Eye-Based Interaction in Graphical Systems: Theory & Practice

Andrew Duchowski
Clemson University

Part I
Introduction to the Human Visual System (HVS)

A Visual Attention

> "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..."
>
> "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.'"*
>
> –William James [Jam81, pp.381-382,p.384]

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. 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 roughly bounded by the "magic" number 7 ± 2 [Mil56]. While listening to an orchestra, it is possible to concentrate on specific instruments, e.g., the flute or the oboe. The brain processes sensory input by concentrating on specific components of the entire sensory realm so that interesting sights, sounds, smells, etc., 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 section, attention is recounted from a historical perspective following the narrative found in [Van92]. The discussion focuses on attentional mechanisms involved in vision, with emphasis on two main components of visual attention, namely the "what" and the "where".

A.1 Chronological Review of Visual Attention

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.

A.1.1 Von Helmholtz's "where"

At the start of the 20^{th} century, Hermann Von Helmholtz 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." [VonH25, p.63]. 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.

A.1.2 James' "what"

In contrast to Von Helmholtz's ideas, William James believed attention to be a more internally covert mechanism akin to imagination, anticipation, or in general, thought [Jam81, XI]. 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.

A.1.3 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, Donald Broadbent performed auditory experiments designed to demonstrate the selective nature of auditory attention [Bro58]. 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.

A.1.4 Deutsch and Deutsch's "importance weightings"

In contrast to the notion of a selective filter, J. Anthony Deutsch and Diana Deutsch proposed that all sensory messages are perceptually analyzed at the highest level, precluding a need for a selective filter [DD63]. 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" which 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", while 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").

A.1.5 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 wholistic 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 [Mar82, p.51]), early eye movement recordings showed that visual recognition is at least partially serial in nature.

Yarbus measured subjects' eye movements over an image after giving subjects specific questions related to the image [Yar67]. Such a picture is shown in Figure 2. Questions posed to subjects ranged from queries specific to the situation, e.g., are the people in the image related,

> 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.

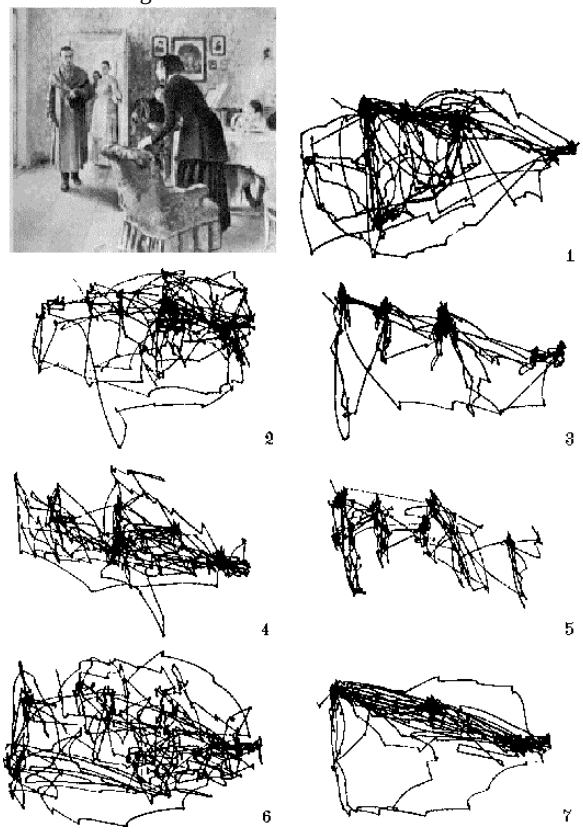


Figure 2: Yarbus' early eye movement recordings. Adapted from [Yar67].

what are they wearing, what will they have to eat, etc. The eye movements Yarbus recorded demonstrated sequential viewing patterns over particular regions in the image.

Noton and Stark performed their own eye movement measurements over images and coined the observed patterns "scanpaths" [NS71a, NS71b]. 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.

A.1.6 Posner's "spotlight"

Contrary to the serial "what" of visual attention, the orienting, or the "where", is performed in parallel [PSD80]. Posner 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's empirical identification of foveal regions of interest. Posner, however, dissociates the spotlight from foveal vision. Instead, the spotlight is an attentional mechanism independent of eye movements. Posner identified two aspects of visual attention: the orienting and the detecting of attention. Orienting of attention may be an entirely central (covert, or mental) aspect of attention, while the detecting aspect 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, the orienting of attention must be done in parallel and must precede the detecting aspect of attention.

A.1.7 Treisman's "glue"

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. Posner 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 with a feature integration theory of visual attention [TG80, Tre86]. In essence, attention provides the "glue" which 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.

A.1.8 Kosslyn's "window"

Recently, Kosslyn proposed a refined model of visual attention [Kos94]. 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 since 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" [Fin89, p.2].

A.2 Summary and Further Reading

An 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 [Van92]. An earlier, but nevertheless very readable introduction to visual processes is a small paperback by Gregory [Gre90]. For a more neurophysiological perspective, see [Kos94]. Another good text describing early attentional vision is [PCGK95].

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 display and interface design since this concept can guide informational content presentation by matching the processing capacity of foveal and peripheral vision. In visual search work, the consensus view is that a parallel, pre-attentive stage acknowledges the presence of four basic features: color, size, orientation, and presence and/or direction of motion [Dol93, Wol93]. Doll et al. suggest that features likely to attract attention include edges, corners, but not plain surfaces [DMS93]. Todd and Kramer 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 [TK93]. There is doubt in the literature that human visual search can be described as an integration of independently processed features [VOD93]. Van Oden 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 section which surveys the relevant neurological literature.

B Neurological Substrate of the HVS

Neurophysiological and psychophysical literature on the human visual system suggests the field of view is inspected *minutatim* through brief fixations over small regions of interest. This allows perception of detail through the fovea. Foveal vision, subtending 5° (visual angle), allows fine scrutiny of 3% of the entire screen (21in monitor at ∼60cm viewing distance). Approximately 90% of viewing time is spent in fixations. When visual attention is directed to a new area, fast eye movements (saccades) reposition the fovea.

The limited information capacity of the Human Visual System (HVS) provides epistemological reasons for visual attention from a phylogenetical standpoint, and is the *raison d'être* of visual attention. The dynamics of visual attention probably evolved in harmony with (or perhaps in response to) the perceptual limitations imposed by the neurological substrate of the visual system. The neural substrate of the human visual system is examined in this section. Emphasis is placed on differentiating the processing capability of foveal and peripheral vision.

B.1 The Eye

Often called "the world's worst camera", the eye, shown in Figure 3, suffers from numerous optical imperfections, e.g.,

- spherical aberrations: prismatic effect of peripheral parts of the lens;
- chromatic aberrations: shorter wavelengths (blue) refracted more than longer wavelengths (red);
- curvature of field: a planar object gives rise to a curved image.

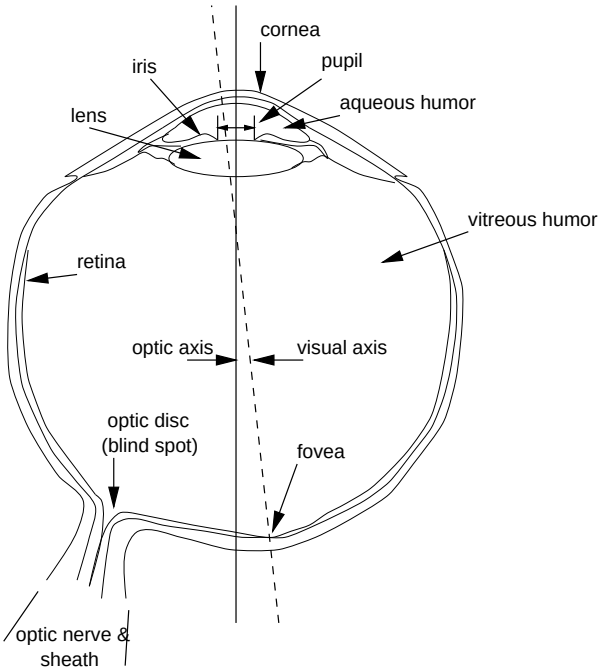


Figure 3: The eye. Adapted from [BL88b, p.34 (Fig. 1)].

However, the eye is also endowed with various mechanisms which reduce degradive effects, e.g.,

- to reduce spherical aberration, the iris acts as a stop, limiting peripheral entry of light rays;
- to overcome chromatic aberration, the eye is typically focused to produce sharp images of intermediate wavelengths;
- to match the effects of curvature of field, the retina is curved compensating for this effect.

The optics of the eye are schematically shown in Figure 4. For an excellent review of physiological optics, and visual perception in general, see [HW97].

B.2 The Retina

The retina contains receptors sensitive to light (photoreceptors) which constitute the first stage of visual perception. Photoreceptors can effectively be thought of as "transducers" converting light energy to electrical impulses (neural signals). Neural signals originating at these receptors lead to cortical visual centers. Photoreceptors are functionally classified into rods and cones. Rods are sensitive to dim and achromatic light (night vision), while cones respond to brighter, chromatic light (daylight vision). The retina contains 120 million rods and 7 million cones, and is is arranged concentrically.

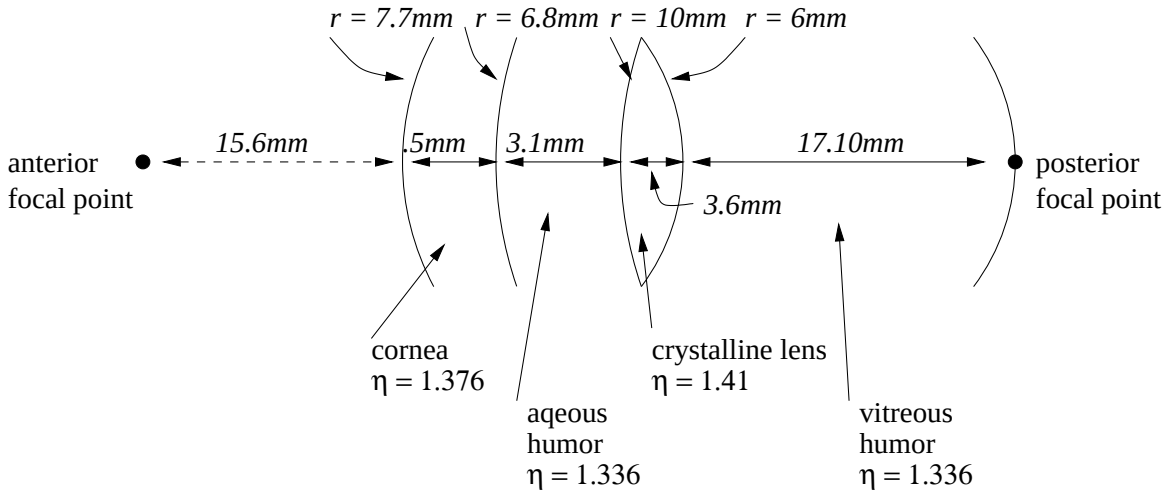


Figure 4: The optics of the eye. Adapted from [HW97, p.3 (Fig. 1.2)].

The retina is composed of multiple layers of different cell types [DD88]. Surprisingly, the "inverted" retina is constructed in such a way that photoreceptors are found at the bottom layer. This construction is somewhat counterintuitive since rods and cones are furthest away from incoming light, buried beneath a layer of cells. The retina resembles a three-layer cell sandwich, with connection bundles between each layer. These connectional layers are called plexiform or synaptic layers.

The retinogeniculate organization is schematically depicted in Figure 5(a). The outermost layer (w.r.t. incoming light) is the outer nuclear layer which contains the photoreceptor (rod/cone) cells. The first connectional layer is the outer plexiform layer which houses connections between receptor and bipolar nuclei. The next outer layer of cells is the inner nuclear layer containing bipolar (amacrine, bipolar, horizontal) cells. The next plexiform layer is the inner plexiform layer where connections between inner nuclei cells and ganglion cells are formed. The top layer, or the ganglion layer, is composed of ganglion cells.

