Chapter 21
Neuroscience and Psychology

A wide assortment of eye tracking studies can be found in the increasingly related fields of neuroscience and psychology. Topics range from basic research in vision science to the investigation of visual exploration in aesthetics (e.g., perception of art). A useful approach to navigating through vast collections of early and contemporary literature is to (for the outset) dissociate high-level cognitive studies from those concerned with a low-level functional view of the brain. In this sense, to use a computational analogy, one can distinguish between the "hardware" (low-level brain circuitry) on which the "software" (high-level cognition) functions. In a complementary view of the apparently disparate disciplines, neuroscience identifies the physiological components that are ultimately responsible for perception. In the context of vision and eye movements, knowledge of the physiological organization of the optic tract as well as of cognitive and behavioral aspects of vision is indispensable in obtaining a complete understanding of human vision.

To illustrate the interdependence of neuroscience and psychology, consider again the scene integration problem (see Chap. 1). Neurophysiological studies clearly identify the visual components involved in dynamic (or active) vision. That is, due to the limited informational capacity of the fovea, the eyes shift from point to point while scanning the visual field. The neuronal organization of retinal cells, which in a sense is the reason for eye movements, is well known. Furthermore, the general organization of foveo–peripheral vision has also been mapped along the magno- and parvocellular pathways leading to deeper regions of the brain and farther still into regions implicated in higher cognitive functions. From psychological observations, we know that humans are aware of a large field of view, even though physiology does not permit a holistic cameralike capture of the entire scene in one exposure. This is the crux of the scene integration problem. Psychologists show us that we are quite adept at maintaining a fairly accurate mental image of the visual scene in front of us. Indeed, various illusory pictures such as the Kanizsa (1976) square show us that we "see" more than what is physically there. The main question of how the brain is able to "piece together" small high-resolution snapshots of the scene remains a mystery.

A.T. Duchowski, *Eye Tracking Methodology*, DOI 10.1007/978-3-319-57883-5_21

Although there are many eye-movement related topics examined in the fields of neuroscience and psychology, current neuroscientific trends related to the study of eye movements are briefly discussed touching on exemplary work investigating illusory contours, attention, and brain imaging. Focus is then shifted to psychological investigation of four applied perceptual examples: the study of eye movements in reading, during scene perception (including perception of art and film), visual search, and in natural tasks.

21.1 Neurophysiological Investigation of Illusory Contours

Neuroscientific investigation of vision and in the specific context of eye movements has been covered to a certain extent in previous chapters. Results of this research have led to the identification of various interconnected components of vision, starting from retinal photoreceptors, and (more or less) ending in the cortical regions implicated in low-level vision. Important concepts of retino-geniculate pathways of vision as well as the important lower- and higher-level visual brain regions have been well established. Augmenting traditional neuroscientific investigation with eye tracking devices can lead to further understanding of perceptually puzzling phenomena such as Kanizsa-type illusions.

Logothetis and Leopold (1995) conducted an experiment investigating the neural mechanisms underlying binocular rivalry, the alternating perceptions experienced when two dissimilar patterns are stereoscopically viewed. Single cells were recorded in visual areas V1, V2, and V4 and monkeys reported the perceived orientation of rivaling sinusoidal grating patterns. Monkeys were trained to perform a fixation and an orientation discrimination task. Both tasks required continuous fixation of a small central spot within a $0.8^\circ \times 0.8^\circ$ window. To confirm the location of fixations, eye movements were measured with a scleral coil technique. Logothetis and Leopold suggest that results of this study indicate that awareness of a visual pattern during binocular rivalry arises through interactions between neurons at different levels of visual pathways, and that the site of suppression is unlikely to correspond to a particular visual area, as often hypothesized on the basis of psychophysical observations. Together with earlier psychophysical evidence, the cell types of modulating neurons and their overwhelming preponderance in higher rather than in early visual areas also suggests the possibility of a common neural mechanism underlying binocular rivalry as well as other bistable percepts, such as those experienced with ambiguous figures (e.g., Kanizsa-type illusions and the Necker cube). That is, perception of illusory contours may be physiologically related to stereoscopic perception.

21.2 Attentional Neuroscience

An important problem peculiar to eye tracking studies is that of the dissociation of visual attention from the point of regard. That is, most of the time, we can assume that one's visual attention is associated with the point of fixation. This is usually

referred to as overt attention, because it is the component of visual attention (when associated with foveal vision) that can be measured overtly (i.e., by an eye tracker). However, it is certainly possible to fix one's gaze at a specific point, and yet move one's attention to a nearby region. Astronomers do this fairly regularly when looking for faint stars or star clusters with the naked eye. The Little Dipper is a good example of star clusters that is found more easily when one looks for it "off the fovea".

The dissociation of attention from ocular fixation poses a problem for eye tracking researchers. When examining a scanpath over a visual stimulus, we can often say that specific regions were "looked at", perhaps even fixated (following analysis of eye movements), however, we cannot be fully confident that these specific regions were fully perceived. There is (currently) simply no way of telling what the brain was doing during a particular visual scan of the scene. Ideally, we would have to not only record the point of one's gaze, but also of one's brain activity. Research that combines eye tracking with traditional neuroscientific paradigms offers the dual benefit of monitoring brain activity as well as oculomotor function.

Investigating neuronal activity related to fixational eye movements, Snodderly et al. (2001) used a double Purkinje image eye tracker (2–3 minarc resolution; 100 Hz sampling rate) to record the position of a macaque monkey's eye when fixating a light-emitting diode for 5 s. Action potentials were recorded from neurons in area V1 of three adult female macaque monkeys that were trained to hold visual fixation. Snodderly et al. show that responses of V1 neurons to fixational eye movements are specific and diverse. Some cells are activated only by saccades, others discharge during drift periods, and most show a mixture of these two influences. Three types of eye movement activation were found: (1) "saccade-activated cells" discharged when a fixational saccade moved the activating region onto the stimulus, off the stimulus, and across the stimulus; (2) "position/drift cells" discharged during the intersaccadic (drift) intervals and were not activated by saccades that swept the activating region across the stimulus without remaining on it; and (3) "mixed cells" fired bursts of activity immediately following saccades and continued to fire at a lower rate during intersaccadic intervals. The patterns of activity reflect the interactions among the stimulus, the receptive-field activating region, the temporal response characteristics of the neuron, and the retinal positions and image motions imparted by eye movements. The diversity of the activity patterns suggests that during natural viewing of a stationary scene some cortical neurons are carrying information about saccade occurrences and directions whereas other neurons are better suited to coding details of the retinal image.

Snodderly et al. report that the two components of fixational eye movements, saccades and drifts, activate different subpopulations of V1 neurons in distinctive ways. Saccade-activated and mixed cells fire bursts of spikes when a fixational saccade moves the activating region onto the stimulus, off the stimulus, or across the stimulus. The sign of contrast (light or dark) is unimportant. These characteristics imply that the burst responses are conveying rather crude information about the details of the stimulus. For example, if an appropriately oriented stimulus contour activated the neuron following a saccade, it would be difficult to determine whether the contour remained in the activating region (saccade moved region onto contour) or was outside

it (saccade moved region across contour). This ambiguity would have particular relevance when viewing complex natural images, because the neuronal discharge could be evoked either by a stimulus feature crossed by the activating region during the saccade, or by a completely different feature on which the region landed at the end of the saccade. Bursts of spikes fired by saccade and mixed cells following fixational saccades may suggest that the bursts are important sources of information about the visual scene. An alternative role for the saccade neurons might be to participate in the suppression of visual input associated with saccades. Snodderly et al. argue that, theoretically, saccade neurons could participate in saccadic suppression by inhibiting other neurons that carry stimulus information, or by adding noise to the signal, thereby raising thresholds and making stimulus events undetectable at the times of saccades. This role would make the saccade neurons the source of saccadic suppression.

According to Snodderly et al., the position/drift neurons must play a quite different set of roles that are complementary to those of the other eye movement classes. Because the position/drift neurons do not respond to saccades, they may be spared the potentially detrimental effects of saccade-related activity. They signal accurately the position of stimulus features on the retina, and in many cases the sign of contrast. Thus, their activity could in principle be the basis for a reconstruction of the image.

Involved in attentional shifting behavior, the prefrontal (PF) cortex has long been thought to be central to the ability of choosing actions appropriate not only to the sensory information at hand but also according to the situation in which it is encountered (Asaad et al. 2000). Recent studies, reviewed by Asaad et al., indicate that the specific sensory, motor, and cognitive demands of the task (the behavioral context) can be an important factor in determining PF neural responses. For example, neural activity to an identical visual stimulus can vary as a function of which portion of that stimulus must be attended or with the particular motor response associated with it. Damage to the PF cortex of humans and monkeys tends to produce impairments when available sensory information does not clearly dictate what response is required. For example, PF lesions impair spatial delayed response tasks in which a cue is briefly flashed at one of two or more possible locations and the subject must direct an eye movement to its remembered location.

