Chapter 1
Visual Attention

In approaching the topic of eye tracking, we first have to consider the motivation for recording human eye movements. That is, why is eye tracking important? Simply put, we move our eyes to bring a particular portion of the visible field of view into high resolution so that we may see in fine detail whatever is at the central direction of gaze. Most often we also divert our attention to that point so that we can focus our concentration (if only for a very brief moment) on the object or region of interest. Thus, we may presume that if we can track someone's eye movements, we can follow along the path of attention deployed by the observer. This may give us some insight into what the observer found interesting, that is, what drew their attention, and perhaps even provide a clue as to how that person perceived whatever scene she or he was viewing.

By examining attention and the neural mechanisms involved in visual attention, the first two chapters of this book present motivation for the study of eye movements from two perspectives: a psychological viewpoint examining attentional behavior and its history of study (presented briefly in this chapter); and a physiological perspective on the neural mechanisms responsible for driving attentional behavior (covered in the next chapter). In sum, both introductory chapters establish the psychological and physiological basis for the movements of the eyes.

To begin formulating an understanding of an observer's attentional processes, it is instructive to first establish a rudimentary or at least intuitive sense of what attention is, and whether the movement of the eyes does in fact disclose anything about the inner cognitive process known as visual attention.

Visual attention has been studied for over a hundred years. A good qualitative definition of visual attention was given by the psychologist William James:

> Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others…

A.T. Duchowski, *Eye Tracking Methodology*, DOI 10.1007/978-3-319-57883-5_1

> When the things are apprehended by the *senses*, the number of them that can be attended to at once is small, *'Pluribus intentus, minor est ad singula sensus.'*
>
> —W. James (1981)

The Latin phrase used above by James roughly translates to "*Many filtered into few for perception*." The faculty implied as the filter is attention.

Humans are finite beings that cannot attend to all things at once. In general, attention is used to focus our mental capacities on selections of the sensory input so that the mind can successfully process the stimulus of interest. Our capacity for information processing is limited. The brain processes sensory input by concentrating on specific components of the entire sensory realm so that interesting sights, sounds, smells, and the like, may be examined with greater attention to detail than peripheral stimuli. This is particularly true of vision. Visual scene inspection is performed *minutatim*, not *in toto*. That is, human vision is a piecemeal process relying on the perceptual integration of small regions to construct a coherent representation of the whole.

In this chapter, attention is recounted from a historical perspective following the narrative found in Van der Heijden (1992). The discussion focuses on attentional mechanisms involved in vision, with emphasis on two main components of visual attention, namely the "what" and the "where".

1.1 Visual Attention: A Historical Review

The phenomenon of visual attention has been studied for over a century. Early studies of attention were technologically limited to simple ocular observations and oftentimes to introspection. Since then the field has grown to an interdisciplinary subject involving the disciplines of psychophysics, cognitive neuroscience, and computer science, to name three. This section presents a qualitative historical background of visual attention.

1.1.1 Von Helmholtz's "Where"

At the second half of the 19th century, Von Helmholtz (1925) posited visual attention as an essential mechanism of visual perception. In his *Treatise on Physiological Optics*, he notes, "We let our eyes roam continually over the visual field, because that is the only way we can see as distinctly as possible all the individual parts of the field in turn." Noting that attention is concerned with a small region of space, Von Helmholtz observed visual attention's natural tendency to wander to new things. He also remarked that attention can be controlled by a conscious and voluntary effort, allowing attention to peripheral objects without making eye movements to that object. Von Helmholtz was mainly concerned with eye movements to spatial locations, or the

"where" of visual attention. In essence, although visual attention can be consciously directed to peripheral objects, eye movements reflect the will to inspect these objects in fine detail. In this sense, eye movements provide evidence of overt visual attention.

1.1.2 James' "What"

In contrast to Von Helmholtz's ideas, James (1981) believed attention to be a more internally covert mechanism akin to imagination, anticipation, or in general, thought. James defined attention mainly in terms of the "what", or the identity, meaning, or expectation associated with the focus of attention. James favored the active and voluntary aspects of attention although he also recognized its passive, reflexive, non-voluntary and effortless qualities.