The fovea's photoreceptors are special types of neurons—the nervous system's basic elements (see Figure 5(b)). Retinal rods and cones are specific types of dendrites. In general, individual neurons can connect to as many as 10,000 other neurons. Comprised of such interconnected building blocks, as a whole, the nervous system behaves like a large neural circuit.

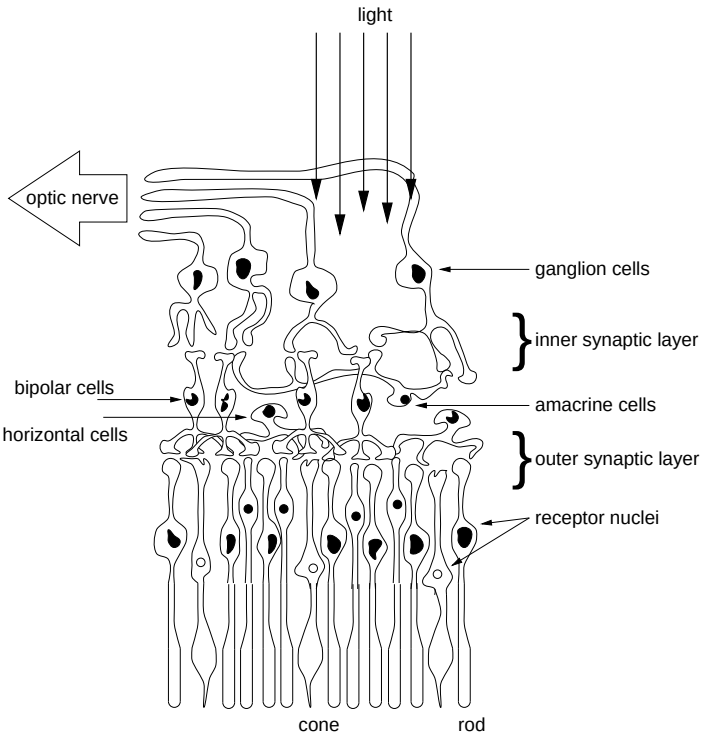
Certain neurons (e.g., ganglion cells) resemble a "digital gate", sending a signal (firing) when the cell's activation level exceeds a threshold. Ganglion cell activation comes as an effect of the potassium-sodium pump. Signals propagate through the axon in a wave of depolarization—the action potential. The action potential (lasting less than 1ms) occurs as sodium ions (Na+) permeate the depolarized neuronal membrane. As sodium ions flow in, potassium (K+) ions flow out restoring resting potential. Inhibitory signals allow inflow of chloride ions (Cl-) preventing depolarization. The myelin sheath is an axonal cover providing insulation which speeds up conduction of impulses. Unmyelinated axons of the ganglion cells converge to the optic disk (an opaque myelin sheath would block light). Axons are myelinated at the optic disk, and connect to the Lateral Geniculate Nuclei (LGN) and the Superior Colliculus (SC).

The Outer Layer Rods and cones of the outer retinal layer respond to incoming light. A simplified account of the function of these cells is that rods provide monochromatic, scotopic (night) vision, and cones provide trichromatic, photopic (day) vision. Both types of cells are partially sensitive to mesopic (twighlight) light levels.

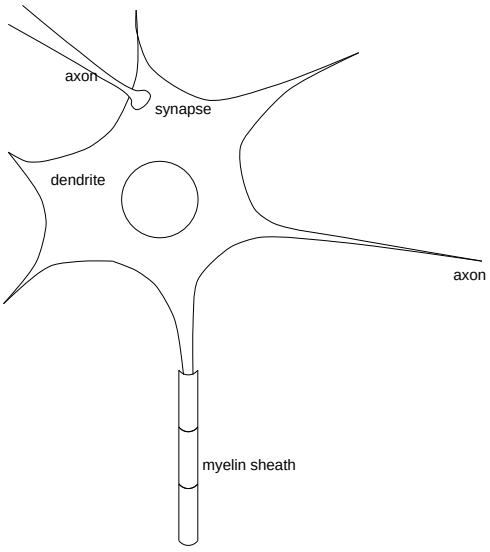
The Inner Nuclear Layer Outer receptor cells are laterally connected to the horizontal cells. In the fovea, each horizontal cell is connected to about 6 cones, and in the periphery to about 30-40 cones. Centrally, the cone bipolar cells contact one cone directly, and several cones indirectly through horizontal or receptor-receptor coupling. Peripherally, cone bipolar cells directly contact several cones. The number of receptors increases eccentrically. The rod bipolar cells contact a considerably larger number of receptors than cone bipolars. There are two main types of bipolar cells, ones that depolarize to increments of light (+), and others that depolarize to decrements of light (-). The signal profile (cross-section) of bipolar receptive fields is a "Mexican hat", or center-surround, with an on-center, or off-center signature.

The Ganglion Layer Ganglion cells form an "active contrast-enhancing system," not a camera-like plate. Centrally, ganglion cells directly contact one bipolar. Peripherally, ganglion cells directly contact several bipolars.

Ganglion cells are distinguished by their morphological and functional characteristics. Morphologically, there are two types of ganglion cells, the α and β cells. Approximately 10% of retinal ganglion cells are α cells possessing large cell bodies and dendrites, and about 80% are β cells with small bodies and dendrites [LWL95]. The α cells project to the magnocellular (M-) layers of LGN while the β cells project to the parvocellular (P-) layers. A third channel of input relays through narrow, cell-sparse laminae between the main M- and P-layers of the LGN. Its origin in the retina is not yet known. Functionally, ganglion cells fall into three classes, the X, Y, and W cells [DD88, Kap91]. X cells respond to sustained stimulus, location and fine detail, and nervate along both M- and P-projections. Y cells nervate only along the



(a) The retinocellular layers.
Adapted from [HH73, p.24 (Fig. 2.15)].



(b) Schematic of the neuron.
Adapted from [BL88a, pp.31-32 (Fig. 2.1, Fig. 2.2)].

Figure 5: The neurophysiology of the retina.

M-projection, and respond to transient stimulus, coarse features, and motion. W cells respond to coarse features, and motion, and project to the superior colliculus.

With about 120 million rods and cones and only about 1 million ganglion cells eventually nervating at the LGN, there is considerable convergence of photoreceptor output. That is, the signal of many (on the order of about 100) photoreceptors are combined to produce 1 type of signal. This intreconnecting arrangement is described in terms of receptive fields.

The receptive fields of ganglion cells are similar to those of bipolar cells (center-surround, on-center, off-center). Center-on and center-off receptive fields are depicted in Figure 6. Plus signs ($+$) denote illumination stimulus, minus signs ($-$) denote lack of stimulus. The vertical bars below each receptive field depict the firing response of the receptive field. This signal characteristic (series of "ticks") is usually obtained by inserting an electrode into the brain. The signal profile of receptive fields resembles the "Mexican hat" operator, often used in image processing.

B.3 The Optic Tract and the Magno- and Parvo-Cellular Visual Channels

Neural signals are transmitted from the retina to the occipital (visual) cortex through the optic tract, crossing in the optic chiasm, making connections to various cortical centers along the way. This physiological optic tract is often described functionally in terms of visual pathways, with reference to specific cells (e.g., ganglion cells). The optic tract is depicted in Figure 7. It is interesting to note the decussation (crossing) of the fibers from the nasal half of the retina at the optic chiasm, i.e., nasal retinal signals cross, temporal signals do not. The reference to the ciliary ganglion and the oculomotor nucleus in the midbrain relates to the pupillary light reflex (see [HW97]).

Along the optic tract, the Later Geniculate Nucleus (LGN) is of particular importance. Like other regions in the thalamus serving similar functions, the LGN is a cross-over point, or relay station, for α and β ganglion cells. The physiological organization of the LGN, with respect to nervations of these cells, produces a visual field topography of great clinical importance. Here, the Magno-Cellular and the Parvo-Cellular ganglionic projections are clearly visible (under microscope), forming junctions within two distinct layers of the LGN, correspondingly termed the M- and P-layers. Thalamic axons from the M- and P-layers of the LGN terminate in are V1 (the primary visual center) of the striate cortex within the occipital lobe.

The functional characteristics of ganglionic projections to the LGN and the corresponding Magno- and Parvo-cellular pathways are summarized in Table 1.

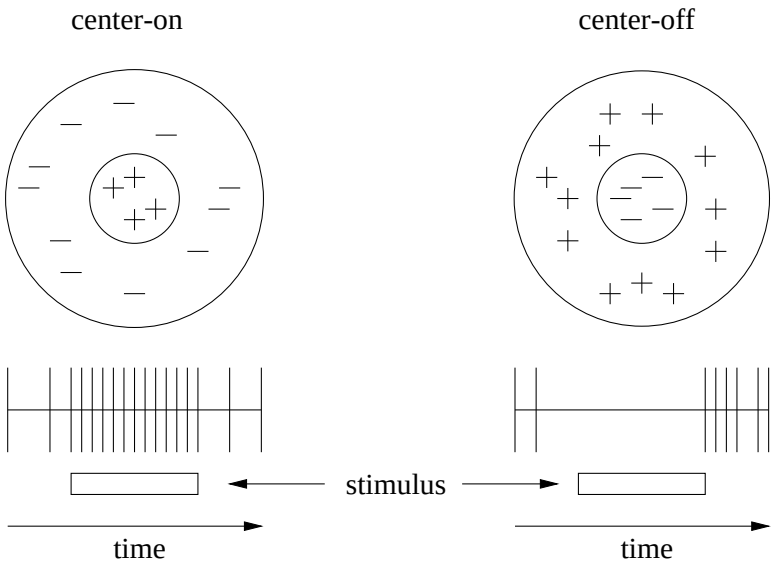


Figure 6: Schematic of receptive-fields.

ON, Oculomotor Nucleus; SC, Superior Colliculus; LGN, Lateral Geniculate Nucleus.

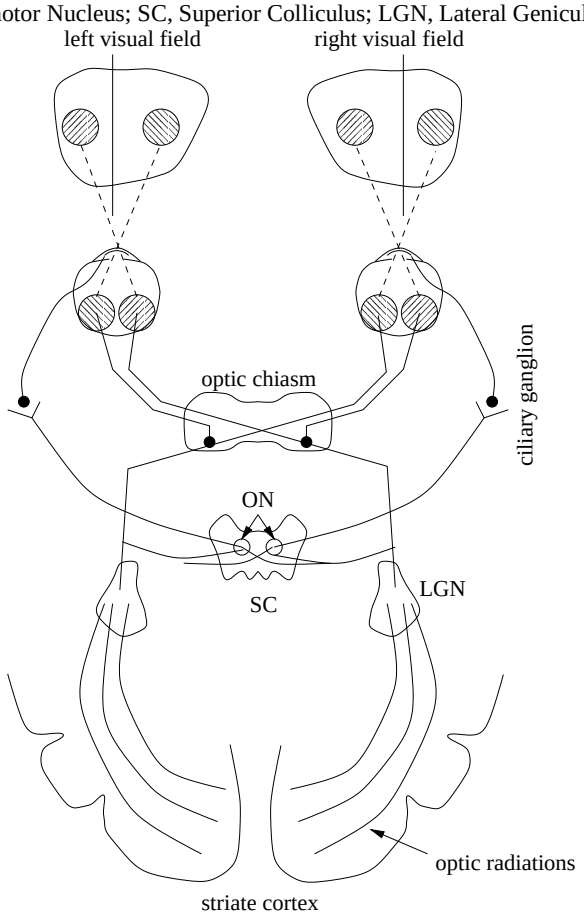


Figure 7: Schematic of the optic tract and radiations (visual pathways). Adapted from [HW97, p.11 (Fig. 1.8)].

Table 1: Functional characteristics of ganglionic projections.

