Chapter 8

TRANSSACCADIC RECOGNITION IN SCENE EXPLORATION

PETER DE GRAEF
Univeristy of Leuven, Belgium
Abstract

Contrary to what is the case in theories of reading, theories of object and scene recognition have shown a remarkable blind spot to the fact that visual perception typically involves integration of information across a number of spatiotemporally discrete fixations. I present two main reasons which I think are responsible for this neglect: the apparent power of single-shot perception and the contested nature of transsaccadic integration. Both reasons are claimed to be flawed, as argued on the basis of a review of existing evidence and the presentation of some new data on transsaccadic preview benefits. It is concluded that a transsaccadic theory of object and scene recognition should be put on the agenda, inspired by new insights on the interaction of feed-forward and re-entrant visual processing streams.
1. Introduction

Visual acuity is not uniform across the retina and not all aspects of a visual stimulus are equally salient or equally pertinent to a viewer’s current task. Inevitably, this leads the visual system to constantly sample different aspects and locations in its environment, which produces a continuous series of fixations and saccades that bring new stimulus aspects or locations into the processing focus. This implies that any visual entity is typically viewed in a series of discrete samples collected at different times and at different retinal and/or spatial coordinates resulting in qualitative differences in the informational content of each of the samples. The logical question therefore is whether and how these samples are integrated to construct a coherent percept.

In reading research, the inherent multi-sample nature of visual processing is at the core of an ever-increasing number of studies. The functional characteristics of eye movements in reading have been studied for well over a century (see Rayner, 1998, for a comprehensive review). With the development of new theoretical frameworks of language processing and technological advances in eye-tracking equipment and software in the 1970s (see Wade & Tatler, 2005, for a history of eye movement measurement), this has culminated in detailed studies of how overt eye movements are related to the ongoing covert processes of text comprehension. An important part of this work has focussed on the question of whether and how text samples acquired on successive fixations are integrated. Integration has been studied across a single fixation–saccade–fixation cycle to determine whether and how a peripheral preview of the word that is about to be fixated could be integrated with the subsequent foveal view of that word on the next fixation and thus facilitate word recognition (e.g., Balota, Pollatsek, & Rayner, 1985; Hyöna, Niemi, & Underwood, 1989; McConkie & Zola, 1979; Pollatsek, Lesch, Morris, & Rayner, 1992; Rayner & Well, 1996). Processes of integration have also been studied across multiple fixation–saccade–fixation cycles to understand how individual words are added into a developing syntactic and semantic representation of the sentence (e.g., Altmann, van Nice, Garnham, & Henstra, 1998; Brysbaert & Mitchell, 1996; Van Gompel, Pickering, Traxler, 2001) or, more globally, into a coherent comprehension of the discourse contained in the text (e.g., Cook & Myers, 2004; Kambe, Rayner, & Duffy, 2001).

Somewhat surprisingly, the great interest of reading research in the integration of information samples across eye movements has not been matched in other flourishing domains of vision science such as object and scene perception. In a sense, most prominent models of object and scene recognition have adopted what could be called an in vitro approach. They outline in great computational detail how object and scene structure can be recovered from the image, but the image is implicitly assumed to be a single, indefinitely sustained, uniform-acuity projection of scene and objects in the optic array (e.g., Biederman, 1987; Edelman, 1999; Oliva & Torralba, 2001; Tarr & Bülthoff, 1998). It has been tacitly ignored that in real-life perception any scene component or object has generally been glimpsed in peripheral vision before it is actually fixated, and that specific scene parts or objects often receive multiple distinct fixations during the first visual inspection of the scene or object. In both cases, the visual system will have at
its disposal multiple distinct samples of the same stimulus, which can all be used in the recognition process.

While sample integration has not been incorporated in the dominant theoretical frameworks in object and scene recognition it would be false to claim that it has not been studied in object and scene perception. Recently, a whole new line of work is emerging aimed at identifying the buildup and stability of scene and object memory representations that are developed over the course of multiple fixation–saccade–fixation cycles (e.g., Hollingworth, 2004; Hollingworth & Henderson, 2002; Melcher, 2001, 2006; Tatler, Gilchrist, & Land, 2005; Tatler, Gilchrist & Rusted, 2003). In addition, for the past 25 years there has been a steady output of papers posing the question whether and how the visual contents that are acquired during a single fixation–saccade–fixation cycle are integrated into one percept. More recently, the notion of transsaccadic perceptual integration is gaining increased acceptance as the neurophysiological underpinnings of this mechanism are starting to become clear (e.g., Khayat, Spekreijse, & Roelfsema, 2004; Melcher & Morrone, 2003; Prime, Niemeier, & Crawford, 2006).

However, in spite of this growing interest in multi-sample integration in visual perception, the concept has not (yet) made it into the mainstream of models accounting for the basic level identification of an object or a scene. In my opinion, there are two main reasons why this is so. First, there is increasingly compelling evidence that single-shot perception is very powerful, that is, that object and scene recognition can be achieved within a single fixation of modal duration (around 220 ms according to an overview presented by Henderson & Hollingworth, 1998). Second, in recent years there has been a strong line of empirical and theoretical work arguing the proposition that transsaccadic integration does not exist. In what follows, I will discuss these two reasons and I will attempt to demonstrate that they are insufficient grounds to further delay the development of a genuine transsaccadic theory of object and scene recognition.

