

1 Visual sensors using eye movements

Oliver Landolt
Agilent Technologies Inc., Palo Alto, CA
oliver_landolt@agilent.com

1.1 Introduction

Visual processing in robotics applications lags far behind the visual systems of even simple animals. Close examination reveals that the operating principles underlying biological visual systems differ fundamentally from conventional electronic image sensing and processing hardware. One important difference, frequently overlooked even within the field of biologically inspired engineering, is that most animals can move their eyes with respect to their head and/or body. Such animals display a repertoire of eye movements designed to improve the visual data acquisition process and save neural hardware. By contrast, the visual axis of electronic visual systems is typically pointed in a fixed direction with respect to the supporting platform, or can move only slowly compared to the time scale of visual events. The present chapter argues that eye movements are an important aspect of visual sensing and should be built into electronic implementations of visual sensing systems. Section 1.2 presents an overview of the major types of eye movements found in humans and discusses their importance to vision. Sections 1.3 and 1.4 describe engineered systems incorporating similar principles. Lastly, the conclusion discusses to which degree the biological domain can inspire the architecture and algorithms of synthetic vision systems.

1.2 Eye movements in humans

Eye movements observed in humans fall into two broad categories, namely saccades and target-holding movements. Saccades consist of fast (on the order of $1000^\circ/\text{s}$), essentially ballistic jumps from one fixation point to another. Humans usually look at visual scenes in an endless succession of saccades and fixations (up to 3 per second) concentrated on areas of interest [1]. The main purpose of saccades is to bring an image location of interest into the high-acuity foveal region of the retina for detailed analysis. Another purpose of saccades is to bring the eyeball rapidly back to the center after drifting away, e.g. in pursuit of a moving object. The most obvious benefit of a moving fovea is economical: to

obtain foveal resolution over a wide field of view without eye movements, a tremendous number of photoreceptors would be needed, as well as a matching amount of cortical visual processing hardware. Similar economical considerations are highly relevant for electronic visual sensors as well.

Unlike saccades, target-holding eye movements are smooth and are designed to stabilize the retinal image. Several categories are distinguished depending on the nature of the stimulus controlling the movement. The vestibulo-ocular reflex (VOR) is driven by inertial measurements of head movements, performed by the vestibular system. The VOR counteracts the effect of head movements by generating opposite eye movements, thereby greatly reducing the velocity of retina image slip. The peak velocity of VOR movements is on the order to 500°/s. The opto-kinetic reflex (OKR) is similar in purpose but controlled by vision itself. The OKR compensates for large-field image slip up to about 80°/s, such as produced by the subject's own body movements with respect to the environment. Smooth pursuit consists of tracking a small moving object in order to keep it in the fovea. It operates up to 40°/s. Target-holding movements are important for the reduction of motion blur [2]. Because of their limited temporal bandwidth, photoreceptors can follow light intensity fluctuations caused by a moving image only up to some velocity, dependent on the spatial features of the image. At higher velocities, the image appears blurred. Total or partial compensation of image slip increases the range of conditions under which a target can be seen clearly.

In addition to the large-scale eye movements described so far, it is known that even during attempts to maintain perfect fixation, humans actually generate small involuntary eye movements [3]. These movements typically alternate between smooth drift and micro-saccades, with a median delay on the order of 0.6s between the onset of successive saccades. The amplitude of these movements is on the order of a few minutes of arc. It has been established that the function of these involuntary movements is to maintain permanent vision. If the image of a target is kept perfectly stable on the retina—by mounting the target directly on the eyeball for instance—then visual perception of the target vanishes in a matter of seconds [3]. Instead, observers report seeing a gray fuzzy field, or sometimes even extreme blackness. In other cases, observers describe intermittent perception of a dim, unsharp or partial image of the target. Reportedly, the target may transiently reappear clearly on events such as illumination flashes, or by applying a movement of sufficient

amplitude (on the order of 2' of arc) to the target. Failure of vision in the case of a perfectly stabilized retina image is explained by the fact that the visual system detects and encodes light intensity fluctuations over time—and space—rather than steady light levels. Therefore, permanent visual perception is possible only if the image keeps moving. Small-amplitude involuntary eye movements provide the conditions for sustained vision.

Although it may seem odd at first that humans rely on a visual system requiring eye movements to operate properly, this architecture turns out to have interesting properties. In natural viewing conditions, spatial patterns which remain perfectly stable on the retina are most likely artifacts from the visual system itself, such as the blind spot, shadows of blood vessels, after-images due to adaptation, or fixed-pattern noise due to photoreceptor mismatch. By detecting only temporal fluctuations caused by eye movements, such artifacts can in principle be discounted. In contrast with biological retinas, most electronic image sensors—notably cameras—measure the DC light intensity received by every pixel and are thus affected by fixed-pattern noise caused by photoreceptor mismatch. To deal with this effect, the fixed pattern must be measured in a calibration step, stored in some type of memory and discounted at image readout. By building electronic visual sensors capable of shifting their optical axis, the benefits of the biological approach can be exploited without sacrificing sustained vision. Another prospective advantage of small-amplitude eye movements is resolution enhancement. Shifting a photoreceptor in space provides access to a continuum of image points instead of a discrete grid. In principle, it is possible to use such scanning movements to detect features much narrower than photoreceptor spacing. It is not likely that humans exploit this potential because photoreceptor spacing in the fovea is so narrow that the effective resolution is actually limited by the imaging optics [3]. However, there exist animals known to rely on scanning for visual data acquisition, presumably for resolution enhancement purposes [4][5].

Given the benefits provided by moving eyes to their animal users, perhaps gaze control could improve electronic visual sensors as well and contribute to bridge the performance gap between biological and electronic visual systems. For the development of practical visual systems exploiting eye movements to the same extent as animals, several technological issues must be addressed.

1. A mechanical/optical solution must be found to the problem of rapidly shifting the visual axis. Device size and power consumption must remain compatible with small robotics applications. For performance

comparable to humans, the device should be capable of executing saccades within about 50-100ms with an accuracy on the order of 1° , over a total angle of $60-90^\circ$.

2. Image sensing principles must be developed, which take advantage of eye movements to achieve capabilities such as fixed-pattern noise rejection or resolution enhancement.
3. High-level control strategies must be developed to achieve the built-in functionality of the human oculo-motor system, such as target-holding reflexes and saccadic exploration. Conflicts between subsystems implementing different functions must be resolved.
4. Efficient computation of coordinate mappings between visual space and motor space must be implemented. Motor space is defined by the degrees of freedom of the device shifting the visual axis. Fast and accurate coordinate mapping is an essential part of the control system of a moving eye.