Asaad et al. performed an eye tracked neurophysiological experiment to explore the role of the PF cortex. Subjects were two rhesus monkeys (with immobilized heads). One animal was implanted with an eye-coil to monitor eye movements, and an infra-red monitoring system from ISCAN was used for the second animal. Eye position was monitored at 100 Hz in both animals. Microsaccades and saccades were detected using a simple velocity threshold set at four times the standard deviation of the signal derived from the fixation period. Neural recording sites were localized using magnetic resonance imaging (MRI). Recording chambers were positioned stereotaxically over the left or right lateral prefrontal cortices of each animal, such that the principal sulcus and surrounding cortex, especially the ventrolateral PF cortex, was readily accessible. Recordings were made using arrays of eight durapuncturing, tungsten microelectrodes. Activity of up to 18 individual neurons could be recorded simultaneously in any given session. Signal from 210 neurons was recorded from the

left lateral PF cortex of one monkey and 95 neurons from the right lateral PF cortex of the other. Monkeys performed an object memory task (delayed match-to-sample), an associative task (conditional visuomotor task), and a spatial memory task (spatial delayed response). The first two tasks shared common cue stimuli but differed in how these cues were used to guide behavior, whereas the latter two used different cues to instruct the same behavior. All three required the same motor responses. The associative task required the animals to associate a foveally presented cue stimulus with a saccade either to the right or left. The object task used the same cue stimulus as the associative task; however, in this case, they needed only to remember the identity of the cue and then saccade to the test object that matched it. Conversely the spatial task used small spots of light to explicitly cue a saccade to the right or left and required the monkeys to remember simply the response direction.

Asaad et al. report that most of the 305 recorded neurons displayed a task-dependent change in overall activity, particularly in the fixation interval preceding cue presentation. Results show that for many PF neurons, activity was influenced by the task being performed. This influence included changes in their baseline firing rates, modulations of neuronal activity related to particular stimuli and behavioral responses, and difference in their firing rate profiles. Asaad et al. suggest that results indicate the formal demands of behavior are represented within PF activity and thus support the hypothesis that one PF function is the acquisition and implementation of task context and the "rules" used to guide behavior.

21.3 Eye Movements and Brain Imaging

Recent investigations of eye movements and functional brain imaging simultaneously examine readings from an eye tracker and from a device that images brain activity. For example, Gamlin and Twieg (1997) have embarked on designing a combined visual display and eye tracking system for high-field fMRI (functional Magnetic Resonance Imaging) studies. The proposed project is a collaboration between a neuroscientist and biomedical engineer. The technological goal is the design, development, and implementation of a combined high-resolution binocular display and eye tracking system for use in fMRI studies of the human brain. The scientific goal of the project is the study of neural control of saccadic eye movements, allowing investigation of stereopsis and depth perception, as well as permitting oculomotor studies of smooth pursuit, optokinetic nystagmus, vergence, and accommodative eye movements.

Systems that marry functional brain imaging with eye tracking can be used to at least corroborate a subject's fixation point while simultaneously recording cortical enhancement during attentional tasks. Presently, possibly due to prohibitive cost, combined eye tracking and brain imaging equipment is not widespread, although such devices are beginning to appear; e.g., see Fig. 21.1.

Employing a combined brain imaging and eye tracking device, possibly similar to the one shown in Fig. 21.1, Özyurt et al. (2001) used an fMRI device to compare the neural correlates of visually guided saccades in the step and gap paradigms. During

Fig. 21.1 Example of eye tracking fMRI scanner. Courtesy of SensoMotoric Instruments (SMI), Needham, MA http://www.smiusa.com. Reproduced with permission

task performance, saccadic eye movements were recorded with an MR-Eyetracker. Subjects viewed stimuli that were projected onto a screen attached to the front of the MR scanner. A block-design fMRI was used with alternating blocks of step, rest, and gap trials. Results from the study by Özyurt et al. indicate significant task-related activity in striate and extrastriate cortex, the frontal eye fields, the supplementary motor area, parietal cortex, and angular gyrus, the frontal operculum, and the right prefrontal area 10. This type of research helps identify functional brain structures implicated in attentional behavior.

21.4 Reading

Perhaps the first well-known applied uses of eye trackers in the study of human (overt) visual attention were those conducted during reading experiments. An excellent book on eye movements was collected by Rayner (1992), an influential researcher in eye movements and reading. Rayner's collection of articles contains exemplary research on reading and on scene perception. Rayner's (1998) article gives a rather

comprehensive survey of eye tracking applications, reviewing studies of eye movements in reading and other information-processing tasks such as music reading, typing, visual search, and scene perception. The major emphasis of the review is on reading as a specific example of cognitive processing.

Rayner (1998) reviews a good deal of previous work on eye movements synthesizing a large amount of information gleaned from over 100 years of research. Although the reader is referred to Rayner's article for the complete review, three interesting examples of eye movement characteristics during reading are given here. First, eye movements differ somewhat when reading silently from reading aloud: mean fixation durations are longer when reading aloud or while listening to a voice reading the same text than in silent reading. Second, when reading English, eye fixations last about 200–250 ms and the mean saccade size is seven to nine letter spaces (see below). Third, eye movements are influenced by textual and typographical variables; e.g., as text becomes conceptually more difficult, fixation duration increases and saccade length decreases. Factors such as the quality of print, line length, and letter spacing influence eye movements. There is of course a good deal more that has been learned, however, here the methodology behind such discoveries is what is of primary interest. Below, three main experimental paradigms used in eye tracking experiments are discussed.

Three experimental paradigms, the *moving window*, *boundary*, and *foveal mask*, have been developed to explore eye movements. Although first developed for reading studies, these paradigms have since been adapted to other contexts such as scene perception (see below). In the moving window paradigm, or *gaze-contingent display change* paradigm, developed by McConkie and Rayner (1975), a window is sized to include a number of characters (e.g., 14) to the left and right of a fixated word. For example, the sentence

```
the quick brown fox jumped over the lazy dog
```

is presented as follows in four subsequent temporal instances (note the change of word `fox` to `cat`; the asterisks indicate fixation locations).

```
xhe quick brown xxxxxxxxxxxxxxxxxxxxxxxxxxxx
       *
xxxxxxxxk brown fox jumxxxxxxxxxxxxxxxxxxxxx
              *
xxxxxxxxxxxxxxxxcat jumped ovexxxxxxxxxxxxxx
                        *
xxxxxxxxxxxxxxxxxxxxxxxxd over the laxxxxxx
                                  *
```

The assumption with this technique is that when the window is as large as the region from which the reader can obtain information, there is no difference between reading in that situation and when there is no window. A related but reverse method, developed

by Rayner and Bertera (1979) (see also Bertera and Rayner 2000), places a foveal mask over a number of fixated characters (e.g., seven):

```
the qxxxxxxxown fox jumped over the lazy dog
        *
the quick brown cat jumpedxxxxxxxhe lazy dog
                                *
```

This situation creates an artificial foveal scotoma and eye movement behavior if the situation is quite similar to the eye movement behavior of patients with real scotomas (Rayner 1998).

In the boundary technique, developed by Rayner (1975), the stimulus changes as fixation crosses a predefined boundary. Rayner used eye movements to investigate reading because he found tachistoscopic methods inadequate: tachistoscopic (strobelike) presentation of letters and words relies on the presentation of material for very brief exposures to exclude the possibility of an eye movement during the presentation. Prior to eye tracked reading studies, this method was often thought of as being analogous to a single fixation during reading. Based on his and others' research, Rayner argued that what subjects report from a tachistoscopic presentation cannot be taken as a complete specification of what they saw. The argument for eye movement recording over tachistoscopic displays carries over to scene perception and is discussed further in the next section.

In an example of the boundary technique below, a single critical target word is initially replaced by another word or by a nonword. The boundary paradigm allows the experimenter to be more diagnostic about what kind of information is acquired at different distances from fixation. In this technique, a word or nonword letter string is initially presented in a target location. However, when the reader's eye movement crosses an invisible boundary location, the initially presented stimulus is replaced by the target word. The amount of time that the subject looks at the target word is influenced by the relationship between the initially presented stimulus and the target word and the distance from the launch site to the target location. The fixation time on the target word thus allows the experimenter to make inferences about the type of information acquired from the target location when it was in parafoveal vision. For example, the word `fox` is changed to `cat` when the eyes cross the boundary.

```
the quick brown fox jumped over the lazy dog
          *     |
the quick brown cat jumped over the lazy dog
                |             *
```

The assumption is that if a reader obtains information from the initially presented stimulus, any inconsistency between what is available on the fixation after crossing the boundary and with what was processed on the prior fixation is registered in the fixation time on the target word.

Table 21.1 Reading strategies/tactics

Tactic	"Careful" strategy	"Risky" strategy
Single fixation	One fixation in the critical region	One fixation in the critical region
Double fixation	One fixation before critical region, one after (or the reverse—one fixation after and one backwards fixation before critical region)	Two fixations in the critical region

Experiments using the above gaze-contingent techniques have shown, generally, (1) that readers typically acquire the visual information necessary for reading during the first 50–70 ms of a fixation, and (2) that a serial scan of letters in foveal vision does not occur. Furthermore, studies using these techniques consistently indicate that the size of the perceptual span is relatively small (e.g., for readers of alphabetical orthographies, such as English, Dutch, or French, the span extends from 3 to 4 letters to the left of fixation to about 14–15 letter spaces to the right of fixation), with a still smaller word identification span (Rayner 1998).