2. The power of single-shot perception

2.1. Recognition at a glance

In the past, it has been demonstrated repeatedly that picture presentations well below the modal fixation duration are sufficient to recognize the general gist of a scene (e.g., Biederman, Mezzanotte, & Rabinowitz, 1982; Intraub, 1981; Potter, 1976) or to identify a depicted object (e.g., Biederman & Ju, 1988). Even when foveal masking dynamically curtailed stimulus presentation times on every fixation during extended scene exploration, viewers only seemed to require 50–70 ms of unmasked fixational content to allow for normal scene

---

1 It is not my intention to extensively review this research, several excellent reviews are already available (e.g., O’Regan, 1992; Pollatsek & Rayner, 1992; Bridgeman, Van der Heijden, & Velichkovsky, 1994; Irwin, 1996; McConkie & Currie, 1996; Verfaillie et al., 2001; Deubel, Schneider & Bridgeman, 2002; Germeys, De Graef & Verfaillie, 2002; Henderson & Hollingworth, 2003a).
exploration (van Diepen, De Graef, & d’Ydewalle, 1995). In recent years, this impressive power of single-shot perception has been explored to its limits in a number of studies trying to detail the chronometry of visual object and scene categorization. In a seminal paper, Thorpe, Fize, and Marlot (1996) presented their participants with a central 20 ms exposure of a natural scene and required participants to release a button if they saw an animal (i.e., a go/no-go task). Event-related potentials (ERPs) were recorded throughout the task and showed a strong frontal negativity about 150 ms after the onset of stimuli for which participants correctly responded that there was no animal (i.e., on no-go trials). Thorpe et al. (1996) interpreted this as showing that 150 ms of processing time was sufficient to analyse the entire stimulus to such an extent that the presence of an animal could be ruled out.

However, one could also argue that 150 ms is a conservative estimate of the speed with which a stimulus can be processed because faster ERP effects have been documented (e.g., Delorme, Rousselet, Macé, & Fabre-Thorpe, 2004) for tasks involving specific image recognition (i.e., of a learned face) rather than overall image categorization (i.e., in animal/non-animal categories). Similarly, Keysers, Xiao, Földiak, and Perrett (2001) demonstrated that 14 ms of undistorted stimulus availability was enough to record target-selective activity in macaque temporal cortex, and that human target-detection performance was above-chance at a presentation rate of one image per 14 ms. Additional converging evidence for very fast stimulus processing comes from MEG studies indicating that 20 ms of natural scene processing is sufficient for above-chance discrimination from a distractor scene (Rieger, Braun, Bülthoff, & Gegenfurtner, 2005).

2.2. But is it recognition?

In view of these benchmarks of the visual system, it appears justified to conclude that all the visual processing that needs to be done on the input encoded during a single fixation can easily be completed within the course of that fixation, which raises the question why object and scene recognition would even need transsaccadic processing of multiple fixation samples? The answer to this question lies in the interpretation of the processes that are reflected in the fast physiological and behavioral effects that were cited above. Most of the tasks for which these effects were demonstrated involve the discrimination of specific target images from visually different distractor images. Therefore, one may wonder whether the results reveal true object and scene recognition or rather a sensitivity to low-level visual dissimilarities between target and distractor. VanRullen and Thorpe (2001) have argued that estimates of object recognition time can only be inferred from effects which reflect a task-dependent categorization of a stimulus as a target rather than a distractor. To obtain such estimates they employed the go/no-go categorization task as used by Thorpe et al. (1996, see above), and presented participants with blocks of trials in which the same two categories of images (animals and vehicles) alternately served as targets and distractors for the categorization task. ERPs recorded in this design showed differential activity for animals and vehicles after 75–80 ms following stimulus onset, regardless of which image category served as target. In addition, a target–distractor
difference was observed at 150 ms poststimulus regardless of which image category served as target. Based on these data, VanRullen and Thorpe (2001) proposed that the early effect reflects a response to visual differences between the image categories, while the later effect provides an upper estimate of image recognition time.

Recently, however, Johnson and Olshausen (2003) argued that even the target–distractor effect at 150 ms poststimulus could be attributed to low-level visual dissimilarities. To ensure that ERP effects truly reflected target or distractor categorization and not a pre-categorical sensitivity to systematic visual differences between target and distractor, they ran the go/no-go categorization task with a different target on every trial. Across trials, specific images were equally often target as they were distractor, allowing for an assessment of the categorization response without the confound of visual target–distractor differences. Under these conditions, a categorization-related ERP effect emerged with a variable latency between 150 and 300 ms poststimulus. When compared to ERPs which Johnson and Olshausen (2003) recorded for a replication of the original animal vs non-animal categorization task, the same late effect showed, along with a faster fixed-latency target–distractor difference at 150 ms poststimulus. Taking into account that a spatial frequency analysis of animal and distractor images showed consistent differences in the power spectrum of the two image categories, Johnson and Olshausen concluded that visual category differences can be picked up within 150 ms, but recognition-based effects take between 150 and 300 ms to surface. Based on a study of priming effects on ERPs recorded in an object-naming task, Schendan and Kutas (2003) derived a similar estimate of the time required to access a stored structural representation of the object that is being viewed, that is 150–250 ms. Large, Kiss, and McMullen (2004) even extended these intervals by showing that differences in the type of object categorization that was required from participants only had ERP correlates after about 300 ms from stimulus onset. Specifically, at that point ERP responses to targets were stronger for superordinate categorizations (e.g., animal vs vehicle) than for basic level (e.g., dog vs cat) and for subordinate categorizations (e.g., beagle vs collie). Large et al. proposed that this shows a gross to detailed progression in object processing, with longer processing times required as a more specific identification is attempted.