The author and his colleagues have been addressing these issues over several years and have proposed and implemented original solutions in the frame of two independent developments of visual sensing systems exploiting eye movements. Custom hardware solutions were favored, which rely on operating principles inspired by neurobiology [6]. This choice is motivated partly by interest in exploring the potential of this approach, and partly by anticipation of strong constraints in terms of cost, size, speed and power consumption in many applications of eye movements. Arguably, bio-inspired approaches should provide superior solutions—on the long run—in classes of problems where biology excels whereas computers do poorly. Indeed, neural structures have been optimized over millions of years of evolution by selection of the best designs for survival.

1.3 Oculo-motor system based on micro-prism gratings

1.3.1 Functionality

An oculo-motor system has been built, which implements two types of eye movements found in humans, namely saccades and smooth pursuit [7]. Rotations of the visual axis are produced by a light deflection device mounted in front of a fixed image sensor. The system otherwise contains four analog Very Large Scale Integrated circuits (VLSI) based on bio-inspired principles. The primary purpose of this system is to demonstrate the validity of novel principles underlying the optical front-end, the related

control strategy and its custom hardware implementation. However, some additional elements are likely to be required before practical applications can be addressed. Particularly, difficult visual processing problems—such as identification of saccade targets in natural scenes or discrimination of moving objects from their background—have been circumvented by assuming visual scenes made of light patches over a dark background, whereby luminance directly identifies objects of interest. The functionality of the oculo-motor system can be extended without alteration of its core by inserting more sophisticated visual processing hardware between the retina chips and the rest of the control system.

1.3.2 Light deflection device

To rotate the visual axis, we have selected an approach consisting of keeping the retina chip and the focusing lens fixed, while deflecting light using refractive elements mounted in front of the lens (Figure 1.1A). The light deflection device consists of two transparent and flat plastic disks with a micro-prism grating on one side, mounted perpendicularly to the optical axis of the lens. The angle of the micro-prisms is 30° and the refraction index is 1.49. The disks have an overall diameter of 7mm and a thickness on the order of 1mm. Each disk can rotate without restriction around this axis, independently from the other. As a whole, each micro-prism grating acts on light essentially like a single bulky prism, except that it takes much less space and weight (Figure 1.1B). Although a single fixed prism cannot have an adjustable deflection angle, with two mobile prisms in the light path, any magnitude and direction of deflection can be obtained within some boundaries. Their contributions may combine either constructively or destructively depending on the relative prism orientations. With the selected material and geometry, the available range extends up to $\pm 45^\circ$ horizontally and vertically. The relationship between prism orientations and the direction and magnitude of deflection has been derived in detail [8]. This relationship turns out to be strongly non-linear and there are generally two combinations of prism orientations achieving the same overall deflection. Therefore, control of this device involves the computation of non-trivial coordinate mappings. However, the decisive advantage of this device over many other solutions is that only two small passive optical elements have to move. This feature enables fast movements with moderately powered motors and avoids electrical connections to moving parts. The measured delay between onset of a saccade and stabilization on the target ranges from 45ms to 100ms depending on the saccade magnitude. Optical aberrations introduced by the prisms degrade the sharpness of the image. The degradation is

negligible at low deflection angles, but at the extreme deflection of 45° , the image of an ideal spot extends over a distance of 300-400 μm . However, when the device is used in conjunction with a typical electronic retina, this degradation is not significant up to deflections of $30\text{-}35^\circ$. The reason is that these image sensors are characterized by large pixel spacing (typically 50-100 μm) due to the presence of focal-plane electronic processing.

1.3.3 Control system architecture

The oculo-motor system consists of three modules as shown on the block schematic in Figure 1.2. The moving eye module contains the light deflection device described in §1.3.2 with its two motors. The orientations of the two micro-prism gratings are monitored by means of coding wheels and optical detectors. The motors are controlled by two independent position servo loops keeping each grating in the orientation specified by the content of a register. These registers are the points of convergence between the saccadic exploration module and the smooth pursuit module. A saccade is generated by loading a new orientation setting in the registers, whereas smooth movements are generated by applying incremental changes to the stored settings. The gratings track these changes within the delay imposed by the dynamics of the low-level control loops.

The light deflection device is mounted on top of a retina chip with a focusing lens dimensioned to achieve a relatively narrow field of view of 12° . Since this narrow field can be moved over a large angle, this retina can be considered as a high-resolution “spotlight” gathering image details sequentially in time. In this respect, it is similar to the fovea found in humans and other animals. The retina chip contains an array of 35 by 35 pixels turning the image into electrical signals encoded in the timing of pulses. The firing rate of a given pixel is proportional to the ratio between the incident light intensity and the local average intensity in the neighborhood of the pixel. This integrated circuit, which is described in detail in [9], was not designed specifically for the oculo-motor system. The pixels respond to steady illumination instead of only temporal transients as the human retina. However, this discrepancy does not significantly affect system operation. The data delivered by this retina chip is used internally for smooth pursuit control. It also constitutes the main output delivered by the oculo-motor system to higher levels of processing.

1.3.4 Smooth pursuit module

This module contains an element detecting the location of a target on the retina image. The target location is transmitted to a so-called incremental control chip implementing non-linear coordinate mapping between visual and motor space. This chip determines the direction and relative magnitudes of prism orientation increments required to slightly shift the image of the target toward the center of the retina. For this purpose, it also takes the current orientations of the two prisms into account, because the appropriate action depends on the present state of the system. The requested actions are executed by the moving eye module, thereby causing the image of the target to shift on the retina. Corrective actions are integrated over time as a result of continuous operation of this visual control loop, whereby a target initially away from the center will progressively shift toward the center of the retina and then remain there. If the target moves, the eye tracks it until the next saccade causes locking onto a new target, or until the object leaves the field accessible to the moving eye.

Information coding and processing within the incremental control chip are based on principles inspired by biological neural structures. Throughout the chip, variables are represented by the location of a patch of activity within a topologically organized array of cells. This coding convention is called *place coding* [8] to emphasize the fact that location—rather than only amplitude—determines the represented quantity. This approach is closely related to population coding [10], which is widespread in biological neural structures. At the input of the incremental control chip, the angular position of the visual target with respect to the center of the retina is encoded by a map consisting of eight nodes. Each of these nodes is related to a particular preferred direction. These directions are uniformly spaced at intervals of 45° . In order to indicate that the target is located at a given arbitrary angle (such as 81°), the closest matching input cells of the chip (45° and 90°) must be stimulated by pulse streams of some frequency. The relative frequencies at which these nodes are driven indicate how closely the represented direction matches their preferred direction (Figure 1.3). The convention chosen for the definition of place coding is that the value represented by a given pattern of stimulation is the centroid of the pattern [8]. A direction of 81° can therefore be represented by applying 20% of the total stimulation to the node related to 45° , whereas 80% of the stimulation is applied to the node whose preferred direction is 90° .