Both views of attention, which are not mutually exclusive, bear significantly on contemporary concepts of visual attention. The "what" and "where" of attention roughly correspond to foveal (James) and parafoveal (Von Helmholtz) aspects of visual attention, respectively. This dichotomous view of vision is particularly relevant to a bottom-up or feature-driven explanation of visual attention. That is, when considering an image stimulus, we may consider certain regions in the image that will attract one's attention. These regions may initially be perceived parafoveally, in a sense requesting further detailed inspection through foveal vision. In this sense, peripherally located image features may drive attention in terms of "where" to look next, so that we may identify "what" detail is present at those locations.

The dual "what" and "where" feature-driven view of vision is a useful preliminary metaphor for visual attention, and indeed it has formed the basis for creating computational models of visual attention, which typically simulate so-called low-level, or bottom-up visual characteristics. However, this view of attention is rather simplistic. It must be stressed that a complete model of visual attention involves high-level visual and cognitive functions. That is, visual attention cannot simply be explained through the sole consideration of visual features. There are higher-level intentional factors involved (e.g., related to possibly voluntary, preconceived cognitive factors that drive attention).

1.1.3 Gibson's "How"

In the 1940s Gibson (1941) proposed a third factor of visual attention centered on intention. Gibson's proposition dealt with a viewer's advance preparation as to whether to react and if so, how, and with what class of responses. This component of attention explained the ability to vary the intention to react while keeping the expectation of the stimulus object fixed, and conversely, the ability to vary the expectation of the stimulus object while keeping the intention to react fixed. Experiments involving ambiguous stimuli typically evoke these reactions. For example, if the viewer

is made to expect words describing animals, then the misprint "sael" will be read as "seal". Changing the expectation to words describing ships or boats invokes the perception of "sail". The reactive nature of Gibson's variant of attention specifies the "what to do", or "how to react" behavior based on the viewer's preconceptions or attitude. This variant of visual attention is particularly relevant to the design of experiments. It is important to consider the viewer's perceptual expectation of the stimulus, as (possibly) influenced by the experimenter's instructions.

1.1.4 Broadbent's "Selective Filter"

Attention, in one sense, is seen as a "selective filter" responsible for regulating sensory information to sensory channels of limited capacity. In the 1950s, Broadbent (1958) performed auditory experiments designed to demonstrate the selective nature of auditory attention. The experiments presented a listener with information arriving simultaneously from two different channels, e.g., the spoken numerals {7, 2, 3} to the left ear, {9, 4, 5} to the right. Broadbent reported listeners' reproductions of either {7, 2, 3, 9, 4, 5}, or {9, 4, 5, 7, 2, 3}, with no interwoven (alternating channel) responses. Broadbent concluded that information enters in parallel but is then selectively filtered to sensory channels.

1.1.5 Deutsch and Deutsch's "Importance Weightings"

In contrast to the notion of a selective filter, Deutsch and Deutsch (1963) proposed that all sensory messages are perceptually analyzed at the highest level, precluding a need for a selective filter. Deutsch and Deutsch rejected the selective filter and limited capacity system theory of attention; they reasoned that the filter would need to be at least as complex as the limited capacity system itself. Instead, they proposed the existence of central structures with preset "importance weightings" that determined selection. Deutsch and Deutsch argued that it is not attention as such but the weightings of importance that have a causal role in attention. That is, attentional effects are a result of importance, or relevance, interacting with the information.

It is interesting to note that Broadbent's selective filter generally corresponds to Von Helmholtz's "where", whereas Deutsch and Deutsch's importance weightings correspond to James' expectation, or the "what". These seemingly opposing ideas were incorporated into a unified theory of attention by Anne Treisman in the 1960s (although not fully recognized until 1971). Treisman brought together the attentional models of Broadbent and Deutsch and Deutsch by specifying two components of attention: the attenuation filter followed by later (central) structures referred to as "dictionary units". The attenuation filter is similar to Broadbent's selective filter in that its function is selection of sensory messages. Unlike the selective filter, it does not completely block unwanted messages, but only attenuates them. The later

stage dictionary units then process weakened and unweakened messages. These units contain variable thresholds tuned to importance, relevance, and context. Treisman thus brought together the complementary models of attentional unit or selective filter (the "where"), and expectation (the "what").