2.3. The need for a second shot

Initial studies of the time course of object and scene recognition seemed to clearly show that the time of one modal duration fixation in scene perception (i.e., about 220 ms) is all that is needed to complete recognition of the fixated object or scene. However, more controlled follow-up work has now indicated that this time interval may not be sufficient to achieve object or scene recognition. When recognition processes are studied in the context of realistic scene exploration, this discrepancy between recognition and fixation time creates a complex puzzle. Consider what happens when a viewer freely explores an everyday scene. Eye recording in such a situation typically reveals a scanpath that consists of an alternating sequence of one or more within-object fixations followed by a between-object saccade (De Graef, Christiaens, & d’Ydewalle, 1990; Henderson, 2003;
Land, Mennie, & Rusted, 1999). Logically, this implies that recognition of an object can be achieved in three different ways. First, the gaze may remain on object \( n \) until it is identified, resulting in multiple within-object refixations. Second, visual information encoding may be completed during fixation of object \( n \), while identification is completed during the subsequent between-object saccade to object \( n + 1 \), or even during the first fixation on \( n + 1 \). Third, processing of object \( n \) may start during fixation of object \( n + 1 \), resulting in a presaccadic peripheral preview of \( n \) which is integrated with its subsequent postsaccadic foveal view, allowing for identification of \( n \) during its first fixation.

In reading research, these three mechanisms for transsaccadic (word) identification have all received ample attention in the study of within-word refixations (e.g., Reichle, Rayner, & Pollatsek, 1999), intrasaccadic and/or between-word spill-over effects (e.g., Irwin, 1998), and transsaccadic preview benefits (e.g., Rayner, Liversedge, & White, 2006) including their extreme manifestation in word skipping (e.g., Brysbaert, Drieghe, & Verh, 2005). Moreover, quantitative models of eye-movement control in reading have provided an integrated account of the interplay between the three transsaccadic processes (Reichle, Rayner, & Pollatsek, 2003). In contrast, in scene perception research, treatment of the various possible types of transsaccadic identification has been much less extensive, primarily because the very notion of transsaccadic information integration has been severely challenged.

3. Transsaccadic information integration in scene exploration: The Pit and the Pendulum

3.1. The Pit and the Pendulum

The study of transsaccadic integration in object and scene perception is perhaps best described as an ongoing scientific mystery novel involving the two main props of Edgar Allan Poe’s 1842 story The Pit and the Pendulum. The mystery is how to solve the discrepancy between the fact that the input to the visual system is a temporally and spatially discontinuous train of “snapshots”, and the phenomenological experience of a stable and continuous visual world. Attempts to solve the mystery can be viewed as pendulum swings across a pit. On the one hand we have the pit: the saccadic period of 30–40 ms between fixations, during which visual information intake is suppressed (e.g., Matin, E., 1974; Matin, L., 1986; Ross, Morrone, Goldberg, & Burr, 2001). And on the other hand, we have the pendulum, which in the present context symbolizes scientific opinion swinging back and forth across the pit, with each new swing changing the view on whether or not there is actually something at the bottom of the pit.

3.2. A mirror at the bottom or no bottom at all?

A first swing across the pit launched the notion of an integrative visual buffer (McConkie & Rayner, 1976) or spatiotopic fusion (Feldman, 1985; Henderson &
Hollingworth, 2003a; Jonides, Irwin, & Yantis, 1982) according to which viewers could spatiotopically superimpose a presaccadic, peripheral stimulus and a postsaccadic, foveal stimulus. According to this scheme, transsaccadic integration of pre- and postsaccadic information samples was complete and visually detailed: the pit contained a faithful reflection of presaccadic fixation contents which are merged on a pixel-by-pixel basis with the postsaccadic fixation contents.

On the second swing across the pit, spatiotopic fusion was disproved in a number of studies (e.g., Bridgeman & Mayer, 1983; Henderson, 1997; Irwin, 1991; Pollatsek & Rayner, 1992). The notion of an intrasaccadic void was further reinforced by a flurry of change blindness studies (for an overview, see Simons, 2000; also see Rensink, 2002). In all these studies, the central proposition was that viewers are blind to change in their visual environment as long as the local transient associated with the change is somehow masked. Such masking can be achieved by simultaneously introducing attention-capturing local transients (O’Regan, Rensink, & Clark, 1999) or a global transient such as a grey-out of the image (Rensink, O’Regan, & Clark, 1997), a blink (O’Regan, Deubel, Clark, & Rensink, 2000), or a saccade (Grimes, 1996; Henderson & Hollingworth, 1999). Thus, change only seemed to be detectable on the basis of a within-fixation transient, and not on the basis of a comparison between pre- and postsaccadic images, which appeared to rule out a putative process of transsaccadic integration: the pit was bottomless. In fact, in two influential papers, O’Regan (1992; O’Regan & Noë, 2001) argued that there simply was no pit. Specifically, O’Regan proposed that there is no functional need to internally represent the outside world across a saccade, because that world is always there to be sampled in a fixation whenever needed. As soon as attentive sampling stops, however, viewers are afflicted with inattentional amnesia, that is, the failure to represent components of the visual world which they are no longer attending to (Rensink, 2000; Wolfe, Klempen, & Dahlen, 2000). This does not mean that we have no recollection of the state of the outside world from one fixation to the next: Whenever we have identified a setting or an object it can be stored in an evolving episodic memory trace of the scene we are currently viewing. O’Regan & Noë’s (2001) point, however, is that this representation is a completely post-categorical and verbal summary, making abstraction of visual detail.

3.3. Construction at the bottom of the pit

On the third and currently ongoing swing across the pit, there appears to be a growing consensus that information gathered on multiple fixations during scene exploration is actively and selectively used to construct a task-relevant representation of the scene and objects that are viewed. This construction process is being studied at two levels.