With a place coding scheme, computation can be performed by networks of passive interconnections—called *networks of links*—which embody the relationship between an input map and an output map [8]. To implement functions of several variables, multiple input maps can be merged by means of an array of fuzzy logic AND gates. The resulting intermediate map can then be projected onto the output map by means of a network of links. Fuzzy logic AND gates can be implemented by an extremely simple circuit consisting of only one metal-oxide-semiconductor (MOS) transistor per input. The incremental control chip is a direct application of this approach. Its input stage consists of three place coding maps representing the target location and the current orientations of the two prisms. These maps receive stimulation encoded as the frequency of brief pulses. Two networks of links project the three input maps onto two intermediate maps (Figure 1.4). A third network of links performs an additional computational step resulting in the final maps. These maps represent control actions which are transmitted to the moving eye module in the form of pulses. The architecture of the chip has been derived manually by studying the properties of the control function to be implemented. The topology of the networks of links has been generated automatically by software on the basis of a behavioral model of the oculo-motor system.

Place coding may at first seem complicated compared to ordinary analog or digital coding, but it actually lends itself well to compact and power-efficient hardware implementation of nonlinear mappings such as the oculo-motor system requires. The core of the incremental control chip (Figure 1.5) uses 2mm^2 of silicon area in a $0.7\mu\text{m}$ CMOS process, and consumes $30\mu\text{W}$ under a supply voltage of 5V. Conversion from digital coding to place coding on a chip is simple and efficient [8]. Besides, it is possible in some cases to get place coding directly from a sensor. In the smooth pursuit module, the input map representing the angular location of the target is derived from the visual data delivered by the retina chip. The surface of the retina is divided into eight radial slices and one central area (Figure 1.6). The pulses generated by all pixels within a radial slice are merged into a single signal, whereby the average frequency of the resulting pulse stream is roughly the sum of spiking frequencies of all pixels within that slice. This pulse stream stimulates one node of the target location map, namely the node whose preferred direction matches the orientation of the radial slice. If a light spot shining on the retina is located away from the center, firing of pixels illuminated by the spot causes stimulation of a small fragment of the map. The location of the most active node provides coarse information about the target location. If the spot is

large enough to overlap two slices, two neighboring nodes will be stimulated at a rate depending on the area of the intersection between the spot and the retina slice. The stimulation pattern received by the chip thereby provides continuous information about the spot location. This representation of the angular position may be slightly distorted, but is accurate enough to guarantee that the control actions applied under permanent feedback cause the spot to ultimately reach the center of the retina.

1.3.5 Saccadic exploration module

The saccadic exploration module contains a second instance of the same retina chip as the moving eye module. However, the visual axis of this retina is fixed and the associated focusing optics is such that the field of view extends over most of the area accessible by eye movements. This retina provides a coarse image of the surroundings for the selection of salient (i.e. interesting) locations. Therefore, it is comparable to the peripheral area of the human retina. Under the simplifying assumption that the visual scene is made of light patches over a dark background, luminance can be used as a measure of saliency. Therefore, the raw image delivered by the retina can be considered as a distribution of saliency without further processing. For practical applications, more sophisticated and possibly task-dependent measures of saliency could be implemented [11][12]. The related processing could be carried out by an additional chip taking the retina image as input and providing a saliency map at its output (as suggested in Figure 1.2).

The distribution of saliency derived from the retina image, encoded as mean firing rates on an array of cells, is applied to a so-called saccadic control chip. This chip has the function of selecting saccade targets based on the distribution of saliency and determining the timing of saccades between locations of interest. The general idea is to select the most salient point at any time as the target of the next saccade. To prevent neglect in the case of multiple salient points, incoming saliency signals are integrated over time, and the next saccade target is the location with the highest integral. Saliency signals originating from the currently attended location and its immediate vicinity are ignored and the corresponding integrators are let to slowly leak instead of letting them keep increasing. Thereby, whenever a given point becomes the target of a saccade, it will remain the most salient location only for a limited amount of time before another location wins the competition. As a result of the integration process, any location with nonzero saliency will become the saccade

target sooner or later. However, the number of visits to more salient locations is more frequent because the corresponding integrals increase at a higher rate. Moreover, fixation tends to be maintained for a longer time at more salient locations. The network generating saccade dynamics has been implemented in analog hardware as part of the saccadic control chip, which is described in detail in [8]. Saccade dynamics have been measured on the chip alone under controlled electrical stimulations. For illustration, statistics obtained by recording saccades over three minutes in the case of four equally salient points are given in Table 1.1. The four points, arbitrarily labeled from A to D, are spaced well apart in the visual field. Similar measurements were performed on the entire oculo-motor system under optical stimulation as well (§1.3.7). Beside generating saliency-driven saccades, the chip also implements coordinate mapping between visual and motor spaces. This mapping differs from the smooth pursuit module, because visual space is fixed instead of being bound to the visual axis of the moving eye. Moreover, the saccadic control chip has to compute absolute prism orientations instead of increments with respect to the present location. The hardware implementation of this coordinate mapping relies on place coding as in the previous module.

1.3.6 Module coordination

Due to the coexistence of two independent control modules, the control system must incorporate some amount of coordination between them to avoid interference. For instance, during a saccade, the smooth pursuit module would attempt to compensate for the observed retina image slip if no mechanism prevented it from doing so. In order to handle this situation, the saccadic exploration module inhibits the smooth pursuit module and resets the incremental control chip during every saccade. After the saccade, the smooth pursuit module starts tracking the new target without memory of its previous state. Another coordination mechanism has been incorporated to handle boundary problems during smooth pursuit. If a tracked target crosses the boundary of the visual field accessible to the moving eye, this situation is detected by the smooth pursuit module, which triggers a signal causing the saccadic control chip to generate a saccade back to the center. This mechanism is represented on Figure 1.2 by an arrow labeled 'nystagmus', by analogy with the reflex found in humans and other animals.