Up to this point, even though Treisman provided a convincing theory of visual attention, a key problem remained, referred to as the scene integration problem. The scene integration problem poses the following question: even though we may view the visual scene through something like a selective filter, which is limited in its scope, how is it that we can piece together in our minds a fairly coherent scene of the entire visual field? For example, when looking at a group of people in a room such as in a classroom or at a party, even though it is impossible to gain a detailed view of everyone's face at the same time, nevertheless it is possible to assemble a mental picture of where people are located. Our brains are capable of putting together this mental picture even though the selective filter of vision prevents us from physically doing so in one glance. Another well-known example of the scene integration problem is the Kanizsa (1976) illusion, exemplified in Fig. 1.1, named after the person who invented it. Inspecting Fig. 1.1, you will see the edges of a triangle, even though the triangle is defined only by the notches in the disks. How this triangle is integrated by the brain is not yet fully understood. That is, although it is known that the scene is inspected piecemeal as evidenced by the movement of the eyes, it is not clear how the "big picture" is assembled, or integrated, in the mind. This is the crux of the scene integration problem. In one view, offered by the Gestalt psychologists, it is hypothesized that recognition of the entire scene is performed by a parallel one-step process. To examine this hypothesis, a visualization of how a person views such an image (or any other) is particularly helpful. This is the motivation for recording and visualizing a viewer's eye movements. Even though the investigation of eye movements dates back to 1907 (Dodge 1907),[1] a clear depiction of eye movements would not be available until 1967 (see Chap. 5 for a survey of eye tracking techniques). This early eye movement visualization, discussed below, shows the importance of eye movement recording not only for its expressive power of depicting one's visual scanning characteristics, but also for its influence on theories of visual attention and perception.

1.1.6 Yarbus and Noton and Stark's "Scanpaths"

Early diagrammatic depictions of recorded eye movements helped cast doubt on the Gestalt hypothesis that recognition is a parallel one-step process. The Gestalt view of recognition is a holistic one suggesting that vision relies to a great extent on the tendency to group objects. Although well-known visual illusions exist to support this view [e.g., subjective contours of the Kanizsa figure; see Fig. 1.1 above Kanizsa

[1] As cited in Gregory (1990).

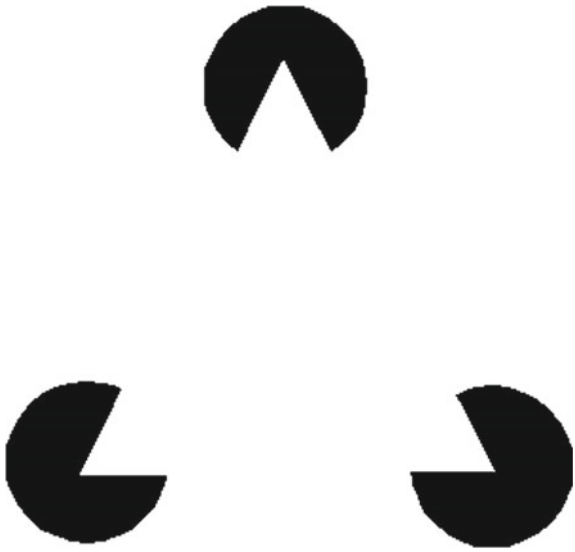

Fig. 1.1 The Kanizsa illusion

(1976)], early eye movement recordings showed that visual recognition is at least partially serial in nature.

Yarbus (1967) measured subjects' eye movements over an image after giving subjects specific questions related to the image. Such a picture is shown in Fig. 1.2. Questions posed to subjects included a range of queries specific to the situation, e.g., are the people in the image related, what are they wearing, what will they have to eat, and so on. The eye movements Yarbus recorded demonstrated sequential viewing patterns over particular regions in the image.

Noton and Stark (1971a, b) performed their own eye movement measurements over images and coined the observed patterns "scanpaths". Their work extended Yarbus' results by showing that even without leading questions subjects tend to fixate identifiable regions of interest, or "informative details". Furthermore, scanpaths showed that the order of eye movements over these regions is quite variable. That is, given a picture of a square, subjects will fixate on the corners, although the order in which the corners are viewed differs from viewer to viewer and even differs between consecutive observations made by the same individual.

In contrast to the Gestalt view, Yarbus' and Noton and Stark's work suggests that a coherent picture of the visual field is constructed piecemeal through the assembly of serially viewed regions of interest. Noton and Stark's results support James' "what" of visual attention. With respect to eye movements, the "what" corresponds to regions of interest selectively filtered by foveal vision for detailed processing.