First, there is the level of the overall scene for which, across multiple fixation-saccade-fixation cycles, an internal, episodic model of the contents and spatial layout of that particular scene is developed (Friedman, 1979; De Graef, 1992; Rayner & Pollatsek, 1992; Chun & Nakayama, 2000; Hollingworth & Henderson, 2002; Melcher, 2006; Tatler et al., 2003). Initially, this representation was assumed to make complete abstraction of
visual detail and to be sparse, primarily constrained by the capacity limits of visual short-term memory (Alvarez & Cavanagh, 2004; Irwin & Andrews, 1996; Irwin & Zelinsky, 2002; Luck & Vogel, 1997). More recently, however, several authors have argued that every fixation on a scene leaves a visually detailed sensory trace which the visual system samples to construct a large-capacity online scene representation which contains as much visual detail as is required to perform the task at hand (Hollingworth, 2004; Melcher, 2006; Tatler, 2002; Tatler et al., 2005). Indeed, information selection and integration in an overall scene representation has been found to reflect the goals and stages of execution of even very complex tasks such as block copying (Hayhoe, Bensinger, & Ballard, 1998; Hayhoe, Karn, Magnuson, & Mruczek, 2001), driving in the real world (Crundall, 2005) or in a simulator (Shinoda, Hayhoe, & Shrivastava, 2001), playing chess (Reingold & Charness, 2005), sandwich- and tea-making (Land & Hayhoe, 2001), or playing ping-pong (Land & Furneaux, 1997).

Second, there is the level of the individual object, identified across a single fixation–saccade–fixation cycle during which object information is sampled and integrated from a sequence of peripheral and foveal glimpses. As was the case for scenes, the dominant view on transsaccadic object perception was that it consisted of postcategorical, conceptual and non-visual priming in the object lexicon between the presaccadically activated node for the extrafoveally previewed saccade target and the postsaccadically activated node for that target once it is foveated (e.g., Gordon & Irwin, 2000). This view was supported by numerous demonstrations of failures to notice or react to intrasaccadic changes of visual object detail such as visible object contours (Henderson, 1997), object position (Pollatsek, Rayner, & Henderson, 1990), object size (Pollatsek, Rayner, & Collins, 1984), object in-depth orientation (Henderson & Hollingworth, 1999), individual object parts (Carlson, 1999), or even object exemplar (Archambault, O’Donnell, & Schyns, 1999). Countering this view, three lines of research can be distinguished which advocate the idea that transsaccadic object perception is not merely a matter of postcategorical priming by a peripherally attended saccade target, but rather involves the mandatory integration of specific visual object features across the saccade, even for objects that were not peripherally attended.

3.4. Previous research indicating transsaccadic integration of visual object detail

In a first line, change-blindness studies are challenged as having very limited relevance for understanding representation as it develops across consecutive fixation–saccade cycles (De Graef et al., 2001). First, change blindness disappears entirely when one is warned in advance about the location and type of change that will occur. In contrast, the failure to note certain intrasaccadic changes is resistant to such advance warning, indicating that very different mechanisms are at work (e.g., Verfaillie, 1997). Second, the finding that a viewer cannot overtly report a change does not mean that the changed information was not represented (e.g., Fernandez-Duque, Grossi, Thornton, & Neville, 2003; Fernandez-Duque & Thornton, 2000). This was recently confirmed for scene exploration by showing that when objects were intrasaccadically changed in a scene and when that change
was not detected, fixation times on the changed object were still elevated relative to a
no-change control (Henderson & Hollingworth, 2003b). In addition, a variety of change-
detection studies using temporary postsaccadic blanking of the visual stimulus have
revealed a richly detailed but maskable and rapidly decaying transsaccadic representation
which is formed after every fixation (De Graef & Verfaillie, 2002; Deubel, Schneider, &
Bridgeman, 2002; Gysen, Verfaillie, & De Graef, 2002). Recently, this earlier criticism of
change blindness as an inappropriate basis for understanding transsaccadic representation
has received eloquent support from some of the authors that launched the systematic
study of change blindness (Simons & Ambinder, 2005; Simons & Rensink, 2005).

A second line of work, which counters the claim that visual object detail is not
mandatorily represented across saccades, addresses the proposal that we have no on-line
representation of those objects which we did not selectively attend to (e.g., Holling-
worth, 2004). Contrary to this view, Germeyns, De Graef, and Verfaillie (2002) demon-
strated location-specific transsaccadic preview benefits on gaze durations for contextual
or bystander objects: that is, objects that were present before, during, and after the saccade
to another object. Although it is clear that, prior to the saccade to the target object, these
bystander objects were abandoned by attention (if they ever were attended in the first
place), Germeyns et al. (2002) found that the bystanders were easier to identify than a new
object at the same location. Importantly, this transsaccadic preview benefit only occurred
when the bystander retained its location throughout the fixation–saccade–fixation cycle.
This rules out an explanation in terms of location-independent postcategorial priming
between stored object representations in a long-term object lexicon and firmly places the
effect at the level of an on-line pre-attentive representation of the current visual stimulus.

The third line of work defending the notion of mandatory transsaccadic object percep-
tion involves transsaccadic object preview studies, aimed at revealing whether postsac-
cadic processing of an object on fixation n is modulated by presaccadic processing of
that object on fixation n–1. To achieve this, the relation between pre- and postsaccadic
object appearance is systematically manipulated in order to identify the object prop-
erties that are integrated across saccades. Specifically, during the critical saccade towards
the target object, the presaccadic object image is replaced by the postsaccadic object
image thus selectively altering or preserving particular features of the presaccadic image.
Importantly, the transient associated with the change itself is imperceptible due to sac-
cadic suppression, and participants are not asked to monitor and detect changes. Instead,
measures of subsequent ease of identification of the postsaccadic object (e.g., naming
latency or gaze durations) are collected to determine whether the (violation of) transsacc-
adic correspondence between pre- and postsaccadic object images has any effect.2