1.3.7 System-level measurements

The complete oculo-motor system (Figure 1.7) consists of three independent printed circuit boards corresponding to the three modules described in §1.3.3. Most experiments were carried out with the system mounted on an optical bench in front of a dark screen punctuated by light-emitting diodes (LED). The direction of the visual axis was permanently monitored by observing internal signals within the moving eye module. Image data from either retina chip was acquired by recording all transitions on the output bus of the chip by means of a logic analyzer as long as memory would allow, then processing the data off-line in order to count the number of spikes emitted by each pixel. Stimulation patterns in portions of the system using place coding were acquired by a similar method. A first round of experiments consisted of qualitative observations of the behavior of the system while turning on either the saccadic exploration module, or the smooth pursuit module, or both. With the saccadic module only, when shown a scene consisting of a single LED turned on, the moving eye points toward this LED most of the time. Brief saccades toward any point in the background occur occasionally. The fraction of time spent watching the background depends on the contrast between target saliency and background saliency [7]. The experimental setup had a dark background intended to keep background saliency low, therefore the system spent only a few percent of the time on the background. It is perfectly appropriate that the oculo-motor system pays some attention to the background if its saliency is not strictly zero. This behavior is due to the temporal integration of saliency built into the saccadic control chip (§1.3.5). With two LEDs turned on, saccades alternate from one target to the other in an essentially periodic fashion, except for infrequent saccades toward background locations as in the previous case. With additional LEDs, all intended targets are visited frequently, but no obvious periodicity can be observed. The system exhibits only minor preferences between targets of the same nominal intensity, consistent with observations made on the saccadic control chip alone (Table 1.1). The rate at which saccades occur depends on the rate at which retina pixels emit spikes. This rate can be tuned by altering bias conditions of the retina chip. In most experiments, the retina was tuned to achieve about 3-5 saccades per second, well within the range supported by the underlying mechanics. Saccade accuracy was measured by activating several targets at various locations on the visual field. Immediately after every saccade, the location of the target was measured by computing the centroid of its image on the retina chip mounted underneath the light deflection device. The distance between the centroid

of the target and the center of the retina was considered to be the error. In all cases, the error is less than 2° of visual angle. The error vector tends to be identical upon successive returns to the same target, suggesting that the accuracy of saccades is limited by built-in errors rather than noise. The major causes are rounding distortions inherent to the operation of networks of links, and map alignment errors due to the fact that the visual axes of the two retinas are not exactly identical.

When the smooth pursuit module is activated alone, the oculomotor system tracks a light source waved manually within its field of view. Because this module lacks a wide-angle acquisition mechanism, tracking is initiated only after the target enters the visual field of the moving eye. Smooth pursuit operates within a range of velocity limited by the bandwidth of the retina chip and the dynamics of the low-level control loops. For bright targets, tracking velocities up to $50^\circ/\text{s}$ have been observed. The dynamics of the feedback loop depends on target contrast because the system uses luminance directly as a feature to identify the target. Lighter targets cause the retina to fire at a higher rate, thereby causing more frequent updates of the control action fed back to the mechanical parts. When both modules are active simultaneously, the observed behavior is essentially a combination of the first two cases. When a fixed scene of LEDs is shown, saccades occur as before, except that targets end up accurately centered on the retina. The reason is that the smooth pursuit module cancels the residual error after saccades under closed-loop visual control. Conversely, in the presence of a slowly moving target within the wide-angle visual field of the system, an initial saccade toward the target is triggered 1-2s after its initial appearance, then the target is tracked by the smooth pursuit module.

1.4 Visual sensor relying on mechanical vibrations

1.4.1 Concept

In Section 1.2, it was argued that small-amplitude movements of the visual axis can be used as a means to overcome fixed-pattern noise inherent to photoreceptors, and as a means to gain access to the image in continuity rather than on a discrete grid. A visual sensor exploiting this technique is described in the present section. The general idea consists of applying continuous oscillatory scanning movements to the image focused on the surface of a photoreceptor array. The amplitude of the oscillation should be on the order of pixel spacing, and the frequency should be high enough

that the image cannot change significantly over a single scanning cycle. A frequency of a few hundred cycles per second is expected to be sufficient in most cases. In principle, the oscillation does not need to be periodical, but periodicity simplifies signal processing. As a result of these oscillations, spatial variations of light intensity in the image turn into temporal fluctuations of light intensity at every photoreceptor. Knowing the pattern of motion applied to the image, elementary spatial image features such as edges or textures can be detected by processing these temporal signals. For instance, by sweeping the image of a thin line—such as a dark cable over the background of a clear sky—over a photoreceptor, an impulse of photocurrent will be observed, even if the line is much thinner than pixel spacing. The temporal signature of a dark line is a sharp transition from light to dark followed soon by another transition from dark to light. In addition to the mere existence of this line, its orientation can also be determined by relating the occurrence of temporal changes in photocurrent with the instantaneous direction of movement applied to the image at the same time. No fluctuation will occur while the photoreceptor is shifted parallel to the line, whereas the steepest transitions will be detected when the scanning path crosses the line perpendicularly. If the photoreceptor scans the same area at different angles, the general orientation of the underlying pattern can be determined by identifying the directions of scanning producing the most or the least intensity fluctuations. Finally, the thickness of the line can be determined by measuring the time elapsed between the falling and the rising transition, and scaling it by the velocity of the scanning movement over this time interval. With this principle, every single pixel on the image sensor has the functionality of a local spatial feature detector, provided that the image is kept in permanent movement and that this movement is known at all times. Obviously, the whole signal processing chain leading from raw photocurrents to a detailed interpretation of the visual scene cannot be incorporated into every pixel of a single chip. Instead, in the system described herein, pixel functionality is limited to the detection of significant temporal transitions. At the occurrence of such a transition, a pixel immediately fires a brief pulse. At the periphery of the pixel array, pulses are tagged by the coordinates of the firing pixel and transmitted off-chip in real time. External hardware is used to identify temporal patterns of pulses within each pixel and relate them to the instantaneous scanning movement applied to the image. Image feature maps resulting from parallel operation of several temporal pattern detectors can be used as the input of higher level processing such as landmark recognition or depth

estimation. The outcome of this process can be used for autonomous vehicle navigation or other machine vision applications.

The format in which visual data is encoded by such a sensor is very different from existing cameras. In fact, rebuilding the original image from pulse timings for display to human observers would be rather resource-consuming, although possible in principle. Instead, the point of this device is rather to act as a front-end to a visual system extracting relevant information from an image in order to provide a machine with visual sensing capabilities. In this respect, the proposed scheme has several interesting properties. First of all, spatial features are detected without relying on comparisons between different photoreceptors. Therefore, the issue of fixed-pattern noise is virtually eliminated. Spatial aliasing is also eliminated because scanning movements are continuous. However, temporal aliasing is still present as in most other image sensors because scanning occurs in discrete cycles. Temporal coding of image data in the timing of spikes lends itself to efficient implementation of feature detectors, either with custom VLSI chips or with off-the-shelf digital hardware. Another property of this coding scheme is that pixels communicate information only to the extent that the image contains significant features in their scanning area. This strategy makes more efficient use of the available communication bandwidth than systematic serial readout [13][14]. On the down side, it is clear that every pixel requires substantially more silicon area than a pixel of a conventional camera because some amount of signal processing is carried out locally. This drawback applies to all other image sensors with focal-plane processing as well. However, in the specific design described herein, this limitation is largely offset by the fact that the effective resolution and the amount of information extracted from the image by a scanning pixel is much larger than a fixed pixel can achieve.

