Title
Roadway and Work Crew Conspicuity

Authors
Barton, Joseph E.
Misener, James A.

Publication Date
2000-12-01

CALIFORNIA PATH PROGRAM
INSTITUTE OF TRANSPORTATION STUDIES
UNIVERSITY OF CALIFORNIA, BERKELEY

Roadway and Work Crew Conspicuity

Joesph E. Barton
James A. Misener

California PATH Research Report
UCB-ITS-PRR-2000-23

This work was performed as part of the California PATH Program of the University of California, in cooperation with the State of California Business, Transportation, and Housing Agency, Department of Transportation; and the United States Department of Transportation, Federal Highway Administration.

Report for MOU 328

December 2000

ISSN 1055-1425

MOU 328 Final Report

Roadway and Work Crew Conspicuity

Joseph E. Barton and James A. Misener

Abstract. The work reported here quantitatively addresses the measurement of conspicuity of highway features and Caltrans work zones -- from the perspective of driver detection. The method focused on acquiring and operating on a computational visual signature analysis tool, but it evolved into evaluating the detection process, then selecting and exercising human perception-acquisition models suitable for development into a tool for conspicuity measurement. A composite, quantitative model of conspicuity was developed, verified, and applied to some sample roadside scenes.

Key Words. Conspicuity, Perception, Visual Performance, Human Vision, Vigilance, Pre-attentive, Pre-cognitive, Contrast Model, Work Zones

Executive Summary. Work under this MOU was slated to employ a state-of-the-art visual signature analysis tool as a means to measure and improve conspicuity of human and roadway hazards on California highways. It was believed – and is still believed – that this tool could serve as a powerful, cost-effective and semi-automated method of "virtual prototyping" in which drivers' perception of increased conspicuousness could be gauged. Notional designs and configurations could be simulated with very little investment, under different geometries, color/illumination combinations and ambient lighting conditions. The "best" design which optimizes some combination of cost-effectiveness and safety could then be confidently built and implemented.

Due to other PATH research needs, and lack of availability of an available existing computational vision tool, the work became an in-house effort to develop a conspicuity model of our own. The technique developed was an objective, quantitative algorithm centered on contrast sensitivity, a known factor in conspicuity. The algorithm was tested using a still image of a roadside scene in which a pedestrian was wearing a Caltrans safety jacket. Additional tests were conducted using a mathematically derived pattern with a region of high conspicuity and two versions of the original roadside scene that had been adjusted to demonstrate more and less conspicuity. In all cases the algorithm provided a scoring grid which highlighted the regions of high conspicuity, thus demonstrating that the algorithm is valid.

Therefore, the technique developed under this MOU provides an objective and quantitative method for testing the conspicuity of roadside scenes. While additional modifications to the algorithm would provide increased precision, the work here provides an excellent first pass for those interested in testing the conspicuity of roadside scenes as well as a method for examining how changes in the scene affect conspicuity.

Table of Contents

I. Introduction

A. Background

This project is a follow-on to MOU 284 "Evaluation of Work Crew and Highway Hazard Conspicuity"[1]. The first project encountered institutional roadblocks to completion as originally intended. The project described in this paper was an attempt to accomplish many of the original goals of the first project, but with a model of human vision devoid of the institutional barriers encountered in the first project.

As originally intended, we have attempted to employ a state-of-the-art visual signature analysis tool as a means to measure and improve conspicuity of human and roadway hazards on California highways. It is believed that this tool can serve as a powerful, cost-effective and semi-automated method of "virtual prototyping" in which drivers' perception of increased conspicuousness could be gauged. Notional designs and configurations could be simulated with very little investment, under different geometries, color/illumination combinations and ambient lighting conditions. The "best" design which optimizes some combination of cost-effectiveness and safety could then be confidently built and implemented.

The primary goal of this project was to develop a quantitative method of evaluating roadside conspicuity with the intent to provide a tool that can eventually be implemented as means of making work zones safer.

The first project (MOU 284) encountered difficulty in acquiring the genesis behind the project: a tool developed by the U.S. Army Tank-Automotive Research, Development and Engineering Center (TARDEC) and applied by the TARDEC National Automotive Center (NAC) and the General Motors Research and Development and Safety and Restraint Centers to detect taillight conspicuity. This tool, the NAC Visual Performance Model (VPM), was to be adapted for highway drivers and its use constituted the main thrust of the original project plan. The original legal vehicle to use this tool was to be a pre-arranged Cooperative Research and Development Agreement (CRADA) between PATH and TARDEC, but in time-consuming deliberations, TARDEC CRADA administrators determined not to go that route. Finally, toward the end of 1997, the VPM was supplied to PATH – too late to do substantial work in the planned area. Additional institutional barriers led to the abandonment of the VPM in favor of one synthesized in-house.

This project does not fulfill the currently-addressed need for quick running and believable human driver model inputs for PATH microsimulation activities; the model described in this paper is too computationally-intensive. However, it does fully address the original goals of the first project, focusing in particular on an accurate methodology to assess relative improvements in conspicuity enhancements for Caltrans work zones. Furthermore, the path to developing a deployable tool is less difficult as the model is not encumbered by as many institutional issues.

B. Overview

Two elements of perception were considered in this study human vision- and cognition-based detection models: *acquisition* (defined for here as proximal obstacle or vehicle detection probability P_d at range x) and *tracking* (defined for here as deceleration x' relative to the driver). In *acquisition*, progresssively higher fidelity models in human vision-based target acquisition were shown, with positive and negative aspects described. The work began with the Bailey-Rand (BR) Contrast Model, then progressed to the Doll-Schmieder (DS) Model, and ended with the originally-proposed focus of work, the National Automotive Center-Visual Performance Model (VPM). These models are shown to sequentially yield higher confidence results in the analysis of more specific driver-assist or work zone implementations and scenarios. In *tracking*, the work here is contained to a mathematical model describing longitudinal time to collision-based perception.

Of the three perception-acquisition models investigated – the Bailey Rand (BR) Contrast Model, the Doll-Schmieder Model and Visual Performance Model – the BR model is determined to be the "best" for short-term application into the SmartAHS microsimulation, due to a combination of believability and its low computational complexity. However, a logical progression and continued checking of computational complexity of salient aspects of the other listed candidate models is recommended in a carefully constructed longer term program to gradually build up the human vision aspects of SmartAHS.

[1] Misener, J. A. (1998). *Evaluation of Work Crew and Highway Hazard Conspicuity* (UCB-ITS-PWP-98-14). Richmond, CA: University of California, Partners for Advanced Transit and Highways.

Additionally, incorporation of perception-tracking models is recommended, starting first with the already-explored longitudinal component and progressing to yet-to-be-explored lateral tracking.

Human vision models are applicable to the PATH SmartAHS microsimulation package mainly because of the impending driver-assist research needs. To illustrate how these models could be used, a collision avoidance case was illustrated. In that case, the need to supplement human detection with driver-assist detection technologies in inclement weather was highlighted. To more fully explore this and the potential ramifications of weather and configuration on Caltrans work zone conspicuity, especially in the context of the originally planned scope of this MOU 284, further elaboration on an probably not-for-SmartAHS tool – the Visual Performance Model – is suggested. This is precisely the focus of the follow-on MOU 328.

The fundamental aspect of human vision that we exploit in this study is *Contrast Sensitivity*. The human visual system is tuned to detect not the absolute luminance levels of various areas in a visual scene, but rather the contrast between the luminance levels of adjacent areas. The signal supplied by the eye to the central visual structures does not give equal weight to all regions of the visual scene. Rather, it emphasizes the regions that contain the most information—namely, the regions where there are differences in luminance. Such regions are where our attention is directed when we look out into the world: they are the most *conspicuous*. Contrast sensitivity, then, underlies our ability to detect objects of interest in the visual scene, and also to discern patterns. This capability is of course the result of the way in which the human visual system has evolved. Appendix I, *The Anatomy of Human Sight*, describes the elements and structure of the human visual system that give rise to contrast sensitivity. The nature of light itself has significantly influenced this evolution, and a complete understanding of contrast sensitivity requires in turn an understanding of the *Physics of Sight*, which is taken up in Appendix II.

The material presented in these appendices is not new, and has been presented in several classic references. We do not wish to duplicate the considerable efforts that went into those works here, and therefore have extracted and reassembled portions from them to present a succinct but comprehensive introduction to the subject. Appendix I has been taken from the text *Neuroscience*[2], Chapters 1 2, 5, 10, and 11; and Wandell[3], Chapters 1-5. Appendix II has been taken from *Physics*[4] (Chapters 19, 28, 34, 41, 43, 45, 46, and 48), *The Eye and Visual Optical Instruments*[5] (Chapters 26, 34, and 35) and *Principles of Optics*[6] (Chapter 8). All of the figures presented in these appendices have also been taken from these references. Other references are noted in the end notes to these appendices.

Special thanks are extended to Professors Theodore E. Cohn, Stanley A. Klein, and to PhD Candidate Laura Walker of the Department of Vision Science, University of California at Berkeley, for their generous advice and assistance.

[2] Purves, et al, *Neuroscience*, Sinauer Associates, Inc., Sinderland, MA, 1997.

[3] Wandell, *Foundations of Vision*, Sinauer Associates, 1995, Sunderland, MA.

[4] Halliday, Resnick, and Crane, *Physics*, John Wiley & Sons, Inc., New York, 1992

[5] Smith and Atchison, *The Eye and Visual Optical Instruments*, Cambridge University Press, Cambridge, UK, 1997

[6] Born and Wolf, *Principles of Optics*, Cambridge University Press, Cambridge, UK, 1999

II. Model of the Contrast Sensitivity Function[1]

A. Overview

The description of the human visual system that we presented in Appendix I is based upon anatomical and physcophysical[2] studies of human, primate, and other mammalian visual systems extending over the past 100 years. One of the main conclusions we draw from the results of these studies is that the human visual system is particularly adept and efficient at detecting alternating patterns of light/dark, red/green, and blue/yellow. We infer in turn that this just be a fundamental aspect of what we call "seeing". How well it can detect a particular pattern, i.e., its *contrast sensitivity*, is a function of that pattern's spatial frequency. To model this aspect of the visual system, we proceed as follows:

1. The given scene is decomposed into three sub-images: a black and white, or light/dark image, an image giving just the reds and greens contained in the scene, and one giving just the blues and yellows.
2. Next, we construct a number of filters, each corresponding to a particular spatial frequency and orientation (rotation) about the line of sight.
3. To each sub-image, we apply each filter, and then weight the result according to the corresponding Contrast Sensitivity Function for that image and filter frequency.
4. Each sub-image is then built back up by pooling these filtered images together.
5. The result of these steps, then***, is what one infers*** the visual system is actually able to discern from the original scene. Each of the three rebuilt sub-images obtained in step 4 will have some average brightness, which is just the average of the brightnesses of all the pixels making up the image.
6. Each sub-image is then divided into squares, and an average brightness for each of these squares is calculated in the same manner as it was for the overall scene.
7. The ratio of the average brightness of each square (Step 6) to that of the overall sub-image (Step 5) is calculated.
8. A *Scoring Grid* for each sub-image is then constructed from the ratios obtained in Step 7.
9. Finally, an *Overall Scoring Grid* is constructed by combining the scoring grids for the three sub-images (Step 8).

The most conspicuous object in the original scene will lie in the Scoring Grid square whose ratio is the greatest.

B. Construction of Subimages

The Black/White, Red/Green, and Blue/Yellow sub-images are next calculated according to the relations

$$\begin{aligned} \text{MonScene} &= \frac{\text{Red} + \text{Blue} + \text{Green}}{3}, \\ \text{RGScene} &= \text{Red} - \text{Green}, \\ \text{BYScene} &= \text{Red} + \text{Green} - 2\ \text{Blue}, \end{aligned} \tag{2.1}$$

Since the individual entries of the RGScene and BYScene arrays can fall outside the 0-1 range as a result of this operation, they are scaled such that their values fall within it. The three sub-images are rendered (drawn) in black and white. For the Red/Green sub-image, white indicates all-red, while black indicates all-green. For the Blue/Yellow sub-image, white indicates all-yellow and black indicates all-blue.

C. Construction and Application of Filters

There is evidence to suggest that the human visual system is tuned to the spatial frequencies and rotational orientations listed in **Table 2-1**, and we will therefore use these in constructing our filters.

Spatial Frequency (Cycles per Degree of Visual Angle)	Rotational Orientation (From Horizontal)
$2^{-1} = 0.5000$	0
$2^{-.5} = 0.7071$	$\pi/6$ rad = 30^{o}
$2^{0} = 1.0000$	$2\pi/6$ rad = 60^{o}
$2^{.5} = 1.4142$	$3\pi/6$ rad = 90^{o}
$2^{1} = 2.0000$	$4\pi/6$ rad = 120^{o}
$2^{1.5} = 2.8284$	$5\pi/6$ rad = 150^{o}
$2^{2} = 4.0000$	$6\pi/6$ rad = 180^{o}
$2^{2.5} = 5.6568$	
$2^{3} = 8.0000$	
$2^{3.5} = 11.3137$	
$2^{4} = 16.0000$	
$2^{4.5} = 22.6274$	

Spatial Frequency and Rotational Orientation

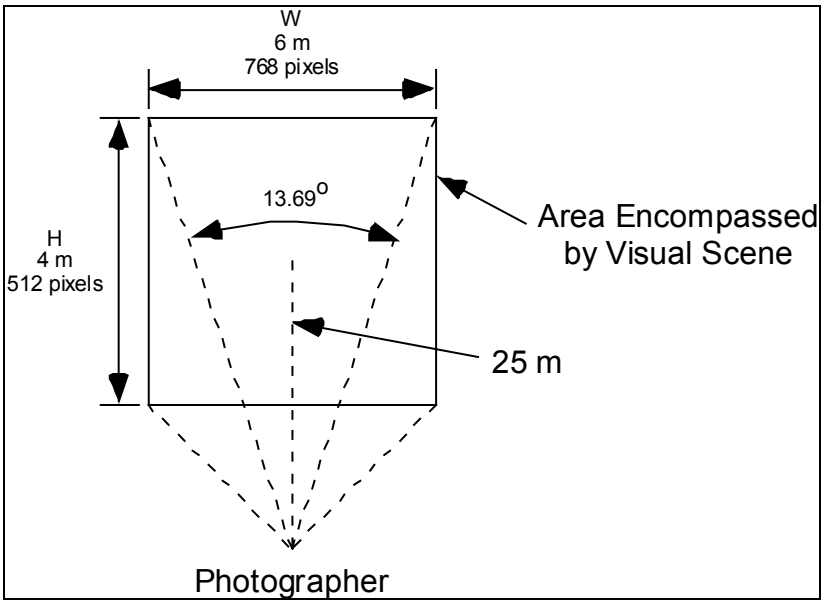
Table 2-1

Figure 2-1 shows the scene (a scanned photograph) that we will be analyzing later. In taking this photograph, the photographer

Original Roadside Scene

Figure 2-1

captured an area whose approximate dimensions are shown in **Figure 2-2**. Using these dimensions, the visual angle subtended by the width of this scene is given by



Dimensions of Scene

Figure 2-2

$$\alpha = \tan^{-1}\left(\frac{W/2}{D}\right) = \tan^{-1}\left(\frac{6/2}{25}\right),$$
$$= 13.69^{o},$$
$$\approx 14^{o}. \tag{2.2}$$

Since the scene is a digitized photograph, its dimensions in both pixels (768 pixels wide by 512 pixels high) and inches (10.667" wide by 7.111" high) are available. Based on these, a square 1^{o} of visual angle to a side would be

$$\frac{768\,\text{pixels}}{13.69^{o}} \approx 56\,\frac{\text{pixels}}{\text{degree}} \tag{2.3}$$

(or .7778") to a side.

For this model we use a variant of the *Gabor Filter*, as described by Watson and Ahumada[3]. Our filter will extend over this 1^{o} square[4], and for this purpose we will map the square onto a Cartesian coordinate frame. The x and y coordinates of this frame will each be normalized between –1 and +1 and will take on intermediate values corresponding to each pixel within the square. Thus we will have an x- and y array, each having 56 elements whose values are equally spaced between –1 and +1. Since it is desirable to have a coordinate corresponding to zero, we will increase our arrays by one, so they will each have 57 elements. The result is a two-dimensional, 57 x 57 element filter. From these coordinates, we can then obtain a set of rotated coordinates x_{rot} and y_{rot} associated

with any counter-clockwise angular rotation ψ (which takes values according to the second column in **Table 2-1**) using the standard coordinate transformation

$$x_{rot} = x\cos\psi + y\sin\psi,$$
$$y_{rot} = -x\sin\psi + y\cos\psi. \tag{2.4}$$

The filter itself takes the form

$$\text{Filter} = e^{-\frac{x_{rot}^2 + y_{rot}^2}{\lambda^2}} \cdot \cos\left(\frac{2\pi\sigma}{dW} x_{rot}\right), \tag{2.5}$$

where dW is the dimension of our 1^{o} square, σ is the spatial frequency (which takes values according to the first column in **Table 2-1**), and where

$$\lambda = \frac{\rho}{\sigma / dW}, \qquad \rho = \frac{3\sqrt{\log_{10}(2)}}{\pi}. \tag{2.6}$$

The basis for these parameters is described in the Watson and Ahumada article referenced above. The units of the cosine's argument in Equ. (2.5) are

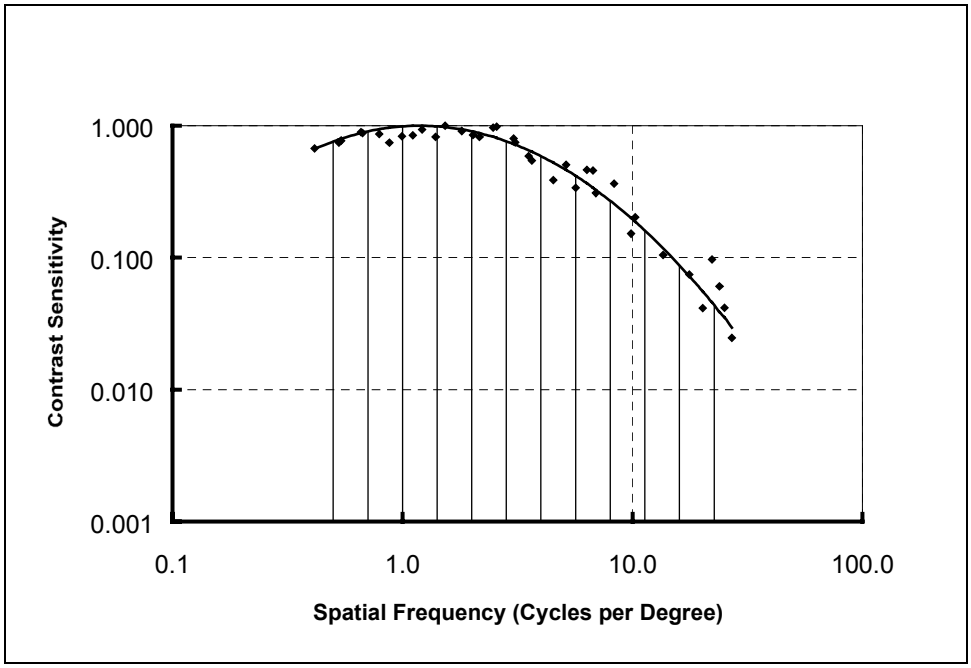
$$\frac{2\pi \frac{\text{radians}}{\text{cycle}} \cdot \sigma \frac{\text{cycles}}{\text{visual degree}}}{dW \frac{\text{inches}}{\text{visual degree}}} x_{rot} \text{ inches} = \frac{2\pi\sigma}{dW} x_{rot} \text{ radians}, \tag{2.7}$$

as required.

The filters are then applied by convolving them with the visual scene. According to the Convolution Theorem, this is equivalent to taking the Fourier Transforms of the Filter and the Scene, multiplying them together, and then taking the Inverse Fourier Transform. This is the underlying algorithim for the Matlab[5] function "**conv2**" (**conv2** performs a two dimensional convolution, which we require since here the filter and scene are functions of the two spatial variables x and y), that we employ in our model. According to **Table 2-1**, we have $12 \times 7 = 84$ different filters, so our algorithm performs 84 convolutions[6].