2 The results of transsaccadic object preview studies have been corroborated by transsaccadic object change
detection studies in which viewers are explicitly instructed to detect intrasaccadic changes by comparing pre-
and postsaccadic object images (e.g., Verfaillie & De Graef, 2000; De Graef, Verfaillie, & Lamote, 2001;
Henderson & Hollingworth, 2003b). However, while transsaccadic integration can be inferred from transsaccadic
change detection, it is also possible that the explicit nature of the task activates a comparison process which is
not a mandatory part of transsaccadic perception.
For instance, Henderson and Siefert (1999; 2001) found increased transsaccadic preview benefits on object-naming latencies and gaze durations when the presaccadic left-right orientation of an object was preserved across a saccade. As mentioned above, Germeys et al. (2002) found shorter object gaze durations for all objects (i.e., saccade targets and saccade bystanders) that maintained their exact spatial position across a saccade. Earlier, the transsaccadic preservation of visual features such as global object shape had already been noted by Pollatsek et al. (1984). Specifically, these authors observed transsaccadic preview benefits on naming latencies for a fixated object that had been preceded by another, visually similar object peripherally previewed at the same location (e.g., a carrot preceding a baseball bat).

Based on these studies, several authors have proposed that transsaccadic object perception in scenes should be regarded as a basic, functional routine that has evolved in our everyday visuo-motor interaction with the world (Hayhoe, 2000; Verfaillie et al., 2001). The purpose of this routine is to take advantage of the fact that generally we are allowed more than one fixation on a given scene. For the objects in that scene this provides us with the opportunity to integrate object information sampled from a sequence of initially peripheral and ultimately foveal glimpses. Thus, foveal object identification can be jump-started by preliminary peripheral processing. In other words, while a single fixation may in principle be sufficient to identify an object, the preferred *modus operandi* is to increase speed and reliability of object identification by transsaccadic integration of foveal and extrafoveal evidence from multiple fixations. In the next section, I will report some recent evidence in support of this claim.

4. Some new data: Transsaccadic object recognition in scenes

To the best of my knowledge, mandatory transsaccadic object preview effects have never been examined in the context of full, realistic scenes, but have always been established with isolated objects or sparse arrays of isolated objects. This is a potentially important problem if one wants to claim that transsaccadic object perception is the default mode for object recognition in scenes. Indeed, because objects in scenes are camouflaged and/or laterally masked by the background and nearby companion objects, the information extracted from a presaccadic, peripheral preview may be of such a low quality that it provides no constraints whatsoever on the subsequent, postsaccadic foveal identification process.

Thus, what was needed was a study in which viewers explore a complex scene in which their attention and subsequent gaze shift are guided towards a predesignated target location. Before they actually make the required saccade, the target location should be occupied by the postsaccadic target object (the *identical preview* condition), or by a different object (different preview condition), or by no object at all (no preview condition).

---

3 Henderson and Hollingworth (2003b) did record gaze durations on intrasaccadically changed objects during full scene exploration, but this was in the context of an explicit change detection task making it difficult to decide whether the observed effects are mandatory or due to the requested comparison process.
During the saccade, the preview should be replaced by the same target object in all preview conditions. Upon fixation of the target, fixation parameters can be recorded in order to infer ease of target processing.

4.1. Recording object-fixation parameters in a non-object search task

To fill in this lacuna, a study was designed in which participants were presented with realistic black-on-white line drawings of real-world scenes from a pool made available by van Diepen and De Graef (1994) and extended by Hollingworth and Henderson (1998). Participants had to scan the scene in order to count the number of non-objects (a task first used in De Graef, Christiaens, & d’Ydewalle, 1990). The notion non-object was borrowed from Kroll and Potter (1984) and refers to a form that resembles real objects in terms of size range and the presence of a clear and closed part-structure but is completely meaningless and nameless (see Figure 3 for an example). By including a variable number of these non-objects in a scene populated with real objects, viewers are encouraged to scan the scene and saccade from one object-like form to the next in search of non-objects. In doing so, they also fixate real objects at which point they only have to determine whether the pictured object matches an object representation in long-term memory and then move on to the next object-like form. This object decision does not require memorization of object characteristics, nor does it require cognitive enhancement of degraded visual object information. In that sense the required object processing approximates what we usually do with most objects in our visual environment, namely register their presence and familiarity without explicitly retrieving their name, identity, or semantics.

During the non-object search task, eye-movement patterns were recorded, allowing the use of object-fixation parameters as a measure of ease of object identification for the real objects, non-object fixation times are never used in the analyses. Of course, there is no consensus on a single, “ideal” oculomotor measure of object perceptibility. In reading research, Inhoff and Radach (1998) found no less than 14 different oculomotor measures which were all claimed to reflect some type of perceptual or cognitive processing. In research on scene perception a similar variety of measures has been used, mostly in an exploratory fashion to determine whether they showed any effects at all (e.g., Antes & Penland, 1981; Henderson, Weeks, & Hollingworth, 1999). For this study, I selected three oculomotor measures based on a simple qualitative model of how ease of object identification could be reflected in the pattern of fixations on that object.

Specifically, it was assumed that when an object is fixated, perceptual encoding increases activation in the object lexicon until one entry reaches a criterion level and triggers object identification. In this view, objects are easier to process when their identification thresholds are reached more quickly due to a higher rate of activation accumulation and/or a lower identification threshold. There is a good deal of evidence that activation rate is directly reflected in fixation duration. Temporarily masking or degrading the stimulus in foveal vision increases fixation durations (Loftus, Kaufman, Nishimoto, & Ruthruff, 1992; Sanders & van Duren, 1998; van Diepen, Ruelens, & d’Ydewalle, 1999;
van Duren & Sanders, 1992). Similarly, higher identification thresholds affect fixation durations. For instance, words with a low frequency in the language (Rayner & Duffy, 1986; Rayner, Sereno, & Raney, 1996) or a low level of predictability from the sentence context (Rayner & Well, 1996; Sereno, 1995) were found to receive longer fixations. Hence, it seems a reasonable hypothesis that object-fixation time reflects the ease with which the fixated object can be encoded and identified.