Characteristics	Magnocellular	Parvocellular
ganglion size	large	small
transmission time	fast	slow
receptive fields	large	small
sensitivity to small objects	poor	good
sensitivity to change in light levels	large	small
sensitivity to contrast	low	high
sensitivity to motion	high	low
color discrimination	no	yes

B.4 The Occipital Cortex and Beyond

The cerebral cortex is composed of numerous regions classified by their function [Zek93]. A simplified representation of cortical regions is shown in Figure 8. The human visual system is functionally described by the connections between the retina and cortical regions, known as

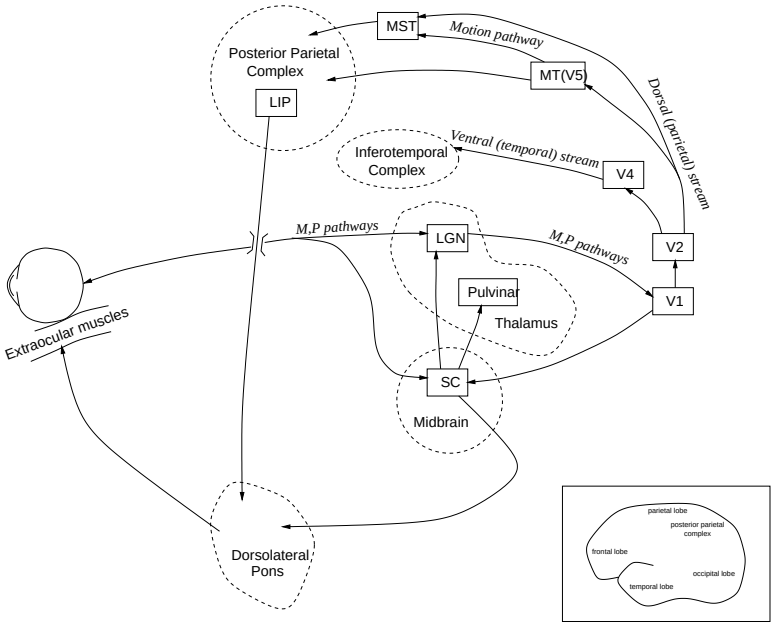


Figure 8: The brain and the visual pathways (with cortical lobe designations inset).

visual pathways. Pathways joining multiple cortical areas involved in common visual functions are referred to as streams. Figure 8 highlights regions and pathways relevant to selective visual attention. For clarity, many connections, particularly involving the Pulvinar, are omitted.

Thalamic axons from the M- and P-layers of the LGN terminate mainly in the lower and upper halves (β, α divisions, respectively) of layer 4C in middle depth of area V1 [LWL95]. Cell receptive field size and contrast sensitivity signatures are distinctly different in the M- and P-inputs of the LGN, and vary continuously through the depth of layer 4C. Unlike the center-surround receptive fields of retinal ganglion and LGN cells, cortical cells respond to orientation-specific stimulus [Hub88, pp.67-71]. Cortical cells are distinguished by two classes: *simple* and *complex*.

The size of a simple cell's receptive field depends on its retinal position, relative to the fovea. The smallest fields are in and near the fovea, with sizes of about $1/4 \times 1/4$ degree. This is about the size of the smallest diameters of the smallest receptive field centers of retinal ganglion or LGN cells. In the far periphery, simple cell receptive field sizes are about 1×1 degree. Simple cells fire only when a line or edge of preferred orientation falls within a particular location of the cell's receptive field. Complex cells fire wherever such a stimulus falls into the cell's receptive field [LWL95]. The optimum stimulus width for either cell type is, in the fovea, about 2 minutes of arc. The resolving power (acuity) of both cell types is the same.

About 10-20% of complex cells in the upper layers of the striate cortex show marked directional selectivity [Hub88, p.77]. Directional selectivity (DS) refers to the cell's response to a particular direction of movement. Cortical directional selectivity (CDS) contributes to motion perception and to the control of eye movements [GN95]. CDS cells establish a motion pathway from V1 projecting to MT and V2 (which also projects to MT) and to MST. In contrast, there is no evidence that retinal directional selectivity (RDS) contributes to motion perception. RDS contributes to oculomotor responses [GSA95]. In vertebrates, it is involved in optokinetic nystagmus, a type of eye movement discussed in Section D.

B.4.1 Significance of Motion-Sensitive Single-Cell Physiology for Perception

There are two somewhat counterintuitive implications of the visual system's motion-sensitive single-cell organization for perception. First, due to motion sensitive cells, fixations are never perfectly still but make constant tiny movements called *microsaccades* [Hub88, p.81]. These movements are more or less spatially random varying over 1 to 2 minutes of arc in amplitude. The counterintuitive fact regarding fixations is that if an image is artificially stabilized on the retina, vision fades away within about a second and the scene becomes blank. Second, due to the response characteristics of single (cortical) cells, the "retinal buffer" representation of natural images is much more abstract than what intuition suggests. An object in the visual field stimulates only a tiny fraction of the cells on whose receptive field it falls [Hub88, pp.85-87]. Perception of the object depends mostly on the response of (orientation-specific) cells to the object's borders. For example, the homogeneously shaded interior of an arbitrary form (e.g., a kidney bean) does not stimulate cells of the visual system. Awareness of the interior shade or hue depends on only cells sensitive to the borders of the object. In Hubel's words, "...our perception of the interior as black, white, gray, or green has nothing to do with cells whose fields are in the interior—hard as that may be to swallow...What happens at the borders is the only information you need to know: the interior is boring." [Hub88, p.87]

B.4.2 Summary of Important Cortical Regions

Of particular importance to visual perception are the following cortical regions, summarized in terms of relevance to attention:

- SC (Superior Colliculus): involved in programming eye movements; also remaps auditory space into visual coordinates (presumably for target foveation).
- Area V1 (primary visual cortex): detection of range of stimuli, e.g., color, motion, orientation.
- Areas V2, V3, V3A, V4, MT: higher-level vision (recognition).
- Area MT (Middle Temporal) and MST (Middle Superior Temporal): furnish large projections to Pons; hence possibly involved in smooth pursuit movements (area MT is thought to be implicated in motion processing).
- Area LIP (Lateral Intra-Parietal): contains receptive fields which are corrected ("reset") before execution of saccadic eye movements.
- PPC (Post Parietal Complex): involved in fixations.

Connections made to these areas from area V1 can be generally divided into two streams: the dorsal and ventral streams. Loosely, their functional description can be summarized as:

- Dorsal stream: sensorimotor (motion, location) processing (e.g., the attentional "where").
- Ventral stream: cognitive processing (e.g., the attentional "what").

In general attentional terms, the three main cortical regions and their functions are [Pal99]:

- Posterior Parietal Cortex (Lobe): disengages attention;
- SC: relocates attention;
- Pulvinar: engages, or enhances, attention.

B.5 Summary and Further Reading

Considerable information may be gleaned from the vast neuroscientific literature regarding the functionality (and limitations) of the Human Visual System. It is often possible to qualitatively predict observed psychophysical results by studying the underlying visual "hardware". For example, visual spatial acuity may be roughly estimated from knowledge of the distribution of retinal photoreceptors. Higher-level characteristics of human vision may also be estimated from the organization of further cortical regions. However, the reader must be cautioned against underestimating the complexity of the visual system which may be suggested by examining the neurological structure. For example, the apparent "what" and "where" dual pathways are most probably not independent functional channels. There is a good deal of interconnection and "cross-talk" between these and other related visual centers which deems the dichotomous analysis overly simplistic. Nevertheless, there is a great deal of invaluable information to be found in the neurological literature as human vision is undoubtedly the most studied human sense.

For an excellent introduction to neuroscience, see Hubel's text [Hub88]. For a more recent description of the cortex with an emphasis on color vision, see [Zek93]. Apart from these texts on vision, several "handbooks" have also been assembled describing current knowledge of the cortex. The field of neuroscience has advanced greatly in recent years, so these texts are rather voluminous (and expensive). Arbib's handbook is one such example [Arb95]. It is an excellent source summarizing current knowledge of the brain, although it is somewhat difficult to read and navigate through. The handbook comes with its own roadmap which facilitates the latter considerably.[1] Another such well organized but rather large text is Gazzaniga's [Gaz00].

C Visual Perception

Given the underlying physiological substrate of the Human Visual System, measurable performance parameters often (but not always!) fall within ranges predicted by the limitations of the neurological substrate. Visual performance parameters, such as visual acuity, are often measured following established experimental paradigms, generally derived in the field of psychophysics (e.g., Receiver Operating Characteristics, or ROC paradigm, is one of the more popular experimental methods).

[1] A new edition of Arbib's book has recently been announced.

Table 2: Common visual angles.

Object	Distance	Angle subtended
thumbnail	arm's length	1.5–2°
sun or moon	–	.5° or 30′ or arc
US quarter coin	arm's length	2°
US quarter coin	85m	1′ (1 minute of arc)
US quarter coin	5km	1″ (1 second of arc)

Unexpected observed performance results are often results of complex visual processes (e.g., visual illusions), or combinations of several factors. For example, the well-known Contrast Sensitivity Function, or CSF, describing the Human Visual System's response to stimuli of varying contrast and resolution, depends not only on the organization of the retinal mosaic, but also on the response characteristics of complex cellular combinations, e.g., receptive fields.

In these notes, the primary concern is visual attention, and so the notes primarily consider the distinction between foveo-peripheral vision. This subject, while complex, is discussed here in a fairly simplified manner, with the aim of illucidating only the most dramatic differences between what is perceived foveally and peripherally. In particular, visual (spatial) acuity is arguably the most studied distinction and is possibly the simplest parameter to alter in eye-based interaction systems (at least at this time). It is therefore the topic covered in greatest detail, in comparison to the other distinctions covered here briefly: temporal and chromatic foveo-peripheral differences.

C.1 Spatial Vision

Dimensions of retinal features are usually described in terms of projected scene dimensions in units of degrees visual angle, defined as

$$A = 2\arctan\frac{S}{2D},$$

where S is the size of the scene object and D is the distance to the object (see Figure 9). Common visual angles are given in Table 2.

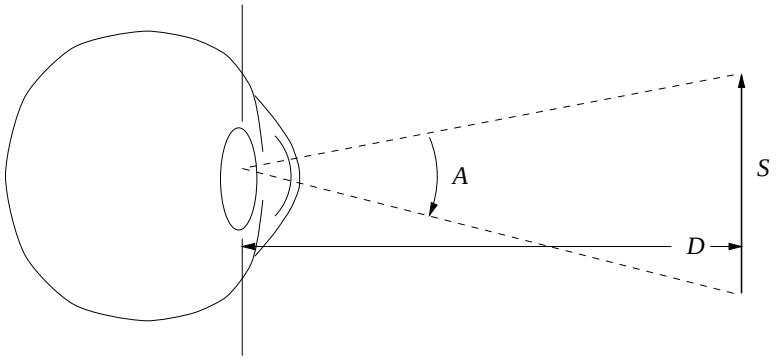
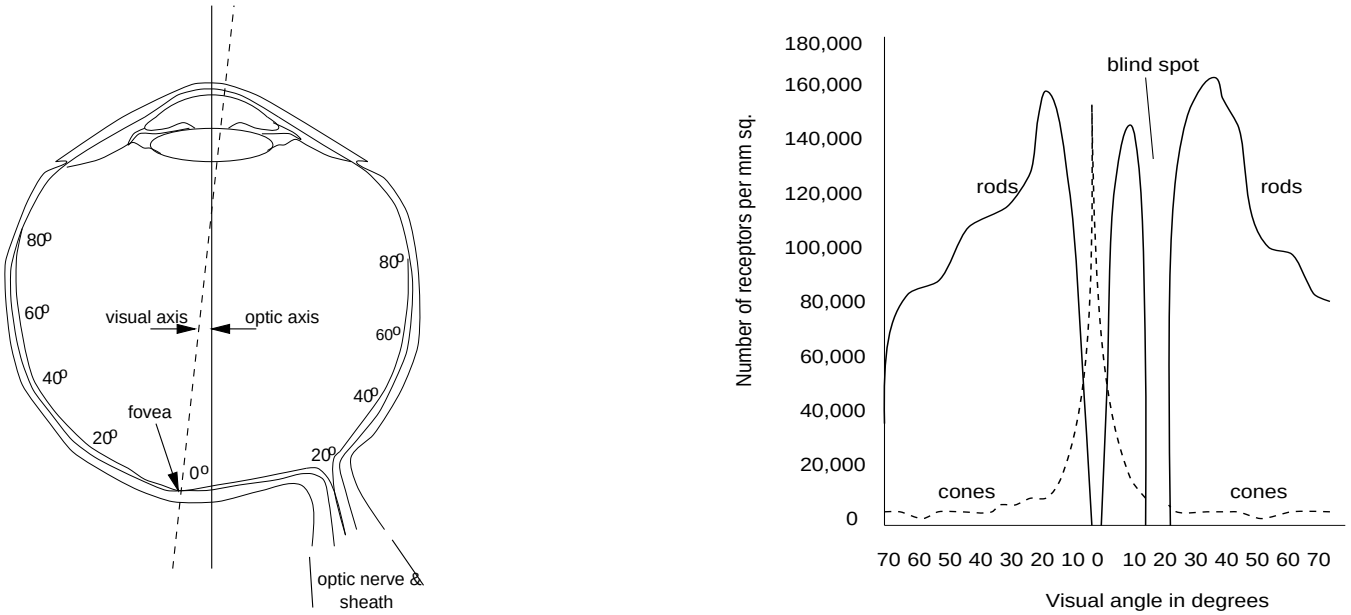


Figure 9: Visual angle. Adapted from [HH73, p.15 (Fig. 2.7)].