An implementation of the approach outlined in this introduction is under development as of this writing. Some design details and intermediate results are presented in the remainder of this section.

1.4.2 Light deflection

Two different methods for applying mechanical oscillations to an image have been considered. Periodic image movements at a constant velocity on a circular path can be achieved by spinning a tilted mirror in front of the focusing optics (Figure 1.8). The mirror must be mounted on the shaft of a motor which should be tilted at an angle of about 45° with respect to the optical axis of the lens. If the mirror is not exactly perpendicular to the

shaft but tilted by a small angle ε , rotation of the motor will cause the reflective surface to wobble, thereby causing the image to travel a circular path with a radius of 2ε in viewing angle. A prototype device using this principle has been designed and built. The measured mirror angle ε was 0.56° . A DC motor spinning up to 19000rpm was selected in order to provide scanning frequencies in excess of 300Hz. A magnetic encoder was coupled to the motor in order to indicate the orientation of the mirror at all times.

The spinning mirror device provides accurate control over the scanning path. However, the size and power consumption of the motor are significant compared to the image sensing chip. For applications where space and power are an issue, an alternative device has been developed, where the scanning pattern is produced by displacements of the lens focusing the image onto the chip. In this device, the lens is mounted on springs allowing lateral X-Y displacements but maintaining constant spacing between the lens and the chip. If the system is mounted onto a vibrating platform such as a vehicle driving on a rough surface, the mechanical energy available in the vicinity of the resonance frequency of the lens/spring system will cause scanning movements. To be effective, the amplitude of these movements must be on the order of pixel spacing on the chip, i.e. a few tens of microns. In applications where the permanent availability of environmental vibrations is not guaranteed, small piezo-electric actuators could be added to the device. These actuators would need to be turned on only when external vibrations are insufficient. The shape of the scanning path will depend on the relative magnitudes and phases of vibrations applied to the X and Y axes, and on the resonance frequency matching between these axes. As the scanning path will vary over time depending on environmental vibratory conditions, it is necessary to monitor the position of the lens and use this information in the interpretation of the signals generated by the visual sensing chip. For this purpose, the lens position is monitored by capacitive measurements between the lens socket and surrounding fixed electrodes. A prototype scanning device operating on the principle described herein has been manufactured (Figure 1.9). Its dimensions are 34.5mm×34.5mm×8.2mm, and its measured resonance frequency is 645Hz. This frequency is reached with the mass of just the lens and its socket. It can be reduced as needed by attaching additional mass to the lens socket.

1.4.3 Image sensor

An integrated circuit implementing an array of 32×32 pixels has been designed and manufactured. A block schematic of a single pixel is shown in Figure 1.10. In the first stage of signal processing, the current delivered by a photodiode is applied to a logarithmic current-to-voltage converter. The same visual scene under different illumination levels produces images differing only by a scaling factor in intensity. After logarithmic transform and differentiation, the temporal waveforms produced by scanning are essentially independent of illumination level. Beside logarithmic compression, this circuit also enhances the temporal bandwidth of the photoreceptor with respect to a passive solution. A large bandwidth is crucial to the operation of this visual sensor because scanning must occur at a much larger frequency than typical fluctuations within the image itself. In addition, photoreceptor bandwidth determines the effective resolution of the visual system. If the photocurrent impulse caused by crossing a thin line is too brief, it will be filtered out and therefore the line will not be detected. It can be shown that at 10KHz or above, the limiting factor of resolution is no longer bandwidth [15].

The signal resulting from logarithmic compression is differentiated with respect to time and half-wave rectified, whereby both the positive and the negative fraction are retained separately. Current signals delivered at both outputs of the rectifier are sent to independent non-leaky integrate-and-fire circuits, where the charge is accumulated over time until the resulting voltage reaches a threshold. At this point, the integrate-and-fire block emits a short pulse, resets its integrator and resumes operation. Whenever the scanning path of a photoreceptor crosses a sharp edge causing an amplitude change exceeding the built-in threshold, at least one spike is reliably generated at this point at every scanning cycle. In another prototypical case where an area of the image contains only a weak intensity gradient instead of a sharp edge, the temporal waveform contains only low amplitude fluctuations proportional to the magnitude of the gradient. In this case, it takes several scanning cycles—in inverse proportion to the gradient magnitude—before a spike can be generated, and this spike may occur any time the intensity is changing. In the general case, it can be shown that the probability of spiking at a particular phase of a periodical scanning cycle is proportional to the gradient of the image at this point.

The integrated circuit has been manufactured using a CMOS process with a feature size of $0.6\mu\text{m}$, two levels of polysilicon and three levels of metal. A single pixel has a size of $68.5\mu\text{m} \times 68.5\mu\text{m}$, including a

photodiode of $10\mu\text{m}\times 10\mu\text{m}$. The entire chip area is about 10mm^2 . The layout of a single pixel is shown in Figure 1.11. With a supply voltage of 3V, the measured steady current consumed by the analog stages of the chip ranges between $22\mu\text{A}$ and $26\mu\text{A}$ depending on incident illumination. In addition, the digital communication bus transmitting pulses off-chip consumes about 1mA when operating at a rate of 1.2Mpulses/s. Consumption of this block is roughly proportional to the data rate. Besides the main pixel array, an additional test pixel was incorporated on the chip, which can be stimulated electrically instead of optically for accurate control over experimental conditions. The bandwidth of the photoreceptor circuit was measured by sweeping the frequency of a sine wave at the electrical input emulating the photodiode current, and observing the waveform at a test point located immediately before the rectifier. The bandwidth depends on the DC photocurrent level. Measurement results within the current range of interest are plotted in Figure 1.12. The bandwidth exceeds 10KHz for photocurrent levels of $4\mu\text{A}$ and above. With the lens built into the devices described in §1.4.2, this photocurrent level is reached in indoor illumination conditions. Another experiment aimed at verifying that spiking probability is proportional to the gradient of the input signal. For this purpose, a sine wave of constant frequency was applied to the input of the test pixel while recording the times at which pulses are emitted at one of its outputs. Timing of zero-crossings of the input sine wave were recorded simultaneously on the same instrument. Time within each period of the input signal was split into 100 bins, and a histogram was built by counting the number of spikes occurring in each bin over many cycles. As expected, the histogram has the shape of a half-wave rectified sine wave (Figure 1.13) with a 90° phase shift with respect to the input signal. Although the chip described in this paragraph has much of the expected functionality, it turned out that intrinsic noise within the pixel circuits caused excessive background firing, thereby making it unfit for acquisition of satisfactory image data. This problem is being addressed by a redesign in progress as of this writing.