However, objects generally are fixated more than once during scene exploration so the question arises which object-fixation time reflects ease of object identification. The first restriction is that only first-pass fixation times will be considered, that is, fixations during the first visit to the object. When the eye re-enters an object it has left before, this can be incidental in which case object processing will be atypically fast because the object is still active in the lexicon, or because processing is terminated as soon as the system detects it was already there. Alternatively, a re-entry can be purposeful to check for change or to resolve inconsistencies between the initial object interpretation and the contextual information gathered from other parts of the scene. In this case, fixation times are bound to be atypically slow and reflect a great deal of post-perceptual integration and comparison.

Even when only first object passes are considered, refixations are quite common and a variety of processing measures can and have been defined: single-fixation durations (i.e., discard all multiple-fixation passes), first-fixation duration, second-fixation or refixation duration, first-gaze or summed-fixation duration, average fixation duration, and number of first-pass fixations. From these measures, I selected first-fixation duration, first-gaze duration, and number of first-pass fixations, henceforth first-pass density. The rationale behind this selection is that on every first pass of an object, the perceptual system enters a fixation decision space which is schematically represented in Figure 1. As activation in the object lexicon starts to accumulate, a continuous evaluation is made of $\Delta a$, the amount of added activation per unit of time. As long as this activation rate stays above a threshold $A$, the ongoing fixation is maintained until eventually the identification threshold ID is reached and the fixation can be terminated. Any subsequent refixations are assumed to reflect post-identification processing. If, however, $\Delta a$ does not reach $A$ or drops below it before ID is reached, the system decides that the ongoing fixation is suboptimal and will refixate in order to achieve identification. Importantly, the decision to refixate is not governed exclusively by the low-level detection of a deviation from some optimal viewing position (as, for instance, stated by the strategy-tactics theory of eye-movement control in reading, O’Regan, 1992b). Instead, the probability of refixation is a function of activation rate which can be influenced by both high-level factors such as object familiarity or object predictability, and low-level factors such as stimulus contrast or perhaps deviation from the object’s center of gravity (Henderson, 1993; Melcher & Kowler, 1999).

Given this model, the most reliable index of ease of object identification would in principle be single-fixation duration because all measures involving refixations may include post-identification components, while first-of-multiple fixations may be terminated prior to identification. However, discarding all multiple-fixation cases would greatly reduce the power of eye-movement studies and eliminate observations that may be just as valid as the
single-fixation cases: In a direct comparison of word frequency effects on word-fixation parameters, Rayner et al. (1996) found the same effects on single-fixation durations, first-of-multiple fixations and first-gaze durations. I therefore opted to use first-fixation duration as a lower-limit estimate of ease of object identification, first-gaze duration as an upper-limit estimate and first-pass density as an index of duration-refixation tradeoffs. Shorter first-fixation durations only indicate easier object identification when they are not combined with higher first-pass densities. In the latter case, they in fact suggest a more difficult object identification.

One might argue that first-pass density is in fact a superfluous measure because refixations will always increase first-gaze duration. Thus, shorter first fixations combined with longer first gazes would indicate that the first fixations were prematurely interrupted because of a sub-criterion activation rate (i.e., an activation rate located in the grey area in Figure 1). However, Henderson (1993) pointed out that object refixations that follow short first fixations often are so short that they sum to gaze durations which are shorter than single-fixation durations. This suggests that there is a redundancy gain in refixating, possibly due to a higher processing rate of visual information during the initial stages of a fixation (Rayner, Inhoff, Morrison, Slowiaczek, & Bertera, 1981; van Diepen, De Graef, & d’Ydewalle, 1995). Therefore, a premature first-fixation cutoff does not always produce longer gaze durations but it does always lead to an increase in first-pass density.

Figure 1. Fixation decision space relating activation in the object lexicon to fixation duration and refixation probability. Activation curves are plotted for 4 objects ranked for identifiability from easy ($l_1$) to difficult ($l_4$). Fixations are maintained until identification threshold (ID) is reached unless activation rate $\Delta a$ drops below $A$ (shaded area). In that case a refixation is planned. Note that lower identifiability can lead to longer single fixations ($l_4 > l_1$), or to a long first-of-multiple fixations ($l_3 > l_1$), or to a short first-of-multiple fixations ($l_2 < l_1$).
4.2. Measuring transsaccadic preview benefits in the non-object search task

In the present study, participants were told that they would take part in one of a series of experiments on how good people are at detecting various kinds of information in images of varying complexity. In this particular experiment, they would see line drawings depicting real-world scenes containing both real objects and non-objects. Each scene would have to be explored, starting at a peripheral position marked by a red rectangle (average eccentricity was 5°). After each trial, participants would have to answer (a) whether the object at the marked location was a non-object, and (b) whether any additional non-objects were present in the scene. All participants were informed that their accuracy in detecting the non-objects would be evaluated in two ways. First, following each stimulus they would have to use response keys to answer the two specified questions (one about the marked object, the other about the rest of the scene). Second, their eye movements would be registered during the entire scene exposure to determine whether they had indeed localized the non-objects in the display or had just guessed. Eye movements were recorded with a Generation 5.5 dual-Purkinje-image eye-tracker with a 1000 Hz sampling frequency and a spatial accuracy of 1 min of arc (Crane & Steele, 1985).