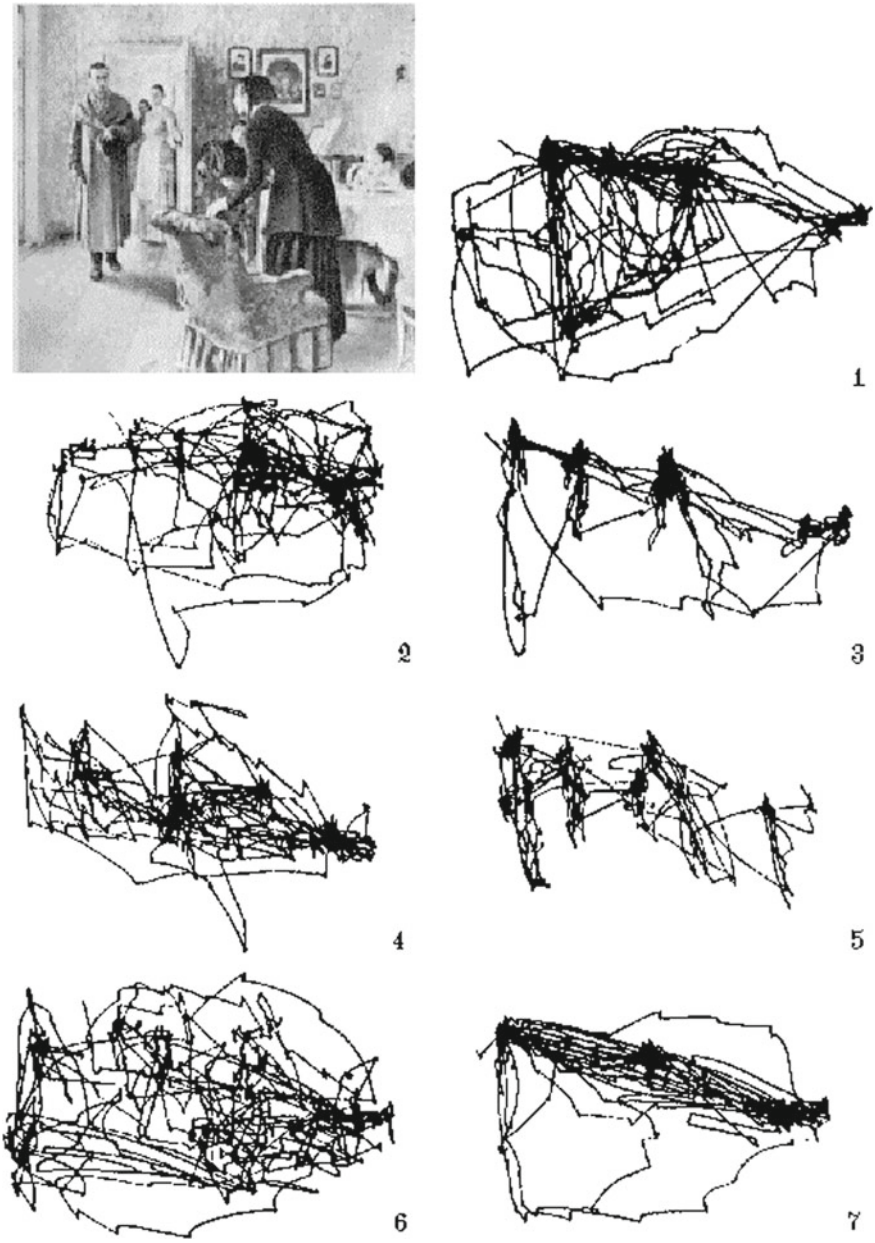

Fig. 1.2 Yarbus' early eye movement recordings. Reprinted from Yarbus (1967) with permission ©1967 Plenum Press. In each of the traces, the subject was asked to: Trace 1, examine the picture at will; Trace 2, estimate the economic level of the people; Trace 3, estimate the people's ages; Trace 4, guess what the people were doing before the arrival of the visitor; Trace 5, remember the people's clothing; Trace 6, remember the people's (and objects') position in the room; Trace 7, estimate the time since the guest's last visit