1.5 Conclusion

Providing robots or vehicles with sufficient visual capabilities to confer them some degree of autonomy in a wide range of environments is a difficult problem. The stumbling block is not light sensing, which is very well mastered in electronics, but rather processing of visual information. Existing cameras provide high quality images fit for display, but when it comes to automatically determining what the image represents, solutions

available today typically require either excessive computing time, or an amount of hardware whose volume, weight and/or power consumption is incompatible with most practical applications. In sharp contrast with this situation, even simple biological creatures possess visual systems powerful enough to let them navigate through their environment and take care of their business. The principles underlying the operation of their visual system are fundamentally different from current mainstream electronic image sensors and associated computational means. The performance gap is such that despite steady progress in integrated circuits manufacturing technology, it is questionable whether the traditional architecture based on a huge pixel array, serial readout and a low number of powerful processors will ever reach the performance level found in biological organisms.

The field of bio-inspired—or neuromorphic—engineering is founded on the premise that principles underlying biological computational structures can be used as the basis of electronic designs and can confer them a comparable level of performance. More than a decade after the emergence of this field [6], it is somewhat embarrassing to admit that few visual sensor designs following this approach have ever reached the stage of a commercial product [16]. It seems appropriate to inquire about the causes of this slow development. The relative modesty of means invested into this approach—compared to mainstream electronics—comes to mind. Other circumstances might also slow down neuromorphic engineering, such as a drop in popularity of analog circuit design compared to digital VLSI, software or communication systems design. However, the most serious limitation might be the current degree of understanding of biological structures. The most common abstractions used today to grasp the functionality of neural structures implementing visual processing might simply miss essential points. For instance, although many neural models consider only the mean firing rate of neurons as relevant, it is now commonly recognized that the detailed timing of spikes also matters in many cases. Without stepping from a mean-firing-rate abstraction to a more detailed description, the operation of some neural circuits simply cannot be understood. In the field of vision, it is possible that exceedingly reduced abstractions hinder the discovery of the key features which really confer biological architectures their power. One of the missing dimensions might be the fact that an eye is not just an array of photoreceptors passively measuring incoming light, but a subsystem in permanent motion whose movements are the basis of a reliable visual data acquisition process. It is the author's belief that the premise of bio-inspired engineering is correct, but that radically new perspectives on the operation

of visual systems are still to be discovered, and that such perspectives will eventually lead to powerful hardware implementations of such systems. The work described in the present chapter is meant to be a modest step in this direction by attempting to consider the electronic and mechanical aspects of visual sensing within the same designs.

1.6 Acknowledgements

The oculo-motor system has been developed at CSEM SA in Neuchâtel, Switzerland with basic research funding from the Swiss government. Besides the author, the project team includes Patrick Debergh, Friedrich Heitger, Stève Gyger, Eduardo Franzini. Additional contributions by Johann Bergqvist, Lorenzo Zago, William Beaudot, Philippe Venier, Alessandro Mortara and Eric Vittoz are gratefully acknowledged.

The vibrating visual sensor is under development at Caltech in Pasadena, USA under funding by DARPA/ONR and the Center for Neuromorphic Systems Engineering, as part of the National Science Foundation Engineering Research Center program. Besides the author, the project team includes Ania Mitros, Theron Stanford and Christof Koch.

The author is very grateful to Ania Mitros for revising the manuscript and suggesting numerous improvements.

1.7 References

- [1] A. Yarbus, *Movements of the eyes*, Plenum Press, New York, 1967
- [2] M. F. Land, "Motion and vision: why animals move their eyes", *Journal of Comparative Physiology A*, Vol. 185, 1999, pp. 341-352
- [3] R. W. Ditchburn, *Eye-movements and visual perception*, Oxford University Press, Oxford, 1973
- [4] M. F. Land, "Mechanisms of orientation and pattern recognition by jumping spiders (Salticidae)", in: R. Wehner (ed), *Information processing in the visual systems of arthropods*, Springer Verlag, Berlin, 1972, pp. 231-247
- [5] R. Hengstenberg, "Eye movements in the housefly *Musca Domestica*", in: R. Wehner (ed), *Information processing in the visual systems of arthropods*, Springer Verlag, Berlin, 1972, pp. 93-96
- [6] C. Mead, *Analog VLSI and Neural Systems*, Addison-Wesley Publishing, Reading MA, 1989

- [7] O. Landolt, S. Gyger, "An oculo-motor system with multi-chip neuromorphic analog VLSI control", in: S. A. Solla, T. K. Leen and K.-R. Muller, *Advances in Neural Information Processing Systems 12*, MIT Press, Cambridge, 2000, pp. 710-716
- [8] O. Landolt, *Place coding in analog VLSI – A neuromorphic approach to computation*, Kluwer Academic Publishers, Dordrecht, 1998
- [9] P. Venier, "A contrast sensitive silicon retina based on conductance modulation in a diffusion network", Proc. 6th International Conference on Microelectronics for Neural Networks and Fuzzy Systems, Dresden, September 1997
- [10] A. Pouget, "Statistically efficient estimation using population coding", *Neural Computation*, Vol. 10, 1998, pp. 373-401
- [11] L. Itti, C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention", *Vision Research*, Vol. 40, No. 10-12, 2000, pp. 1489-1506
- [12] L. Itti, *Models of bottom-up and top-down visual attention*, Ph.D. thesis, California Institute of Technology, Jan 2000
- [13] A. Mortara, E. Vittoz, P. Venier, "A communication scheme for analog VLSI perceptive systems", *IEEE Journal of Solid-State Circuits*, Vol. 30, No. 6, June 1995, pp. 660-669
- [14] K. Boahen, "Retinomorph vision systems II: communication channel design", Proc. IEEE International Symposium on Circuits and Systems (ISCAS'96), Atlanta, May 1996
- [15] O. Landolt, A. Mitros, C. Koch, "Visual sensor with resolution enhancement by mechanical vibrations", Proc. 19th Conference on Advanced Research in VLSI, Salt Lake City, March 2001
- [16] X. Arreguit, F.A. van Schaik, F. Bauduin, M. Bidiville, E. Raeber, "A CMOS motion detector system for pointing devices", Proc. International Solid-State Circuits Conference 96, San Francisco, February 1996