The course of a trial in this study is illustrated in Figure 2. First, the probe question appeared in the center of the screen (i.e., “any non-objects?”). When the participant had read this question, a button press replaced the question with a central fixation cross. If this was properly fixated for at least 200 ms, one of the preview displays appeared. Participants were instructed to always make their first eye movement to the center of the red rectangle in the preview display. As soon as the first eye movement was detected a display change command was initiated (typically 4–5 ms after saccade onset) and the preview display was replaced by a target display. The display change was started in mid-scan and was completed in one refresh (20 ms). Thus, the time required to intrasaccadically switch from preview to target display was about 25 ms, well below the duration of the viewers’ saccades from the display center to the marked location (i.e., 45 ms). Following the display change, the eye-tracker monitored the position of the red rectangle for a period of 100 ms. If no fixation was detected inside the rectangle within that period, the trial was interrupted and participants received a text message instructing them to start scene exploration at the red rectangle. If a proper fixation was detected, then the target display remained on the screen until 4 s had elapsed since the onset of the preview display. Then, the target display was replaced by the first response screen with the question “In the red rectangle?” and the response alternatives “left = yes, right = no”. Thus, participants were reminded that they had to decide whether the object in the rectangle was a non-object. Following a response, the second response screen appeared with the follow-up question “In the rest of the scene?” and the response alternatives “left = yes, right = no”. Following the second response, a new trial was initiated.

To study transsaccadic preview benefits, the correspondence between pre- and post-saccadic contents of the location marked by the red rectangle was manipulated as shown in Figure 3. Presaccadically, the rectangle either contained no object (no preview), or it already contained the postsaccadic object (target preview), or it contained an object that
was semantically consistent with the scene and the postsaccadic target object (consistent companion preview), or it contained an object that was semantically inconsistent with the scene and the postsaccadic target object (inconsistent companion preview). Postsaccadically, the marked location always contained the target object, or a non-object while the rectangle had disappeared. Target-fixation parameters were recorded to infer ease of target object processing and to study whether that was affected by the type of preview. Specifically, if having a target preview helped postsaccadic target identification then target fixations should be shorter and fewer in the preview condition than in the no-preview condition. The consistent and inconsistent companion previews were used to examine whether transsaccadic object perception in scenes is (partly) mediated by episodic priming. Henderson and Anes (1994) and Gernseys et al. (2002) demonstrated that transsaccadic preview benefits for isolated objects are partly based on pre-saccadic access to the preview’s long-term representation in the object lexicon. Henderson (1992) argued for an automatic spreading of activation between episodically related individual object representations in the object lexicon, a claim recently supported by Bar (2004). The logical combination of these lines of research is that transsaccadic preview benefits in scene exploration are at least partly based on access to the preview’s episodic
Figure 3. Four types of pre-saccadic preview display are orthogonally combined with two types of post-saccadic target display. Red rectangles appeared in the actual stimulus and indicate the target location for the participant’s first saccade.

membership, that is, on knowledge about the preview’s plausible companion objects in real-world settings. By including consistent and inconsistent companion previews, it could be determined whether a consistent preview would facilitate subsequent target processing relative to an inconsistent preview.

As can be seen in Figure 4, clear differences between preview conditions emerged, with fastest processing for the target (identical) preview and slowest for the inconsistent companion preview. A reliable transsaccadic preview benefit was found in the comparison
Figure 4. Target object fixation parameters. Means and standard errors for first-fixation duration (top), gaze duration (middle) and first-pass density (bottom) as a function of preview type.
between target preview and no preview, with a longer first fixation \( (p < 0.0001) \), a shorter gaze duration \( (p < 0.02) \) and fewer first-pass refixations \( (p < 0.0001) \) when the target was preceded by an identical preview. The pattern of a sustained first fixation followed by shorter and fewer additional fixations is compatible with a higher rate of information extraction as argued above (see Figure 1).

Of course, one could argue that the pattern is perhaps an artifact produced by the unavailability of a proper saccade target in the no-preview condition. Specifically, in that condition participants are saccading toward an empty rectangle rather than toward an object and therefore the postsaccadic deviation between object center and landing position may be systematically higher in the no-preview condition, causing a short first fixation followed by a quick corrective saccade and a refixation. Because postsaccadic deviation from the target’s center did indeed prove to be higher in the no-preview than in the identical preview condition \( (0.93^\circ) \) than in the other conditions \( (0.68^\circ) \), this alternative explanation deserves further consideration. A first argument against it is that the same pattern of a longer first fixation, shorter gaze and fewer refixations is evident in the comparison between target preview and inconsistent companion preview. In this comparison, both preview objects were presaccadically presented at exactly the same location and no differences in postsaccadic deviation resulted, yet the same difference in fixation parameters was present. A second argument follows from an additional analysis: To correct for possible effects of postsaccadic deviation form the object's center, this variable was included as a covariate. Although greater postsaccadic deviation reliably decreased first-fixation duration \( (p < 0.0001) \), decreased gaze duration \( (p < 0.009) \), and increased first-pass density \( (p < 0.04) \), the pattern and reliability of preview effects were not changed.

In addition to a transsaccadic preview benefit, the data for consistent targets also showed transsaccadic priming: Targets preceded by an episodically related consistent companion preview were fixated shorter \( (p < 0.0001 \) for first-fixation duration and \( p < 0.02 \) for gaze) than targets preceded by an unrelated inconsistent companion preview. First-pass density showed the same pattern but non-reliable \( (p = 0.14) \).