1.1.7 Posner's "Spotlight"

Contrary to the serial "what" of visual attention, the orienting, or the "where", is performed in parallel (Posner et al. 1980). Posner et al. suggested an attentional mechanism able to move about the scene in a manner similar to a "spotlight." The spotlight, being limited in its spatial extent, seems to fit well with Noton and Stark; Noton and Stark's empirical identification of foveal regions of interest. Posner et al., however, dissociate the spotlight from foveal vision and consider the spotlight an attentional mechanism independent of eye movements. Posner et al. identified two aspects of visual attention: the orienting and the detecting of attention. Orienting may be an entirely central (covert or mental) aspect of attention, whereas detecting is context-sensitive, requiring contact between the attentional beam and the input signal. The orienting of attention is not always dependent on the movement of the eyes; that is, it is possible to attend to an object while maintaining gaze elsewhere. According to Posner et al., orientation of attention must be done in parallel and must precede detection.

The dissociation of attention from foveal vision is an important point. In terms of the "what" and the "where", it seems likely that the "what" relates to serial foveal vision. The "where", on the other hand, is a parallel process performed parafoveally, or peripherally, which dictates the next focus of attention.

1.1.8 Treisman's "Glue"

Posner et al. and Noton and Stark advanced the theory of visual attention along similar lines forged by Von Helmholtz and James (and then Broadbent and Deutsch and Deutsch). Treisman once again brought these concepts together in the feature integration theory of visual attention (Treisman and Gelade 1980; Treisman 1986). In essence, attention provides the "glue" that integrates the separated features in a particular location so that the conjunction (i.e., the object) is perceived as a unified whole. Attention selects features from a master map of locations showing *where* all the feature boundaries are located, but not *what* those features are. That is, the master map specifies where things are, but not what they are. The feature map also encodes simple and useful properties of the scene such as color, orientation, size, and stereo distance. Feature Integration Theory, or FIT, is a particularly important theory of visual attention and visual search. Eye tracking is often a significant experimental component used to test FIT. Feature integration theory, treated as an eye tracking application, is discussed in more detail in Chap. 21.

1.1.9 Kosslyn's "Window"

Recently, Kosslyn (1994) proposed a refined model of visual attention. Kosslyn describes attention as a selective aspect of perceptual processing, and proposes an attentional "window" responsible for selecting patterns in the "visual buffer". The window is needed because there is more information in the visual buffer than can be passed downstream, and hence the transmission capacity must be selectively allocated. That is, some information can be passed along, but other information must be filtered out. This notion is similar to Broadbent's selective filter and Treisman's attenuation filter. The novelty of the attentional window is its ability to be adjusted incrementally; i.e., the window is scalable. Another interesting distinction of Kosslyn's model is the hypothesis of a redundant stimulus-based attention-shifting subsystem (e.g., a type of context-sensitive spotlight) in mental imagery. Mental imagery involves the formation of mental maps of objects, or of the environment in general. It is defined as "…the mental invention or recreation of an experience that in at least some respects resembles the experience of actually perceiving an object or an event, either in conjunction with, or in the absence of, direct sensory stimulation" (Finke 1989). It is interesting to note that the eyes move during sleep (known as Rapid Eye Movement or REM sleep). Whether this is a manifestation of the use of an internal attentional window during sleep is not known.

1.2 Visual Attention and Eye Movements

Considering visual attention in terms of the "what" and "where", we would expect that eye movements work in a way that supports the dual attentive hypothesis. That is, vision might behave in a cyclical process composed of the following steps.

1. Given a stimulus, such as an image, the entire scene is first seen mostly in parallel through peripheral vision and thus mostly at low resolution. At this stage, interesting features may "pop out" in the field of view, in a sense engaging or directing attention to their location for further detailed inspection.
2. Attention is thus turned off or disengaged from the foveal location and the eyes are quickly repositioned to the first region that attracted attention.
3. Once the eyes complete their movement, the fovea is now directed at the region of interest, and attention is now engaged to perceive the feature under inspection at high resolution.

This is a bottom-up model or concept of visual attention. If the model is accurate, one would expect to find regions in the brain that correspond in their function to attentional mechanisms. This issue is further investigated in Chap. 2.

The bottom-up model is at least correct in the sense that it can be said to be a component of natural human vision. In fact, the bottom-up model forms a powerful basis for computational models of visual search. Examples of such models are

presented later in the text (see Chap. 21). The bottom-up view of visual attention is, however, incomplete. There are several key points that are not addressed. Consider these questions:

1. Assuming it is only the visual stimulus (e.g., image features) that drives attention, exactly what types of features attract attention?
2. If the visual stimulus were solely responsible for attracting attention, would we ever need the capability of making voluntary eye movements?
3. What is the link between attention and eye movements? Is attention always associated with the foveally viewed portion of the visual scene?