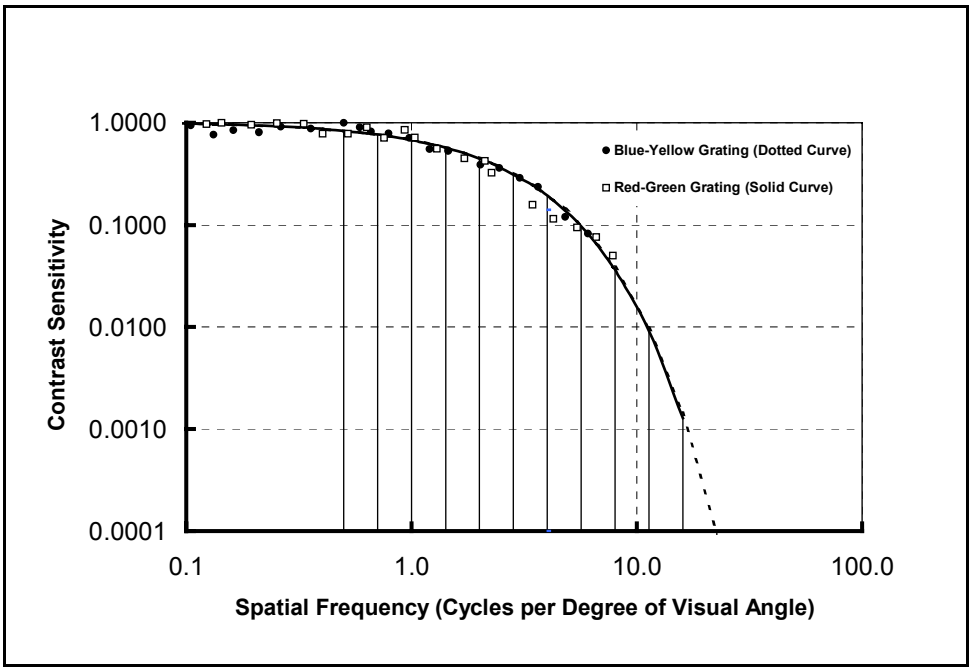
D. Weighting and Pooling of Filtered Images

The Contrast Sensitivity Functions used in this analysis have been taken from Mullen[7], and are shown in **Figures 2-3** and **2-4**.



Monochromatic Contrast Sensitivity Function

Figure 2-3



Chromatic Contrast Sensitivity Functions

Figure 2-4

Each filtered sub-image is weighted by the value of the corresponding contrast sensitivity function for that spatial frequency. These are then *pooled* according to the relation

$$\text{Pooled Sub-Image}_i = \left(\sum_{\sigma} \left[\text{CSF}_i(\sigma) \;\; \text{Filtered Sub-Image}_i(\sigma) \right]^{\frac{1}{p}} \right)^p , \tag{2.8}$$

where

$$i = \begin{cases} 1 & \text{Monochrome Sub-Image} \\ 2 & \text{Red/Green Sub-Image} \\ 3 & \text{Blue/Yellow Sub-Image} \end{cases} , \tag{2.9}$$

and where $CSF_i(\sigma)$ is the value of the corresponding Contrast Sensitivity Function at the spatial frequency σ, and p is the *pooling factor*, here taken to be 4. The three pooled sub-images represent the visual information actually taken in at the retina and sent back to the central visual centers in the brain.

E. Identification of Conspicuous Elements in the Visual Scene

The scenes we will be dealing with are *bitmapped images*. Such images are represented in Matlab as three two-dimensional arrays (y pixels high by x pixels wide) overlayed on top of one-another. The elements of each array are unsigned, 8-bit binary integers. Such a number can range from 0 to 11111111 binary, or 0 to 255 decimal. The first array indicates the amount of red each pixel in the final image contains, and we refer to it simply as *R*. The second array (*G*) indicates the amount of green, and the third array (*B*) the amount of blue. Thus, each pixel can have 256 gradations of red, green, and blue. A value of zero indicates no red (or green or blue), and a value of 255 indicates the brightest or most luminant shade of red (or green or blue). If all three values associated with a particular point are zero, its color would be black (the absence of all color), and if they were all 255, its color would be white (the presence of all color). If a point's three values were all equal but between 0 and 255, its color would be a shade of gray ranging from black to white. The brightness of a particular point in the picture can thus be obtained by averaging its three *RGB values*. The brightness of the entire scene can be obtained in the same way, by averaging all the values of all the points making it up, and similarly for any portion of the scene. Such a file is rendered in Matlab with the "**image**" command. Since Matlab cannot perform arithmetic operations on unsigned, 8-bit integers, these arrays are converted to arrays whose elements consist of real numbers ranging from zero to

one. Matlab can accommodate such arrays, and internally converts them back to unsigned, 8-bit integer arrays when it renders them.

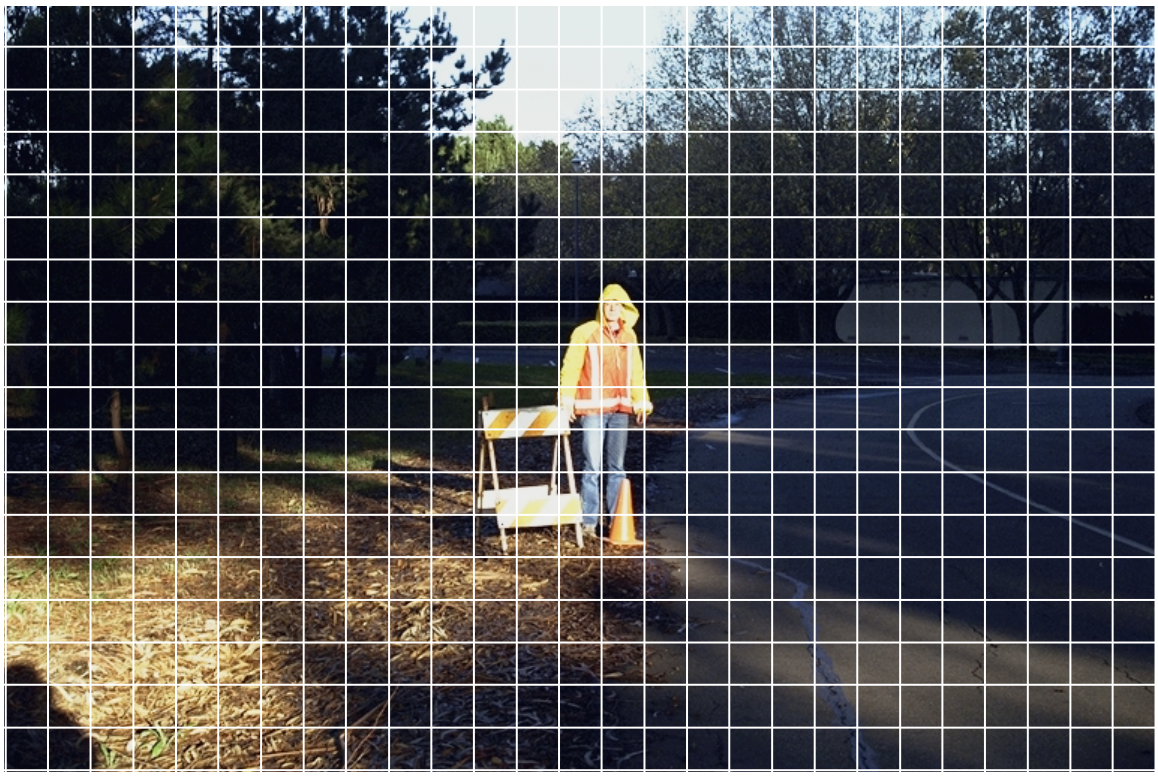
In **Figure 2-5** we've overlayed a grid of squares onto our original roadside scene. We'll define the Contrast Ratio[8] of the i^{th} square as

$$\text{Contrast Ratio}_{\text{Square i}} = \frac{\text{Brightness}_{\text{Square i}}}{\text{Brightness}_{\text{Entire Scene}}}. \tag{2.10}$$

The Contrast Ratio can take on values from zero to infinity. A value of one indicates that a particular square is as bright as the overall scene, and hence does not contrast with it (is not conspicuous). A Contrast Ratio near zero indicates that a square is darker than the overall scene, while a Contrast Ratio larger than one indicates the square is brighter. The farther a Contrast Ratio is from one, the more it contrasts with the surrounding scene and the more conspicuous it is. We'd like to construct from this array of Contrast Ratios a *Scoring Grid*, similar in appearance to that of **Figure 2-6**, that we can quickly scan to identify the most conspicuous elements of a particular scene. The lighter an individual square, the more conspicuous it is, and the darker a square, the less conspicuous it is. We can't accomplish this directly with the Contrast Ratios we calculated above, but we can develop from them a *Modified Contrast Ratio*
to arrive at the desired result. Let CR be our original array of Contrast Ratios. We'll first compute an "intermediate" array CR' from

$$CR' = |CR - 1| \tag{2.11}$$

Looking first $CR - 1$, we see that $-1 \le CR - 1 \le \infty$, and $CR - 1 = 0$ corresponds to no contrast. Thus, $0 \le CR' \le \infty$ and $CR' = 0$ corresponds to no contrast. At this point we've lost one piece of information: CR' cannot tell us whether a particular square is lighter or darker than the overall scene, only that it is different. Since we're primarily concerned with the difference, though, we're willing to sacrifice this. We can still not render CR', since it is not within the range of zero to one. We can, however, calculate from it a normalized array CR_{norm} that does fall within this range, according to the relation



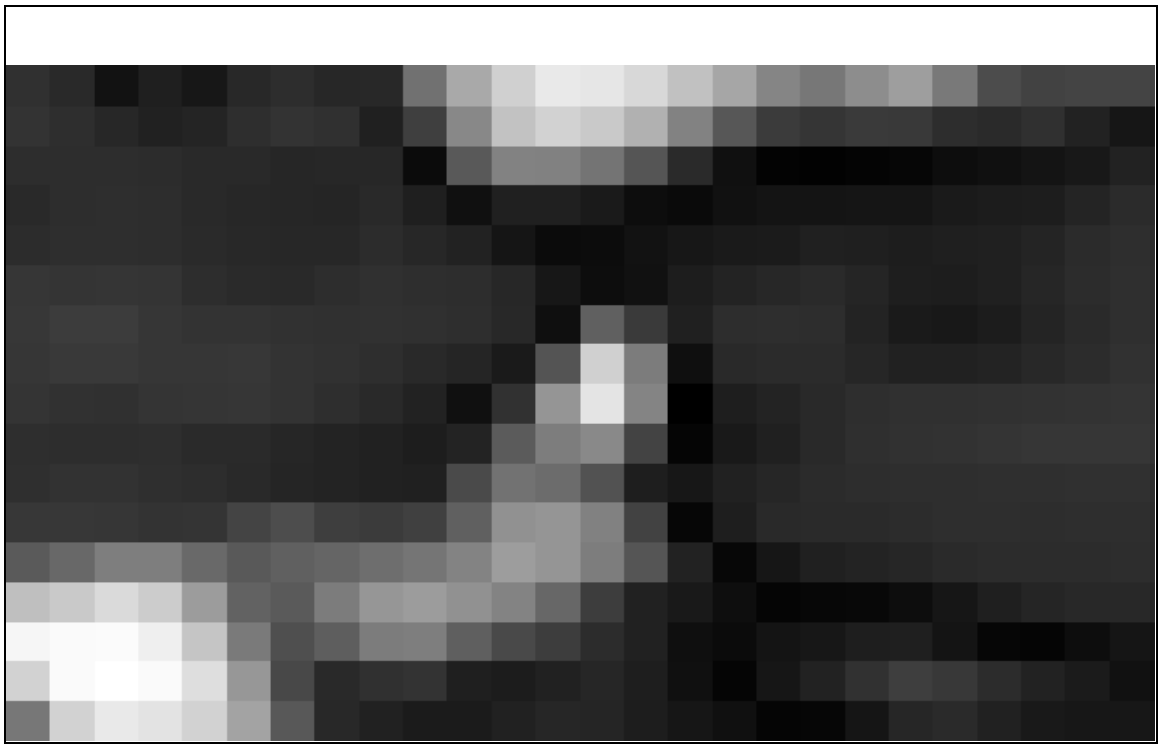
Subdivided Scene

Figure 2-5

$$CR_{norm} = \frac{CR' - \min(CR')}{\max\left[CR' - \min(CR')\right]} . \tag{2.12}$$

In words, we first shifted CR' so that its smallest element is zero, if it wasn't already [the numerator of Equ. (2.12)]. We then divided by the maximum element of this resulting array so that now the largest element is one [the denominator of Equ. (2.12)].

The final step in constructing the scoring grid is to make all the elements associated with a particular square equal to CR_{norm} for that square. The squares will render as varying shades of gray, as shown in **Figure 2-6**. There will always be at least on completely black square, corresponding to the portion of the overall scene having the *least* [not necessarily zero, because of Equ. (2.12)] contrast. Likewise, there will always be at least one completely white square, corresponding to the portion of the overall scene having the greatest contrast.



Scoring Grid

Figure 2-6

Though the Scoring Grid allows us to quickly identify the most conspicuous elements in a particular scene, it is not without its shortcomings:

1. As we noted, we can no longer tell whether an element is conspicuous because it is much darker or much brighter than the overall scene.
2. The calculation of CR_{norm} carries with it the potential to greatly compress the contrast data, as a result of the need to scale CR'. We can identify the most conspicuous element in a scene, but we don't know how *greatly* it contrasts with the scene.
3. For this reason, we can not in general compare the Scoring Grids for two different scenes and infer that the most conspicuous element in one possesses more or less contrast than the most conspicuous element in the other. This would arise, for instance, if we altered an element in a particular scene in various ways, and then sought to compare the resulting Scoring Grids with the goal of identifying the alteration that gave the greatest contrast. To accomplish this, we'd have to normalize the CR' arrays *together*, taking the minima and maxima across all the arrays for use in Equ. (2.12). (This, then is not an insurmountable problem, it just takes more effort to address.)

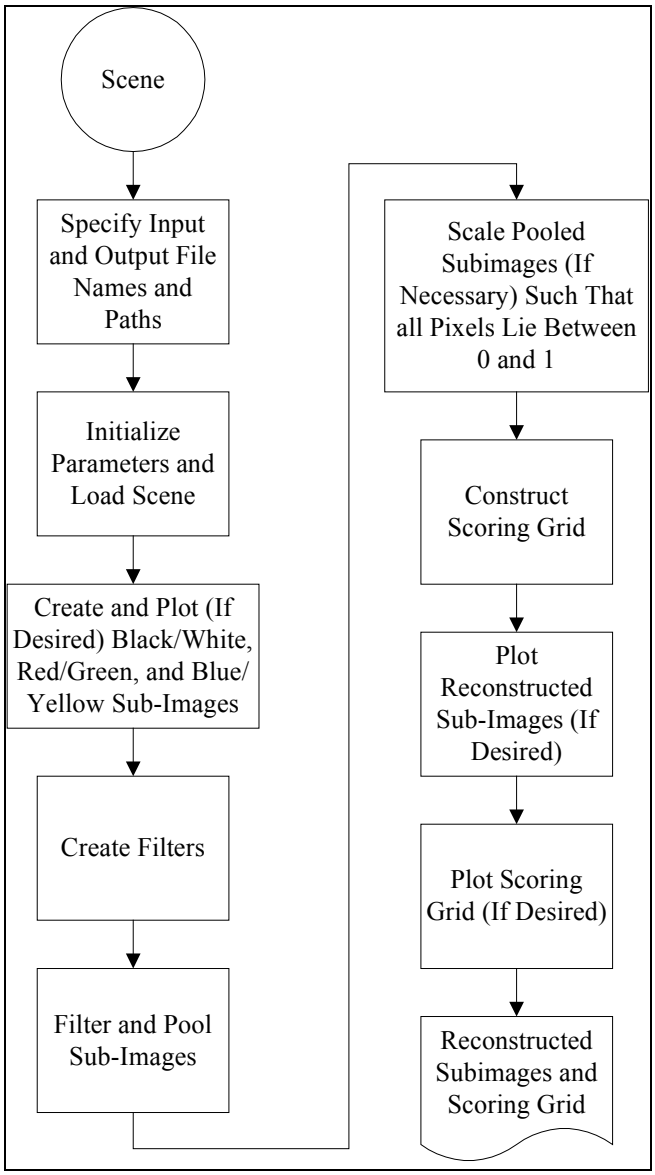
We'll construct a scoring grid for each of the three sub-images, and then combine these into an overall scoring grid, which

we'll use to identify the most conspicuous element in the scene. Because of the data compression problem mentioned in item 2 above, we'll normalize the three sub-images together, as described in item 3, and then take their average to arrive at the overall scoring grid.

Unlike the filters that we described in **Section C**, there is no known physiological basis for the Scoring Grid concept presented here. We have created it with the goal of methodically and consistently picking out the most conspicuous element(s) in a given scene. The object of interest can occupy any portion of a visual scene—from a very small part to a very large part. If we size our squares too small our scoring grid may identify many different "most conspicuous objects" scattered throughout the scene. If we size the squares too large, the most conspicuous object may get "averaged" into the background when the intensity of the square containing it is calculated. Of course, we know from our own experience that the human visual system can usually pick the pertinent object out with ease. This suggests that if anything like a scoring grid is actually being implemented in the human visual system, then it must be capable of varying the square size until some optimization criteria is met. Here we'll assume that this adjustment process has already occurred and we'll manually set the square size appropriate to the scene we're analyzing. (The development of an algorithm to do this would be a logical and desirable extension to the present effort, but is beyond the scope of our currently available resources.)

F. Computer Algorithm of the CSF Model

Our model of the Contrast Sensitivity Function has been implemented via the Matlab file **CSF01**, which is listed in Appendix 3. Its organization is described by the flow chart shown in **Figure 2-7**. The variables used by CSF01 fall into three categories: Main Variables, Progress Variables, and Temporary Variables. These are also listed and described in the three tables in **Appendix 3**.



Flow Chart of CSF01

Figure 2-7

---

[1] The model that we present in Chapters II – IV is based on the physiological and theoretical concepts presented in Appendices I and II. Though these appendices are not a prerequisite to the chapters, the reader my nonetheless wish to read them in order to gain a more comprehensive understanding of the model's basis.

[2] Anatomical studies are, as one might guess, *invasive*. They typically involve the sacrifice and dissection a laboratory animal, or the dissection of a human who has donated his or her remains for this purpose. Psychophysical experiments are *noninvasive* experiments, in which a test subject is presented with a stimulus and his or her response observed. Psychophysical experiments can be and have been performed on humans chimpanzees, cats, and many other types of living things. We could even think of performing such an experiment on an individual cell.

[3] A.B. Watson and A.J. Ahumada, *Model of Human Visual-Motion Sensing*, Journal of the Optical Society of America, 1985, Volume 2, No.2, pp. 328-329.

[4] We have chosen such a small dimension for our filter because it is representative of *foveal vision*. An alternative, equally valid, approach would be to consider *extra-foveal vision*, and to employ correspondingly larger filters if various sizes. In selecting the former case we have implicitly assumed that our observor is *attentive*, and looking "forward" towards the

visual scene. A *distracted* observor would first have to be drawn to focus on the scene by his or her peripheral, or extra-foveal vision. This would be a logical follow-on to this analysis.

[5] The Matlab computing environment has proven to be a particularly convenient one for analyses of this type, and we have therefore implemented this model in it. The Matlab programming language is (with minor modification) the C-programming language.

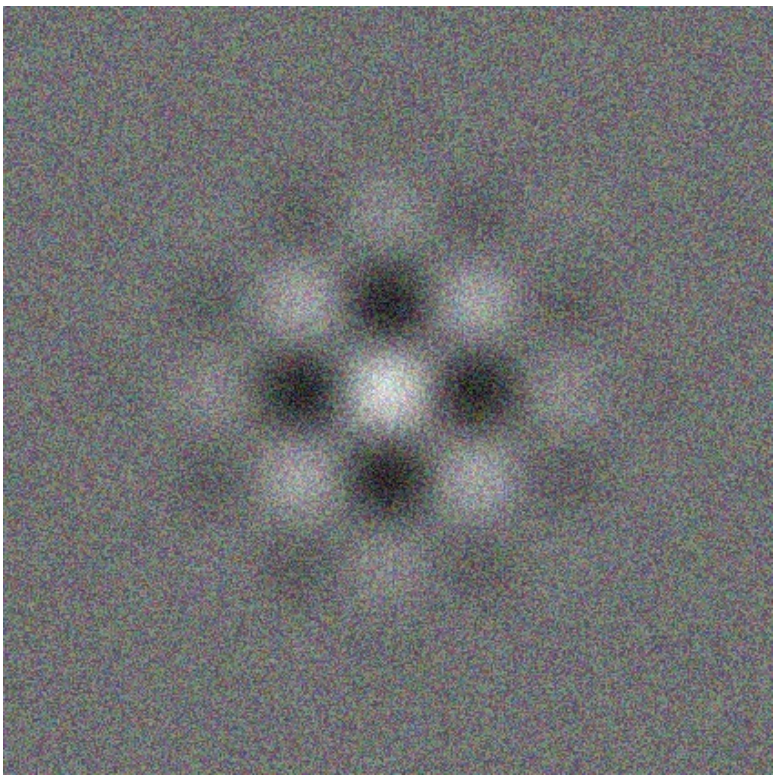
[6] Applying **conv2** in this case is a time-consuming operation. A personal computer with an 890 MHz processor takes approximately one hour to perform these 84 convolutions. Time did not permit researching this subject to identify faster convolution algorithms. Again, this should be taken up as part of a follow-on study.