Experiments in reading can lead to descriptions of individuals' reading strategies, or perhaps even to suggestions for improvement of one's reading strategy. Previous findings have shown that when an orienting visual response (eye movement) is made to a target pair consisting of two neighboring but separated elements, the first saccade lands in between the two elements. This was called the *global effect* by Findlay (1992).[1] In examining the global effect in reading, O'Regan (1992) suggests that an "optimal" viewing position may exist, where the time it takes to recognize the word is minimal. Due to oculomotor constraints or scanning strategies, the eye does not always land at the optimal spot. O'Regan discusses a combination of two observed strategies and tactics for reading, as shown in Table 21.1. A double fixation involves a second fixation that is inversely proportional (in duration) to the first one. If far from optimal position, the fixation is very short (100–150 ms). Single fixations may be long (up to 300 ms). The optimal viewing position phenomenon is weak in reading, but strong in single word experiments. That is, it is not known what the "special phenomenon" is in reading that accounts for the weakening of the optimal viewing position phenomenon. It may depend on the reader's strategy: risky or careful. It is not known whether readers can choose to read faster or slower.

The various types of gaze-contingent word display change techniques have had a major impact on reading research and scene perception because they allow the experimenter to control the information available to the subject at any point in time. Such experimental control over the nature of timing of the available information has enabled researchers to draw interesting conclusions about on-line aspects of reading and scene perception.

[1] The effect may also be referred to as the *center of gravity* effect.

21.5 Scene Perception

Although certain reading patterns are easily recognized (e.g., left-to-right, top-to-bottom for English readers, or right-to-left for Hebrew), no apparent strategies for scene viewing have been easily discerned. Contrary to reading, there appears to be no canonical scanpath for particular objects (i.e., there is no particular "right way" to look at objects; Kennedy 1992). Kennedy suggests that the reading task is composed almost exclusively of saccades, whereas picture viewing is composed of shifts, pursuits, and drifts. There may be context differences at play. Continuing the debate about context effects for scenes and sentences, Kroll (1992) states that although there may be similarities between the two tasks, the tasks are very different. Eye movements in reading are to a large extent driven by the well-known, practiced task, however, we don't know what the "glue" is that holds the scene together. Kroll states,

> One of the common problems in this research is to develop a set of tasks that will allow us to uniquely locate the interaction of context with object recognition over time.

Rayner (1998) recounts the traditionally held belief that examining the fine details of eye movements during scene perception is a high-cost, low-yield endeavor. Experiments using tachistoscopic presentations and eye movement recordings have led to the conclusion that participants get the gist of a scene very early in the process of looking, sometimes even from a single brief exposure. Thus it has been advocated that the gist of the scene is abstracted on the first few fixations, and the remainder of the fixations on the scene are used to fill in details. Such findings give rise to the question of the value gained from information obtained from detailed eye movement analyses as people look at scenes. Rayner reviews several findings that support the contention that important conclusions about temporal aspects of scene perception can be obtained from eye movement analysis. He recounts the argument put forth by Loftus (1981) and Rayner and Pollatsek (1992) stating that tachistoscopic studies have not shown conclusively that they reveal a perceptual effect rather than the outcome of memory processes or guessing strategies.

Loftus (1981) presented results of a masked tachistoscopic study which suggest a model of picture encoding that incorporates the following propositions: (a) a normal fixation on a picture is designed to encode some feature of the picture; (b) the duration of a fixation is determined by the amount of time required to carry out the intended feature encoding; and (c) the more features are encoded from a picture, the better the recognition memory will be from the picture. A major finding of Loftus' experiments was that with exposure time held constant, recognition performance increased with increasing numbers of fixations. When eye fixations are simulated tachistoscopically and their durations experimentally controlled, all traces of this phenomenon disappear. Moreover, Loftus' experiments suggest that within a fixation, visual information processing ceases fairly early; that is, acquired information reaches asymptote soon after the start of a fixation.

An eye fixation has the very salient property that it shifts the gaze to a new place in the picture. The problem identified with tachistoscopic exposures, which were meant to simulate fixations to new places, is that there is no guarantee that this

occurred; it is entirely possible that subjects were simply holding their eyes steady throughout all tachistoscopic flashes. From an eye tracking experiment, Loftus draws the argument that given more places to look at in the picture, more information can be acquired from the picture. Additional (tachistoscopic) flashes are only useful insofar as they permit acquisition of information from additional portions of the picture. Information pertinent to subsequent recognition memory seems to be acquired only from the small $2^\circ \times 3^\circ$ foveal region during a given fixation, and a fixation is useful only to the degree that it falls on a novel place in the picture.

The fixational perceptual span in scene perception mirrors that for reading with one important difference: meaningful information can be extracted much farther from fixation in scenes than in text (Rayner 1998). Objects located within about 2.6° from fixation are generally recognized, but recognition depends to some extent on the characteristics of the object. Qualitatively different information is acquired from the region 1.5° around fixation than from any region farther from fixation. At high eccentricities, severely degraded information yields normal performance. This suggests that low-resolution information is processed in the more peripheral parts of the visual field, whereas high-resolution information is processed in foveal vision. High spatial frequency information is more useful in parafoveal and peripheral vision than low spatial frequency information.

Rayner and Pollatsek (1992) concede that much of the global information about the scene background or setting is extracted on the initial fixation. Some information about objects or details throughout the scene can be extracted far from fixation. However, if identification of an object is important, it is usually fixated. The work discussed in Rayner and Pollatsek's paper indicates that this foveal identification is aided significantly by the information extracted extrafoveally. Rayner and Pollatsek conclude that it is necessary to study eye movements to achieve a full understanding of scene perception. They argue that if the question of interest is how people process scenes in the real world, understanding the pattern of eye movements will be an important part of the answer.

Rayner (1998) summarizes a number of other findings and claims, some controversial, eventually asserting that given the existing data, there is fairly good evidence that information is abstracted throughout the time course of viewing a scene. Rayner concludes that whereas the gist of the scene is obtained early in viewing, useful information from the scene is obtained after the initial fixations.

According to Henderson and Hollingworth (1998), there are at least three important reasons to understand eye movements in scene viewing. First, eye movements are critical for the efficient and timely acquisition of visual information during complex visual-cognitive tasks, and the manner in which eye movements are controlled to service information acquisition is a critical question. Second, how we acquire, represent, and store information about the visual environment is a critical question in the study of perception and cognition. The study of eye movement patterns during scene viewing contributes to an understanding of how information in the visual environment is dynamically acquired and represented. Third, eye movement data provide an unobtrusive, on-line measure of visual and cognitive information processing. Henderson and Hollingworth list two important issues for understanding eye movement control

during scene viewing: *where* the fixation position tends to be centered during scene viewing, and *how long* the fixation position tends to remain centered at a particular location in a scene.

Henderson and Hollingworth (1998) review past results indicating that the positions of fixations within a scene are nonrandom, with fixations clustering on informative scene regions. However, the specific effect of semantic informativeness beyond that of visual informativeness on fixation position is less clear. Several metrics can be used to evaluate the relative informativeness of scene regions: at a macro-level analysis, *total time* that a region is fixated in the course of scene viewing (the sum of the durations of all fixations in that region); this measure is correlated with the number of fixations in that region. At a micro-level analysis, several commonly used measures include *first fixation duration* (the duration of the initial fixation in a region), *first pass gaze duration* (the sum of all fixations from first entry to first exit in a region), and *second pass gaze duration* (the sum of all fixations from second entry to second exit in a region). Generally, first pass gaze durations are longer for semantically informative (i.e., inconsistent) objects. Semantically informative objects also tend to draw longer second pass and total fixation durations. The influence of semantic informativeness on the duration of the very first fixation on an object is less clear. That is, scene context has an effect on eye movements: fixation time on an object that belongs in a scene is less than fixation time on an object that does not belong (Rayner 1998). However, it is not clear whether the longer fixations on objects in violation of the scene reflect longer times to identify those objects or longer times to integrate them into a global representation of the scene (it could also reflect amusement of the absurdity of the violating object in the given context).

Henderson (1992) asks what the influence is of the contextual constraint provided by a predictive scene on the identification of its constituent objects. For example, does the context serve to facilitate identification procedures for objects consistent with that scene? Concretely, can a cow be identified more accurately and/or more quickly if it is viewed in a farm scene rather than in a kitchen scene? Henderson summarizes the predominant view of the relation between object and scene identification as the *schema hypothesis*. Although there are several variations on the schema theme, several commonalities define the hypothesis. According to the schema hypothesis, a memory representation of a prototypical scene is quickly activated during scene viewing, and is used to develop expectations about likely objects. These expectations then influence the object identification processes. For example, under the schema hypothesis, the identification of a cow in a farm scene involves (1) quickly recognizing that the scene is an exemplar of the category "farm scene", (2) accessing from memory the schema for a farm scene, (3) using the information stored with the schema to generate "cow" and other object candidates likely to be found in a farm scene (and possibly their canonical spatial relationships), and (4) using the knowledge that a cow is likely in such a scene to aid object identification processes when the cow is encountered. The above description suggests a serial model, however, this need not be a central assumption of the schema hypothesis. Henderson cautions against the schema hypothesis and suggests the use of eye tracking to search for a better model of scene perception.