The innermost region is the fovea centralis (or foveola) which measures 400μm in diameter and contains 25,000 cones. The fovea proper measures 1500μm in diameter and holds 100,000 cones. The macula (or central retina) is 5000μm in diameter, and contains 650,000 cones. One degree visual angle corresponds to approximately 300μm distance on the human retina [DD88, p.48]. The foveola, measuring 400μm subtends 1.3° visual angle, while the fovea and macula subtend 5° and 16.7°, respectively (see Figure 10(a)). Figure 10(b) shows the retinal distribution of rod and cone receptors. The fovea contains 147,000 cones/mm^2 and a slightly smaller number of rods. At about 10° the number of cones drops sharply to less than 20,000 cones/mm^2 while at 30° the number of rods in the periphery drops to about 100,000 rods/mm^2 [HH73].

The entire visual field roughly corresponds to a 23,400 square degree area defined by an ellipsoid with the horizontal major axis subtending 180° visual angle, and the minor vertical axis subtending 130°. The diameter of the highest acuity circular region subtends 2°, the parafovea (zone of high acuity) extends to about 4° or 5°, and acuity drops off sharply beyond. At 5°, acuity is only 50% [Irw92]. The so-called "useful" visual field extends to about 30°. The rest of the visual field has very poor resolvable power and is mostly used for perception of ambient motion. With increasing eccentricity the cones increase in size, while the rods do not [DD88]. Cones, not rods, make the largest contribution to the information going to deeper brain centers, and provide most of the fine-grained spatial resolvability of the visual system.

The Modulation Transfer Function (MTF) theoretically describes the spatial resolvability of retinal photoreceptors by considering the cells as a finite array of sampling units. The 400μm-diameter rod-free foveola contains 25,000 cones. Using the area of a circle, $25000 = \pi r^2$, approximately $2\sqrt{25000/\pi} = 178.41$ cones occupy a 400μm linear cross-section of the foveola with an estimated average linear inter-cone spacing of 2.24μm. Cones in this region measure about 1μm in diameter. Since one degree visual angle corresponds to approximately 300μm distance on the human retina, roughly 133 cones are packed per degree visual angle in the foveola. By the sampling theorem, this suggests



(a) Degrees visual angle.

(b) Approximate rod/cone density.

Figure 10: Retinotopic receptor distribution. Adapted from [HH73, p.25 (Fig. 2.16)].

a resolvable spatial Nyquist frequency of 66 c/deg. Subjective resolution has in fact been measured at about 60 c/deg [DD88, pp.46-53]. In the fovea, a similar estimate based on the foveal diameter of 1500μm and a 100,000 cone population, gives an approximate linear cone distribution of $2\sqrt{100000/\pi} = 356.82$ cones per 1500μm. The average linear inter-cone spacing is then 71 cones/deg suggesting a maximum resolvable frequency of 35 cycles/deg, roughly half the resolvability within the foveola. This is somewhat of an underestimate since cone diameters increase two-fold by the edge of the fovea suggesting a slightly milder acuity degradation. These one-dimensional approximations are not fully generalizable to the two-dimensional photoreceptor array although they provide insight into the theoretic resolution limits of the eye. Effective relative visual acuity measures are usually obtained through psychophysical experimentation.

At photopic light levels (day, or cone vision), foveal acuity is fairly constant within the central 2°, and drops approximately linearly from there to the 5° foveal border. Beyond the 5°, acuity drops sharply (approximately exponentially). At scotopic light levels (night, or rod-vision), acuity is poor at all eccentricities. Figure 11 shows the variation of visual acuity at various eccentricities and light intensity levels. Intensity is shown varying from 9.0 to 4.6 log micromicrolamberts, denoted by $\log \mathrm{mmL}$ (9.0 log micromicrolamberts = 10^9 micromicrolamberts = 1 mL, see [Dav80, p.311]). The correspondence between foveal receptor spacing and optical limits generally holds in foveal regions of the retina, but not necessarily in the periphery. In contrast to the approximate 60 c/deg resolvability of foveal cones, the highest spatial frequencies resolvable by rods are on the order of 5 c/deg, suggesting poor resolvability in the relatively cone-free periphery. Although visual acuity correlates fairly well with cone distribution density, it is important to note that synaptic organization and later neural elements (e.g., ganglion cells concentrated in the central retina) are also contributing factors in determining visual acuity.

C.2 Temporal Vision

Human visual response to motion is characterized by two distinct facts: the *persistence of vision* and the *phi phenomenon* [Gre90]. The former essentially describes the temporal sampling rate of the HVS, while the latter describes a threshold above which the HVS detects *apparent movement*. Both facts are exploited in television, cinema, and graphics to elicit perception of motion from successively displayed still images.

Persistence of vision describes the inability of the retina to sample rapidly changing intensities. A stimulus flashing at about 50-60Hz (cycles per second) will appear steady (depending on contrast and luminance conditions and observers). This is known as the Critical Fusion Frequency (CFF).[2] A stylized representation of the CFF, based on measurements of response to temporal stimuli of varying contrast, i.e., a temporal contrast sensitivity function, is shown in Figure 12. Incidentally, the curve of the CFF resembles the shape of the curve of the Contrast Sensitivity Function (CSF) which describes retinal spatial frequency response. The CFF explains why flicker is not seen when viewing a sequence of (still) images at a high enough rate. The CFF illusion is maintained in cinema since frames are shown at 24 frames per second (fps, equivalent to Hz), but a three-bladed shutter raises the flicker rate to 72Hz (three for each picture). Television also achieves the CFF by displaying the signal at 60 fields per second. Television's analog to cinema's three-bladed shutter is the interlacing scheme: the NTSC television frame rate is 30 frames per second, but only the even or odd scanlines (fields) are shown per cycle. Although the CFF explains why flicker is effectively eliminated in motion picture (and computer) displays, it does not fully explain why motion is perceived.

The second fact that explains why movies, television, and graphics work is the phi phenomenon, or *stroboscopic motion*, or *apparent motion*.

[2] Also sometimes referred to as the Critical Flicker Frequency.

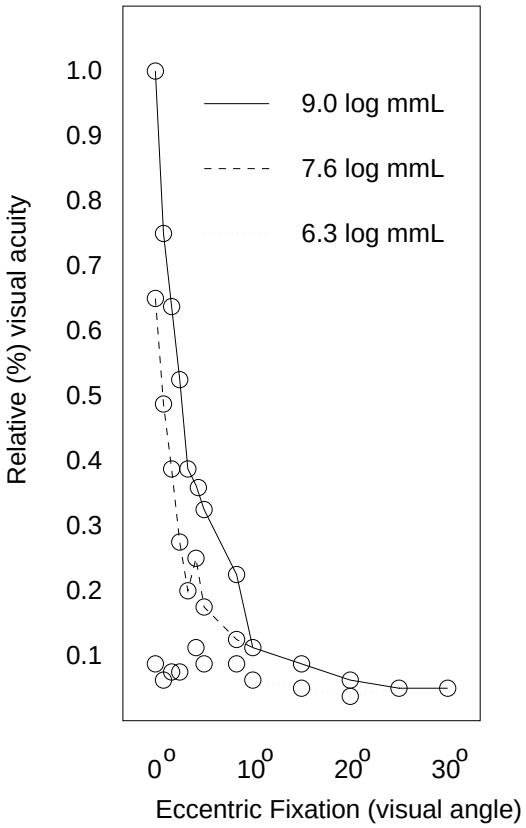


Figure 11: Visual acuity at various eccentricities and light levels. Adapted from [Dav80, p.311 (Fig. 13.1)].

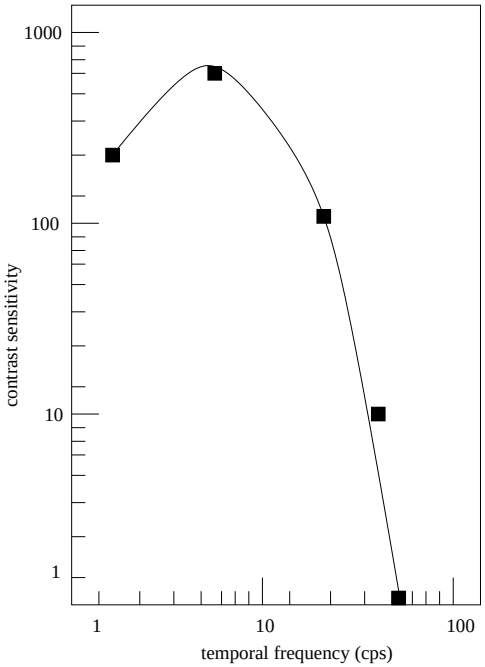


Figure 12: Critical Fusion Frequency. Adapted from [Bas95, p.25.29 (Fig. 14)].

This fact explains the illusion of old-fashioned moving neon signs whose stationary lights are turned on in quick succession. This illusion can also be demonstrated with just two lights, provided the delay between successive light flashes is no than about 62Hz [Bri99]. Inverting this value gives a rate of about 16fps which is considered a bare minimum to facilitate the illusion of apparent motion.

C.2.1 Perception of Motion in the Visual Periphery

In the context of visual attention and foveo-peripheral vision, the temporal response of the HVS is not homogeneous across the visual field. Sensitivity to target motion decreases monotonically with retinal eccentricity for slow and very slow motion (1 cycle/deg) [BL88b]. That is, the velocity of a moving target appears slower in the periphery than in the fovea. Conversely, a higher rate of motion (e.g., frequency of rotation of grated disk) is needed in the periphery to match the apparent stimulus velocity in the fovea. At higher velocities, the effect is reversed.

Despite the decreased sensitivity in the periphery, movement is more salient there than in the central field of view (fovea). That is, the periphery is more sensitive to moving targets than to stationary ones. It is easier to peripherally detect a moving target than it is a stationary one. In essence, motion detection is the periphery's major task; it is a kind of early warning system for moving targets entering the visual field.

C.2.2 Sensitivity to Direction of Motion in the Visual Periphery

The periphery is approximately twice as sensitive to horizontal-axis movement as to vertical-axis movement [BL88b]. Directional motion sensitivity is show in Figure 13.