Figure captions

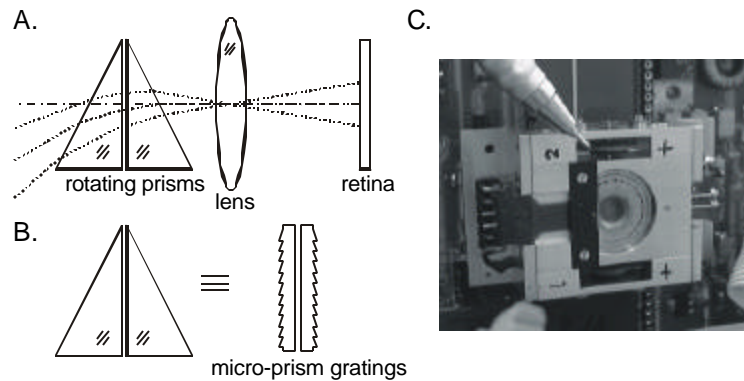


Figure 1.1

A. Light deflection device principle. B. Replacement of conventional prisms by micro-prism gratings. C. Photograph of a mechanical device including the gratings, orientation sensors and motors. The overall device size is 56mm × 34mm × 16mm.

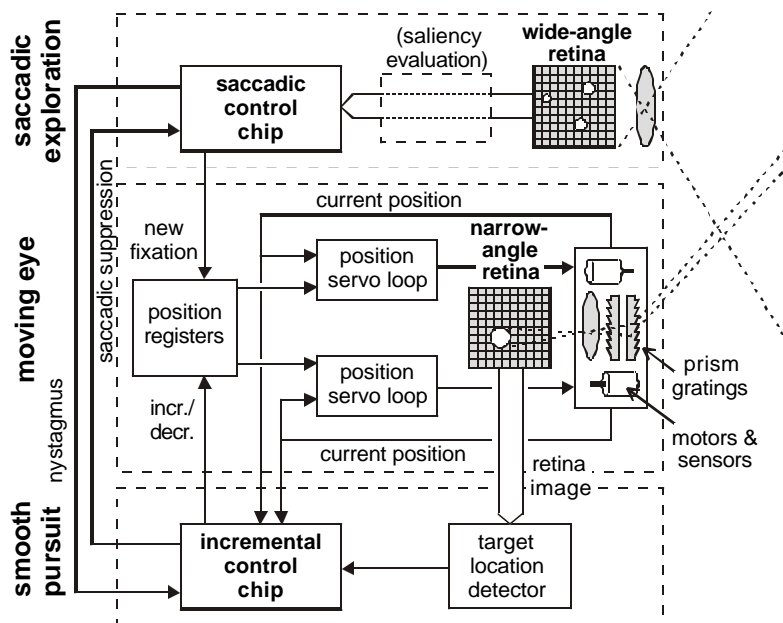


Figure 1.2

Oculo-motor system architecture

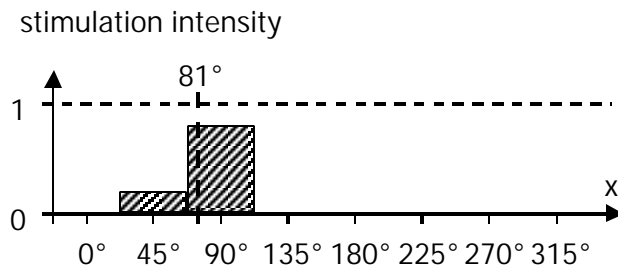


Figure 1.3
Distribution of stimulation intensity encoding an angle of 81° on a map of 8 nodes whose preferred directions are uniformly distributed at intervals of 45°

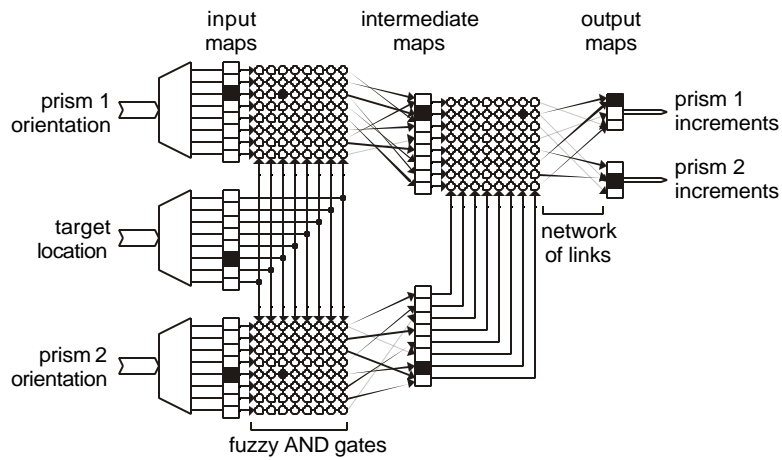


Figure 1.4
Architecture of the incremental control chip

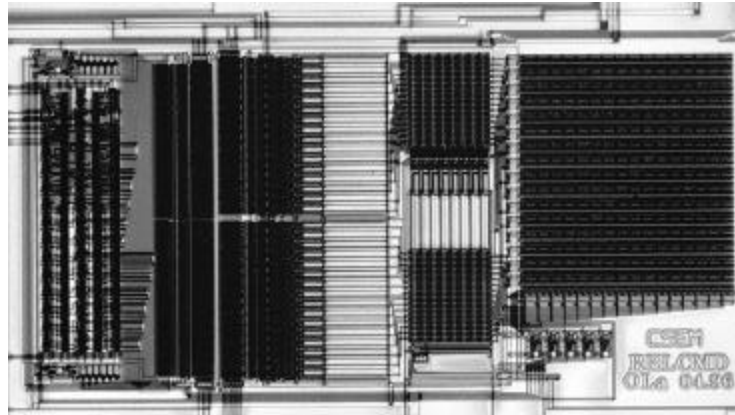


Figure 1.5
Photograph of the incremental control chip. The shown area is 2mm×1mm.