Henderson (1992) offers two criticisms of the eye movement paradigm. First, it is likely that global measures of fixation time, such as the total time spent on an object during the course of scene viewing, and the gaze duration on an object (the time of all initial fixations on an object prior to leaving that object for the first time) reflect postidentification processes. Thus, it is likely that gaze duration in scene processing reflects other processes beyond object identification. Henderson suggests that the preferred fixation measure is the true first fixation duration, or the duration of time from the initial landing of the eyes on an object until the eyes move to any other location, including another location on the object. Second, the basic premise of the eye movement paradigm is that the results will reflect normally occurring visual-cognitive processes because subjects can view scenes in a natural manner. However, unlike reading, where the overall task is arguably transparent, subjects must be given an orienting task when they view a scene. Unfortunately, viewing behavior and eye movement patterns change as a function of the viewing task given to the subject (Yarbus 1967). One way to address the orienting task issue would be to give subjects a task that did not force the creating of a coherent memory representation for the scene, and look for similar scene context effects on fixation time across tasks.

21.5.1 Perception of Art

A particularly interesting subset of scene perception studies is the examination of gaze over a specific set of contextual images, namely art. As observed by Yarbus (1967), a viewer's intent influences eye movements and fixations over a scene. Eye movements over art have further refined the scope of these studies, examining differences in how trained viewers search for meaning and aesthetic qualities in fine art pieces.

The first systematic exploration of fixation positions in scenes was reported by Buswell (1935) (as cited by Henderson and Hollingworth 1998), who asked 200 participants to look at 55 pictures of different types of artwork under a variety of viewing instructions. An important result was that fixation positions were found to be highly regular and related to the information in the pictures (Henderson and Hollingworth 1998). For example, viewers tended to concentrate their fixations on people rather than on background regions when examining *Sunday on the Island of La Grande-Jatte* by Georges Seurat. These data thus provided some of the earliest evidence that eye movement patterns during complex scene perception are related to the information in the scene, and by extension, to perceptual and cognitive processing of the scene. Buswell concluded that:

> Eye movements are unconscious adjustments to the demands of attention during a visual experience. The underlying assumption in this study is that in a visual experience the center of fixation of the eyes is the center of attention at a given time.

Another example of eye movement over art studies, by Molnar (1981) (as reported by Solso 1999), shows small differences in scanpaths between groups of subjects viewing artwork for its semantic meaning or for its aesthetic appeal. Remarkably,

however, both sets of scanpaths are very similar in terms of fixated image features. In Rembrandt's *Anatomy Lesson*, fixations made by two groups of fine art students, given different sets of questions pertaining to the painting, appear to coincide on important elements even though the order of fixations may differ. In this case, the important elements are those of the faces of the anatomy professor and his students in the painting.

Further analyses of eye movements collected over art pieces reported by Molnar (as cited in Solso 1999) seem to suggest a general rule: more complex pictures produce shorter eye fixations than less complex forms. For example, in comparison of classical works of the high Renaissance to those of the mannerist and baroque periods, classical art produces eye movements that are large and slow, reflecting the expansive nature of that style, whereas baroque paintings involve small and quick eye movements, reflecting the dense animated character of that form. Baroque paintings were judged by art experts as more complex than classical paintings. It may be that complex art, which is densely packed with details, demands that attention be given to a large number of visual elements. This demand can be satisfied by allocating shorter fixation times to each feature. Simpler works may contain far fewer features vying for attention, and so may allow more time allocation per feature, resulting in longer fixations per feature.

A recent large-scale eye tracking study of art was conducted by Wooding (2002). An automated eye tracker was left running in a room of the National Gallery, London, as part of the millennium exhibit: "Telling Time." In three months, eye movements were successfully collected from 5638 subjects while they viewed digitized images of paintings from the National Gallery collection. Because a composite representation of eye movements from so many subjects posed a problem, Wooding devised a *fixation map* method of analysis, which might be descriptively termed a landscape or terrain map of fixations, and is in fact similar to the *landscape map* developed independently

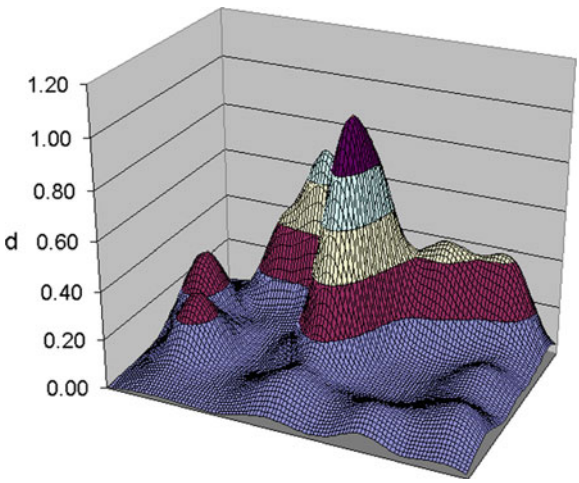

Fig. 21.2 Fixation map from 131 subjects viewing Paolo Veronese painting *Christ Addressing a Kneeling Woman*. From Wooding (2002) © 2002 ACM, Inc. Reprinted by permission

(a) Original image.

(b) Visualization of important regions.

Fig. 21.3 Sample fixations from 131 subjects viewing Paolo Veronese *Christ Addressing a Kneeling Woman* © National Gallery, London, with annotations © IBS, University of Derby, UK. From Wooding (2002) © 2002 ACM, Inc. Reprinted by permission

by Velichkovsky et al. (1996). The value at any point on the map indicates the height or amount of a particular property at that point (e.g., the number of fixations). An example of a fixation map and subsequent visualization of composite fixated regions are shown in Figs. 21.2 and 21.3.

Future eye tracking studies over art, be they comparative or of a mass scale, will undoubtedly lead to further insights of human perception of this particularly pleasing set of images. In a similar goal, but being approached from a different starting point, eye movement and general perceptual principles are currently being applied to the generation or creation of art by computers. A recent collaborative gathering of computer graphics and other scientists, such as representatives from perceptual and cognitive fields, met in Utah (McNamara and O'Sullivan 2001). There, perceptual principles and eye tracking methodologies were discussed as a possible aid to the creation of more aesthetically pleasing or more interactive works of computer graphics and art.

21.5.2 Perception of Film

Another interesting form of artistic media is film. In a sense, film is a dynamic form of art. An interesting example of an eye tracking film study is given by d'Ydewalle et al. (1998). d'Ydewalle et al. distinguish three levels of film editing errors in sequencing successive shots. First-order editing errors refer either to small displacements of the camera position or to small changes of the image size, disturbing the perception of apparent movement and leading to the impression of jumping. Second-order editing errors follow from a reversal of the camera position, leading to a change of the left–right position of the main actors (or objects) and a complete change of the background. With third-order editing errors, the linear sequence of actions in the narrative story is not obeyed. d'Ydewalle et al.'s experiment shows that there is an increase of eye movements from 200 to 400 ms following both second- and third-order editing errors. Such an increase is not obtained after a first-order editing error, suggesting that the increase of eye movements after second- and third-order editing errors is due to postperceptual cognitive effects.

21.6 Visual Search

The question of how humans perceive the visual scene through the movements of the eyes, in the context of more natural or free tasks such as picture viewing (which is significantly different from the task of reading), can generally be modeled by the process known as visual search. Visual search, in general, refers to the process of visually scanning a scene and forming a conceptual "image" or notion of the scene as assembled by the brain. In comparison to reading, there have not been nearly as many studies dealing with visual search (Rayner 1998).

When eye movements are recorded during extended search, fixations tend to be longer than in reading. However, there is considerable variability in fixation time and saccade length as a function of the particular search task (Rayner 1998). Specifically, visual search tasks vary widely, and tasks in which eye movements have been monitored consist of at least the following: search (a) through text or textlike material, (b) with pictorial stimuli, (c) with complex arrays such as X-rays, and (d) with randomly arranged arrays of alphanumeric characters or objects. Because the nature of the search task influences eye movement behavior, any statement about visual search and eye movements needs to be qualified by the characteristics of the search task.

Henderson and Hollingworth (1998) list factors that may vary from study to study: image size (usually measured in visual angle), viewing task (e.g., later recognition memory task, image preference, counting nonobjects, visual search, "free viewing" (it is well known that viewers place their fixations in a scene differently depending on viewing task), viewing time per scene (ranging from very short, tachistoscopic to longer durations, e.g., on the order of 50 ms–10 s or even 30 min (Yarbus 1967)), image content (e.g., artwork, "natural scenes", human faces), and image type (e.g., highly artificial and regular such as sine wave gratings, color, monochrome, grayscale, or full-color computer graphics imagery). These factors could each produce main effects and could also interact with each other in complex ways to influence dependent measures of eye movement behavior such as saccadic amplitudes, fixation positions, and fixation durations.