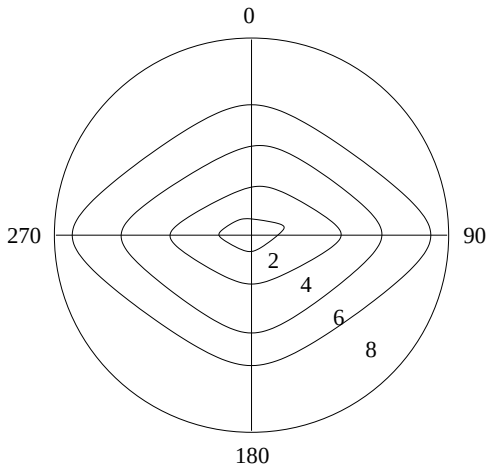


Figure 13: Absolute threshold isograms for detecting rotary movement in the periphery. Numbers are rates of pointer movement in revolutions per minute. Adapted from [BL88b, Section 5.206, p.922 (Fig. 1)].

C.3 Color Vision

Foveal color vision if facilitated by the three types of retinal cone photoreceptors. Spectral sensitivity curves for cone receptors is sketched in Figure 14. A great deal is known about color vision in the fovea, however, relatively little is known about peripheral color vision. Of the 7 million cones, most are packed tightly into the central 30° region of the fovea with scarcely any cones found beyond. This cone distribution suggests that peripheral color vision is quite poor in comparison to the color sensitivity of the central retinal region. Visual fields for monocular color vision are shown in Figure 15. Fields are shown for the right eye; fields for the left eye would be mirror images of those for the right eye. Blue and yellow fields are larger than the red and green fields; no chromatic visual fields have definite border, instead, sensitivity drops off gradually and irregularly over a range of 15-30° visual angle [BL88b].

Quantification of perceptual performance is not easily found in the literature. Compared to investigation of foveal color vision, only a few experiments have been performed to measure peripheral color sensitivity. Two studies, of particular relevance to peripheral location of color CRTs in an aircraft cockpit environment, investigated the chromatic discrimination of peripheral targets.

In the first study, Doyal concludes that peripheral color discrimination can approximate foveal discrimination when relatively small field sizes are presented (e.g., 2° at 10° eccentricity, and less than 4° at 25°) [Doy91]. While this sounds encouraging, color discrimination was tested at limited peripheral eccentricities (within the central 30°).

In the second, Ancman tested color discrimination at much greater eccentricities, up to about 80° visual angle. She found that subjects wrongly identified the color of a peripherally located, 1.3° circle displayed on a CRT 5% of the time if it was blue, 63% of the time if red, and 62% of the time if green [Anc91]. Furthermore, blue could not be seen further than 83.1°off the fovea (along the *x*-axis); red had to be closer than 76.3° and green nearer than 74.3° before subjects could identify the color.

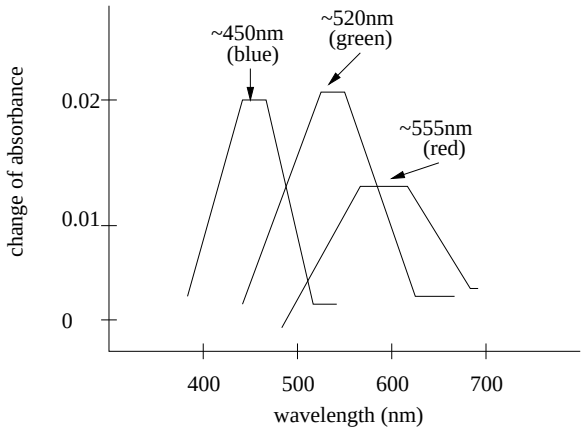


Figure 14: Approximate spectral sensitivity curves for the three types of cone photoreceptors. Adapted from [HW97, p.8 (Fig. 1.5)].

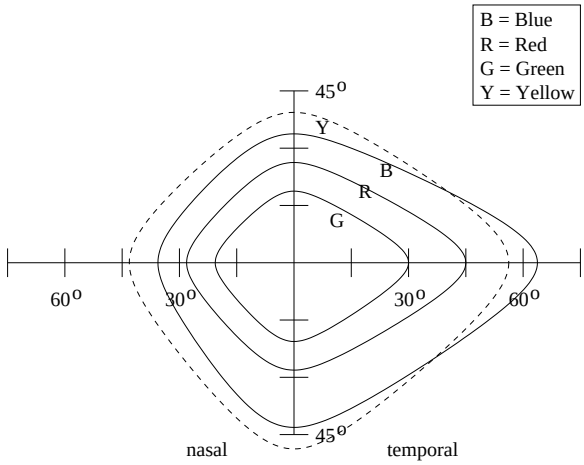


Figure 15: Visual fields for monocular color vision (right eye). Adapted from [BL88b, Section 1.237, p.108 (Fig. 1)].

There is much yet to be learned about peripheral color vision. Being able to verify a subject's direction of gaze during peripheral testing would be of significant benefit to these experiments. This type of psychophysical testing is but one of several research areas where eye tracking studies could play an important supporting role.

C.4 Implications for Attentional Design of Visual Displays

Both the structure and functionality of human visual system components place constraints on the design parameters of a visual communication system. In particular, the design of a gaze-contingent system must distinguish the characteristics of foveal and peripheral vision.

The parvocellular pathway in general responds to signals possessing the following attributes: high contrast (the parvocellular pathway is less sensitive to luminance), chromaticity, low temporal frequency, and high spatial frequency (due to the small receptive fields). Conversely, the magnocellular pathway can be characterized by sensitivity to the following signals: low contrast (the magnocellular pathway is more sensitive to luminance), achromaticity, moderate-to-high temporal frequency (sudden onset stimuli), and low spatial frequency (due to the large receptive fields). In terms of motion responsiveness, Koenderink et al. provide support that the foveal region is more receptive to slower motion than the periphery, although motion is perceived uniformly across the visual field [KDG85].

M and P ganglion cells in the retina connect to M and P channels, respectively. Zeki suggests the existence of four functional pathways defined by the M and P channels [Zek93]: motion, dynamic form, color, and form (size and shape). Furthermore, it is thought that fibers reaching the superior colliculus represent retinal receptive fields in rod-rich peripheral zones, while the fibers reaching the LGN represent cone-rich areas of high acuity [BL88a]. It seems likely that the M ganglion cells correspond to rods, mainly found in the periphery, and the P cells correspond to cones, which are chromatic cells concentrated mainly in the foveal region. A *visuotopic* representation model for imagery based on these observations is proposed:

1. **Spatial Resolution** should remain high within the foveal region and smoothly degrade within the peripheral, matching human visual acuity. High spatial frequency features in the periphery must be made visible "just in time" to anticipate gaze-contingent fixation changes.
2. **Temporal Resolution** must be available in the periphery. Sudden onset events are potential attentional attractors. At low speeds, motion of peripheral targets should be increased to match apparent motion in the central field of view.
3. **Luminance** should be coded for high visibility in the peripheral areas since the periphery is sensitive to dim objects.
4. **Chrominance** should be coded for high exposure almost exclusively in the foveal region, with chromaticity decreasing sharply into the periphery. This requirement is a direct consequence of the high density of cones and parvocellular ganglion cells in the fovea.
5. **Contrast** sensitivity should be high in the periphery, corresponding to the sensitivity of the magnocellular ganglion cells found mainly outside the fovea.

Special consideration should be given to sudden onset, luminous, high frequency objects (i.e., suddenly appearing bright edges).

A gaze-contingent visual system faces an implementational difficulty not yet addressed: matching the dynamics of human eye movement. Any system designed to incorporate an eye-slaved high resolution of interest, for example, must deal with the inherent delay imposed by the processing required to track and process real-time eye tracking data. To consider the temporal constraints that need to be met by such systems, the dynamics of human eye movements must be evaluated.

C.5 Summary and Further Reading

Psychophysical information may be the most usable form of literature for the design of graphical displays, attentional in nature or otherwise. Some introductorary texts may include function plots of some aspect of vision (e.g., acuity) which may readily be used to guide the design of visual displays. The reader should be cautioned, however, in that one needs to read and evaluate this kind of information carefully. Most psychophysical data is based on some kind of empirical study. One needs to evaluate the experimental design used in the study to determine the generalizability of reported results. Furthermore, similar caution should be employed as in reading neurological literature: psychophysical results may often deal with a certain specific aspect of vision, which may or may not be readily applicable to display design. For example, visual acuity may suggest the use of relatively sized fonts on a web page (larger font in the periphery), but acuity alone may not be sufficient to determine the required resolution in something like an attentional image or video display program. For the latter, one may need to piece together information concerning the visual contrast sensitivity function, temporal sensitivity, etc. Furthermore, psychophysical studies may involve relatively simple stimuli (sine wave gratings), the results of which may or may not generalize to more complex stimuli such as imagery.

For a good introductory book on visual perception, see [HW97]. This text includes a good introductory chapter on the neurological basis of vision. Another good introductory book which also includes an interesting perspective on the perception of art is [Sol99]. For a somewhat terse but fairly complete psychophysical reference, see the USAF Engineering Data Compendium [BL88b]. This is an excellent, "quick" guide to visual performance.

D Taxonomy and Models of Eye Movements

Almost all normal primate eye movements used to reposition the fovea result as combinations of five basic types: saccadic, smooth pursuit, vergence, vestibular, and physiological nystagmus (miniature movements associated with fixations) [Rob68, p.1033]. Vergence movements are used to focus the pair of eyes over a distant target (depth perception). Other movements such as adaptation and accommodation refer to non-positional aspects of eye movements (i.e., pupil dilation, lens focusing). With respect to visual display design, positional eye movements are of primary importance.

D.1 The Extra-Ocular Muscles, and The Oculomotor Plant

In general, the eyes move within six degrees of freedom: three translations within the socket, and three rotations. There are six muscles responsible for movement of the eyeball: the *medial* and *lateral recti* (sideways movements), the *superior* and *inferior recti* (up/down movements), and the *superior* and *inferior obliques* (twist) [Dav80]. These are shown in Figure 16. The neural system involved in generating eye

Left (view from above): 1, superior rectus; 2, levator palbebrae superioris; 3, lateral rectus; 4, medial rectus; 5, superior oblique; 6, reflected tendon of the superior oblique; 7, annulus of Zinn. *Right (lateral view):* 8, inferior rectus; 9, inferior oblique.

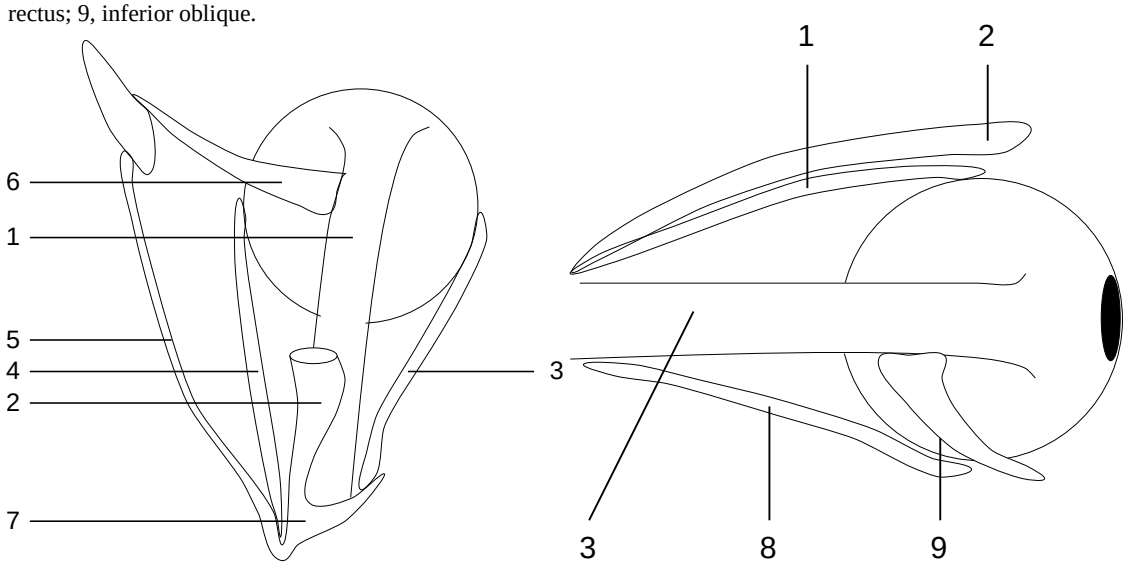


Figure 16: Extrinsic muscles of the eye. Adapted from [Dav80, p.385 (Fig. 16.2), p.386 (Fig. 16.3)].

movements is known as the oculomotor plant. The general plant structure and connections are shown in Figure 17 and described in [Rob68]. Eye movement control signals emanate from several functionally distinct regions. Areas 17, 18, 19, and 22 are areas in the occipital cortex thought to be responsible for high-level visual functions such as recognition. The superior colliculus bears afferents emanating directly from the retina, particularly from peripheral regions conveyed through the magno-cellular pathway. The semicircular canals react to head movements in three-dimensional space. All three areas, i.e., the occipital cortex, the superior colliculus, and the semicircular canals convey efferents to the eye muscles through the mesencephalic and pontine reticular formations. Classification of observed eye movement signals relies in part on the known functional characteristics of these cortical regions.