[7] K. T. Mullen, *The Contrast Sensitivity of Human Colour Vision to Red-Green and Blue-Yellow Chromatic Gratings*, Journal of Physiology, 1985, Vol. 359, pp. 381-400

[8] Here we're comparing each square with the *entire* scene. As with our selection of filter size, we could, alternatively, compare each square with the squares in its immediate vicinity. Our rationale here is that a human observor will focus first on the most conspicuous element in the entire scene, and then scan the rest of the scene (conditions permiting), focusing on any other elements that stand out from their immediate surroundings. We want our target to be that first element the observor's attention is drawn to.

III. Verification of the Model

Before turning our attention to the roadside scene of **Figure 2-1**, we'll first apply our algorithm to the scene shown in **Figure 3-1**. This consists of a monochrome, sinusoidally varying pattern that is damped exponentially in both the horizontal and vertical



Gabor Pattern Superimposed on a Uniform Field of Noise
Figure 3-1

directions, superimposed on a uniform field of "background noise". It was constructed based on the dimensions shown in **Figure 3-2.** The pattern is a variant of a two-dimensional Gabor Function (described in Watson and Ahumada article referenced in the previous section), given by

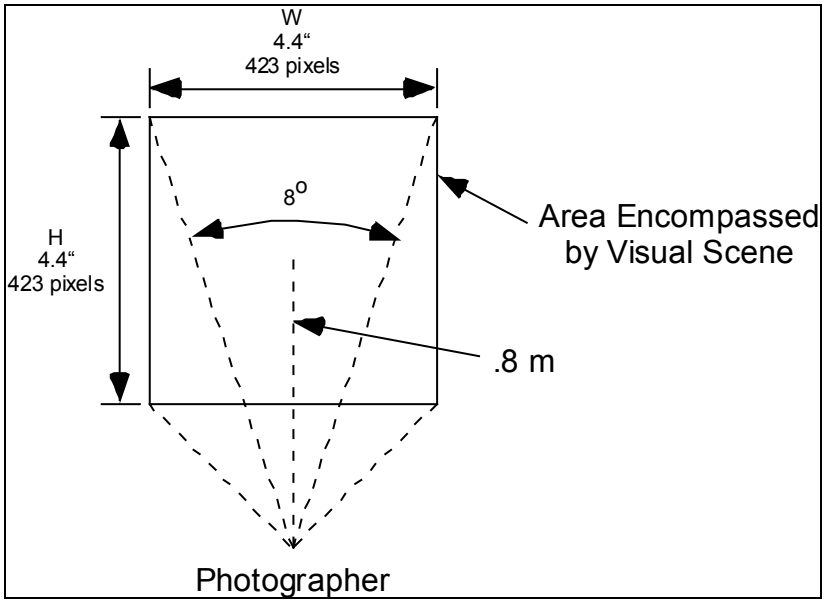
$$G(x,y)=I\left\{e^{-\left(\frac{x}{\lambda}\right)^2}\cos\left[2\pi\left(\frac{\sigma}{dW}\right)x\right]\right\}\left\{e^{-\left(\frac{y}{\lambda}\right)^2}\cos\left[2\pi\left(\frac{\sigma}{dW}\right)y\right]\right\}, \tag{3.1}$$

where the amplitude I and the spatial frequency σ are chosen to be

$$I=1,$$
$$\sigma=.5\frac{\text{Cycles}}{\text{Degree of Visual Angle}}, \tag{3.2}$$

and where ρ and λ are as defined in the previous section. As before, dW is the dimension (in this case 53 pixels) of a square 1^{o} of visual angle to a side, as viewed on a Personal Computer monitor by an observor approximately .8 meters in front of it. x and y (not to be confused with x and y as used in the previous section) are the coordinates (also in pixels) associated with the scene, adjusted so that the point (0,0) corresponds to its center. This construction has been implemented in the Matlab file CSF02.m, which is listed in **Appendix 3**.

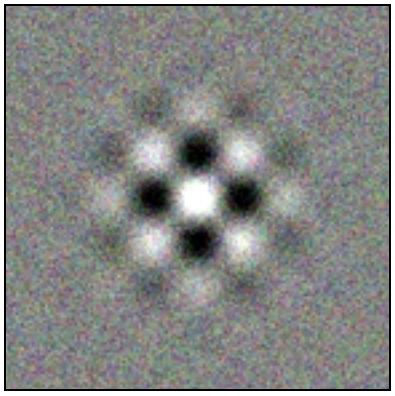
Because the form of this pattern corresponds to that of our filters, we should very nearly recover the monochrome sub-image in our analysis. [We won't recover it exactly because the noise will be filtered out. Note too that because of the way we deconstructed the



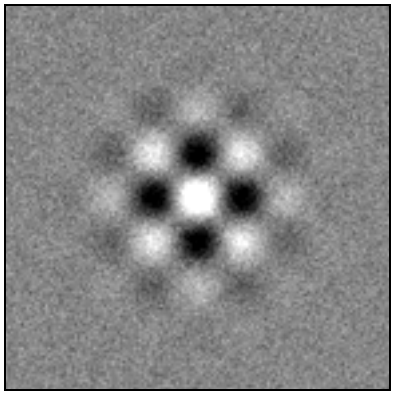
Dimensions of Scene

Figure 3-2

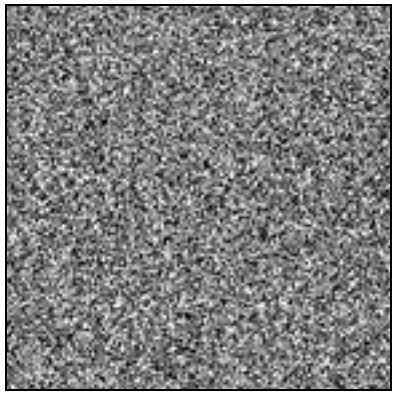
original image—into B/W, R/G, and B/Y sub-images according to Equs. (4.12), it is not possible to combine their filtered and reconstructed counterparts together to arrive at an overall reconstructed image.] As we can readily see, the area of greatest contrast lies at the center of the figure. This of course corresponds to the central peak of our pattern. The adjacent peaks exhibit progressively diminished contrast, and our Scoring Grids should reflect this. Application of our model gives the results shown in **Figure 3-3.** The first row shows the B/W, R/G, and B/Y sub-images. Note that the R/B and B/Y Sub-Images show no pattern. This is because a monochrome input pattern, such as we have, is added to each of the R, G, and B arrays, and then subtracted out [per Equ. 4.12)] when the sub-images are created. It only shows up, then, in the B/W sub-image. The second row shows the filtered, pooled, and reconstructed sub-images, and the third row shows the Scoring Grids. From the Scoring Grids we see that our predictions are confirmed: we have recovered the original monochrome sub-image without the noise, and according to both the Bl;ack and White and Overall Scoring Grids the most conspicuous part of the scene lies at the very center, with progressively less conspicuous areas fanning out from it and located at the sites of the subsequent Gabor peaks.



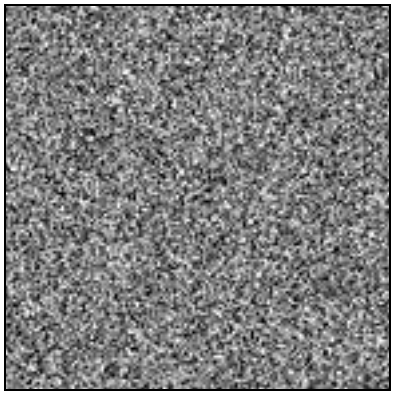
Original Scene



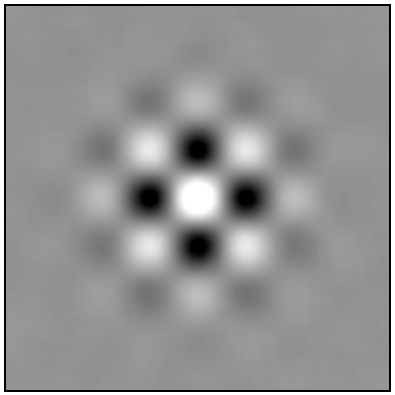
B/W Sub-Image



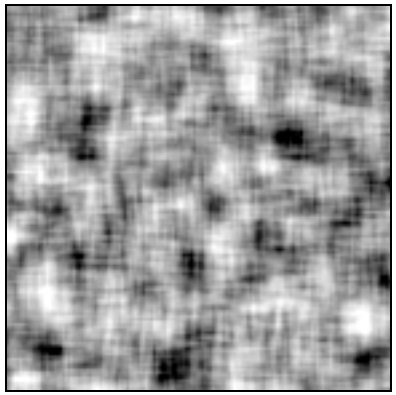
R/G Sub-Image



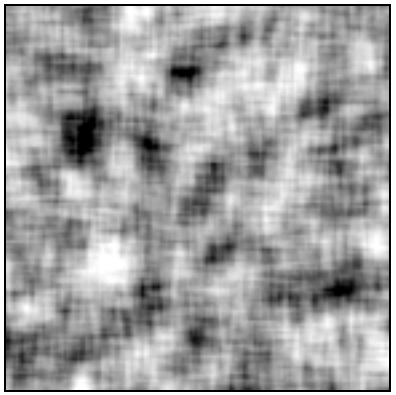
B/Y Sub-Image



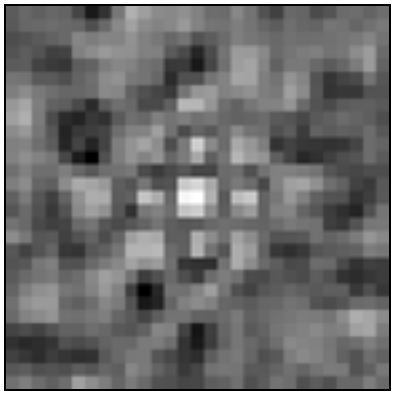
Reconstructed B/W Sub-Image



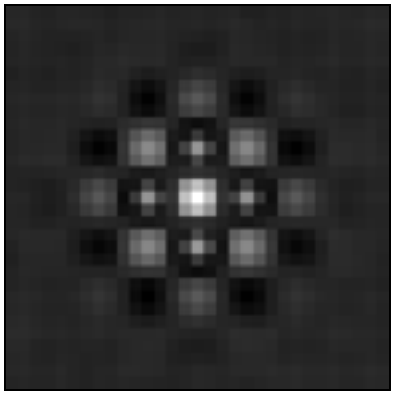
Reconstructed R/G Sub-Image



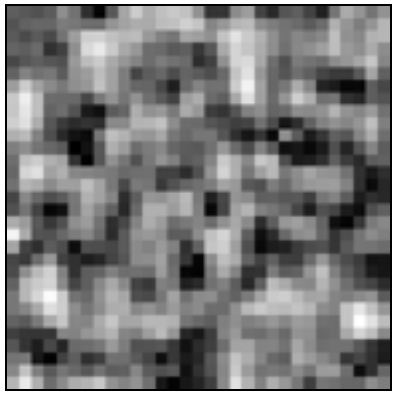
Reconstructed B/Y Sub-Image



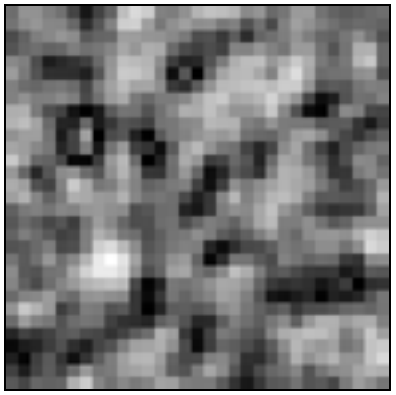
Overall Scoring Grid



B/W Scoring Grid



R/G Scoring Grid



B/Y Scoring Grid

Analysis of a Two-Dimensional Gabor Function

Figure 3-3

IV A Roadside Scene

The roadside scene of **Figure 2-1** is analyzed in the same manner as the Gabor Function of the previous section, with the result shown in **Figure 4-3** on the next page. From the Scoring Grid, we see that the workman's vest is the most conspicuous element in the scene, as we would expect. We'll repeat this analysis, but this time we'll introduce a bright lime-green circle into the scene, as shown in **Figure 4-1**. We now obtain the result shown in **Figure 4-4**. The scoring grids indicate that the circle is more conspicuous than the

workman's vest, a result we can "sense" just by glancing quickly at **Figure 4-1**. We can get a better idea of why this is by looking at the individual scoring grids. The circle is considerably brighter than the vest in both the Red/Green and Blue/Yellow sub-images. (Neither it nor the vest seem to be able to compete with the sky in the Black/White sub-image, and here they appear to be equally bright.) One insight into making an object more conspicuous, then, is to increase its conspicuity in all three of the visual "channels"—the Black/White, Red/Green, and Blue/Yellow. Lime-green appears to do a better job of this than orange.

Going in the other direction, we'll make the vest the same color as the shaded grass immediately behind the workman. This case is shown in **Figure 4-2**, and gives the result shown in **Figure 4-5**. Now the workman almost disappears from the scene, and the most conspicuous elements in the scene become the sky above the workman and the lighted ground in front and to the left of him. Again, this result is easily confirmed by glancing quickly at **Figure 4-2**.

It's interesting, though on second thought not surprising, that the sky competes so effectively for our attention. It is for all intents and purposes a light source, while all the other objects in the scene merely reflect the light it is generating. This again offers us an insight into how to make objects more conspicuous—by having the workman's vest itself generate light in some fashion, or by making it out of fluorescent material.