As demonstrated by Yarbus (1967) and then by Noton and Stark (1971a, b), eye tracked scanpaths strongly suggest the existence of a serial component to the picture viewing process. However, serial scanpaths do not adequately explain the brain's uncanny ability to integrate holistic representations of the visual scene from piecemeal (foveal) observations. That is, certain perceptual phenomena are left unaccounted for by scanpaths, including perception of illusory images such as the Kanizsa (1976) figures or the Necker (1832) cube. Although scanpaths cast doubt on a purely Gestalt view of visual perception, it would seem that some sort of holistic mechanism is at work which is not revealed by eye movements alone. Models of visual search attempt to answer this dilemma by proposing a parallel component, which works in concert with the serial counterpart exhibited by eye movements.

In visual search work, the consensus view is that a parallel preattentive stage acknowledges the presence of four basic features: color, size, orientation, and presence and/or direction of motion (Doll 1993; Wolfe 1993). Todd and Kramer (1993) suggest that attention (presumably in the periphery) is captured by sudden onset stimuli, uniquely colored stimuli (to a lesser degree than sudden onset), and bright and unique stimuli. There is doubt in the literature whether human visual search can be described as an integration of independently processed features (Van Orden and DiVita 1993).

Visual search, even over natural stimuli, is at least partially deterministic, rather than completely random (Doll et al. 1993). Determinism stems from either or both of two kinds: the observer's strategy determines the search pattern (as in reading), and/or

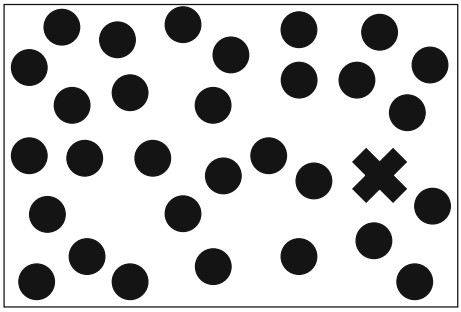

Fig. 21.4 Example of "pop-out" effect

the direction of the next saccade is based on information gained through peripheral vision about surrounding stimuli. Features that are likely to be fixated include edges, corners, spatially high-frequency components, but not plain surfaces.

A theory that consolidated the serial and parallel counterparts of visual attention was put forth by Treisman (1986) (see also Treisman and Gelade 1980). Treisman and her colleagues reported on the "pop-out" effect; that is, a visual target pops out from a field of similar but distinct distractors, seemingly drawing attention to itself (see Fig. 21.4). This observation led to the proposition that the visual system can search for certain stimulus properties in parallel over the whole visual field, a preattentive process. Other target items that do not pop out must be searched for by the observer in some (usually serial) order-focused attentional processing. The difference between preattentive and postattentive modes of processing can be demonstrated experimentally via paradigms designed to elicit visual search. In a typical visual search task, the observer is shown an array of items and asked to search for a specific target among the field of distractors. Reaction time is usually measured when reporting the presence or absence of a target. If reaction time does not vary with the number of items, the whole array is thought to be searched in parallel, or preattentively. If reaction time increases with the number of items, the array is most likely being scanned item by item in a serial process that is depicted by scanpaths. Treisman proposed the feature integration theory (FIT) unifying these processes.

The main tenet of FIT is that certain features pop out because they are explicitly represented by low-level vision. This is consistent with known retinotopic maps of cortical cells tuned to basic stimulus properties such as color, orientation, and size. Conjunctions of features, however, require serial search because they do not form part of this initial representation. Thus, according to this model, visual processing can be described by three broad stages: (1) feature extraction, (2) feature binding, and (3) object representation.

Feature integration theory is a theory of how elementary visual features are attentionally bound together to construct unitary perceptual objects. The first (parallel) stage of processing involves multiple, retinotopic master maps of locations of elementary features, showing where all feature boundaries are located, but not what those features are. Maps correspond to feature attributes such as color, orientation,

etc. According to FIT, perceptual objects are constructed by attention to a specific location, binding together the simultaneous activity at the location in all feature maps. Thus, attention can be thought of as a "glue" that integrates separated feature attributes at a particular location so that the conjunction is perceived as a unified whole.

FIT is a useful theory because it adequately explains both serial and parallel components of visual search. Parallel search is supported by FIT because targets defined by elementary properties are available in feature maps. Serial search is supported ince search for conjunctions requires focused attention to bind together features in separate maps. However, although FIT can explain search performance fairly well, particularly over simple stimuli (e.g., search for Q in a field of Os), it is not clear whether FIT generalizes to more complex stimuli such as natural imagery (e.g., aerial photographs).

Feature integration theory is currently held to be a useful heuristic starting point for theoretical treatments of visual search although it is widely recognized to be oversimplified (Findlay and Gilchrist 1998). Searches for more sophisticated accounts have taken various forms. Wolfe (1994) has pointed out the weakness of the assumption that search must be either serial or parallel and developed a model of the way in which interactions between serial and parallel processes could occur. A particular problem of FIT is that it suggests that although preattentive feature processes can perform feature searches in parallel, attention from item to item is required for all other (serial) searches. Conjunction searches, however, appear to be too efficient to be explained as purely serial attentive searches. Wolfe's Guided Search (GS) model accounts for this efficiency by proposing that preattentive feature processes could guide the deployment of attention in conjunction searches (Wolfe 1993, 1994). No preattentive process can identify a particular conjunction, however, two different preattentive processes (e.g., a color and an orientation process) can cooperate to mark conjunctive targets. If the output of these two preattentive processes is combined into an attention-guiding activation map, attention will be guided to conjunction items (e.g., black vertical bars in a field of white or black horizontal or vertical bars). Wolfe's model is discussed further in Sect. 21.6.1.

Relatively few studies have addressed the relationship between eye movements and the search process (Findlay and Gilchrist 1998). Findlay and Gilchrist argue that the tradition in search research to pay little attention to eye movements and instead to use the concept of covert visual attention movements (redirecting attention without moving the eyes; an important component of FIT) is misguided. The authors demonstrate that when viewers are free to move their eyes, no additional covert attentional scanning occurs. They show that unless instructions explicitly prevent eye movements, subjects in a search task show a natural propensity to move their eyes, even in situations where it would be more efficient not to do so. The authors suggest that the reason for this preference is that in naturally occurring search situations, eye movements form the most effective way of sampling the visual field.

Findlay (1997) recorded eye movements during tasks of a simple feature search and a color shape feature conjunction search. Eye movements were recorded by having the subject wear a contact lens-type search coil positioned at the center of

two large Helmholtz field coils. The induced currents in the eye coils measured eye position in space in a way that minimized head movement artifact, measuring eye position with an accuracy of 10 min arc or better following calibration. Using a single feature (color) search task, in the homogeneous distractor condition, only 0.5% of first saccades were directed at a nontarget and in the heterogeneous distractor condition the percentage of misdirected saccades was under 2%. This accuracy was achieved with no cost in the time needed to program the saccade. This provided an impressive confirmation that search for a prespecified color target can be carried out in parallel. When two targets are presented simultaneously in neighboring positions, the first saccade is directed toward some "center of gravity" position. Results suggest that the control of the initial eye movement during both simple and conjunction searches is through a spatially parallel process.

Results from a conjunction search experiment (color and shape) are particularly relevant because this is a situation where serial scanning would be expected according to the classical search theory of Treisman and Gelade (1980). If a rapid serial scanning with covert attention could occur before the saccade is initiated, it is not clear why incorrect saccades would occur as frequently as observed in the experiment (three subjects were able to locate targets with a single saccade on 60–70% of occasions in the inner ring of a two-ring concentric display, and 16–40% in the outer ring). Moreover, Findlay argues that the data place constraints on the speed of any hypothetical serial scanning process because it would be necessary for a number of locations to be scanned before the target is located, given the accuracy obtained. Alternative accounts of the visual search process have appeared that assign much more weight to the parallel processes and avoid the postulation of rapid serial scanning. The results of the conjunction search are consistent with a search model that limits parallel scanning to about eight items but requires serial search for displays of larger number.

Bertera and Rayner (2000) had viewers search through a randomly arranged array of alphanumeric characters (i.e., letters and digits) for the presence of a target letter. They used both the moving window technique and the moving mask technique. The number of items in the array was held constant, but the size of the display varied ($13° \times 10°$, large, $6° \times 6°$, medium, and $5° \times 3.5°$, small).

In the moving window study, search performance reached asymptote when the window was 5°. The moving mask had a deleterious effect on search time and accuracy, indicating the importance of foveal vision during search; the larger the mask, the longer the search time. For the window conditions, six different window sizes were used. The window was 1.0°, 2.3°, 3.7°, 5.0°, or 5.7°; in addition, a control condition was run in which the entire array was presented (i.e., there was no window). Six different mask sizes were also used: 0.3°, 1.0°, 1.7°, 2.3°, 3.0°, or no mask present. A Stanford Research Institute Dual Purkinje Eye Tracker was used to record eye movements, with five subjects participating in the study.

Under the window condition, search performance improved (i.e., search time decreased) as the window size increased. In general, search performance reached asymptote when the window was 5.0°. There was no effect of either window size or array size on accuracy; the subjects successfully located the target letter 99% of

the time across the different conditions. Under the mask condition, as mask size increased, search time increased. A large mask was more detrimental to search than a small window. Unlike window size, the size of the mask had a significant effect on accuracy. The accuracy values were 100, 99, 94, 73, 58, and 39% for mask sizes 0 (no mask), 0.3°, 1.0°, 1.7°, 2.3°, and 3.0°, respectively.