Two pertinent observations regarding eye movements can be drawn from the oculomotor plant's organization:

1. The eye movement system is, to a large extent, a feedback circuit.
2. Signals controlling eye movement emanate from cortical regions which can be functionally categorized as voluntary (occipital cortex), involuntary (superior colliculus), and reflexive (semicircular canals).

The feedback-like circuitry is utilized mainly in the types of eye movements requiring stabilization of the eye. Orbital equilibrium is necessitated for the steady retinal projection of an object, concomitant with the object's motion and movements of the head. Stability is maintained by a neuronal control system.

D.2 Saccades

Saccades are rapid eye movements used in repositioning the fovea to a new location in the visual environment. The term comes from an old French word meaning "flick of a sail" [Gre90, p.64]. Saccadic movements are both voluntary and reflexive. The movements can be voluntarily executed or they can be invoked as a corrective optokinetic or vestibular measure (see below). Saccades range in duration from 10ms to 100ms, which is a sufficiently short duration to render the executor effectively blind during the transition [SF83]. There is some debate over the underlying neuronal system driving saccades. On the one hand, saccades have been deemed ballistic and stereotyped. The term stereotyped refers to the observation that particular movement patterns can be evoked repeatedly. The term ballistic refers to the presumption that saccade destinations are pre-programmed. That is, once the saccadic movement to the next desired fixation location has been calculated (programming latencies of about 200ms have been reported), saccades cannot be altered. One reason behind this presumption is that, during saccade execution, there is insufficient time for visual feedback to guide the eye to its final position [Car77, p.57]. One the other hand, a saccadic feedback system is plausible if it is assumed that instead of visual feedback, an internal copy of head, eye, and target position is used to guide the eyes during a saccade [LR86, FKS85]. Due to their fast velocities, saccades may only appear to be ballistic [ZOCR+76, p.251].

Various models for saccadic programming have been proposed [Fin92]. These models, with the exception of ones including "center-of-gravity" coding (see for example [HK89]), may inadequately predict unchangeable saccade paths. Instead, saccadic feedback systems based

CBT, corticobular tract; CER, cerebellum; ICTT, internal corticotectal tract; LG, lateral geniculate body; MLF, medial longitudinal fasciculus; MRF, mesencephalic and pontine reticular formations; PT, pretectal nuclei; SA, stretch afferents from extraocular muscles; SC, superior colliculi; SCC, semicircular canals; T, tegmental nuclei; VN, vestibular nuclei; II, optic nerve; III, IV, and VI, the oculomotor, trochlear, and abducens nuclei and nerves; 17, 18, 19, 22, primary and association visual areas, occipital and parietal (Brodmann); 8, the frontal eye fields.

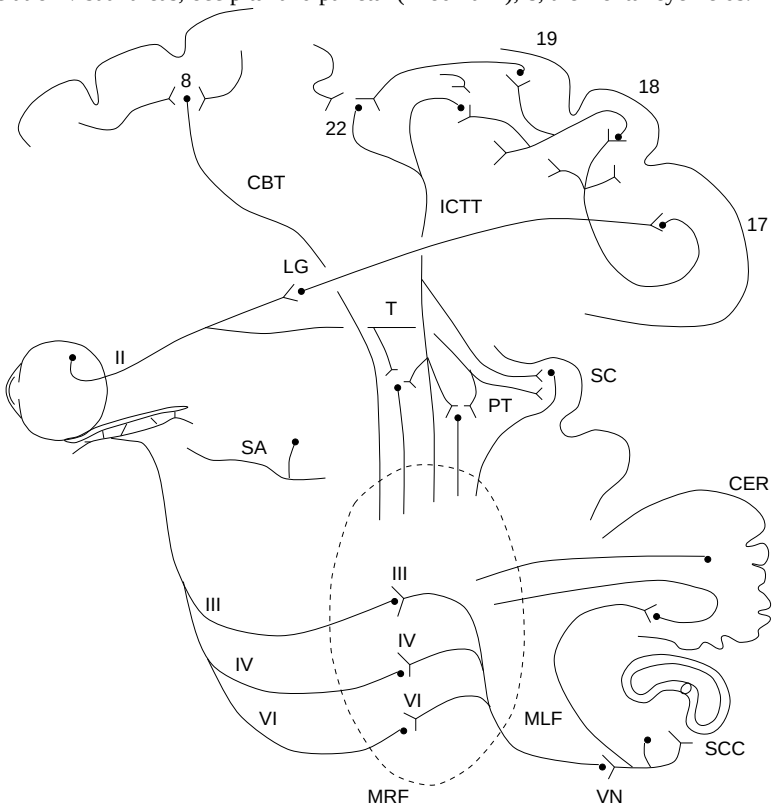


Figure 17: Schematic of the major known elements of the oculomotor system. Adapted from [Rob68, p.1035 (Fig. 2)].

on an internal representation of target position may be more plausible since they tend to correctly predict the so-called double-step experimental paradigm. The double-step paradigm is an experiment where target position is changed during a saccade in mid-flight. Scudder et al. proposed a refinement of Robinson's feedback model which is based on a signal provided by the superior colliculus and a local feedback loop [FKS85]. The local loop generates feedback in the form of motor error produced by subtracting eye position from a mental target-in-space position. Sparks and Mays cite compelling evidence that intermediate and deep layers of the SC contain neurons that are critical components of the neural circuitry initiating and controlling saccadic movements [SM90]. These layers of the SC receive inputs from cortical regions involved in the analysis of sensory (visual, auditory, and somatosensory) signals used to guide saccades. The authors also rely on implications of Listing's and Donder's Laws which specify an essentially null torsion component in eye movements, requiring virtually only two degrees of freedom for saccadic eye motions [Dav80, SM90]. According to these laws, motions can be resolved into rotations about the horizontal *x*- and vertical *y*-axes.

Models of saccadic generation attempt to provide an explanation of the underlying mechanism responsible for generating the signals sent to the motor neurons. Although there is some debate as to the source of the saccadic program, the observed signal resembles a pulse/step function [SM90, p.315]. The pulse/step function refers to a dual velocity and position command to the extraocular muscles [LZ91, p.180]. A possible simple representation of a saccadic step signal is a differentiation filter. Carpenter suggests such a possible filter arrangement for generating saccades coupled with an integrator [Car77, p.288]. The integrating filter is in place to model the necessary conversion of velocity-coded information to position-coded signals [LZ91, p.182]. A perfect neural integrator converts a pulse signal to a step function. An imperfect integrator (called leaky) will generate a signal resembling a decaying exponential function. The principle of this type of neural integration applies to all types of conjugate eye movements. Neural circuits connecting structures in the brain stem and the cerebellum exist to perform integration of coupled eye movements including saccades, smooth pursuits, and vestibular and optokinetic nystagmus (see below) [LZ91, p.183].

A differentiation filter can be modeled by a linear moving average filter as shown in Figure 18. In the time domain, the moving average filter

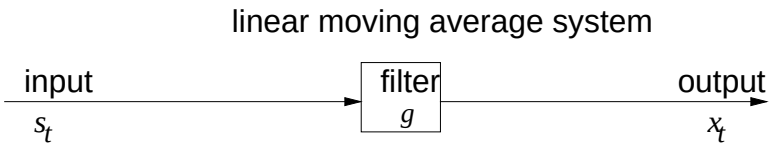


Figure 18: Block diagram of a simple linear moving average system modeling saccadic movements.

is modeled by the following equation

$$\begin{aligned} x_t &= g_0 s_t + g_1 s_{t-1} + \cdots \\ &= \sum_{k=0}^{\infty} g_k s_{t-k}, \end{aligned}$$

where s_t is the input (pulse), x_t is the output (step), and b_k are the moving average filter coefficients. To ensure differentiation, the filter coefficients typically must satisfy properties which approximate mathematical differentiation. An example of such a filter is the Haar filter with coefficients $\{1,-1\}$. Under the *z*-transform the transfer function $X(z)/S(z)$ of this linear filter is

$$\begin{aligned} x_t &= g_0 s_t + g_1 s_{t-1} \\ x_t &= (1)s_t + (-1)s_{t-1} \\ x_t &= (1)s_t + (-1)z s_t \\ x_t &= (1-z)s_t \\ X(z) &= (1-z)S(z) \\ \frac{X(z)}{S(z)} &= 1-z. \end{aligned}$$

The Haar filter is a length-2 filter which approximates the first derivate between successive pairs of inputs.

D.3 Smooth Pursuits

Pursuit movements are involved when visually tracking a moving target. Depending on the range of target motion, the eyes are capable of matching the velocity of the moving target. Pursuit movements provide an example of a control system with built-in negative feedback [Car77, p.41]. A simple closed-loop feedback loop used to model pursuit movements is shown in Figure 19, where s_t is the target position, x_t is the (desired) eye position, and h is the (linear, time-invariant) filter, or gain of the system [Car77, LZ91]. Tracing the loop from the feedback start point gives the following equation in the time domain

$$h(s_t - x_t) = x_{t+1}.$$

Under the *z*-transform the transfer function $X(z)/S(z)$ of this linear system is

$$H(z)(S(z) - X(z)) = X(z)$$

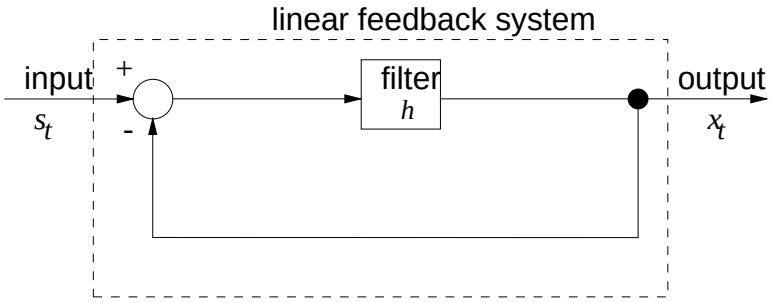


Figure 19: Block diagram of a simple linear feedback system modeling smooth pursuit movements.

$$\begin{aligned} H(z)S(z) &= X(z)(1+H(z)) \\ \frac{H(z)}{1+H(z)} &= \frac{X(z)}{S(z)}. \end{aligned}$$

Signals from visual receptors constitute the error signal indicating needed compensation to match the target's retinal image motion.