A More Conspicuous Target
Figure 4-1

A Less Conspicuous Target
Figure 4-2

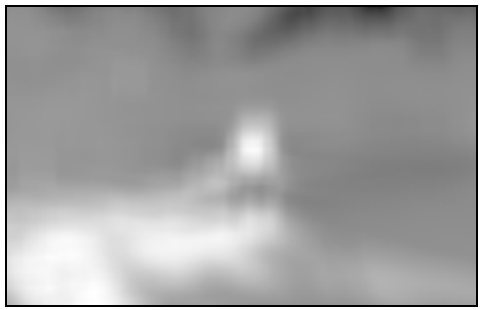
Original Scene

B/W Sub-Image

R/G Sub-Image

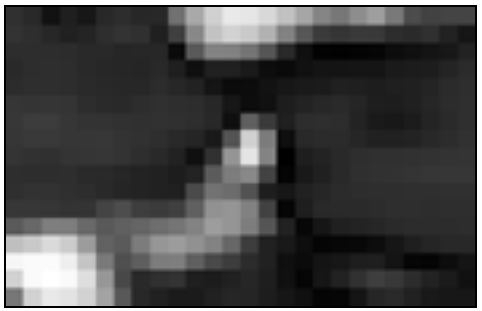
B/Y Sub-Image

Filtered B/W Sub-Image

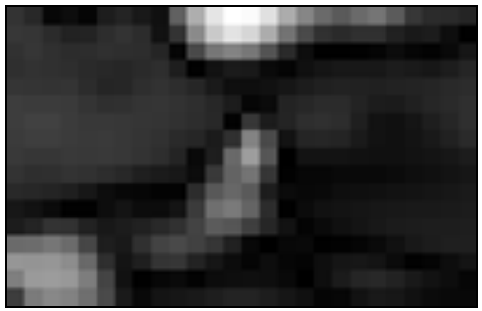


Filtered R/G Sub-Image

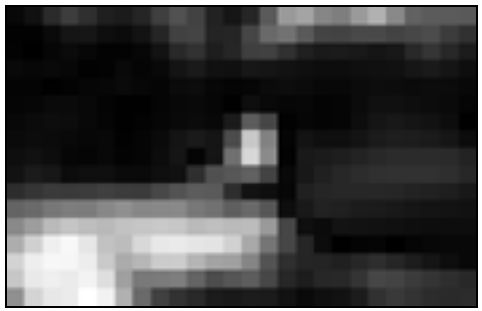
Filtered B/Y Sub-Image



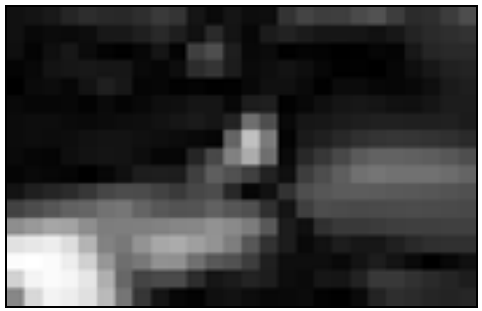
Overall Scoring Grid



B/W Scoring Grid



R/G Scoring Grid



B/Y Scoring Grid

Original Roadside Scene

Figure 4-3

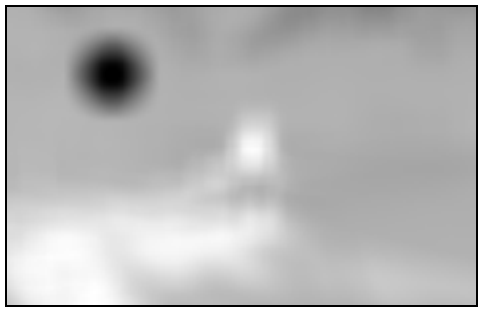
Original Scene

B/W Sub-Image

R/G Sub-Image

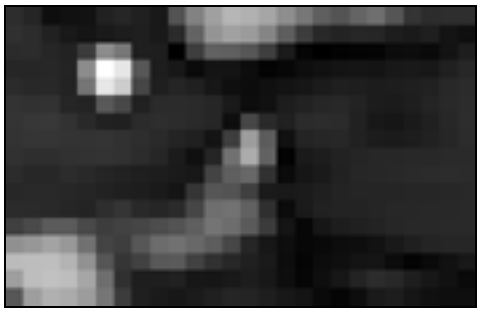
B/Y Sub-Image

Filtered B/W Sub-Image

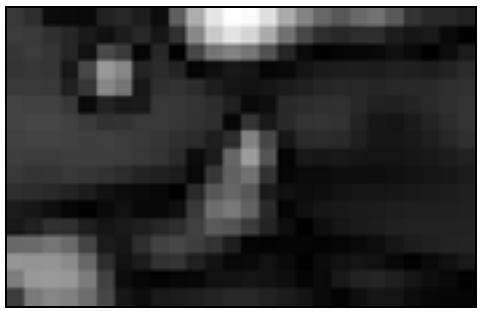


Filtered R/G Sub-Image

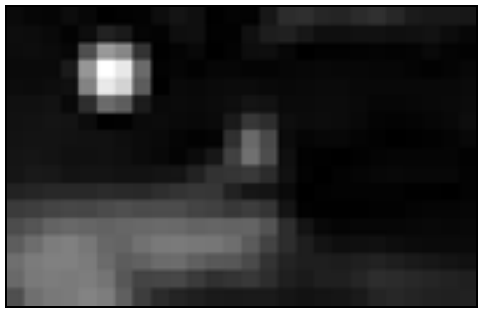
Filtered B/Y Sub-Image



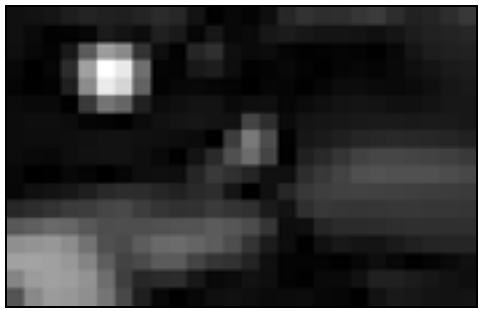
Overall Scoring Grid



B/W Scoring Grid



R/G Scoring Grid



B/Y Scoring Grid

A More Conspicuous Target

Figure 4-4

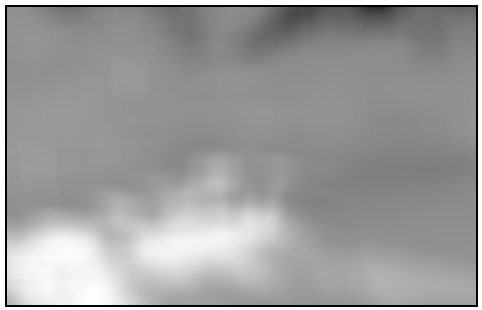
Original Scene

B/W Sub-Image

R/G Sub-Image

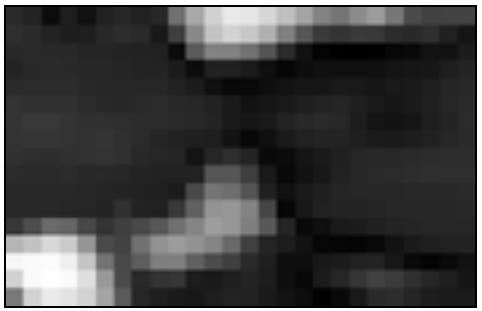
B/Y Sub-Image

Filtered B/W Sub-Image

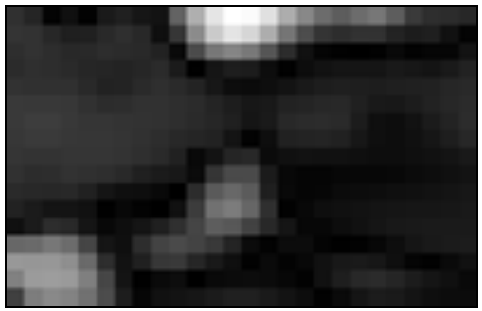


Filtered R/G Sub-Image

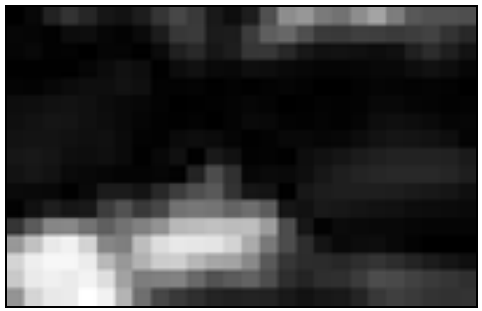
Filtered B/Y Sub-Image



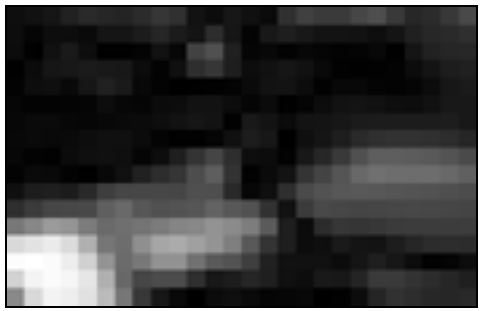
Overall Scoring Grid



B/W Scoring Grid



R/G Scoring Grid



B/Y Scoring Grid

A Less Conspicuous Target

Figure 4-5

V. Conclusions and Next Steps

We have developed an algorithm to identify the conspicuous elements in a visual scene, based on the human visual system's contrast sensitivity function; and verified its performance using a photograph of an roadway construction scene. It is implemented as a Matlab "M-File" in the Matlab computing environment. In its present form the M-File can read a digitized photograph of a construction scene and identify the most conspicuous elements in it. A "scoring grid" has been devised to quickly and easily view the results of the computations. Actual, numerical values associated with the conspicuity of any element in the scene are also available for additional analysis.

The work to date should only be viewed as a first step. A number of refinements and extensions are possible, as discussed in the previous sections, and summarized here:

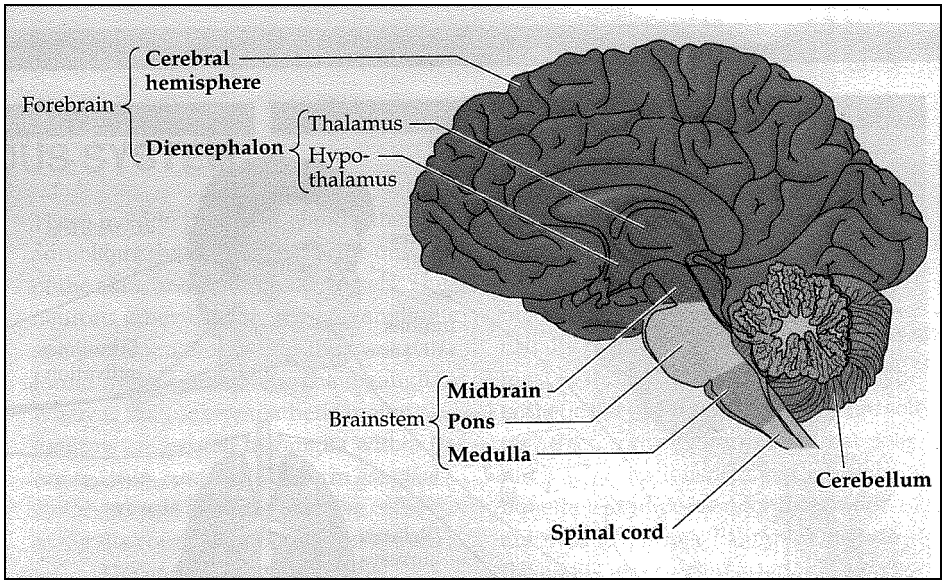
1. Here we assumed an "attentive" driver, and modeled foveal vision. By increasing the size of our filters and adding additional logic we can model extra-foveal (peripheral) vision, and test the ability of various visual schemes to alert inattentive drivers.
2. In quantifying conspicuity, we have compared the elements in a visual scene against a background consisting of the entire scene. This represents an observor's "first glance" at a visual scene, where we assume the driver will focus on the most conspicuous element overall. It would also be desirable to model the observor's "second pass" at the scene, in which the driver quickly scans the scene for other elements that stand out from their *immediate* surroundings. Depending upon what else is in the scene, this activity could enhance or detract from the driver's efforts.
3. Investigate the effect of elements that are made out of fluorescent material or generate light in some fashion.
4. Here we limited our attention to still scenes. *Contrast Sensitivity* also includes sensitivity to movement, which is modeled in a manner quite similar to the still case. By incorporating sensitivity to movement, we can then assess the conspicuity of schemes that add movement (or a sense of movement) to important elements in the scene.
5. Work has been done recently at the NASA-Ames National Laboratory to develop a "spatial standard observer". In particular, this observer or incorporates a "Tyler plus Oblique" contrast sensitivity function along with Minkowski Pooling. (These are described in the NASA-Ames Web site—http://vision.arc.nasa.giv/modelfest/) Our model is already quite similar to this, but it would be desirable to make it more consistent.
6. Finally, on a more mechanical note, we mentioned that it would be desirable to speed our algorithm up. Additional research to identify a faster convolution routine with an eye to making real-time sensing possible would be highly desirable.

While these next steps would permit a more precise measurement of conspicuity, we have shown that it is possible to develop a model for a good first pass. Thus, the algorithm shown here can be used to assess and improve the conspicuity of construction equipment and personnel on California Highways.

Appendix I: The Anatomy of Human Sight[1]

A. Overview of the Central Nervous System

The central nervous system is usually considered to include seven basic parts: the *spinal cord*, the *medulla*, the *pons*, the *cerebellum*, the *midbrain*, the *diencephalon*, and the *cerebral hemispheres*. These are shown in **Figure 1-1**. The medulla, pons, and midbrain are collectively called the *brainstem*; the diencephalon and cerebral hemispheres are collectively called the *forebrain*. This divisional scheme is actually related to the embrionic subdivisions apparent at the earliest stages of neural development. We note, though, that the categorization is somewhat artificial, in that neurons often span the boundaries of these subdivisions, and most neural functions depend on more than one of these components.



The Major Divisions of the Central Nervous System

Figure 1-1

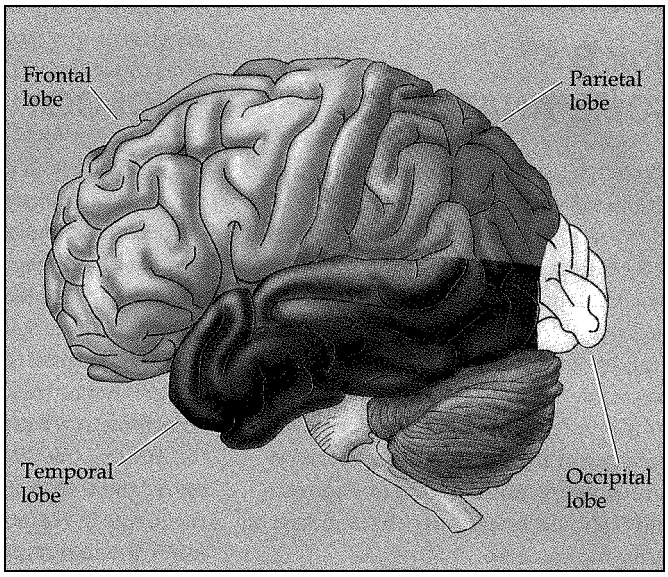
When the human brain is viewed from the side, as in **Figure 1-1**, only the brainstem, the cerebellum, and the cerebral hemispheres are visible. The latter are so large that they hide the rest of the brain's subdivisions from view. In addition to their large size (the cerebral hemispheres represent 85% of the brain by weight), an obvious feature of the hemispheres is their highly convoluted surface. The ridges are known as *gyri* (singular *gyrus*), and the valleys are called *sulci* (singular *sulcus*), or, if they are especially deep, *fissures*. The entire convoluted surface of the hemispheres comprises a laminated rind of neurons and supporting cells about 2 mm thick, called the *cerebral cortex*. The reasons for cerebral *sulcation* are not entirely clear, but the infolding of the brain obviously allows a great deal more cortical surface area (2.2 m^2 on average) to exist within the confines of the skull, or *cranium*. Also, a sulcus or fissure often corresponds to a boundary between two functionally distinct areas; thus, the mechanism of sulcation probably involves the differential growth of distinct cortical regions.

Each hemisphere is conventionally divided into four *lobes*, as shown in **Figure 1-2**, named for the bones of the skull that overlie them. These are the *frontal*, *parietal*, *temporal*, and *occipital lobes*. A particularly important feature of the frontal lobe is the precentral gyrus, whose cortex is referred to as the motor cortex because it contains neurons whose axons project (i.e., extend) to the motor neurons in the brainstem and spinal cord that innervate the skeletal muscles. The temporal lobe contains cortex concerned with audition. The parietal lobe harbors cortex that is concerned with bodily sensation. Finally, the *striate* or *visual cortex* of the occipital lobe (only a small part of which is apparent from this side-view of the brain), is concerned with vision. In addition to their role in primary and sensory processing , each lobe of the cerebral hemispheres has

[1] The material in this Appendix and all of the figures have been taken from:
Purves, et al., *Neuroscience*, Sinauer Associates, 1997, Sunderland, MA (Chapters 1 2, 5, 10, and 11).
Wandell, *Foundations of Vision*, Sinauer Associates, 1995, Sunderland, MA(Chapters 1-5).

characteristic cognitive functions. Thus, the frontal lobe is critical in planning behavior, the temporal lobe in recognizing objects and faces, the parietal lobe in attending to important stimuli, and the occipital lobe in visual analysis.

As we shall see, the signals from the eye pass to the occipital lobe by way of the diencephalon. Here most of these signals are processed in the *lateral geniculate nucleus* of the thalmus before being sent on. Those that aren't are concerned with pupillary light reflex, circadian rhythms (the "day/night" cycle), and head and eye movements. They project to structures within the diencephalon and midbrain.



The Four Lobes of the Brain

Figure 1-2

B. The Anatomy of the Neuron

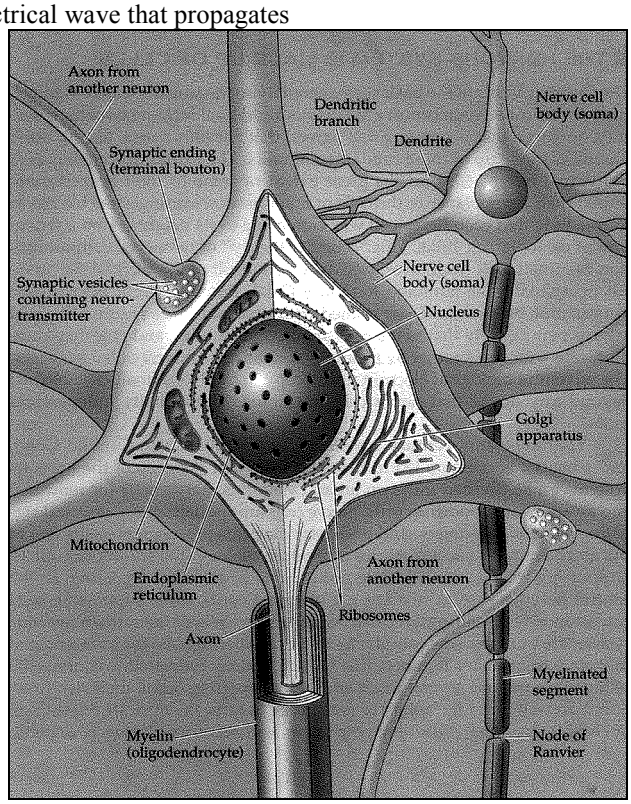
The fact that cells are the basic elements of living organisms was recognized early in the nineteenth century. It was not until well into the twentieth century, however, that neuroscientists agreed that nervous tissue, like all other organs, is made up of these fundamental units. The extraordinary shapes of individual nerve cells and the great extent of some of their branches obscured their resemblance to the cells of other tissues. We now know, however, that nervous tissue is indeed made up of cells that communicate with one another by means of specialized contacts called *synapses*. The cells of the nervous system can be divided into two broad categories: *nerve cells* (or *neurons*) and a variety of *supporting cells*. Nerve cells are specialized for electrical signaling over long distances (relative to the cell diameter). (Supporting cells, in contrast, are not capable of electrical signaling.) Although the cells in the human nervous system are in many ways similar to those of other organs, they are unusual in their extraordinary numbers (the human brain is estimated to contain 100 billion neurons and several times as many supporting cells), their rich functional diversity, and the neurons' ability to form intricate ensembles, or circuits, on which sensation, perception, and behavior ultimately depend.

A prototypical neuron and its component parts is shown in **Figure 1-3**. Like other cells, each neuron has a *cell body* containing a *nucleus*, *endoplasmic reticulum*, *ribosomes*, *Golgi apparatus*, *mitochondira*, and other organelles that are essential to its function. As mentioned before, however, neurons are highly specialized for intercellular communication. This specialization is perhaps most apparent in their bizarre and fascinating geometries. The most salient morphological feature is the elaboration of the *dendrites* (also called *dendritic branches* or *dendritic processes*) that arise from the neuronal cell body. Most neurons have multiple dendrites, which are typically short and highly branched. The dendrites, together with the cell body itself, provide sites for the synaptic contacts made by the terminals of other nerve cells and can thus be regarded as specialized for receiving information.

The spectrum of neuronal geometries ranges from a small minority of cells that lack dendrites altogether to neurons with dendritic aborizations that rival the complexity of a mature tree, as shown in **Figure 1-4**. The number of inputs that a particular neuron receives depends on the complexity of its dendritic arbor: neurons that lack dendrites are innervated by just one or a few other neurons, whereas those with elaborate dendrites are innervated by a commensurately larger number of other neurons. Since a fundamental purpose of neurons is to integrate information from other neurons, the number of inputs

received by each neuron (which in the human nervous system ranges from 1 to about 100,000) is an especially important determinant of neuronal function.

The information from the inputs that impinge on the dendrites is "read out" at the origin of the *axon*, the portion of the neuron specialized for signal conduction. (See again **Figure 1-4**.) The axon is a unique extension from the neuronal cell body that may ravel a few hundred micrometers or much further, depending on the type of neuron and the size of the species. The majority of neurons in the human brain have axons only a few millimeters long, and a few have no axons at all (for example, the *amacrine* cell, which we'll discuss later). Such neurons transmit information locally. In contrast, the axons that run from the retina of the eye to the midbrain region are about five centimeters long and those that run from the human spinal cord to the foot are about a meter long. The axonal mechanism that carries signals over such distances is called the *action potential*, a self-regenerating electrical wave that propagates

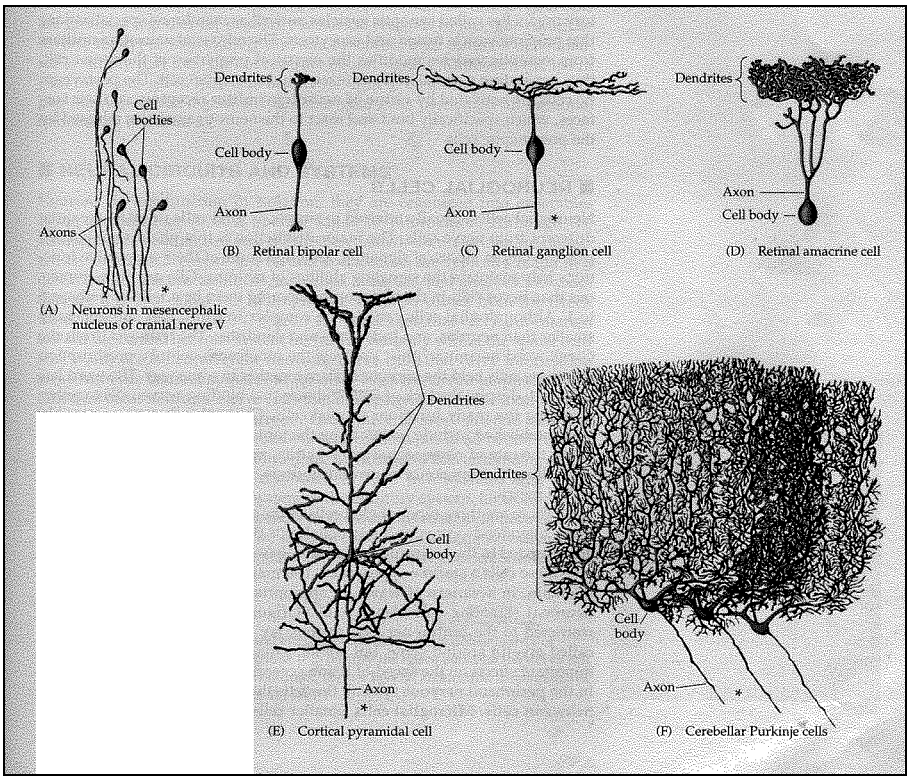


A Neuron and its Component Parts

Figure 1-3

from its point of initiation at the cell body (called the *axon hillock*, not shown in **Figure 1-3**, which is approximately at the location where the axon meets the cell body) to the terminus of the axon.

The information encoded by the action potentials is passed on to the next cell in the pathway by means of *synaptic transmission*. Accordingly, axon terminals are highly specialized to convey this information to target cells—which include other neurons in the brain, spinal cord, autonomic ganglia, and muscles and glands throughout the body. These terminal specializations are called *synaptic endings* (or *terminal boutons*), and the contacts they make with the target cells are called *chemical synapses*, also shown in **Figure 1-3**. (So called to distinguish them from *electrical synapses*, which are not present in the visual system, and which we therefore won't discuss.) A single neuron can receive many thousands of synaptic endings, and can contact as many as a thousand other cells. Each synaptic ending contains secretory organelles called *synaptic vesicles*, as well as membrane specializations that allow these vesicles to transmit their contents to the target cells. The release of *neurotransmitters* from the synaptic vesicles modifies the electrical properties of the target cell, thus generating a signal in that (postsynaptic) cell. The postsynaptic cells are activated by virtue of the *neurotransmitter receptors* on their surfaces, which specifically bind and react to the neurotransmitters released by the presynaptic cells.