Bertera and Rayner's study shows that a useful perceptual span in visual search extends to about 5°. The moving window paradigm for scene perception and visual search is discussed further as an instance of gaze-contingent display technology in Sect. 24.2.

Recently, Greene and Rayner (2001) showed that familiarity with distractors around an unfamiliar target facilitates visual search. Eye movements were recorded (right eye only) by an SMI EyeLink head-mounted tracker. Eye positions were sampled at 250 Hz by an infra-red video-based system that also compensated for head movements. Gaze positions were accurate within 0.5°. A saccade was recorded when eye velocity exceeded 35°/s or eye acceleration exceeded 9500°/s^2. Results indicated comparably long, but fewer fixations when distractors were familiar, discounting the theory that unfamiliar distractors need longer processing.

Results from search studies begin to call into question the role of covert attention in the search process (Findlay and Gilchrist 1998). Findlay and Gilchrist advocate that a reappraisal of the role of covert attention in vision is in order because visual search does not provide a rationale for the existence of the covert attentional mechanism. Search situations in which the use of covert attention is advantageous are artificial and unusual. Most search tasks will be served better with overt eye scanning, guiding the eye as well as possible from the information that is being processed in parallel over the central regions of the visual field. The authors felt able to reject with some confidence the possibility that a fast covert scan of attention takes place during fixations in visual search. Covert attention might play no role, with a purely parallel process leading to saccade destination selection. Under this assumption, the saccade would be directed to the point of highest salience in some hypothetical "salience map."

21.6.1 Computational Models of Visual Search

Recent advances in the research of low-level visual processes have led to the emergence of sophisticated computational models of visual search. These models, operating on a host of still (and sometimes moving) images, model the human visual system's processes from the cornea to the striate cortex. An early example of such a model was proffered by Doll et al. (1993). The model contained six modules:

1. A pattern perception module (simulates HVS from cornea to visual cortex)
2. A visual search module (takes clutter into account)
3. A detection module (calculates probability of target detection given information on fixation locations from visual search module)

4. A decision module (predicts tradeoff between detections and false alarms)
5. An output/executive module (calculates cumulative probability of target acquisition and "missed" acquisitions; false alarms to clutter objects)
6. A target tracking module accounting for luminance contrast.

Doll et al.'s model at the time only handled luminance contrast; it did not contain modules to process chromatic contrast and motion.

Wolfe's Guided Search has undergone several revisions and "upgrades" (the current version is 3.0). Previous versions of GS successfully modeled a wide range of laboratory search tasks. The most recent revisions to the model were made in an effort to handle natural images (e.g., aerial photographs). The previous version of GS ignored two important factors: first, eye movements of human observers and the interplay of covert attentional movements with overt movements of the eyes; second, the anisotropic characteristics of the retina (visual processing is much more detailed at the fovea). The newest version of GS incorporates eye movements and eccentricity effects.

The simulation of Wolfe's Guided Search starts with a greyscale image as input (see Fig. 21.5). The image is processed by an array of On- and Off-Center Units, approximating retinal ganglion cell response. These provide input to Pre-Attentive Feature Maps for brightness and orientation, which model the representation of the stimulus in V1 by a complex log transform. Next, the Attention Gate allows feature information from only one object at a time to reach the higher processes such as Object Recognition. The Attentional Gate is under the control of the Activation Map. The Activation Map is a winner-take-all network that converges on a winner about every 50 ms. Feedback from the identification state to the Activation Map selects or inhibits the selected item, permitting new items to eventually win access to the identification stage.

To simulate eye movements, the Saccade Generation stage creates a saccade map. In GS, the Saccade Generation module is the analog of the superior coliculus. Activity in the saccade map is a blurred version of activity in the activation map. The GS model leads to a cooperative relationship between eye movements and attentional deployments. The simulation produces data that mimic human data on a number of tasks and GS eye movements resemble those recorded in human subjects.

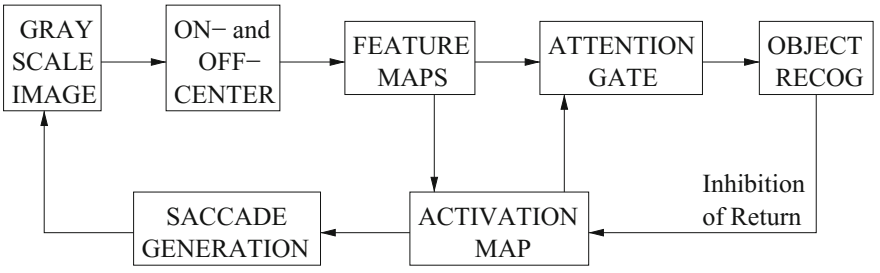

Fig. 21.5 Architecture of Guided Search 3.0. Adapted from Wolfe and Gancarz (1996) with permission © 1996 Kluwer Academic/Plenum Press

Wolfe's Guided Search appears to be restricted to monochrome (greyscale) still images. A similar attentional model has recently been proposed by Osberger and Maeder (1998). Osberger and Maeder's algorithm uses an importance map (IM) to predict the perceptual importance of segmented image regions. The IM is built by considering both low- and high-level visual attributes of segmented regions. Low-level factors include: contrast, size, and shape. High-level factors include the region's location, and foreground/background classification. Osberger and Maeder's algorithm is designed to be flexible to allow easy accommodation of future contributing modules to the IM, including one analyzing color and motion.

An attentional model that considers color, as well as other familiar visual attributes, has been proposed by Itti et al. (1998). Building on biologically plausible architectures of the human visual system, the model is related to Treisman's feature integration theory. The model architecture is shown in Fig. 21.6. Starting with an input image, it is progressively low-pass filtered and subsampled to yield nine dyadic spatial scales. The multiscale image representation is then decomposed into a set of topographic feature maps. Each feature is computed via a set of linear "center-surround" operations akin to visual receptive fields. Different spatial locations then compete for saliency within each map, such that only locations which locally stand out from their surround can persist. All feature maps feed, in a purely bottom-up manner, into a master "Saliency Map" (SM). Feature maps are combined into three "conspicuity maps," for intensity, color, and orientation at each scale of the saliency map. The model's saliency map contains internal dynamics that generate attentional shifts. The SM feeds into a biologically plausible "winner-take-all" (WTA) network. The Focus Of Attention (FOA) is shifted to the winning region once it has been identified by the WTA network. The WTA is then reset by a combination of local and global inhibition mechanisms that allow the selection of a future region to which the FOA is shifted in turn.

The model has been tested on a variety of artificial and natural images, and appears to be very robust, particularly to the addition of noise. Interestingly, the model was able to reproduce human performance for a number of pop-out tasks. When a target differed from an array of surrounding distractors by its unique orientation, color, intensity, or size, it was always the first attended location, irrespective of the number of distractors. Furthermore, when the target differed from the distractors by a conjunction of features, the search time necessary to find the target increased linearly with the number of distractors. This performance is in general consistent with observations in humans and is consistent with Treisman's feature integration theory.

Of course to test any of the above visual attention models, one very attractive methodology is to compare the sequence of Regions Of Interest (ROIs) identified by an attentional model to those actually selected by human observers. A recent study by Privitera and Stark (2000) presents just such a methodology for comparing algorithmic ROIs, or aROIs to those selected by humans, hROIs. The authors present a statistical and computational platform to perform the comparison between aROIs and hROIs. The comparison algorithm relies on two important processes: one of clustering of ROIs for comparison of loci of ROIs and a subsequent step of assembling the temporal sequences of ROIs into ordered strings of points for comparison of

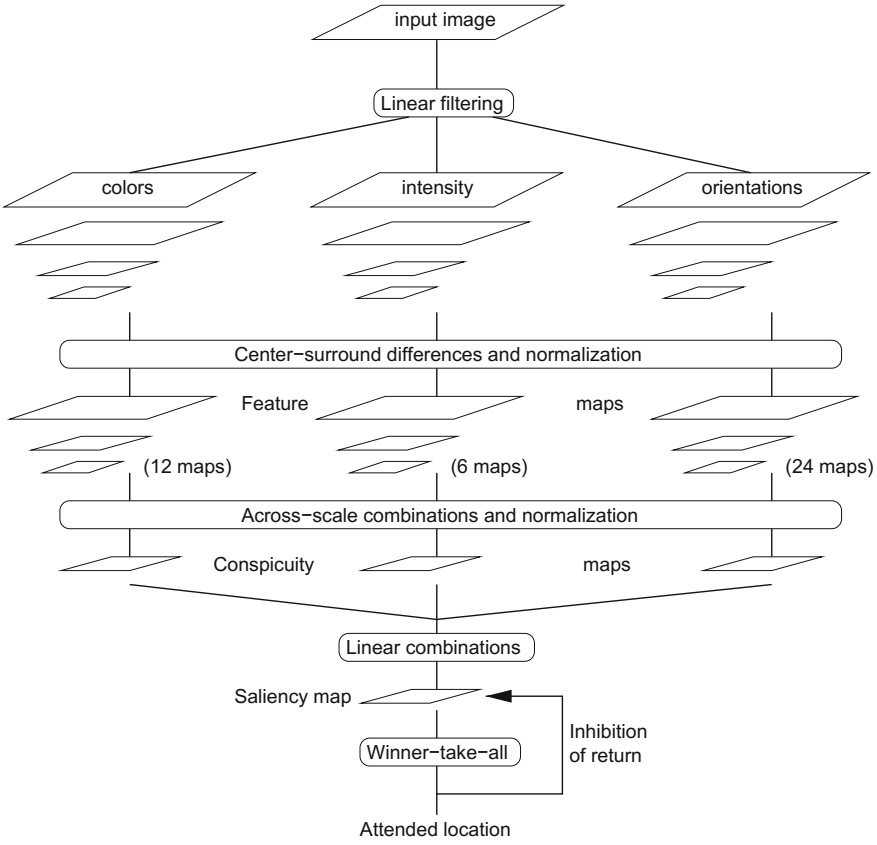

Fig. 21.6 Architecture of Itti et al.'s visual attention system. Adapted from Itti et al. with permission © 1998 IEEE