D.4 Fixations

Fixations are eye movements which stabilize the retina over a stationary object of interest. It seems intuitive that fixations should be generated by the same neuronal circuit controlling smooth pursuits with fixations being a special case of a target moving at zero velocity. This is probably incorrect [LZ91, pp.139-140]. Fixations, instead, are characterized by the miniature eye movements: tremor, drift, and microsaccades. These eye movements are considered noise present in the control system (possibly distinct from the smooth pursuit circuit) attempting to hold gaze steady. This noise appears as a random fluctuation about the area of fixation, typically no larger than 5° visual angle [Car77, p.105]. Although the classification of miniature movements as noise may be an oversimplification of the underlying natural process, it allows the signal to be modeled by a feedback system similar to the one shown in Figure 19. The additive noise in Figure 19 is represented by $e_t = s_t - x_t$, where the (desired) eye position x_t is subtracted from the steady fixation position s_t at the summing junction. In this model, the error signal stimulates the fixation system in a manner similar to the smooth pursuit system, except that here e_t is an error-position signal instead of an error-velocity signal (see [LZ91, p.150]). The feedback system modeling fixations, using the noisy "data reduction" method, is in fact simpler than the pursuit model since it implicitly assumes a stationary stochastic process [Car77, p.107]. Stationarity in the statistical sense refers to a process with constant mean. Other relevant statistical measures of fixations include their duration range of 150ms to 600ms, and the observation that 90% of viewing time is devoted to fixations [Irw92].

D.5 Nystagmus

Nystagmus eye movements are conjugate eye movements characterized by a sawtooth-like time course (time series signal) pattern. Optokinetic nystagmus is a smooth pursuit movement interspersed with saccades invoked to compensate for the retinal movement of the target. The smooth pursuit component of optokinetic nystagmus appears in the slow phase of the signal [Rob68]. Vestibular nystagmus is a similar type of eye movement compensating for the movement of the head. The time course of vestibular nystagmus is virtually indistinguishable from its optokinetic counterpart [Car77].

D.6 Eye Movement Analysis

From the above discussion, two significant observations relevant to eye movement analysis can be made. First, based on the functionality of eye movements, only three types of movements need be modeled to gain insight into the overt localization of visual attention. These types of eye movements are fixations, smooth pursuits, and saccades. Second, based on signal characteristics and plausible underlying neural circuitry, all three types of eye movements may be approximated by a linear, time-invariant (LTI) system, i.e., a linear filter.

The primary requirement of eye movement analysis, in the context of gaze-contingent system design, is the identification of fixations, saccades, and smooth pursuits. It is assumed that these movements provide evidence of voluntary, overt visual attention. This assumption does not preclude the plausible involuntary utility of these movements, or conversely, the covert non-use of these eye movements (e.g., as in the case of parafoveal attention). Fixations naturally correspond to the desire to maintain one's gaze on an object of interest. Similarly, pursuits are used in the same manner for objects in smooth motion. Saccades are considered manifestations of the desire to voluntarily change the focus of attention.

Eye movement signals can be approximated by linear filters. Fixations and pursuits are driven by a relatively simple neuronal feedback system. In the case of fixations, the neuronal control system is responsible for minimizing fixation error. For pursuit movements, the error is similarly measured as distance off the target, but in this case the target is non-stationary. Fixations and pursuits may be detected by a simple linear model based on linear summation.

The linear approach to eye movement modeling is an operational simplification of the underlying nonlinear natural processes [Car77, p.44]. The linear model assumes that position and velocity is processed by the same neuronal mechanism. The visual system processes these quantities in different ways. The position of a target is signaled by the activation of specific retinal receptors. The velocity of the target, on the other hand, is registered by the firing rate (amplitude) of the firing receptors. Furthermore, nonlinearities are expected in most types of eye movements. Accelerational and decelerational considerations alone suggest the inadequacy of the linear assumption. Nevertheless, from a signal processing standpoint, linear filter analysis is sufficient for the localization of distinct features in eye movement signals. Although this approach is a poor estimate of the underlying system, it nonetheless establishes a useful approximation of the signal in the sense of pattern recognition.

The goal of eye movement signal analysis is to characterize the signal in terms of salient eye movements, i.e., saccades and fixations (and possibly smooth pursuits). Typically, the analysis task is to locate regions where the signal average changes abruptly indicating the end of a fixation and the onset of a saccade and then again assumes a stationary characteristic indicating the beginning of a new fixation. A hypothetical plot of an eye movement time course is shown in Figure 20. The graph shows the sought points in the signal where a saccade begins and

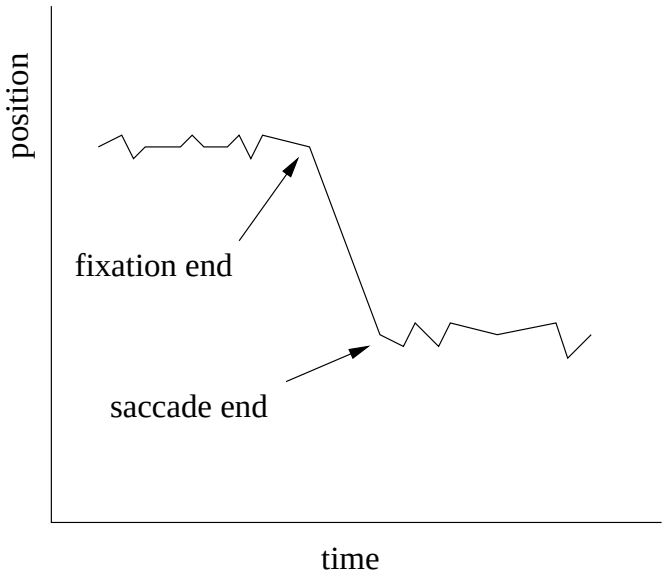


Figure 20: Hypothetical eye movement signal.

ends. Essentially, saccades can be thought of signal edges in time.

Two main automatic types of approaches have been used to analyze eye movements: one based on summation (averaging), the other on differentiation.[3] In the first, the temporal signal is averaged over time. If little or no variance is found, the signal is deemed a candidate for fixation. Furthermore, the signal is classified as a fixation, provided the duration of the stationary signal exceeds a predetermined threshold. This is the "dwell-time" method of fixation determination. In the second, assuming the eye movement signal is recorded at a uniform sampling rate, successive samples are subtracted to estimate eye movement velocity. The latter type of analysis is gaining favor, and appears more suitable for real-time detection of saccades. Fixations are either implicitly detected as the portion of the signal between saccades, or the portion of the signal where the velocity falls below a threshold.

Thresholds for both summation and differentiation methods are typically obtained from empirical measurements. The seminal work of Yarbus is often still referenced as the source of these measurements [Yar67].

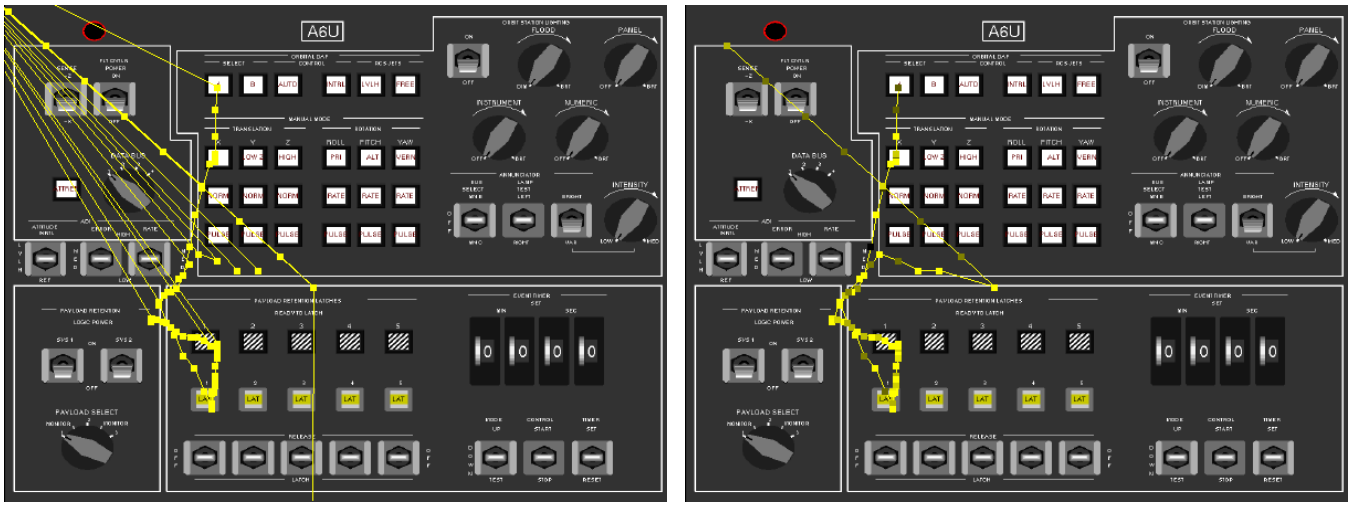
D.6.1 Signal Denoising

Before (or during) signal analysis, care must be taken to eliminate excessive noise in the eye movement signal. Inevitably, noise will be registered due to the inherent instability of the eye, and worse, due to blinks. The latter, considered to be a rather significant nuisance, generates a strong signal perturbation, which (luckily) may often be eliminated, depending on the performance characteristics of the available eye movement recoding device. It is often the case that either the device itself has capabilities for filtering out blinks, or that it simply return a value of $(0, 0)$ when the eye tracker "loses sight" of the salient features needed to record eye movements.

In practice, eye movement data falling outside a given rectangular range can be considered noise and eliminated. Using a rectangular region to denoise the (2D) signal also addresses another current limitation of eye tracking devices: their accuracy typically degrades in extreme peripheral regions. For this reason (as well as elimination of blinks), it may be sensible to simply ignore eye movement data falling outside the "effective operating range" of the device. This range will often be specified by the vendor in terms of visual angle. An example of signal

[3] A third type of analysis, requiring manual intervention, relies on slowly displaying the time course of the signal (the scanpath), either in 1D or 2D, one sample at a time and judging which sample points lie outside the mean. This "direct inspection" method is rather tedious, but surprisingly effective.

denoising is shown in Figure 21, where an interior rectangular region 10 pixels within the image borders defined the operating range. Samples falling outside this constrained interior image boundary were removed from the record (the original images measured 600×450 pixels).



(a) Raw eye movement data.

(b) Eye movement data after eliminating samples falling outside an interior region 10 pixels inside the image borders.

Figure 21: Eye movement signal denoising. Courtesy of Wesley Hix, Becky Morley, and Jeffrey Valdez.

D.6.2 Dwell-Time Fixation Detection

The dwell-time fixation detection algorithm depends on two characterization criteria:

1. identification of a stationary signal (the fixation), and
2. size of time window specifying an acceptable range (and hence temporal threshold) for fixation duration.

An example of such an automatic saccade/fixation classification algorithm, suggested by Anliker, determines whether M of N points lie within a certain distance D of the mean (μ) of the signal [Anl76]. This strategy is illustrated in Figure 22(a) where two N-sized windows are shown. In the second segment (positioned over a hypothetical saccade), the variance of the signal would exceed the threshold D indicating a rapid positional change, i.e., a saccade. The values of M, N, and D are determined empirically. Note that the value of N defines an *a priori* sliding window of sample times where the means and variances are computed. Anliker denotes this algorithm as the *position-variance method* since it based on the fact that a fixation is characterized by relative immobility (low position variance) whereas a saccade is distinguished by rapid change of position (high position variance).