4.3. Transsaccadic object perception in scenes: Conclusions

The test for transsaccadic preview benefits in non-object search yielded two main results. First, object perception in scenes is inherently transsaccadic. When a viewer is given an extrafoveal preview of an object, subsequent foveal processing of that object is faster than when no preview was available. This finding is new because it shows that even when peripheral figure-ground segregation of objects is complicated by the presence of a realistic scene background, preview information is sufficient to mandatorily influence subsequent foveal processing. In other words, the findings in full scenes have now been shown to be in line with earlier work demonstrating transsaccadic preview benefits for both words (e.g., Rayner & Morris, 1992) and isolated objects (e.g., Pollatsek et al., 1984) and provides yet another argument against the view that perception starts anew on every new fixation (O’Regan, 1992).
Second, transsaccadic facilitation of postsaccadic, foveal object processing by a presaccadic, extrafoveal object preview is partly mediated by inter-object connections in long-term memory that code episodic relatedness. Specifically, having a presaccadic preview of an object that is likely to appear in the same real-world settings as the postsaccadic target object facilitates target recognition relative to a situation where the presaccadic object does not belong to the target’s episodic category. This transsaccadic priming effect is at odds with earlier failures to find transsaccadic object priming between semantically related, isolated objects (Henderson, 1992; Pollatsek et al., 1984). The present study suggests that for these episodic inter-object priming effects to operate the presence of a coherent scene setting is required. Bar (2003, 2004) has recently outlined a possible neural substrate for such a mechanism. Specifically, a coarse analysis of background and objects during the very first glance at a scene is assumed to project in parallel to the parahippocampal cortex (PHC) where it activates scene schemas, and to the prefrontal cortex where it activates a set of “initial guesses” about the possible identity of objects present in the scene. Both the activated schema and the set of possible objects project to inferior temporal (IT) cortex where they modulate the activation of stored object representations and thus facilitate the object recognition process. Because the episodic associations between objects are not stored in the object lexicon itself (IT) but in the contextual associations stored in PHC, activation of those associations by the global setting information contained in a full scene is required to observe episodic priming effects.

5. Conclusion: Time to put a transsaccadic theory of recognition on the agenda

Earlier in this chapter, I have attempted to clarify why, in my opinion, current prominent theories of object and scene recognition provide no account of how recognition is achieved on the basis of information gathered from multiple distinct samples of the viewed stimulus, which are collected on spatially and temporally disparate fixations. Two reasons were identified why transsaccadic information integration is neglected in models of visual recognition.

First, there is the widespread conviction that the human visual system is so powerful that one short glance, well below the modal fixation duration, is more than enough to recognize a scene or an object. In a brief review of the existing evidence for that claim, I have tried to show that from a strictly chronometric point of view, single-fixation perception is sufficient to make low-level discriminations between object and scene categories, but may quite often fall short of true recognition. This would necessitate a second fixation if recognition is what the visual system tries to achieve before it moves on the next stimulus or stimulus location. Naturally, identification may not always be the goal of scene exploration (e.g., in single feature search, only the feature value is important not the spatio-temporal entity it is bound to). However, when identification is task-relevant then taking a second shot is likely to yield more reliable results than single-shot perception.
Second, for many years the very notion of transsaccadic information integration has been severely challenged by studies demonstrating (transsaccadic) change blindness and research disproving transsaccadic spatiotopic fusion. This has culminated in the prominent theoretical position that transsaccadic integration is a non-functional concept, because a visual on-line representation of the outside world is superfluous, given that the world itself can continuously be sampled anew on every fixation. In reply to this line of reasoning, I have reviewed evidence that shows that transsaccadic change blindness can easily be undone to reveal the presence of a visual on-line representation. In addition, I have reported new evidence indicating that during scene exploration, viewers routinely use the object information provided in a presaccadic, extrafoveal preview to speed up subsequent processing of that object during its postsaccadic foveation.

In conclusion, I hope to have shown that there are both chronometric and functional arguments to defend the claim that theories of object and scene recognition should make it their business to find out how the visual system exploits the advantage of having multiple extrafoveal and foveal samples of a stimulus in order to speed up and enhance the reliability of stimulus recognition. How exactly presaccadic identity hypotheses may influence postsaccadic foveal processing is outside the present scope, but answers to this question are bound to emerge from a rapidly growing set of studies on reverse hierarchy theory (Ahissar & Hochstein, 1997; Hochstein & Ahissar, 2002), the distinction between feed-forward and recurrent processing modes in vision (Lamme & Roelfsema, 2000; Ullman, 1996), and the notion of re-entrant visual processes (DiLollo et al., 2000). Despite some variations in naming, all these theories are centered around the claim that bottom-up and top-down processing streams can be distinguished behaviorally and neurophysiologically but are not functionally segregated. Specifically, all these theories propose that visual perception involves two streams of processing: A rapid, automatic, and pre-attentive feed-forward sweep of activation through the informational and cortical hierarchy which activates high-level, categorical stimulus representations; and a slower, attention-modulated, re-entrant or recurrent stream of processing originating in the higher-level representations and with a modulating effect on ongoing bottom-up processing of visual input.

The strength of this framework is twofold. First, it is sufficiently precise to outline testable mechanisms of how high-level categorical representations are integrated with lower-level, precategorical representations of the visual input, and this both at the computational level (e.g., Di Lollo et al., 2000; Di Lollo, Enns, & Rensink, 2002) and at the level of specific cortical mechanisms (e.g., Bar, 2003). Second, it is based on both psychophysical effects such as object substitution masking (e.g., Enns & DiLollo, 1997, 2001; Jiang & Chun, 2001; Lleras & Moore, 2003) and on neurophysiological data such as the existence of massive backprojections in the cortical hierarchy (Salin & Bullier, 1995) or the finding that over the course of their response, V1 neurons change their tuning from simple to more complex stimuli (Lamme, Zipser, & Spekreijse, 2002). By applying this framework to our eye-movement paradigms which were developed to study perception within the spatial and temporal information-integration constraints imposed by the oculomotor system, we should be able to develop a detailed account of the mechanisms involved in transsaccadic recognition of objects and scenes.
Acknowledgement

The writing of this chapter was supported by research grant GOA 2005/03 of the Research Fund K. U. Leuven and conventions G.0583.05 and 7.005.05 of the Fund for Scientific Research-Flanders.

References