Nerve Cell Morphologies

Figure 1-4

C. Neural Signaling

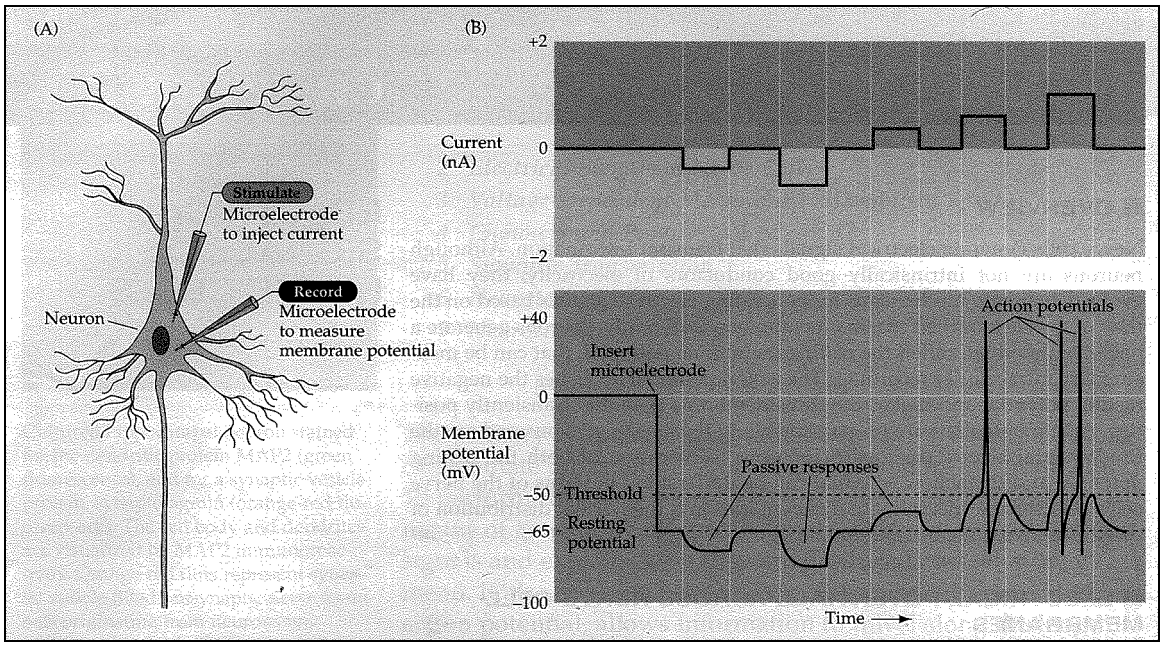
Neurons generate signals in the form of electrical potentials that transmit information. Although they are not intrinsically good conductors of electricity, they have evolved elaborate mechanisms for generating these signals based on the flow of ions across their plasma membranes. Neurons generate two basic types of potentials: *graded potentials* and *action potentials*[1]. Graded potentials are generated in sensory receptors and in dendrites. They are usually sustained in nature, lasting as long as the stimulus, with an amplitude that is proportional to the strength o the stimulus. Graded potentials are local potentials (that is, they are not *transmitted*, which we discuss below), and their amplitudes diminish as they move away from their site of generation. They add together, and they may be either positive of negative. Graded potentials have no *threshold* (also discussed below); for example, a single photon of light absorbed by a photoreceptor generates a small, graded potential. Action potentials, often called *impulses* or *spikes*, are relatively large (.1 volts), transient, (1 to 2 milliseconds in duration), "all—or—nothing" potentials that are generated along axons. They signal the strength of a stimulus by their frequency rather than their amplitude. Action potentials are usually transmitted, that is, they are continually regenerated along an axon. Hence, an action potential at the end of an axon is identical in amplitude to one at the beginning. Action potentials also have a threshold; for example, for an action potential to be generated, the voltage across the cell membrane (the *membrane potential*) must be changed by about 15 mV. Some neurons with short axons may function with graded potentials only. Some have no axons and therefore also function without action potentials. Most neurons, however, have both kinds of potentials. The graded potentials in the dendrites sum to change the membrane potential of the neuron by a sufficient amount to trigger the firing of action potentials in the axons. Thus, graded potentials generate action potentials in a neuron and for this reason are sometimes *generator potentials*.

As we mentioned above, the use of electrical signals—as in sending electricity over wires to provide power or information—presents a fundamental problem for neurons: neuronal axons, which can be quite long (recall that a spinal motor neuron can extend for a meter or more), are not good electrical conductors. Although neurons and wires are both capable of passively conducting electricity, the electrical properties of neurons compare poorly to even the most ordinary wire. Action potentials constitute a "booster system" that allows them to conduct electrical signals over great distances despite their intrinsically poor electrical characteristics. Underlying both graded and action potentials is a *resting potential*, which is the membrane potential of both unexcited neurons and other cells. When it is at rest, the electrical potential across a neuron's plasma membrane is typically –40 to –90 mV: the cell's inside charge is negative relative to the outside charge. Action potentials are elicited when electrical current is passed across the membrane of the neuron. Under normal

circumstances, such a current is generated by another neuron (at the synapse between the two nerve cells), or by the transduction of an external stimulus in sensory neurons. If the current delivered is such as to make the membrane potential more negative (*hyperpolarization*), nothing very dramatic happens. The membrane potential simply changes in proportion to the magnitude of the injected current. If current of the opposite polarity is delivered, so that the membrane potential becomes more positive than the resting potential (*depolarization*), then at a certain point, called the *threshold potential*, an action potential occurs. The effects of hyperpolarizing and depolarizing currents is summarized in **Figure 1-5**. Generation of both the resting potential and the action potential can be understood in terms of the nerve cell's selective permeability to different ions and the normal distribution of these ions across the cell membrane.

The actual communication of the information encoded in the action potential occurs through a *synapse*, the functional contact between neurons. Although there are many synaptic subtypes within the brain, they can be divided into two general classes: electrical synapses and chemical synapses. Electrical synapses, which are a distinct minority, permit direct, passive flow of electrical current from one neuron to another. Chemical synapses enable communication via the secretion of neurotransmitters; in this case, chemical agents released by the presynaptic neurons produce secondary current flow in postsynaptic neurons by activating specific receptor molecules. The generalized structure of a chemical synapse is shown schematically in **Figure 1-6**. The separation between the pre- and postsynaptic neurons is called the *synaptic cleft*. The key feature of all chemical synapses is the presence of small, membrane-bounded organelles called *synaptic vesicles* within the presynaptic terminal. These spherical organelles are filled with one or more types of *neurotransmitters*—the chemical signals secreted from the presynaptic neuron. The use of such chemical agents as messengers between the communicating neurons gives this type of synapse its name.

Transmission at chemical synapses is based on the elaborate sequence of events depicted in **Figure 1-7**. The process is initiated when an action potential is transmitted to the terminal of the presynaptic neuron. The change in membrane potential associated with the arrival of the action potential causes the opening of voltage-gated calcium channels in the presynaptic membrane. Because of the steep concentration gradient of Ca^{2+} across the presynaptic membrane (the external Ca^{2+} concentration is approximately 10^{-3} M, whereas the internal Ca^{2+} concentration is approximately 10^{-7} M), the opening of these channels causes a rapid influx of Ca^{2+} into the presynaptic terminal, which in turn causes the Ca^{2+} concentration of the cytoplasm in the terminal to rise from its normally low level to a much higher value. Elevation of the presynaptic Ca^{2+} concentration allows synaptic vesicles to fuse with the plasma membrane of

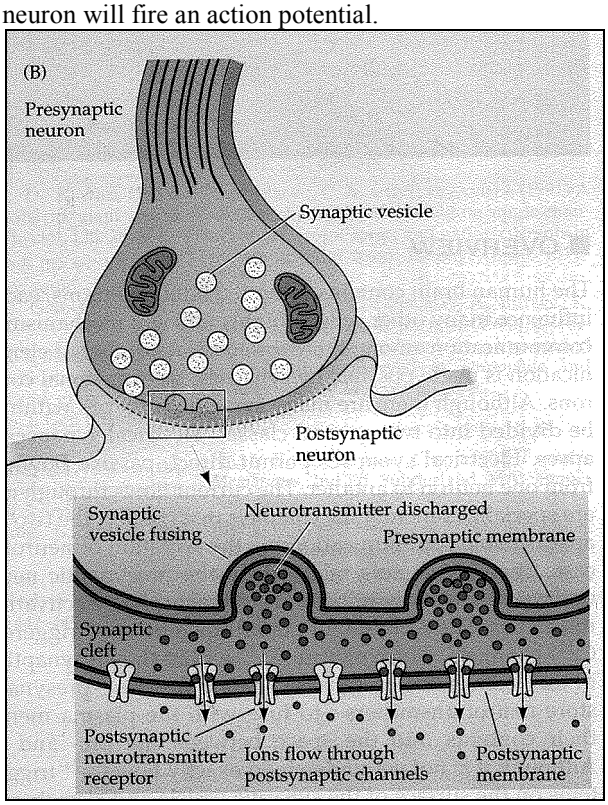


Passive and Active Electrical Signals in a Nerve Cell

Figure 1-5

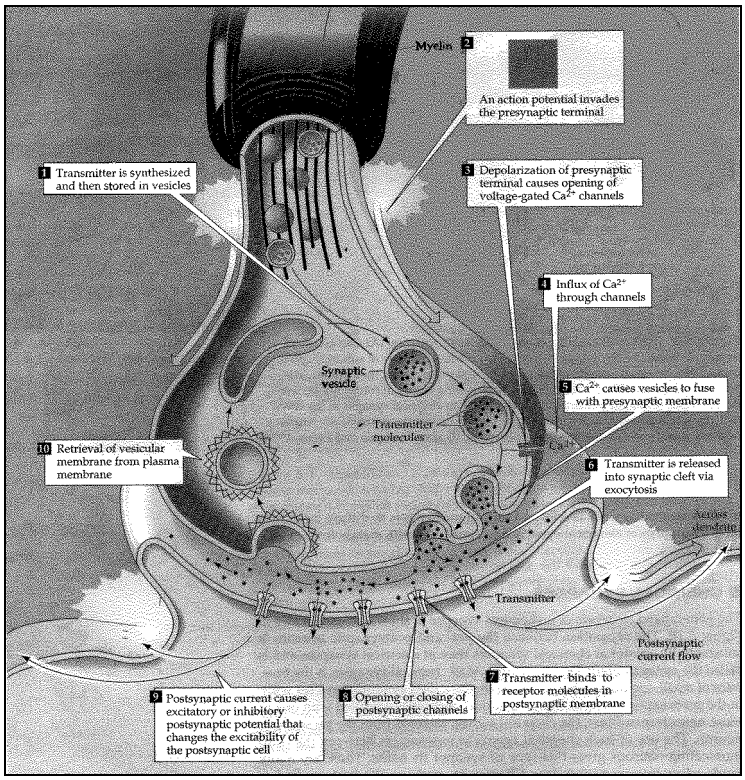
the presynaptic neuron, since this process is Ca^{2+} dependent. The fusion of synaptic vesicles causes their contents, most importantly neurotransmitters, to be released into the synaptic cleft. Following this *exocytosis*, transmitters diffuse throughout the synaptic cleft and bind to specific receptors on the membrane of the postsynaptic neuron. The binding of

neurotransmitter to the receptors causes channels in the postsynaptic membrane to open (or sometimes to close). The resulting neurotransmitter-induced current flow alters the membrane potential of the postsynaptic neuron, increasing or decreasing the probability that the neuron will fire an action potential.



The Chemical Synapse

Figure 1-6

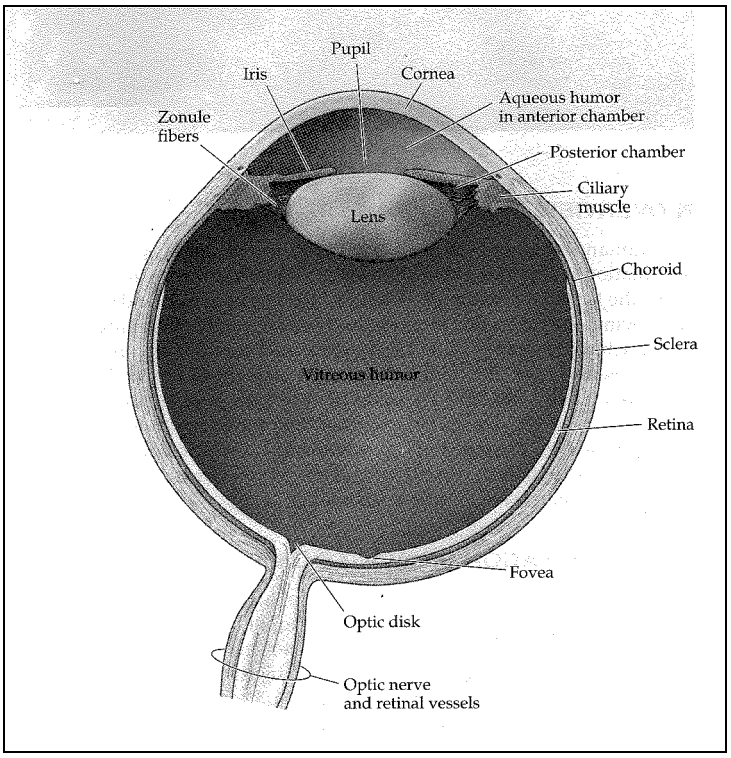


Signal Transmission at a Typical Chemical Synapse

Figure 1-7

D. The Formation of Images on the Retina

Figure 1-8 gives a schematic representation of the imaging components of the eye. The formation of focused images on the photoreceptors of the retina depends in large part on the refraction (bending) of light by the *cornea* and the *lens*. The cornea is responsible for most of the necessary refraction. The lens has considerably less refractive power than the cornea; however, the refraction supplied by the lens is adjustable, allowing objects that lie at various distances from the observer to be brought into sharp focus on the retinal surface.



Imaging Components of the Human Eye

Figure 1-8

The dynamic changes in the refractive power of the lens are referred to as *accommodation*. When viewing distant objects, the lens is made relatively thin and flat and has the least refractive power. For near vision, the lens becomes thicker and rounder, and has the most refractive power. The lens is held in place by radially arranged connective tissue bands (called *zonule fibers*) that are attached to the *ciliary muscle* that runs circumferentially near the inner surface of the eye. The shape of the lens is determined by two opposing forces: the elasticity of the lens, which tends to keep it rounded up (removed from the eye, the lens becomes spheroidal), and the force exerted by the zonule fibers, which tends to flatten it. Under normal conditions, the force of the zonule fibers is greater than the elasticity of the lens and the lens assumes the flatter shape that allows focusing on distant objects. Focusing on closer objects requires relaxing the tension in the zonule fibers, allowing the inherent elasticity of the lens to increase its curvature. This relaxation is accomplished by the contraction of the ciliary muscle. Because the ciliary muscle forms a ring, the attachment points of the zonal fibers move toward the center of the eye when the muscle contracts, thus reducing the tension on the lens.

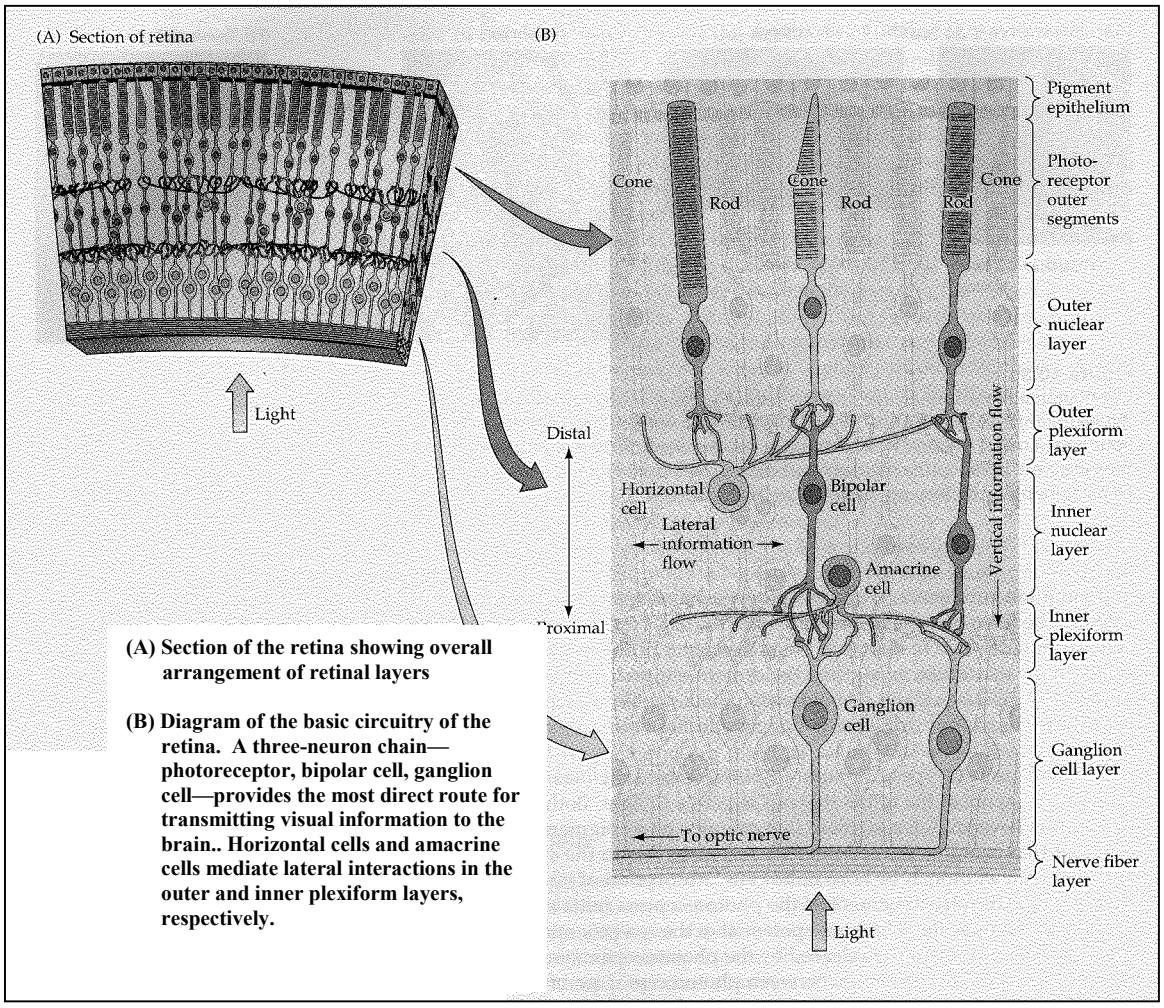
Adjustments in the size of the *pupil* (which is not a physical component of the eye but instead refers to the circular opening in the iris) also contribute to the clarity of images formed on the retina. Like the images formed by other optical instruments, those generated by the eye are not perfect; spherical and chromatic aberrations can cause blurring of the retinal image, a problem that is greatest for the light rays that pass farthest from the center of the lens. Narrowing the pupil therefore reduces both spherical and chromatic aberrations, just as closing the iris diaphragm on a camera lens improves the sharpness of a photographic image. Reducing the size of the pupil also increases the depth of field—that is, the range of distances over which objects can be moved toward or away from the eye without appearing blurred. However, a small pupil also limits the amount of light that reaches the retina; under conditions of dim illumination, visual acuity becomes limited by the number of

available photons rather than by optical aberrations. An adjustable pupil thus provides an effective compromise: it limits optical aberrations and maximizes depth or field as different levels of illumination permit.

The spaces in front of and behind the lens account for most of the eye's volume, and are filled with two different types of fluid. *Aqueous humor*, a clear, watery liquid, fills the space between the lens and the cornea (the *anterior chamber*) and supplies nutrients to both of these structures. Aqueous humor is produced in the *posterior chamber* (the region between the lens and the iris) and flows into the anterior chamber through the pupil. A specialized meshwork of cells that lies at the junction of the iris and the cornea is responsible for its uptake. Under normal conditions, the rates of aqueous humor production and uptake are in equilibrium, ensuring a constant intraocular pressure. The space between the back of the lens and the surface of the retina is filled with a thick gelatinous substance called the *vitreous humor*. In addition to maintaining the shape of the eye, the vitreous humor contains cells that remove blood and other debris that might interfere with light transmission.

E. The Anatomy of the Retina

Despite its peripheral location, the *retina* is actually part of the central nervous system. Consistent with this status, the retina comprises a complex neural circuitry that converts the chemical responses of the photoreceptors into graded electrical activity, and finally into action potentials. There are five types of neurons within the retina: *photoreceptor*s, *bipolar cells*, *ganglion cells*, *horizontal cells*, and *amacrine cells*. The bodies and processes (dendrites and axons) of these retinal neurons are stacked in five alternating layers, shown in **Figure 1-9**. The cell *bodies* are located in the *inner nuclea*r, *outer nuclear*, and *ganglion cell layers*;



(A) Section of the retina showing overall arrangement of retinal layers

(B) Diagram of the basic circuitry of the retina. A three-neuron chain—photoreceptor, bipolar cell, ganglion cell—provides the most direct route for transmitting visual information to the brain.. Horizontal cells and amacrine cells mediate lateral interactions in the outer and inner plexiform layers, respectively.

Structure of the Retina

Figure 1-9

while the processes and synaptic contacts are located in the *inner plexiform* and outer *plexiform layers*. (The terms "inner" and "outer" refer to relative distances from the center of the eye. Thus, "inner" means closer to the center of the eye and "outer" means farther from the center. In **Figure 1-9**, light is shown travelling from the center of the eye so the *nerve fiber layer* is the *innermost* part of the retina, and the *pigment epithelium* the *outermost* part.)