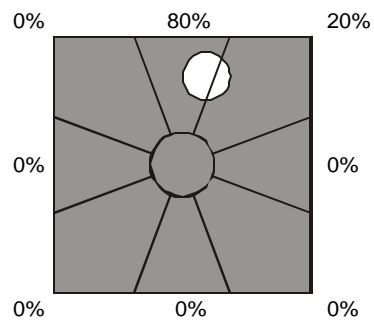


Figure 1.6
Division of the retina surface into radial slices resulting in place coding of the angular position of the target

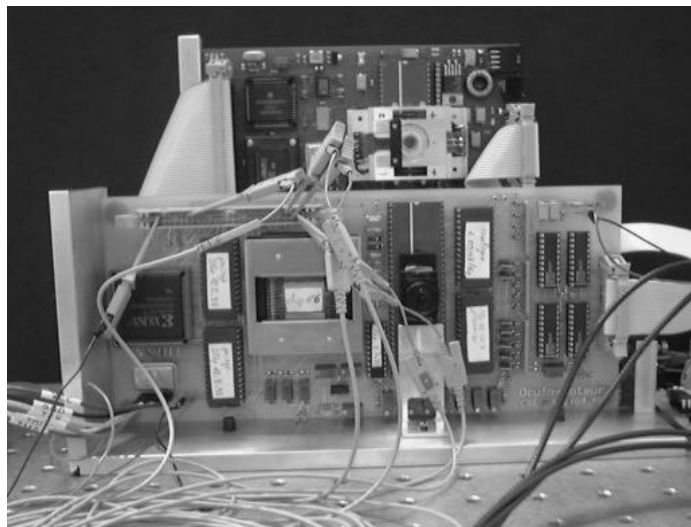


Figure 1.7

Photograph of the oculo-motor system. The length of the largest printed circuit board is about 25cm.



Figure 1.8

Photograph of a device producing circular scanning by spinning a tilted mirror. The device size is 38mm×38mm×38mm.

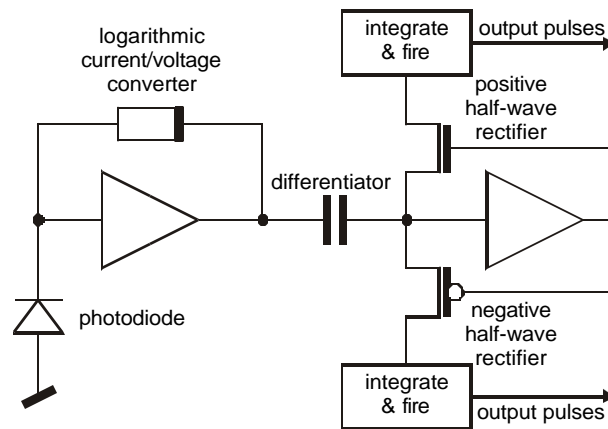
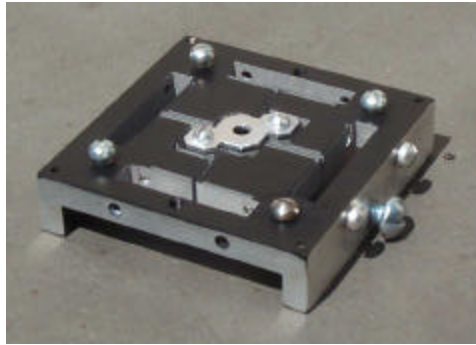


Figure 1.9
 Photograph of a scanning device powered by environmental vibrations.
 The device size is 34.5mm×34.5mm×8.2mm.

Figure 1.10
 Block schematic of a single pixel

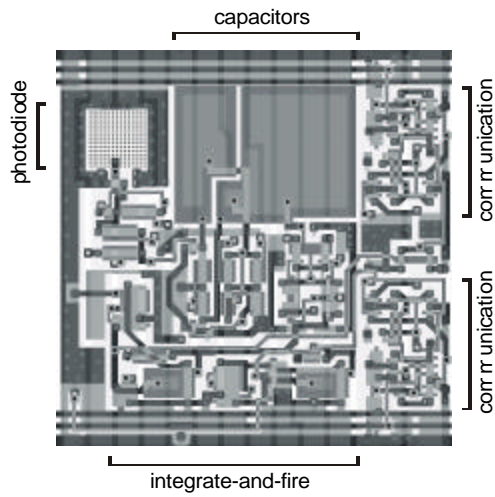


Figure 1.11
Layout of a single pixel ($68.5\mu\text{m} \times 68.5\mu\text{m}$)

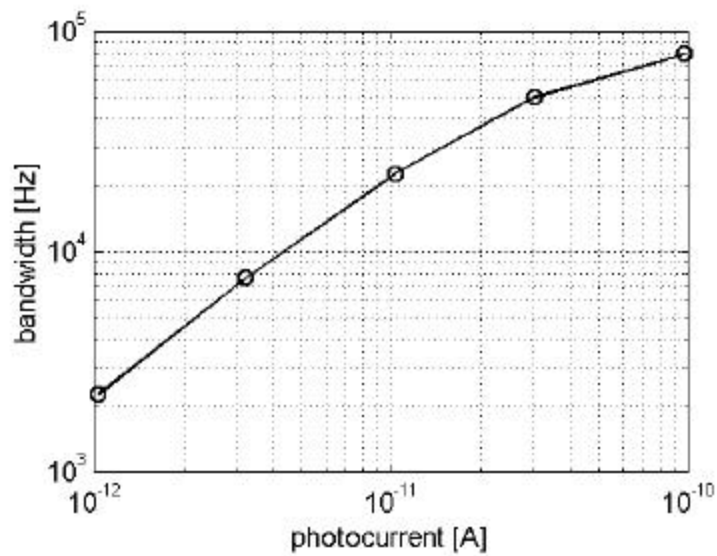


Figure 1.12
Measured bandwidth of the photoreceptor circuit versus photocurrent intensity

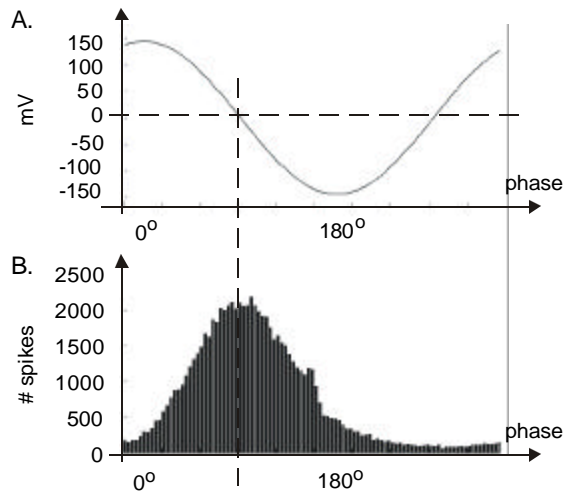


Figure 1.13
 Measured pixel firing histogram with a sinusoidal input. A. Input signal.
 B. Number of spikes versus phase of the input signal.

Tables

visual location	number of hits	average fixation time
A	436	133ms
B	443	88.2ms
C	418	106.4ms
D	434	86.9ms

Table 1.1
 Saccade statistics in the presence of four equally salient points after 3
 minutes of recording