D.6.3 Velocity-Based Saccade Detection

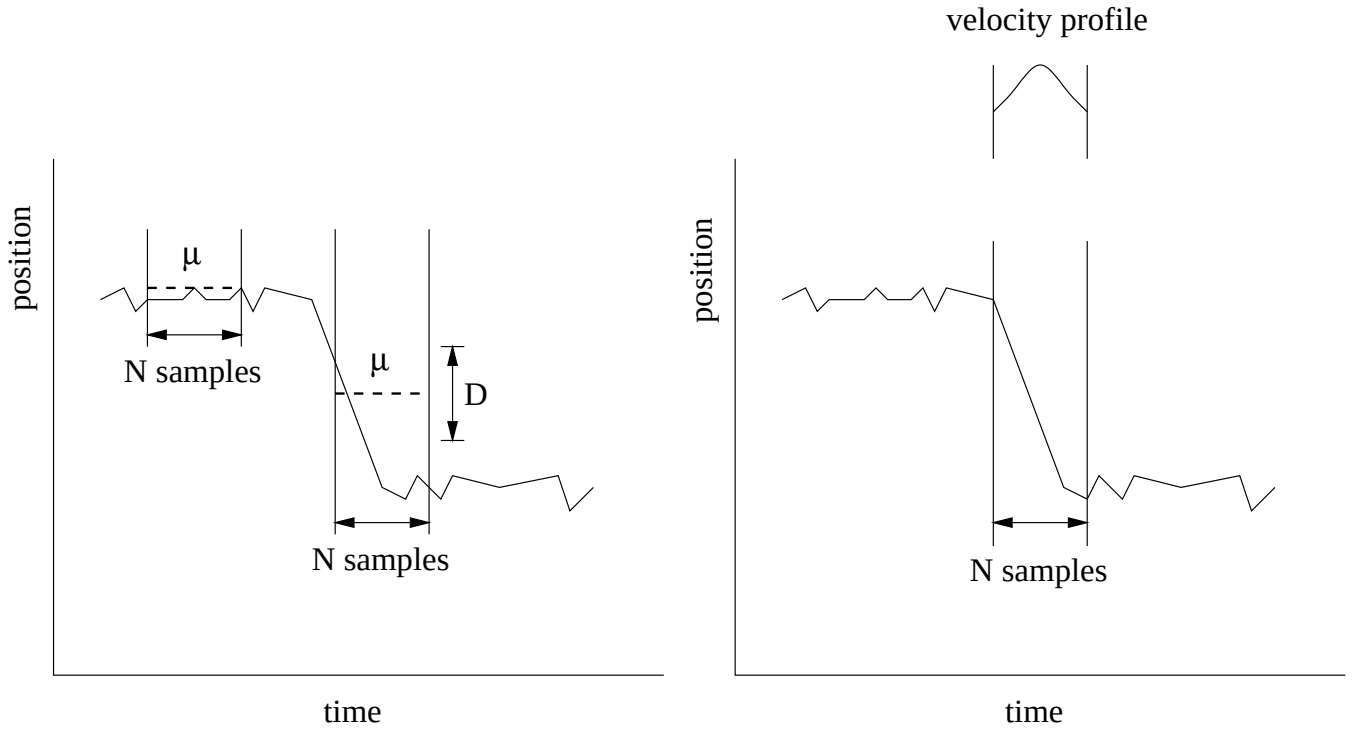
An alternative to the position-variance method is the *velocity detection method* [Anl76]. In this velocity-based approach, the velocity of the signal is calculated within a sample window, and compared to a velocity threshold. If the sampled velocity is smaller than the given threshold, then the sample window is deemed to belong to a fixation signal, otherwise it is a saccade. The velocity threshold is specified empirically. Figure 22(b) shows the hypothetical eye movement time course with the sample window centered over the saccade with its velocity profile above.

Noting Yarbus' observation that saccadic velocity is nearly symmetrical (resembling a bell curve), a velocity-based prediction scheme can be implemented to approximate the arrival time and location of the next fixation. The next fixation location can be approximated as soon as the peak velocity is detected. Measuring elapsed time and distance traveled, and taking into account the direction of the saccade, the prediction scheme essentially mirrors the left half of the velocity profile (up to its peak) to calculate the saccade's end point.

The position-variance and velocity-based algorithms give similar results, and both methods can be combined to bolster the analysis by checking for agreement. The velocity-based algorithm offers a slight advantage in that often short-term differential filters can be used to detect saccade onset, decreasing the effective sample window size. This speeds up calculation and is therefore more suitable for real-time applications. The image in Figure 21(b) was processed by examining the velocity,

$$v = \frac{\sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2}}{dt},$$

between successive sample pairs and normalizing against the maximum velocity found in the entire (short) time sequence. The normalized value, subtracted from unity, was then used to shade the sample points in the diagram. Slow moving points, or points approaching the speed of fixations are shown brighter than their faster counterparts.



(a) Position-variance method.

(b) Velocity-detection method.

Figure 22: Saccade/fixation detection.

Tole and Young suggest the use of short FIR (Finite Impulse Response) filters for saccade and fixation detection matching idealized saccade signal profiles [TY81]. An idealized (discrete) saccade time course is shown in Figure 23(a). A sketch of the algorithm proposed by Tole and Young is presented in Figure 23(b), with corresponding FIR filters given in Figure 24. The algorithm relies on four conditions to detect a saccade:

$$|I_1| > A \quad (1)$$

$$|I_2| > B \quad (2)$$

$$Sgn(I_2) \neq Sgn(I_1) \quad (3)$$

$$T_{\min} < I_2 - I_1 < T_{\max} \quad (4)$$

If the measured acceleration exceeds a threshold (condition 1), the acceleration buffer is scanned forward to check for a second peak with opposite sign to that of I_1 (condition 3) and greater in magnitude than threshold B (condition 2). Amplitude thresholds are derived from theoretical values, e.g., corresponding to expected peak saccade velocities of 600°/s (visual angle). Condition 4 stipulates minimum and maximum durations for the forward search, also based on theoretical limits, e.g., saccade durations in the range 120ms–300ms. If all conditions are met, a saccade is detected.

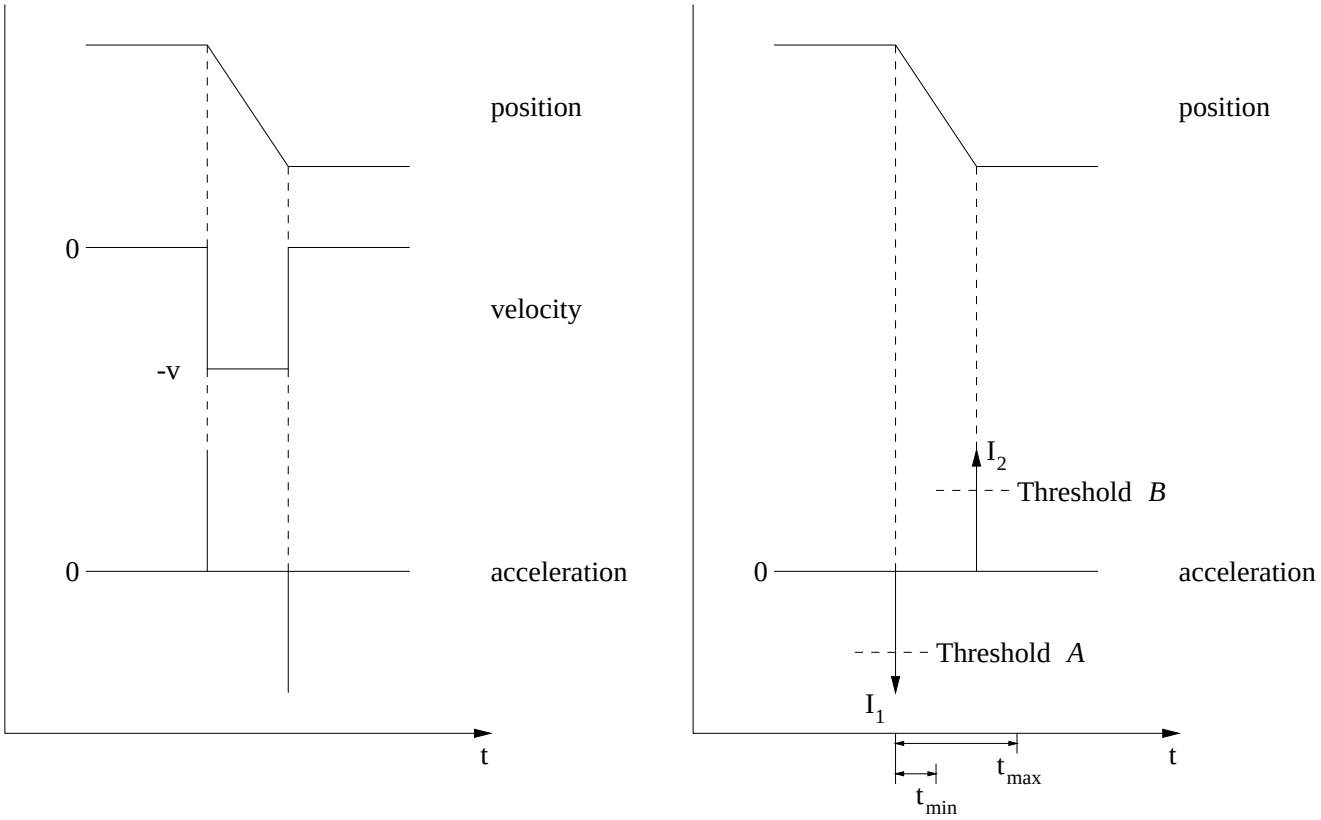
Enhancements to the above algorithm, offered by Tole and Young, include adaptive threshold estimation, adjusted to compensate for recently observed signal noise. This is achieved by reformulating the thresholds A and B as functions of an RMS estimate of noise in the signal:

$$\begin{aligned} \text{Threshold } A &= 4000 \text{ deg/sec}^2 + \text{Accel/RMS} + \text{Accel/DC noise} \\ \text{Threshold } B &= 4000 \text{ deg/sec}^2 + \text{Accel/RMS}. \end{aligned}$$

The term

$$\text{Accel/DC noise} = \frac{2\,|\,\Delta p\,|}{\Delta t^2}$$

estimates noise in acceleration, where Δt is the sampling interval and Δp is the peak-to-peak estimate of noise in position. This term can be estimated by measuring the peak-to-peak amplitude of fluctuations on the input signal when the average eye velocity over the last two seconds is less than 4 deg/sec. This estimate can be calculated once, for example, during calibration when the signal to noise ratio could be expected to remain constant over a test session, e.g., when the subject is known to be fixating a test target. This term may also be updated dynamically whenever velocity is low.



(a) Velocity and acceleration profiles.

(b) Finite Impulse Response (FIR) filter algorithm.

Figure 23: Idealized saccade detection.

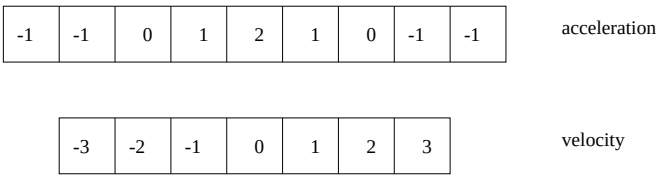


Figure 24: Finite Impulse Response (FIR) filters for saccade detection.

The remaining estimate of noise in the output filter,

$$\text{Accel/RMS} \quad = \quad \sqrt{\frac{1}{T}\sum_{i=1}^{T}(\text{Accel}^2(i) - \overline{\text{Accel}}^{\,2})},$$

assumes mean acceleration is zero for time windows greater than 4 seconds, i.e., $\overline{\text{Accel}} \rightarrow 0$ as $T > 4$s. The noise estimation term can be further simplifed to avoid the square root,

$$\text{Accel/RMS} \quad \leq \quad \frac{1}{T}\sum_{i=0}^{T} |\ \text{Accel}(i)\ |, \quad T > 4\text{s}.$$

since $\sqrt{\sum \text{Accel}^2} \ \leq \ \sum |\ \text{Accel}\ |$. This adaptive saccade detection method is reported to respond well to a temporarily induced increase in noise level, e.g., when the subject is asked to grit their teeth.

D.7 Summary and Further Reading

With the exception of Carpenter's widely referenced text [Car77], there appears to be no single suitable introductory text discussing eye movements exclusively. Instead, there are various texts on perception, cognition, and neuroscience which often include a chapter or section on the topic. There are also various collections of technical papers on eye movements, usually assembled from proceedings of focused symposia or conferences. A series of such books was produced by John Senders et al. in the seventies and eighties (see for example [MS76, FMS81]).

A large amount of work has been performed on studying eye movements in the context of reading. For a good introduction to this literature, see [Ray92].