sequences based on *string editing*. Clustered ROIs within an image, either viewed by human observers, or selected by an attentional algorithm, are assigned character labels. Assembled strings of ROIs are compared by string editing, which, defined by an optimization algorithm, assigns unit cost to three different character operations: *deletion*, *insertion*, and *substitution*. Characters are manipulated to transform one string to another, and character manipulation costs are tabulated to yield a sequence similarity index S_s. An example is given in Fig. 21.7. A positional, or loci, similarity index S_p can be found for two strings by comparing the characters of the second string to those of the first. For the strings in Fig. 21.7, because all the characters of $s_2 = afbffdcdf$ are present in $s_1 = abcfeffgdc$, the two strings yield a loci similarity index of $S_p = 1$. Similarity coefficients are then sorted and stored in a table, named the Y-matrix, having as many rows and columns as the number of different sequence ROIs to be considered.

Given two strings $s_1 = abcfeffgdc$ *and* $s_2 = afbffdcdf$*, the second can be made to equal the first by applying the following operations:*

$s_1 = abcfeffgdc$		
$s_2 = afbffdcdf$	*start*	*cost 0*
$s_1 = abcfeffgdc$		
$s_2 = afeffdcdf$	*substitution of first b by e*	*cost 1*
$s_1 = abcfeffgdc$		
$s_2 = abcfeffdcdf$	*insertion of bc after first a*	*cost 2*
$s_1 = abcfeffgdc$		
$s_2 = abcfeffdc$	*deletion of last df*	*cost 2*
$s_1 = abcfeffgdc$		
$s_2 = abcfeffgdc$	*insertion of g*	*cost 1*

The total combined cost of deletions, insertions, and substitutions is 6. Relative to the original string length (9), the total cost yields a sequence similarity index between the two strings of $S_s = (1 - 6/9) = 0.34$.

Fig. 21.7 String editing example. From Privitera and Stark (2000) with permission © 2000 IEEE

Using string editing similarity measures, Privitera and Stark evaluated six different attentional algorithms, some sharing similarities with the above models of Wolfe, Osberger and Maeder, and Itti et al.. Although the algorithms tested were not the actual ones proposed by the latter group of authors, it appears that a multiresolutional strategy, such as that of Itti et al., seems to be very efficient for several classes of images. In general, although the set of algorithms picked by Privitera and Stark was only a small representative sample of many possible procedures, this set could indeed predict eye fixations.

The problem of computationally modeling human eye movements, and indeed human visual search, is far from being solved. No current model of visual search is as yet complete. Recent progress in algorithmic sophistication is encouraging. Models such as those of Wolfe, Osberger and Maeder, and Itti et al., show definite promise. Certainly the capability of tracking human eye movements has yet to play a crucial role in the corroboration of any model of visual search. Carmi and Itti (2006) recently evaluated the saliency of dynamic visual cues, noting that these cues play a dominant causal role in attracting attention. Peters and Itti (2006) suggest heuristics sensitive to dynamic events as predictors of human attention in next-generation immersive virtual environments and games.

Privitera and Stark's approach to the comparison of human and algorithmic scanpaths is one of the first methods to appear to quantitatively measure not only the loci of ROIs but also the order of ROIs. Undoubtedly future evaluation of human and artificial scanpaths will play a critical role in the investigation of visual search.

Indeed, scanpath comparison has recently received a good deal of attention. New algorithms are quickly being developed and applied with interesting insights. For example, West et al.'s (2006) *eyePatterns* was recently announced, a scanpath comparison technique inspired by bioinformatics techniques used for DNA sequence comparison. West et al.'s approach is similar to Privitera and Stark's string editing, but it makes use of the Needleman–Wunsch string similarity metric rather than Levenshtein's string distance metric. The underlying dynamic programming methods are duals of each other, with Needleman–Wunsch providing significant flexibility through its use of a user-defined similarity scoring matrix (Waterman 1989). Guan et al. (2006) relied on Needleman–Wunsch string editing to validate the Retrospective Think-Aloud (RTA) usability testing protocol. By comparing verbalized and fixated Areas Of Interest (AOIs), they found stimulated RTA to be valid and reliable (stimulated RTA refers to retrospection cued by playback of recorded eye movements, a technique particularly suitable for gaze tracked usability testing; see Chap. 24).

21.7 Natural Tasks

A host of useful factual information has been derived through psychophysical testing (e.g., spatial acuity, contrast sensitivity function, etc.). These types of studies often rely on the display of basic stimuli, e.g., sine wave gratings, horizontal and vertical bars, etc. Although certainly central to the development of such theories as feature integration, one criticism of these artificial stimuli is their simplicity. As discussed above, studies of visual search are expanding to consider more complex stimuli such as natural scenery (Hughes et al. 1996). However, viewing of pictures projected on a laboratory display still constitutes something of an artificial task. Recent advancements in wearable and virtual displays now allow collection of eye movements in more natural situations, usually involving the use of generally unconstrained eye, head, and hand movements.

Important work in this area has been reported by Land et al. (1999) and Land and Hayhoe (2001). The aim of the first study was to determine the pattern of fixations during the performance of a well-learned task in a natural setting (making tea), and to classify the types of monitoring action that the eyes perform. Results of this study indicate that even automated routine activities require a surprising level of continuous monitoring. A head-mounted eye-movement video camera was used, which provided a continuous view of the scene ahead, with a dot indicating foveal direction with an accuracy of about 1°. Foveal direction was always close to the object being manipulated, and very few fixations were irrelevant to the task. Roughly a third of all fixations on objects could be definitely identified with one of four monitoring functions: locating objects used later in the process, directing the hand or object in the hand to a new location, guiding the approach of one object to another (e.g., kettle and lid), and checking the state of some variable (e.g., water level). Land et al. (1999) conclude that although the actions of tea-making are "automated" and proceed with

little conscious involvement, the eyes closely monitor every step of the process. This type of unconscious attention must be a common phenomenon in everyday life.

Investigating a similar natural task, Land and Hayhoe (2001) examined the relations of eye and hand movements in extended food preparation tasks. The paper compares the task of tea-making against the task of making peanut butter and jelly sandwiches. In both cases the location of foveal gaze was monitored continuously using a head-mounted eye tracker with an accuracy of about 1°, and the head was free to move. In the tea-making study the three subjects had to move about the room to locate the objects required for the task; in the sandwich-making task the seven subjects were seated in one place, in front of a table. The eyes usually reached the next object in the sequence before any sign of manipulative action, indicating that eye movements are planned into the motor pattern and lead each action. The eyes usually fixated the same object throughout the action upon it, although they often moved on to the next object in the sequence before completion of the preceding action. Specific roles of individual fixations were found to be similar to roles in the tea-making task (see above). Land and Hayhoe argue that, at the beginning of each action, the oculomotor system is supplied with the identity of the required objects, information about its location, and instructions about the nature of the monitoring required during the action. The eye movements during this kind of task are nearly all to task-relevant objects, and thus their control is seen primarily top-down, and influenced very little by the "intrinsic salience" of objects. General conclusions provided by Land and Hayhoe are that the eyes provide information on an "as needed" basis, but that the relevant eye movements usually precede the motor acts they mediate by a fraction of a second. Eye movements are thus in the vanguard of each action plan, and are not simply responses to circumstances. Land and Hayhoe conclude that their studies lend no support to the idea that the visual system builds up a detailed model of the surroundings and operates from that. Most information is obtained from the scene as it is needed.

A good deal of additional work on eye movement measurement during natural tasks has also been performed by a group of researchers seemingly co-located around Rochester, NY, mainly at the University of Rochester and the Rochester Institute of Technology. This group of researchers has been investigating the relationship among eye, head, and hand movements for some time (among other topics).

Ballard et al. (1995) investigated the use of short-term memory in the course of a natural hand–eye task. The investigation focused on the minimization of subjects' use of short-term memory by employing deictic primitives through serialization of the task with eye movements (e.g., using the eyes to "point to" objects in a scene in lieu of memorizing all of the objects' positions and other properties). The authors argue that a deictic strategy in a pick-and-place task employs a more efficient use of a frame of reference centered at the fixation point, rather than a viewer-centered reference frame that might require memorization of objects in the world relative to coordinates centered at the viewer. Furthermore, deictic strategies may lead to a computational simplification of the general problem of relating internal models to objects in the world. Sequential, problem-dependent eye movements avoid the general problem of associating many models to many parts of the image simultaneously. Thus, in a pick

and place task, the steps required to perform the task of picking up a certain object from a group of objects can employ deictic references by fixating the object to be picked up next, without having to internalize a geometric reference frame for the entire set of objects.