There are two types of photoreceptors, *rods* and *cones*, and these are the only elements of the retina that are sensitive to light. Both types of photoreceptors have an outer segment that contains photopigment and an inner segment that contains the cell nucleus and gives rise to the synaptic terminals that contact bipolar and horizontal cells. Absorption of light by the photopigment in the outer segment of the photoreceptors initiates a cascade of events that changes the membrane potential of the receptor and therefore the amount of neurotransmitter released by the photoreceptor synapses onto the cells they contact. The synapses between photoreceptor terminals and bipolar (and horizontal) cell processes occur in the outer plexiform layer. The cell bodies of photoreceptors make up the outer nuclear layer, whereas the cell bodies of bipolar cells lie in the inner nuclear layer. The axonal processes of bipolar cells make synaptic contacts in turn on the dendritic processes of ganglion cells in the inner plexiform layer. The axons of the ganglion cells form the *optic nerve*, which carries information about retinal stimulation to the rest of the central nervous system.

The two other types of neurons in the retina, horizontal cells and amacrine cells, have their bodies in the inner nuclear layer and are primarily responsible for lateral interactions within the retina. For example, lateral interactions between receptors, horizontal cells, and bipolar cells in the outer plexiform layer are largely responsible for the visual system's sensitivity to luminance contrast. The processes of amacrine cells, which extend laterally in the inner plexiform layer, are postsynaptic to bipolar cell terminals and presynaptic to dendrites of ganglion cells, shown in **Figure 1-9**. There are numerous subclasses of amacrine cells that can be distinguished by the particular neuropeptide transmitter they contain. Amacrine cells remain the least understood of the retinal neurons.

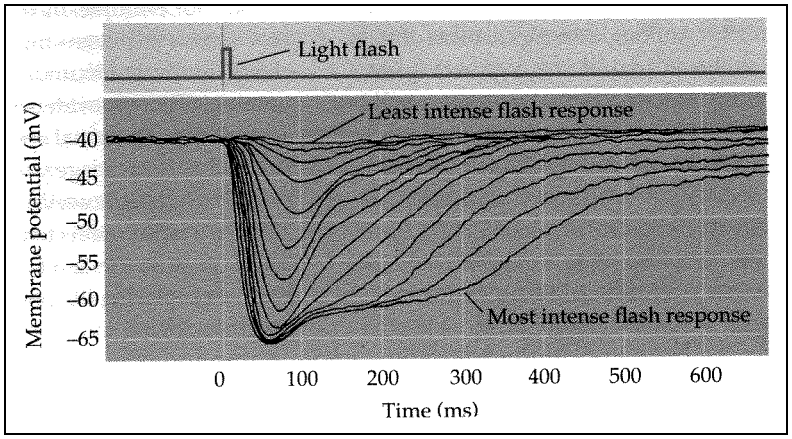
At first glance, the organization of the cellular layers in the retinal seems counterintuitive, since light rays must pass through all of the neural circuitry of the retina (not to mention the retinal vasculature) before striking the outer segments of the photoreceptors. This is certainly not the way an engineer would have designed things. However, this peculiar arrangement allows the tips of the outer segments of the photoreceptors to contact the pigment epithelium. The outer segments contain membranous disks that house the photopigment and other proteins involved in the transduction process. The disks are continuously formed near the inner segment and pushed toward the tip of the outer segment, where they are shed. The pigment epithelium plays an essential role in removing the expended receptor disks—no small task, since all the disks in the outer segments are replaced every 12 days. The pigment epithelium also plays a second important role—it reduces the back-scattering of light that has already passed through the photoreceptors. It is presumably the life cycle of the photoreceptor that explains why photoreceptors are found in the outermost rather than the innermost layer of the retina

F. Phototransduction

In most sensory systems, activation of a receptor by the appropriate stimulus causes the cell membrane to depolarize (i.e., assume a more positive potential), stimulating transmitter release and ultimately a postsynaptic potential in the neurons it contacts. Thus, it may come as a surprise to learn that shining light on a photoreceptor, either a rod or a cone, leads to membrane *hyperpolarization* rather than depolarization, as demonstrated in **Figure 1-10**. In the dark, the receptor is in a depolarized state, with a membrane potential of roughly –40 mV. Progressive increases in the intensity of illumination cause the potential across the receptor membrane to become more negative, a response that saturates when the membrane potential reaches about –65 mV. Transmitter release from the synaptic terminals of the photoreceptor, like that from other neuron, is dependent on the potential difference across the terminal membrane. Thus, the depolarized photoreceptors continually release transmitter in the dark; when they are hyperpolarized by light, the level of transmitter release is *reduced*. Although this arrangement may seem odd, the only logical requirement for subsequent visual processing is a consistent relationship between luminance changes and the activity of photoreceptors. In any event, the reason for this apparent "sign reversal" in the activation of photoreceptor cells is not known.

The depolarized state of photoreceptors in the dark depends on the presence of ion channels in the outer segment membrane that permit sodium, calcium, and magnesium ions to flow into the cell, thus reducing the degree of inside negativity. This is shown schematically in **Figure 1-11**. The probability of these channels in the outer segment being open or closed is regulated by the levels of the nucleotide *cyclic guanosine monophosphate (cGMP)*. In darkness, high levels of cGMP in the outer segment keep the channels open. In light, cGMP levels drop and some of the channels close, leading to hyperpolarization of the outer segment membrane.

The absorption of a photon by a molecule of photopigment in the disks of the photoreceptor outer segment initiates a biochemical cascade that ultimately decreases intracellular levels of cGMP. The photopigment in the receptor discs contains a light absorbing component (*11-cis retinal*) coupled to one of a variety of proteins (*opsins*) that tunes the molecule's absorption of light to a particular

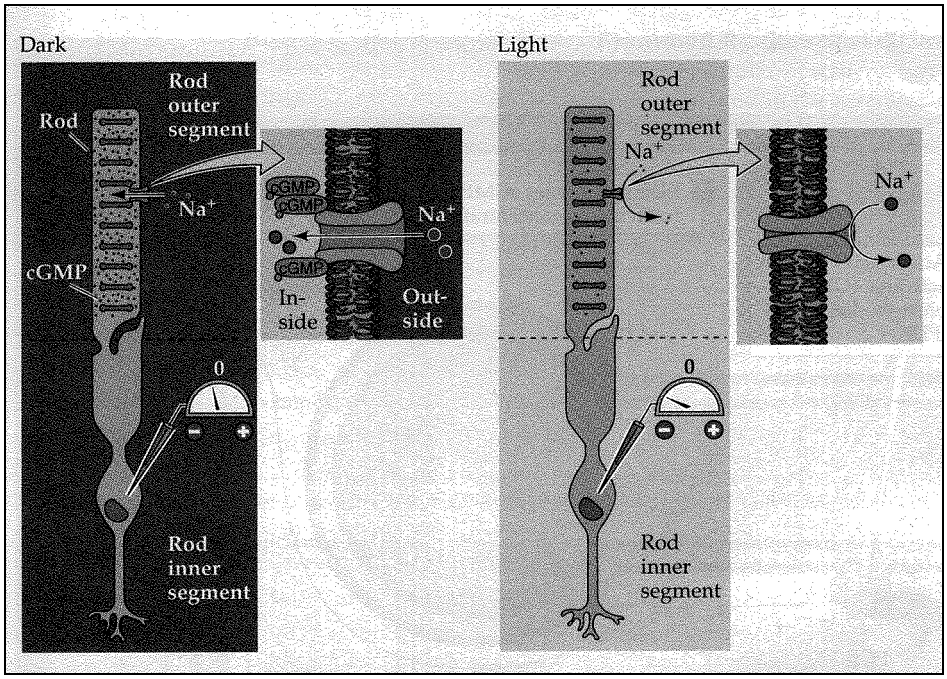


Signal Transmission at a Typical Chemical Synapse

Figure 1-10

region of the spectrum. The different protein component of the photopigment in rods and cones contributes critically to the functional specialization of the two receptor types. Most of what is known about the molecular events of phototransduction has been gleaned from experiments in rods, in which the photopigment is *rhodopsin*. When the retinal moiety in rhodopsin absorbs a photon of light, its configuration changes and triggers a series of alterations in the protein component of the molecule. These changes lead, in turn, to the activation of an intracellular messenger called transducin, which activates a phosphodiesterase that hyudrololyzes cGMP. Thus, absorption of light results in structural changes in rhodopsin, the activation of transducin, the activation of a cGMP phosphodiesterase, and finally the breakdown of cGMP.

This complex cascade provides enormous amplification. A single light-activated rhodopsin molecule can activate hundreds of transducin molecules, which in turn can lead to the hydrolysis of hundreds of cGMP molecules. It has been estimated that the absorption of a single photon by a rhodopsin molecule results in the closure of 300 ion channels, or about 3% of the number of channels in each rod that are open in the dark.

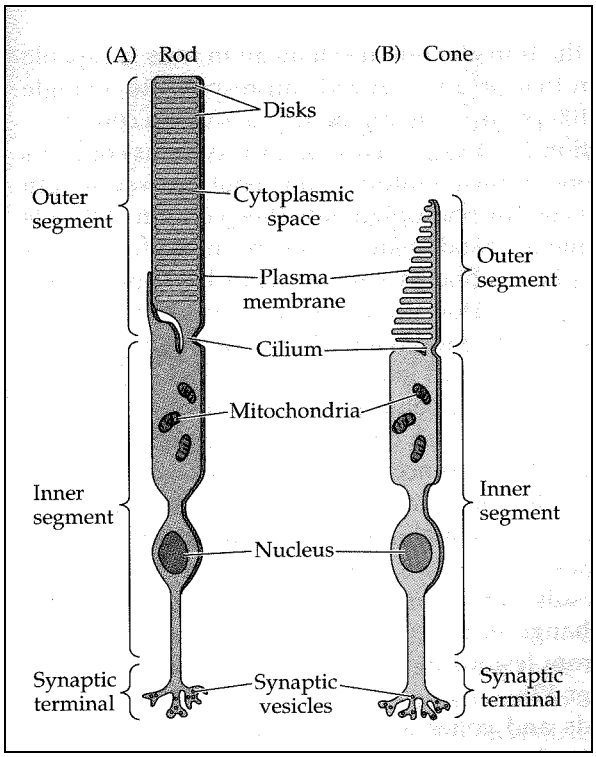


The Role of cGMP in Regulating the Electrical State of Photoreceptors in the Dark

Figure 1-11

G. Specializations of the Rod and Cone Systems

Longitudinal slices of the two types of photoreceptors, rods and cones, are shown in **Figure 1-12**. (When viewed from above, these cells have a circular cross section. Thus, the outer segment of a rod is cylindrically shaped, while that of a cone is conical—hence their names.) They are distinguished by their shape, the type of photopigment they contain, their distribution across the retina, and their pattern of synaptic connections. These properties reflect the fact that the rod and cone *systems* (by which we mean the receptors and their connections within the retina) are specialized for different aspects of vision. The rod system has very low spatial resolution , but is extremely sensitive to light; it is therefore specialized for sensitivity at the expense of resolution. In contrast, the cone system has very high spatial resolution but is relatively insensitive to light (the amplification effect described in the previous section applied to rods—a rod can respond to a single photon, whereas more than 100 photons are required to activate a cone); it is therefore specialized for acuity at the expense of sensitivity. The cone system also allows us to see color.



Structural Differences Between Rods and Cones

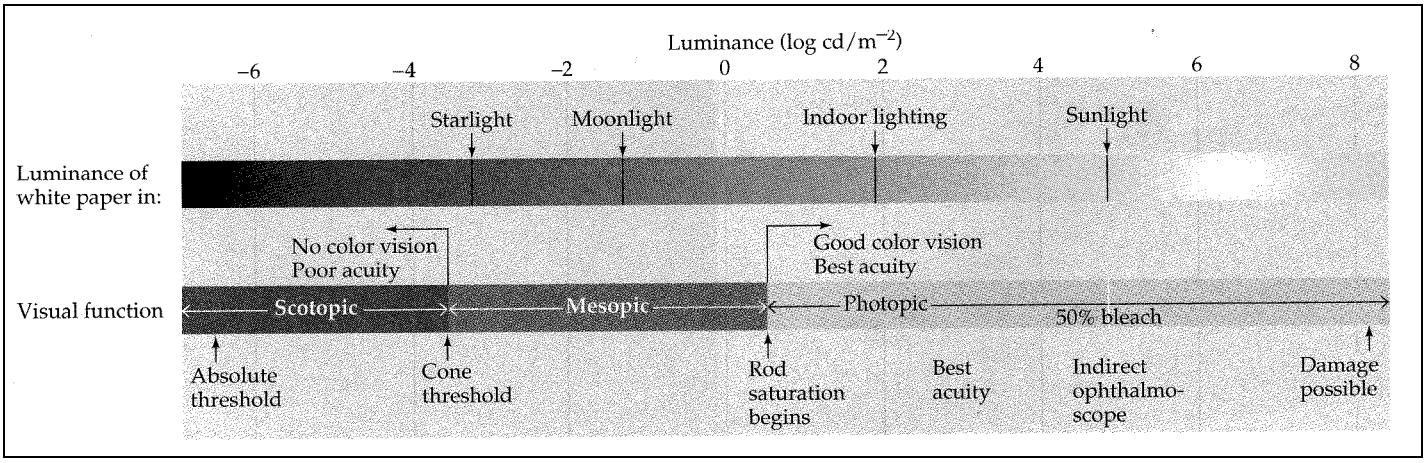
Figure 1-12

The contributions of the rod and cone systems to the range of illumination over which the visual system can operate are shown in **Figure 1-13**. At the lowest levels of illumination (below the level of starlight), rods are the only receptors activated; such rod-mediated perception is called *scotopic vision*. Everyone is familiar with the difficulty of making visual discriminations under very low light conditions, where only the rod system is active. (The problem is the poor acuity of this system, and the fact that all perception of color is lost.) Cones begin to contribute to visual perception at about the level of starlight, and they are the only receptors that function under relatively bright conditions such as normal indoor lighting or sunlight. Cone-mediated vision, called *photopic vision*, occurs at high levels of illumination because the response of rod photoreceptors to light saturates—that is, the membrane potential of individual rods no longer varies as a function of illumination. (See **Figure 1-13**.) Finally, *mesopic vision* occurs at light levels in which both rods and cones contribute. From these considerations it should be clear that most of what we think of as seeing is mediated by the cone system, and that loss of cone function is devastating. Individuals who have lost cone function are legally blind, whereas those who have lost rod function only experience difficulty seeing at low levels of illumination (called night blindness).

The factors that contribute to the functional differences in the rod and cone systems are several. First, differences in the structure of rods and cones, including the amount of photopigment and the shape of the outer segment, make rod receptors more sensitive. (See again **Figure 1-12**.) Rods are longer and contain more photopigment than cones, enabling them to

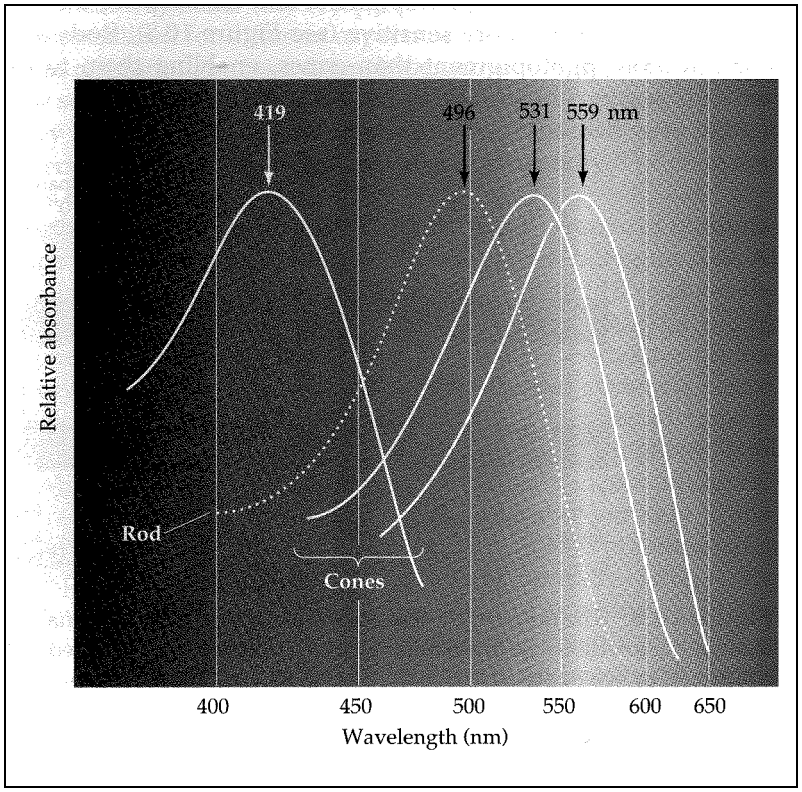
capture more light. Even more important, the transduction mechanism in rods (as mentioned above and in the previous section) is capable of greater amplification than cones.

Another critical distinction between the two receptor systems concerns color vision. Both rods and cones transmit information about the wavelength of light as a function of the types of the photopigments they contain. All rods contain the same photopigment—rhodopsin—whereas individual cones contain one of three different photopigments, collectively called *cone opsins*, that have different but overlapping absorption spectra. This is shown in **Figure 1-14**. The relative activity of these three sets of cones [referred to as S, M, and L cones for the short (blue), middle (green), and long (red) wavelengths] generates retinal signals that ultimately give rise to the sensation of color. Besides it aesthetic appeal, color vision makes it possible to distinguish objects that might be difficult to



The Visible Spectrum

Figure 1-13

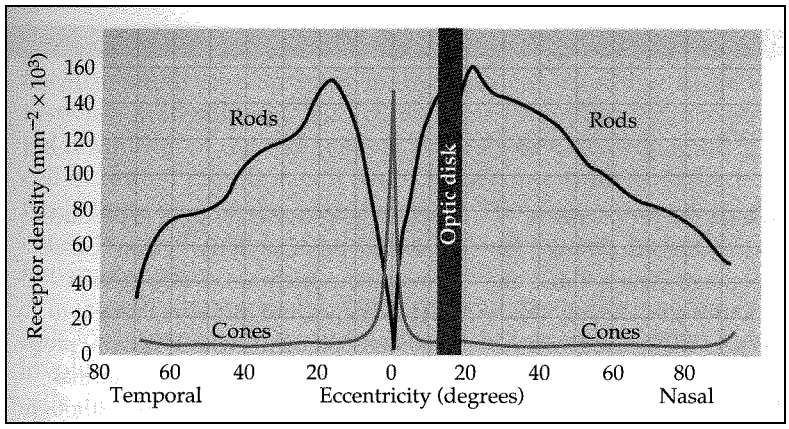


The Absorption Sptectra of the Four Photopigments in the Human Retina

Figure 1-14

contrast with their surroundings. Seeing color is not essential, however; many mammals lack this ability, and color blindness in humans is a relatively minor problem.

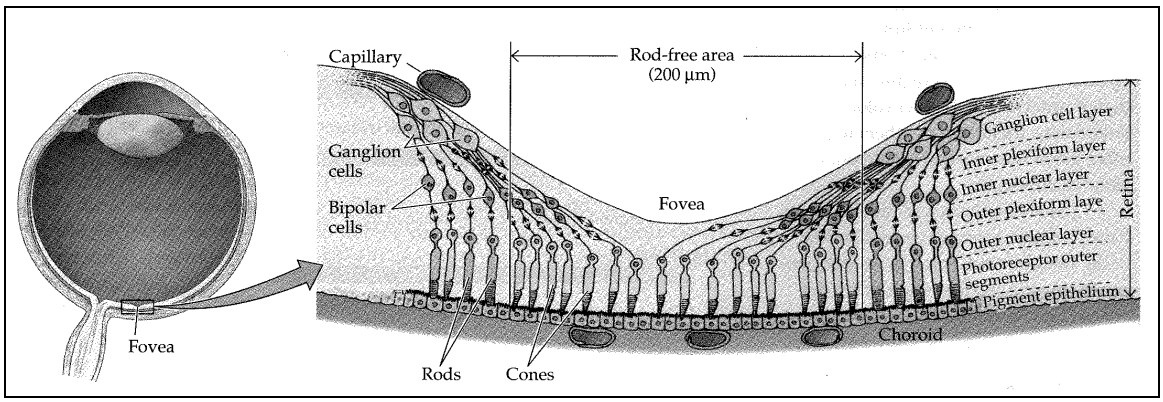
There are approximately 100 million rods and 5 million cones in each eye. Their distribution across the surface of the retina varies markedly, as **Figure 1-15** shows. Cones are the only photoreceptors located in the *fovea*. In this specialized region of the retina, shown in **Figure 1-16**, the layers of cell bodies and processes that generally overlie the photoreceptors are displaced, so light rays are subjected to a minimum of scattering before they strike the receptors. Another potential source of distortion, blood vessels, are also absent from the small region in the center of the fovea (called the *foveola*); because this region lacks a capillary bed, it is dependent on the underlying choroid and pigment epithelium for sustenance.