To gain insight into the first question, we must examine how our physical visual mechanism (our eyes and brain) responds to visual stimulus. To attempt to validate a model of visual attention, we would need to be able to justify the model by identifying regions in the brain that are responsible for carrying out the functionality proposed by the model. For example, we would expect to find regions in the brain that engage and disengage attention as well as those responsible for controlling (i.e., programming, initiating, and terminating) the movements of the eyes. Furthermore, there must be regions in the brain that are responsible for responding to and interpreting the visual stimuli that are captured by the eyes. As shown in the following chapters, the Human Visual System (HVS) responds strongly to some types of stimuli (e.g., edges), and weakly to others (e.g., homogeneous areas). The following chapters show that this response can be predicted to a certain extent by examining the physiology of the HVS. In later chapters we also show that the human visual response can be measured through a branch of psychology known as psychophysics. That is, through psychophysics, we can fairly well measure the perceptive power of the human visual system.

The bottom-up model of visual attention does not adequately offer answers to the second question because it is limited to mostly bottom-up, or feature-driven aspects of attention. The answer to the second question becomes clearer if we consider a more complete picture of attention involving higher-level cognitive functions. That is, a complete theory of visual attention should also involve those cognitive processes that describe our voluntary intent to attend to something, e.g., some portion of the scene. This is a key point that was briefly introduced following the summary of Gibson's work, and which is evident in Yarbus' early scanpaths. It is important to reiterate that Yarbus' work demonstrated scanpaths which differed with observers' expectations; that is, scanpath characteristics such as their order of progression can be *task-dependent*. Based on what they are looking for, people will view a picture differently. A complete model or theory of visual attention is beyond the scope of this book, but see Chap. 21 for further insight into theories of visual search, and also for examples of the application of eye trackers to study this question.

Considering the third question opens up a classical problem in eye tracking studies. Because attention is composed of both low-level and high-level functions (one can loosely think of involuntary and voluntary attention, respectively), as Posner and others have observed, humans can voluntarily dissociate attention from the foveal direction of gaze. In fact, astronomers do this regularly to detect faint constellations

with the naked eye by looking "off the fovea." Because the periphery is much more sensitive to dim stimulus, faint stars are much more easily seen out of the "corner" of one's eye than when they are viewed centrally. Thus the high-level component of vision may be thought of as a covert component, or a component which is not easily detectable by external observation. This is a well-known problem for eye tracking researchers. An eye tracker can only track the overt movements of the eyes, however, it cannot track the covert movement of visual attention. Thus, in all eye tracking work, a tacit but very important assumption is usually accepted: we assume that attention is linked to foveal gaze direction, but we acknowledge that it may not always be so.

1.3 Summary and Further Reading

A historical account of attention is a prerequisite to forming an intuitive impression of the selective nature of perception. For an excellent historical account of selective visual attention, see Van der Heijden (1992). An earlier and very readable introduction to visual processes is a small paperback by Gregory (1990). For a more neurophysiological perspective, see Kosslyn (1994). Another good text describing early attentional vision is Papathomas et al. (1995).

The singular idioms describing the selective nature of attention are the "what" and the "where". The "where" of visual attention corresponds to the visual selection of specific regions of interest from the entire visual field for detailed inspection. Notably, this selection is often carried out through the aid of peripheral vision. The "what" of visual attention corresponds to the detailed inspection of the spatial region through a perceptual channel limited in spatial extent. The attentional "what" and "where" duality is relevant to eye tracking studies because scanpaths show the temporal progression of the observer's foveal direction of gaze and therefore depict the observer's instantaneous overt localization of visual attention.

From investigation of visual search, 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 and that features likely to attract attention include edges and corners, but not plain surfaces (see Chap. 21). There is some doubt, however, whether human visual search can be described as an integration of independently processed features (Van Orden and DiVita 1993). Van Orden and DiVita suggest that "…any theory on visual attention must address the fundamental properties of early visual mechanisms." To attempt to quantify the visual system's processing capacity, the neural substrate of the human visual system is examined in the following chapter which surveys the relevant neurological literature.