Ballard et al. tested the above deictic reference assumption to see whether humans in fact use their eye movements in a deictic fashion in the context of natural behavior. A head-mounted eye tracker was used to measure eye movements over a three-dimensional physical workplace block display, divided into three areas, the model, source, and workspace. The task assigned to subjects was to move and assemble blocks from the source region to the workspace, arranging the blocks to match the arrangement in the model area. An example of the setup is shown in Fig. 21.8. By recording eye movements during the block pick-and-place task, the authors were able to show that subjects frequently directed gaze to the model pattern before arranging blocks in the workspace area. This suggests that information is acquired incrementally during the task and is not acquired in toto at the beginning of the tasks. That is, subjects appeared to use short-term memory frugally, acquiring information just prior to its use, and did not appear to memorize the entire model block configuration before making a copy of the block arrangement.

In a similar block-moving experiment, Smeets et al. (1996) were able to show that horizontal movements of gaze, head, and hand followed a coordinated pattern. A shift of gaze was generally followed by a movement of the head, which preceded the movement of the hand. This relationship is to a large extent task-dependent. In goal-directed tasks in which future points of interest are highly predictable, the

Fig. 21.8 Eye tracking in a natural pick-and-place task. Courtesy of Jeff Pelz, Visual Perception Laboratory, Carlson Center for Imaging Science, Rochester Institute of Technology http://www.cis.rit.edu/people/faculty/pelz/research/ASL_tracker.html. Reproduced with permission

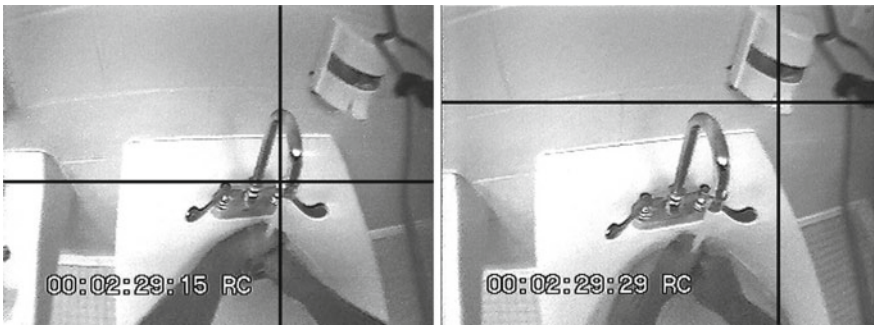

Fig. 21.9 Eye tracking in a natural hand-washing task. From Pelz et al. (2000) © 2000 ACM, Inc. Reprinted by permission

authors hypothesize that although gaze and head movements may decouple, the actual position of the hand is a likely candidate for the next gaze shift.

A recent example of such intentional visual attention was demonstrated by Pelz et al. (2000). Using a wearable eye tracker, Pelz was able to show intentionally based eye movements, which he termed "lookahead" eye movements, during a simple handwashing task. In this experiment, a subject donned a wearable computer and eye tracking rig (the computer was worn in a backpack). During a simple handwashing task, recorded eye movements showed that gaze moved to a location (soap dispenser) prior to the movement of the hands to the same location (see Fig. 21.9).

To further examine issues raised by observations of natural behavior, Hayhoe et al. (2002) have recently begun using complex virtual environments that can be manipulated by the experimenter at critical points during task performance. In a virtual environment where subjects copy toy models, the authors show that regularities in the spatial structure are used by subjects to control eye movement targeting. Other experiments in a virtual environment with haptic feedback show that even simple visual properties such as size are not continuously available or processed automatically by the visual system, but are dynamically acquired and discarded according to the momentary task demands.

21.8 Eye Movements in Other Information Processing Tasks

Eye movements have been recorded and studied in a host of information processing tasks. Rayner (1998) provides a comprehensive review of eye movement work from multiple domains. The reader is referred to Rayner's article for the full account, the remaining classes of eye tracking applications not discussed above are listed here in point form. Unless otherwise noted, the information on this research comes from Rayner (1998).

Auditory Language Processing

In this paradigm, eye movements are recorded as people listen to a story or follow instructions regarding an array at which they are looking. Cooper (1974) introduced this method and found that when people are simultaneously presented with spoken language and a visual field containing elements semantically related to the informative items of speech, they tend to spontaneously direct their line of sight to those elements that are most closely related to the meaning of the language currently heard. Cooper observed three main types of visual behavior: (1) a visual–aural interaction mode, in which fixation of targets was correlated with the meaning of concurrently heard language; (2) a free-scanning mode, in which the subject continually altered his direction of gaze in a manner independent of the meaning of concurrently heard language; (3) a point-fixation mode, in which the subject continued to fixate the same location independent of the meaning of concurrently heard language. It was frequently the case that subjects would vacillate between more than one of these modes during the presentation of a single story or comprehension test. Informal evidence based upon subjects' postexperimental verbal reports suggests that these three types of visual behaviors may be related to their distribution of attention between the visual and auditory modalities.

The eye movement paradigm has also been applied to auditory language processing by Allopenna et al. (1998). These authors used a paradigm in which participants followed spoken instructions to manipulate either real objects or pictures displayed on a computer screen while their eye movements were monitored using an Applied Science Laboratories (ASL) E4000 eye tracker which features a lightweight camera mounted on a headband. Allopenna et al. found that eye movements to objects in the workspace are closely time-locked to referring expressions in the unfolding speech stream, providing a sensitive and nondisruptive measure of spoken language comprehension during continuous speech.

Allopenna et al. (1998) addressed two important methodological issues with the eye tracking paradigm. First, they showed that the use of a restricted set of lexical possibilities does not appear to artificially inflate similarity effects. In particular, no evidence for rhyme effects was found with successive gating, which is a task that emphasizes work-initial information. Second, the authors provided clear evidence in support of a simple linking hypothesis between activation levels and the probability of fixating on a target. The predicted probability that an object would be fixated over time closely corresponded to the behavioral data. The availability of a mapping between hypothesized activation levels and fixation probabilities that can be used to generate quantitative predictions means that eye movement data can be used to test detailed predictions of explicit models.

Allopenna et al. argue that the sensitivity of the response measure coupled with a clear linking hypothesis between lexical activation and eye movements indicates that this methodology will be invaluable in exploring questions about the microstructure of lexical access during spoken word recognition. Eye movement methodology should be especially well suited to addressing questions about how fine-grained acoustic information affects word recognition. Allopenna et al. believe that a

particularly exciting aspect of the methodology is that it can be naturally extended to issues of segmentation and lexical access in continuous speech under relatively natural conditions.

Mathematics, Numerical Reading, and Problem Solving

In this area of investigation, eye movements are recorded as participants solve math and physics problems, as well as analogies. Not surprisingly, more complicated aspects of the problems typically lead to more and longer fixations.

Eye Movements and Dual Tasks

This methodology involves examination of eye movements when viewers are engaged in a dual-task situation; for example, a speeded manual choice response to a tone is made in close proximity to an eye movement. Although there is some slowing of the eye movement, the dual-task situation does not yield the dual-task interference effect typically found.

Face Perception

When examining faces, people tend to fixate on the eyes, nose, mouth, and ears. Fixations tend to be longer when comparisons have to be made between two faces rather than when a single face is examined.

Illusions and Imagery

These studies deal with illusions, such as the Necker cube or ambiguous figures.

Brain Damage

Brain damage studies have examined eye movements of patients with scotomas and visual neglect as they engage in reading, visual search, and scene perception.

Dynamic Situations

Eye movements have been examined in a host of dynamic situations such as driving, basketball foul shooting, golf putting, table tennis, baseball, gymnastics, walking in uneven terrain, mental rotation, and interacting with computers. Some of these applications are covered in the next chapter. Studies in which eye–hand coordination is important, such as playing video games, have revealed orderly sequences in which people coordinate looking and action.

21.9 Summary and Further Reading

This chapter presented eye tracking-related work mainly related to neuropsychology. Neuroscientific investigation linking functional brain mapping and eye tracking may seem a touch futuristic, however, functional brain imaging devices combined with

eye trackers are already available. Currently such devices are undoubtedly expensive, however, it is certain that in due time these devices will be used to investigate visual attentional phenomena from entirely new perspectives.

The chapter focused mainly on traditional psychological investigations of vision featuring eye tracking: vision during reading, visual search, perception of art, and vision in natural and virtual environments. As can be seen, there is a good deal of potential for collaboration among computer scientists, eye tracking researchers, and scientists investigating visual perception. Providing the capability of recording eye movements over new and complex stimuli will undoubtedly extend our knowledge of vision and visual perception.

Good sources of information for keeping track of this type of work are scientific journals and conferences dealing with vision, psychology, and eye tracking. Examples include *Vision Research*, *Behavior Research Methods, Instruments, and Computers (BRMIC)*, and the proceedings of the European Conference on Eye Movements (ECEM), and the U.S.-based Eye Tracking Research & Applications (ETRA).