Distribution of Rods and Cones in the Human Retina

Figure 1-15

Although cones are not restricted to the fovea, their lower density outside the fovea, as well as the lower density of the ganglion cells that they supply, explains why visual acuity declines so markedly as a function of eccentricity. Indeed, the greater acuity of foveal vision is the main reason humans spend so much time moving their eyes (and heads) around—in effect directing the foveas of the two eyes to objects of interest. Acuity is reduced by 75% just 6º eccentric to the line of sight, a fact that can readily be appreciated by trying to read the words on any line of this page away from the word being fixated on. Conversely, the exclusion of rods from the fovea and their presence in high density away from the fovea, explain why the threshold for detecting a light stimulus is much lower outside the region of central vision. It is easier to see a dim object (such as a faint star) by looking away from it, so that the starlight stimulates the region of the retinal that is rich in rods.



Diagrammatic Cross-Section Through the Human Fovea

Figure 1-16

Another distinction between the rod and cone systems is the degree of receptor convergence onto other cell types of the retina. The rod system is highly convergent: many rods synapse on a single bipolar cell, and many bipolar cells that receive rod input converge on the same ganglion cell. In contrast, the cone system is much less convergent. In the center of the

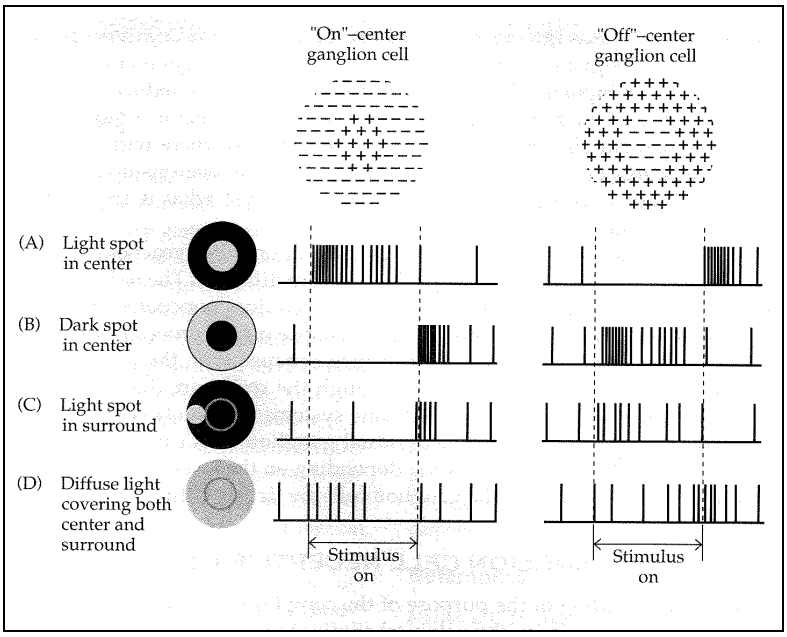
fovea, a single cone may contact only one bipolar cell, which in turn contacts a single ganglion cell. Convergence makes the rod system a good detector of light because small signals from many rods can combine to generate a larger response in the bipolar cell. At the same time, convergence reduces the spatial resolution of the rod system, since the source of a signal in a rod bipolar cell or retinal ganglion cell could have come from anywhere within a relatively large area of the retinal surface. The one-to-one relationship of foveal cones to bipolar and ganglion cells is, of course, just what is required to maximize acuity.

Finally, the routes by which rod and cone information reaches ganglion cells for transmission to central visual targets in the brain are different. The rod pathway involves a class of bipolar cells (*rod bipolars*) that does not contact ganglion cells directly; instead, these bipolar cells synapse on the processes of a specialized class of amacrine cells that in turn synapse with the processes of ganglion cells. Although the routes through the retina are distinct, extrafoveal information from the rod and cone systems ultimately converges on the same ganglion cells. Thus, individual ganglion cells can display both rod- and cone-driven characteristics, depending on the level of illumination. Within the fovea, however, the ganglion cells are driven entirely by cones.

As a result of this convergence of many photoreceptors onto single retinal ganglion cells, the approximately 105 million signals from the photoreceptors in each eye are reduced to approximately 1 million signals—the number of retinal ganglion cell axons making up each optic nerve. We've already described one type of convergence, that of the rods, and the reason for it (to detect faint light signals). In the next section we'll consider a far more important—and interesting—combination of photoreceptors.

H. Retinal Ganglion Cell Receptive Fields

Some understanding of the purpose of the complex synaptic interactions in the retina has come from physiological studies in which small spots of light are used to examine the responses of individual retinal neurons. Stephen Kuffler pioneered this approach in the 1950's by characterizing the responses of single ganglion cells in the cat retina. He found that each ganglion cell responds to stimulation of a small, restricted, circular patch of the retina, which defines the cell's *receptive field*. Based on these responses, Kuffler distinguished two classes of ganglion cells, *"on"-center* and *"off"-center*. Turning a spot of light on in the center of an *on-center ganglion cell* receptive field produces a burst of electrical activity (an "on" response, as indicated in **Figure 1-17**). Turning the light on in the center of an *off-center ganglion cell* has the opposite effect: the spontaneous rate of firing decreases, and when the spot of light is turned off, the cell responds with a burst of action potentials (an "off" response). On- and off- center ganglion cells are present in roughly equal numbers. The receptive fields have overlapping distributions, so that every point on the retinal surface (that is, every part of visual space) is analyzed by several on-center and off-center ganglion cells. The significance of these two distinct types of retinal ganglion cells was further demonstrated by Peter Schiller and his colleagues, who examined the effects of pharmacologically inactivating on-center ganglion cells on a monkey's ability to detect a variety of visual stimuli. After silencing the on-center ganglion cells, the animals showed a dramatic and specific deficit in their ability to detect stimuli that were brighter than the background; however, they could still see objects that were darker than the background.

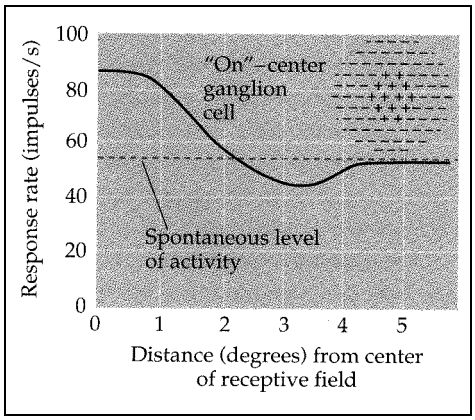


Responses of *On-Center* and *Off-Center* Retinal Ganglion Cells

Figure 1-17

These several observations suggest that information about increases or decreases in luminance (perceived as brightness or darkness, respectively) is carried separately to the brain by these two types of retinal ganglion cells. Having two separate luminance channels means that changes in light intensity are always conveyed to the brain by an excitatory process, rather than relying on decreases in activity below some set resting level to signal diminished luminance. For example, a rapid increase in the firing rate of on-center cells, which have low rate of firing in dim illumination, unambiguously signals a rapid increase in luminance, as shown in **Figure 1-17A**. However, these cells could not reliably signal a rapid decrease in luminance from the original level, as in **Figure 1-17B**. The identification of two distinct classes of retinal ganglion cells, and the demonstration that their activity conveys different types of information to central visual structures, illustrates one strategy for coping with the wealth of information in the visual scene: the emergence of *parallel systems* for analyzing different features of the visual stimulus. Other examples of the parallel processing of different categories of visual information (such as color and motion) will be discussed later.

Kuffler's work called attention to another important feature of visual processing. Retinal ganglion cells do not act as simple photodetectors; in fact, most ganglion cells are relatively poor at signaling differences in the level of diffuse illumination. Instead, they are sensitive to *differences* between the level of illumination that falls on the receptive field center and the level of illumination that falls on the surround—that is, *luminance contrast*. Kuffler noticed that the center of a ganglion cell receptive field is surrounded by a concentric region that, when stimulated, antagonizes the response to stimulation of the receptive field enter. For example, as a spot of light is moved from the center of the receptive field of an on-center cell towards its periphery, the response of the cell to the spot of light decreases, as in **Figure 1-18**. When the spot falls completely outside the center (that is, in the surround), the response of the cell falls below its resting level; the cell is effectively inhibited until the distance from the center is so great that the spot no longer falls on the receptive field at all, in which case the cell returns to its resting level of firing. Off-center cells also show an antagonistic surround. Light stimulation of the surround of an off-center cell increases the firing rate of the cell, a response that opposes the decrease in firing rate that occurs when the center is stimulated (**Figure 1-17C**). Because of their antagonistic surrounds, ganglion cells respond much more vigorously to small spots of light confined to their receptive field centers than to large spots or uniform illumination (**Figure 1-17D**).

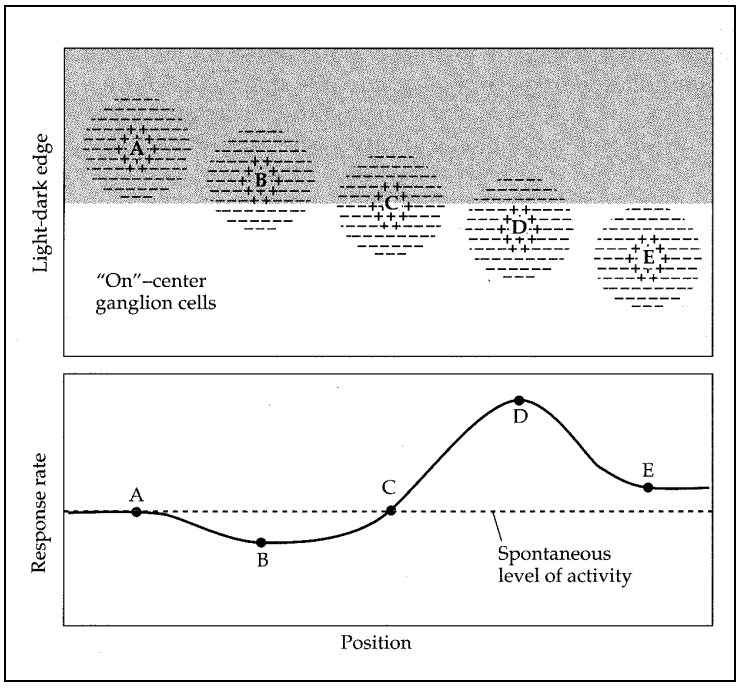


Rate of Discharge of an *On-Center* Ganglion Cell as a Function of Distance From the Center of the Receptive Field

Figure 1-18

To appreciate how center-surround antagonism helps to detect luminance contrast, consider the activity levels in a hypothetical population of on-center ganglion cells whose receptive fields are distributed across a retinal image of a light-dark edge, as shown in **Figure 1-19**. The neurons whose firing rates are most affected by this stimulus—either increased (neuron D) or decreased (neuron B)—are those with receptive fields that lie along light/dark border; those with receptive fields completely illuminated (or completely darkened) remain relatively unaffected (neurons A and E). Thus, the signal supplied by the retina to central visual structures does not give equal weight to all regions of the visual scene; rather it emphasizes the regions that contain the most information—namely, the regions where there are differences in luminance.

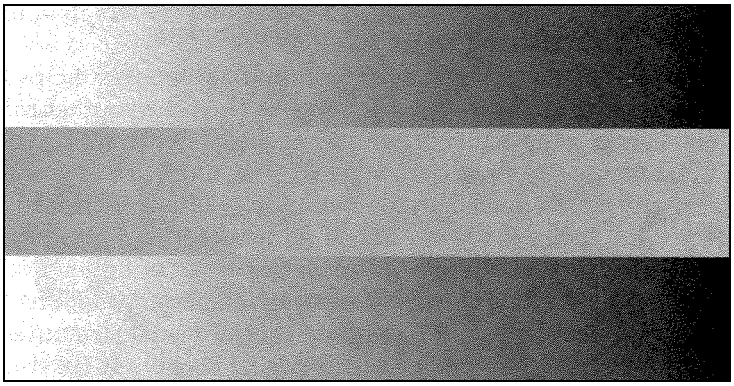
This property of retinal ganglion cells explains why our perception of the brightness or darkness of a given region in the visual scene is influenced so strongly by the luminance of adjacent regions. For example, the middle panel in **Figure 1-20** reflects the same amount of light on the left side as on the right; yet the part on the right appears significantly brighter than the part on the left. (You can convince yourself that the middle panel is really equiluminant by using two pieces of paper to block out the regions above and below it.) The *perceived* brightness of the panel is computed on the basis of the activity of those retinal ganglion cells whose receptive fields intersect its borders. Their activity in turn depends on *contrast*—the difference in the amount of light that falls on their receptive field centers and surrounds.



Responses of a Population of *On-Center* Ganglion Cells to a Light-Dark Edge

Figure 1-19

The special sensitivity of retinal ganglion cells to contrast rather than absolute levels of luminance also explains why the perceived brightness of objects remains constant over a wide range of lighting conditions. Increases or decreases in the overall level of illumination have equal effects on the center and surround of each ganglion cell's receptive field, and thus do little to affect the cell's level of activity. In bright sunlight, for example, the print on this page reflects considerably more light to the eye than it does in room light. In fact, the *print* reflects more light in sunlight than the white *paper* reflects in room light; yet, the print looks just as black (and the page just as white, for that matter) indoors or out. The signal sent to the brain from the retina therefore downplays the background level of illumination while enhancing the salient features of a visual stimulus—in particular, its contrast with the surroundings.



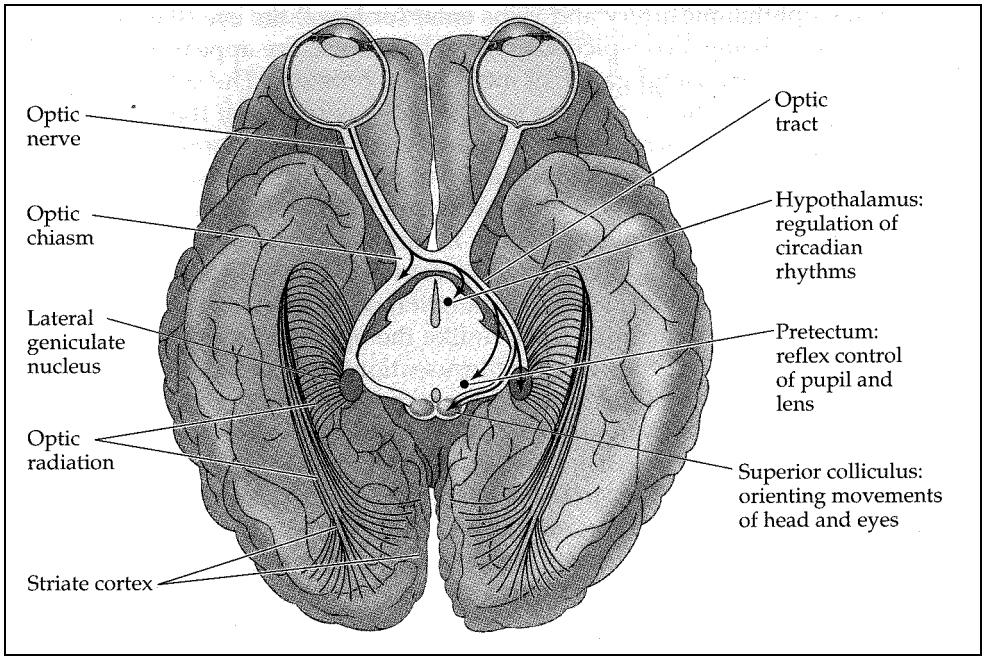
The Effect of Background Contrast on the Perception of Brightness

Figure 1-20

I. Central Projections of Retinal Ganglion Cells

All the visual pathways to the brain arise from ganglion cell axons that exit the retina through a circular region in its nasal part called the *optic disk*, where they bundle together to for the *optic nerve* (See again **Figure 1-8**). (Because this region of the retina has no photoreceptors, it is insensitive to light and produces the perceptual phenomenon known as the *blind spot*.) Ganglion cell axons in the optic nerve run a straight course to the optic chiasm at the base of the diencephalon. In humans, about 60% of the fibers cross the chiasm, while the other 40% continue toward the brain on the same side. Once past the chiasm, the ganglion cell axons on each side form the *optic tract*. Unlike the optic nerve the optic track contains fibers from both eyes. The partial crossing, or *decussation*, of the ganglion cell axons at the optic chiasm allows information from corresponding points on the two retinas to be processed by approximately the same cortical site in each hemisphere.

The ganglion cell axons in the optic tract reach a number of structures in the diencephalon and midbrain, as shown in **Figure 1-21**. The major target in the diencephalon is the *lateral geniculate nucleus* of the thalmus. Neurons in the lateral geniculate nucleus send their axons to the cerebral cortex via the internal capsule. These axons pass through a portion of the internal capsule called the *optic radiation* and terminate in the *primary visual (*or *striate) cortex*. This pathway, referred to as the *primary visual pathway*, is responsible for most of what is thought of as seeing. There are several other targets for retinal ganglion cells. These are the *pretectum* (reflex control of the pupil and lens), the *suprachiasmatic nucleus* of the *hypothalamus* (regulation of circadian rhythms—i.e., the day/night cycle), and the *superior colliculus* (coordination of head and eye movements).



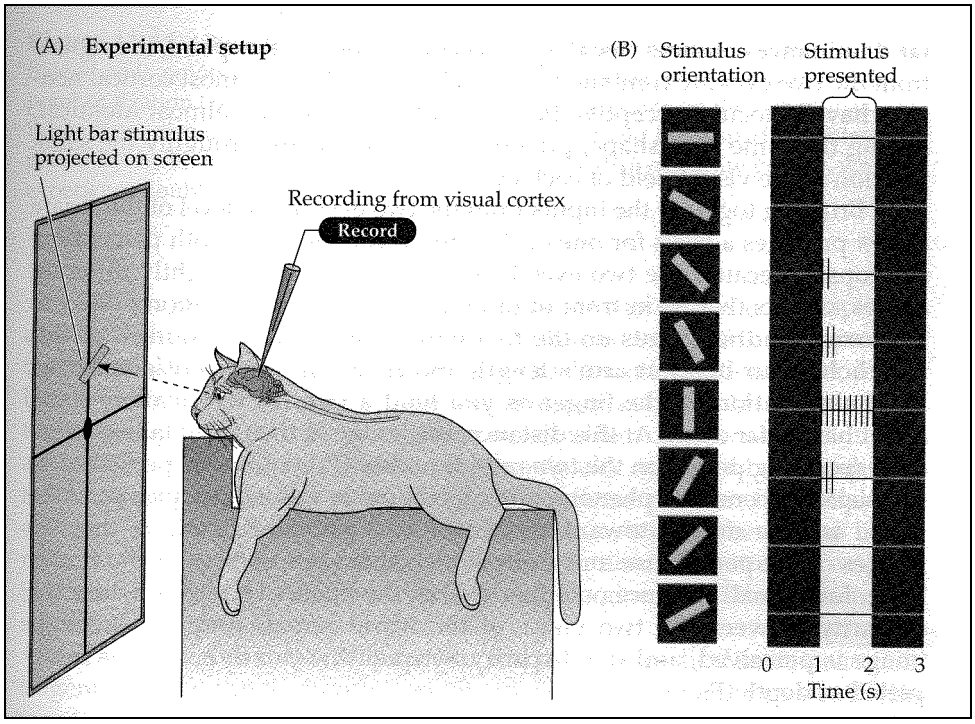
Central Projections of Retinal Ganglion Cells

Figure 1-21

J. The Functional Organization of the Striate Cortex

The discovery by Stephen Kuffler that retinal ganglion cell receptive fields have a center-surround structure that conveys luminance contrast information to the brain led David Hubel and Torsten Wiesel to consider the sorts of information extracted by more central visual structures. An analysis of the receptive field properties of neurons in the lateral geniculate nucleus showed surprisingly little difference from what had been found in the retina. In the striate cortex, however, the small spots of light that were so effective at stimulating neurons in the retina and lateral genidulate nucleus were largely ineffective. Instead, cortical neurons in cats and monkeys responded vigorously to light-dark bars or edges, and only if these bars were presented at a particular orientation within the cell's receptive field, as indicated in **Figure 1-22**. Moreover, each cortical cell responded maximally to a narrow range of edge orientations (the cell's preferred orientation). Thus, all the orientations present in visual scenes appear to be encoded in the activity of distinct populations of orientation-selective neurons.

