloyce, J. (2006). Eye movements during scene inspection: A test of the saliency map hypothesis. *European Journal of Cognitive Psychology, 18*, 321-342.

Werner, S., & Thies, B. (2000). Is "Change Blindness" attenuated by domain-specific expertise? An Expert-Novices comparison of change detection in football images. *Visual Cognition*, *7(1,2,3)*, 163-173.

Yantis, S. (1998). Control of visual attention. In H. Pashler (Ed.), *Attention* (pp. 233-256). Hove, UK: Psychology Press.

Appendix: Scene Samples