Hubel and Wiesel also found that within a class of neurons that preferred the same orientation, there were subtly different subtypes. For example, the receptive fields of some cells, which they called *simple cells*, were composed of spatially separate *on* and *off* response zones, as if the on and off centers of the retinal ganglion cells that supplied these neurons were arrayed in separate parallel bands. Other neurons, referred to as *complex cells*, exhibited mixed on and off responses throughout the receptive field, as if they received their inputs from a number of simple cells. Further analysis uncovered cortical neurons sensitive to the length of the bar of light that was moved across their receptive field, decreasing their rate of response when the bar exceeded a certain length. Hubel an Wiesel called such neurons *hypercomplex* (or *end-stopped*) *cells*. Still other cells responded selectively to the direction in which an edge moved across their receptive field. Although the mechanisms responsible for generating these selective responses are still not fully understood, there is little doubt that the specificity of the receptive field properties of neurons in the striate cortex (and beyond) is essential for perceiving different aspect of the visual scene.



Responses of Neurons in the Striate Cortex to Edges of Different Orientations

Figure 1-22

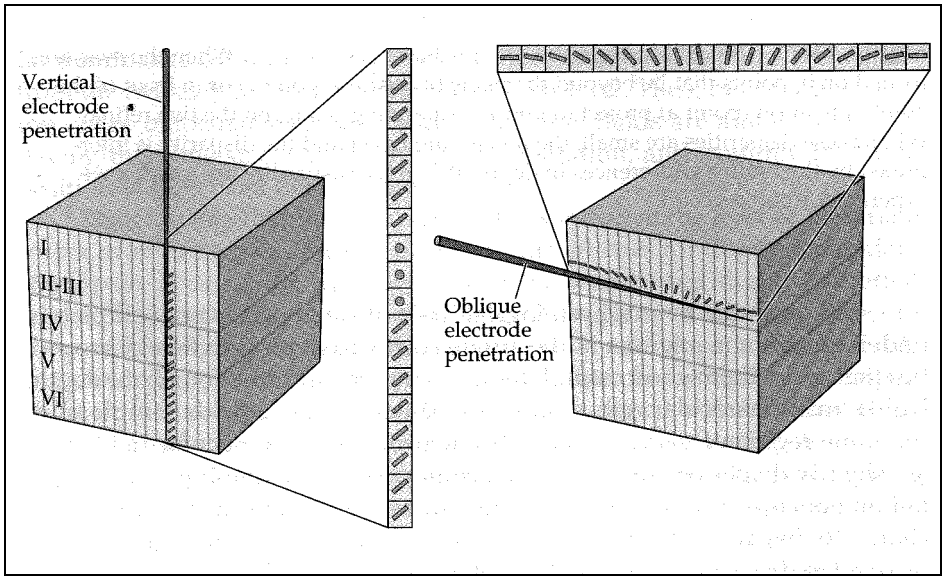
K. The Columnar Organization of the Striate Cortex

A special feature of the functional organization of the striate cortex is the grouping together of neurons that have similar response properties into radial arrays that span the thickness of the cortex. For example, all the neurons encountered in a vertical penetration of the striate cortex (that is, perpendicular to the surface) have similar preferred orientations. Furthermore, electrode penetrations made tangential to the cortical surface show a systematic shift in preferred orientations, such that movements across the surface of about a millimeter are sufficient to encounter a series of neurons that cover the full range of orientations, shown in **Figure 1-23**.

The columnar organization of the striate cortex is equally apparent in the binocular responses of cortical neurons. Although most neurons in the striate cortex respond to both eyes, the relative strength of the inputs from the two eyes varies from neuron to neuron. At the extremes of this continuum are neurons that respond almost exclusively to the left or right eye; in the middle are those that respond equally well to both eyes. As with orientation preference, vertical electrode penetrations encounter neurons with similar ocular preference (or ocular dominance, as it is usually called), and tangential penetrations show gradual shifts in ocular dominance as the electrode moves across the plane of the cortical surface (**Figure 1-24**).

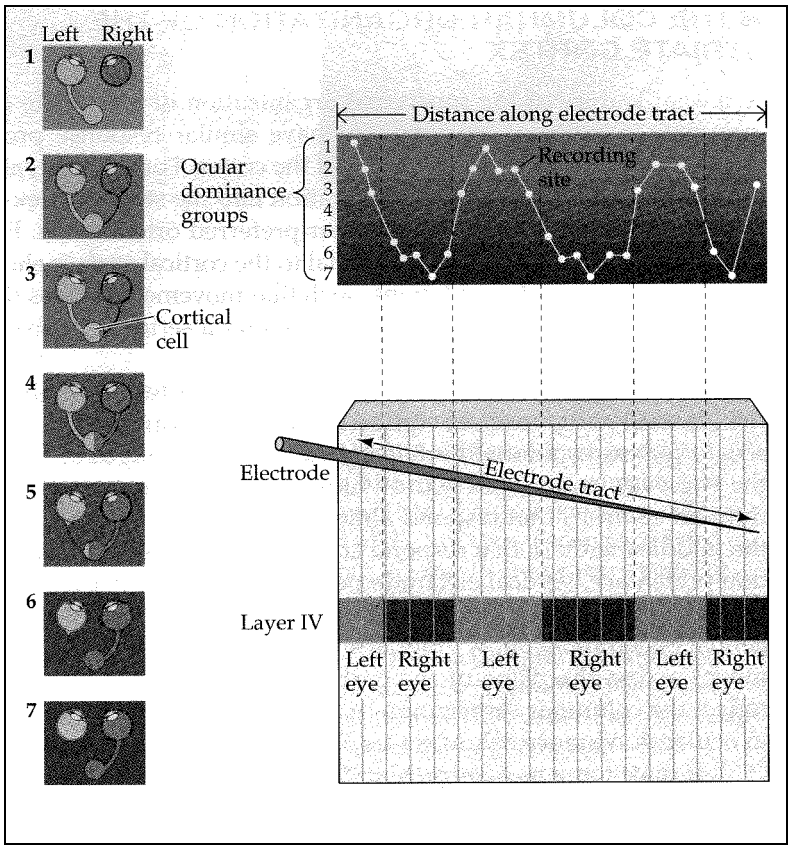
It may come as a surprise to learn that within the retinotopic map of visual space, response properties other than location are represented in an orderly fashion. The map of visual space is relatively gross; on a finer scale, each small region in visual space is represented within the receptive fields of neurons that are distributed over several millimeters of the cortical surface. This is more than enough cortical area to accommodate the columns of cells that are required to cover the complete range orientation preferences and ocular dominance values.

Although the modular arrangement of the striate cortex was first recognized on the basis of orientation and ocular dominance columns, further work has shown that other stimulus features, such as color, direction of motion, and spatial frequency are also distributed in iterated patterns that are systematically related to each other (for example, orientation columns tend to intersect ocular dominance columns at right angles). Thus the striate cortex is composed of repeating units, or modules, that contain all the neuronal machinery to analyze a small region of visual space for a variety of different stimulus attributes.



Columnar Organization of Orientation Selectivity In the Monkey Striate Cortex

Figure 1-23



Columnar Organization of Ocular Dominance

Figure 1-24

[1]Taken from J.E. Dowling, *The Retina: An Approachable Part of the Brain*, Belknap/Harvard University Press, 1987, Cambridge, MA, (pp. 242-243).

Appendix II: The Physics of Sight[1]

A. Wave Motion

Ocean waves travel thousands of miles across the ocean, but the particles of water do not make that journey. We are familiar with energy and momentum being transported from one place to another through the motion of particles; wave motion provides an alternative way for energy and momentum to move from one place to another without material particles making that journey. Water waves, sound waves, and the oscillation of a stretched string are examples of *mechanical waves*, which are waves that travel through a deformable or elastic medium, They originate when some portion of the medium is displaced from its normal position and released. Because of the elastic properties of the medium, the disturbance propagates through the medium. On the microscopic level, such mechanical properties as the forces between atoms are responsible for the propagation of mechanical waves. In this section we'll concentrate on the study of mechanical waves, a particularly simple type of which involves the oscillation of a stretched string. This will allow us to illustrate some of the general properties of waves without incurring too much mathematical complexity. Light waves are not mechanical waves, but our discussion here will be applicable to them as well.

As a wave reaches a particle in the medium, it sets that particle in motion and displaces it, thus transferring both kinetic and potential energy to it. Not only energy but also information about the nature of the wave source can be transmitted over considerable distances by wave motion. We can regard the particles of the medium as moving by only small distances about their previous positions as the wave passes. For example, in waves on the surface of water the particles of water move slightly, both up and down and back and force, tracing out small elliptical paths as the water waves move by. After the wave passes, the particle is more or less where it started before the wave passed.

Types of Waves

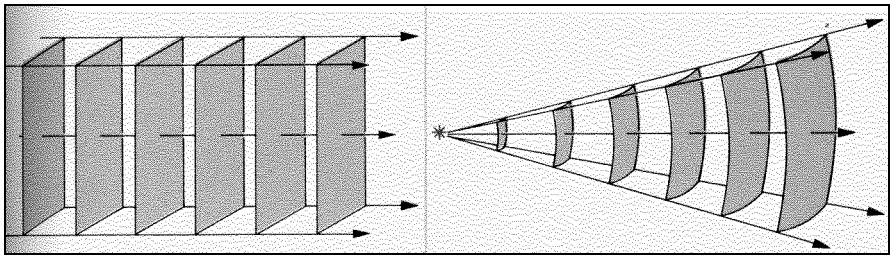
We can distinguish different kinds of mechanical waves by considering how the direction of motion of the particles of matter is related to the direction of propagation of the wave. If the motion of the particles is perpendicular to the direction of propagation of he wave itself, we have a *transverse* wave. A string under tension that is set oscillating back and forth at one end is an example of a transverse wave. The disturbance moves along the string but the string particles vibrate at right angles to the direction of propagation of the disturbance. (Light waves are also transverse waves.) If, however, the motion of the particles in a mechanical wave is back and forth along the direction of propagation, we have a *longitudinal* wave. Sound waves in a gas are longitudinal waves. Some waves are neither purely transverse nor purely longitudinal. Waves on the surface of water, as we pointed out above, trace out elliptical paths (up and down and back and forth) as the waves pass by.

Waves can also be classified as one-, two-, and three-dimensional, according to the number of dimensions in which they propagate energy. Waves moving along a string are one dimensional. Light waves traveling radially outward from a small source are three-dimensional. Waves may be classified further according to how the particles of the medium move in time. For example, we can produce a *pulse* traveling down a stretched string by applying a single sidewise movement at its end. Each particle remains at rest until the pulse reaches it, then it moves during a short time, then it again remains a rest. If we continue to move the end of the string back and forth, we produce a *train of waves* traveling along the string. If our motion is periodic, we produce a *periodic train of waves* in which each particle o the string has a periodic motion. The simplest special case of a periodic wave is a *harmonic wave*, in which each particle undergoes simple harmonic (sinusoidal) motion.

Imagine a stone dropped in a still lake. Circular ripples spread outward from the point where the stone entered the water. Along a given circular ripple, all points are in the same state of motion. Those points define a surface called a *wavefront*. If the medium is of uniform density, the direction of motion of the waves is at right angles to the wavefront. A line normal to the wavefronts, indicating the direction of motion of the waves, is called a *ray*. Wavefronts can have many shapes. A point source at the surface of water produces two-dimensional waves with circular wavefronts and rays that radiate outward from the point of the disturbance. On the other hand, a very long stick dropped horizontally into the water would produce (near its center) disturbances that travel as straight lines, in which the rays are parallel lines. The three-dimensional analogy, in which the disturbances travel in a single direction, is the *plane wave*. At a given instant, conditions are the same everywhere on any plane perpendicular to the direction of propagation. The wavefronts are planes, and the rays are parallel straight lines, as

[1] The material in this Appendix and all of the figures have been taken from:
Halliday, Resnick, and Crane, *Physics*, John Wiley & Sons, Inc., New York, 1992 (Chapters 19, 28, 34, 41, 43, 45, 46, and 48);
Smith and Atchison, *The Eye and Visual Optical Instruments*, Cambridge University Press, Cambridge, UK, 1997 (Chapters 26, 34, and 35)
Born and Wolf, *Principles of Optics*, Cambridge University Press, Cambridge, UK, 1999 (Chapter 8).

shown in **Figure 2-1a**. The three-dimensional analogy of circular waves is spherical waves. Here the disturbance is propagated outward in all directions from a point source of waves. The wavefronts are spheres, and the rays are radial lines leaving the point source in all directions, as shown in **Figure 2-1b**. Far from the source the spherical wavefronts have very small curvature, and over a limited region they can often be regarded as planes. Many other possible wavefront shapes are of course possible.



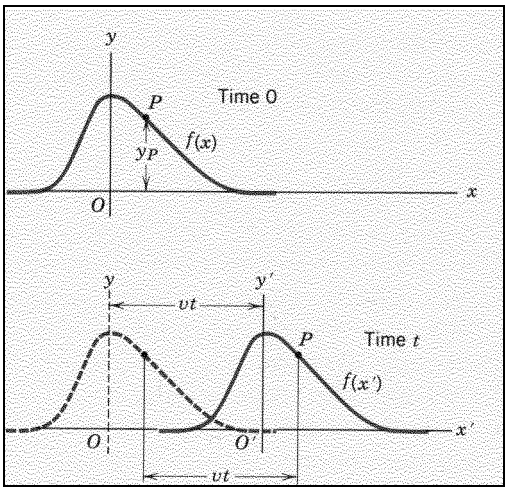
a **b**

A Plane Wave and a Spherical Wave

Figure 2-1

Traveling Waves

Consider again a transverse waveform that travels on a long stretched string. We assume an "ideal" string, in which the disturbance, whether it is a pulse or a train of waves, keeps its form as It travels. For this to occur, frictional losses and other means of energy dissipation must be negligibly small. The disturbance lies in the xy plane and travels in the x direction. The upper plot in **Figure 2-2** shows an arbitrary waveform at t=0, which we can consider to be a pulse traveling along the string. let the pulse move in the



A Transverse Pulse Traveling Along a String at Times 0 and t.

Figure 2-2

positive x direction with speed *v*. At a later time t, the pulse has moved a distance *v*t, as noted in the lower plot of **Figure 2-2**. Note that the waveform is the same at t=0 as it is at later times. the coordinate y indicates the transverse displacement of a particular point on a string. This coordinate depends on both the position x and the time t. we indicate this dependence on tow variables as y(x,t). We can represent the waveform of **Figure 2-1a** as

$$y(x,0)=f(x), \tag{2.1}$$

where f is a function that describes the shape of the wave. At time t, the waveform must still be described by the same function f, because we have assumed that the shape does not change as the wave travels. relative to the origin O' of a reference frame that travels with the pulse, the shape is described by the function f(x'), as indicated in the lower part of **Figure 2-2**. The relationship between the x-coordinates in the two reference frames is

$$x' = x - vt, \tag{2.2}$$

as can be seen from **Figure 2-2**. Thus, at time t, the wave is described by

$$y(x,t) = f(x') = f(x - vt). \tag{2.3}$$

That is, the function f(x-*v*t) has the same shape relative to the point x=*v*t at time t that the function f(x) has relative to the point x=0 at time t=0. To describe the wave completely, we must specify the function f. Later, we shall consider harmonic waves, for which f is a sine or cosine function.

Equs. (2.1) and (2.3) together indicate that we can change a function of any shape into a wave traveling in the positive x direction by merely substituting the quantity x-*v*t for x everywhere that it appears in f(x). For example, if $f(x)=x^2$, then $f(x-vt)=(x-vt)^2$. Furthermore, a wave traveling in the positive x direction must depend on x and t *only* in the combination x-*v*t; thus $x^2-(vt)^2$ does not represent such a traveling wave. Let us follow the motion of a particular part (or *phase*) of the wave, such as that of location P of the waveform of **Figure 2-2**. If the wave is to keep its shape as it travels, then the y coordinate y_p of P must not change. We see from Equ. (2.3) that the only way this can happen is for the x coordinate of P to increase as t increases in such a way that the quantity x-*v*t keeps a fixed value. That is, evaluating the quantity x-*v*t gives the same result at P in the lower part of **Figure 2-2** and at P in the upper part of **Figure 2-2**. this remains true for any location on the waveform and for all times t. Thus for the motion of any particular phase of the wave we must have

$$x - vt = \text{constant}. \tag{2.4}$$

We can verify that Equ. (2.4) characterizes the motion of the phase of the waveform by differentiating with respect to time, which gives

$$\begin{aligned} \frac{dx}{dt} - v &= 0, \\ \frac{dx}{dt} &= v. \end{aligned} \tag{2.5}$$

the velocity dx/dt describes the motion of the phase of the wave, and so it is known as the *phase velocity*. We take *v* to be a positive constant, independent of any property of the wave but possibly (as we shall see) depending on the properties of the medium. If the wave moves in the *negative* x direction, we only need to replace *v* by –*v*. In this case we would obtain

$$y(x,t) = f(x + vt), \tag{2.6}$$

where once again f(x) represents the shape at t=0. That is, substituting if f(x) the quantity x+*v*t in place of x gives a wave that would move to the left in **Figure 2-2**. the motion of any phase of the wave would then be characterized by the requirement that

$$x + vt = \text{constant}, \tag{2.7}$$

and by analogy with Equ. (2.5) we can show that

$$\frac{dx}{dt} = -v, \tag{2.8}$$

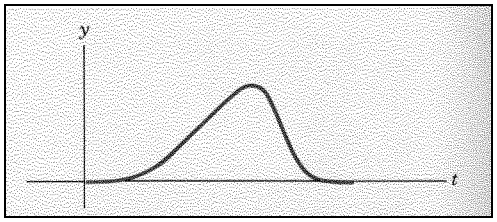
indicating that the x component of the phase velocity in this case is indeed negative.

The function f(x) contains the complete description of the shape of the wave and its motion. At any particular time, say t_1, the function $y(x,t_1)$ gives y as a function of x, which defines the curve; this curve represents the actual shape of the string at that time and can be regarded as a "snapshot" of the wave. On the other hand, we can consider the motion of a particular point on the string, say at the fixed coordinate x_1. The function $y(x_1,t)$ then tells us the y coordinate of that point as a function of the time. **Figure 2-3** shows how a point on the x axis might move with time as the pulse of **Figure 2-2** passes, moving in the positive x direction. At times near t=0, the point is not moving at all. It then begins to move gradually, as the leading edge of the pulse of **Figure 2-2** arrives. After the peak of the wave passes, the displacement of the point drops rapidly back to zero as the trailing edge passes. Note that the form of **Figure 2-3** appears to be reversed from the shape of the waveform in **Figure 2-2**. It is, since the leading edge of the traveling pulse arrives at the point of interest at the earliest times.

Sinusoidal Waves

The above description is quite general. It holds for arbitrary wave shapes, and it holds for transverse as well as longitudinal waves. Let us consider, for example, a transverse waveform having a sinusoidal shape, which has particularly important applications. suppose that at the time t=0 we have a wavetrain along a string given by

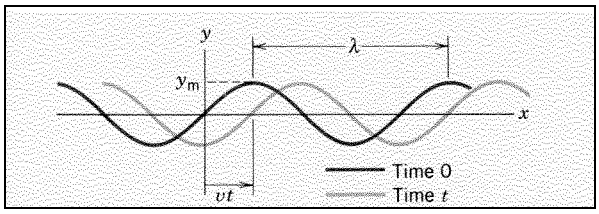
$$y(x,0)=y_m \sin\left(\frac{2\pi}{\lambda}x\right). \tag{2.9}$$



Displacement of x_1 Due to the Traveling Pulse of Figure 2-2

Figure 2-3

This wave shape is shown in **Figure 2-4**. The maximum displacement y_m is called the *amplitude* of the sine curve. The transverse



A Sinusoidal Wavetrain

Figure 2-4

displacement y ahs the same value at any x as it does at x=λ, x+2λ, and so on. The symbol λ represents the *wavelength* of the wavetrain and indicates the distance between two adjacent points in the wave having the same phase. If the wave travels in the +x direction with phase speed *v*, then the equation of the wave is

$$y(x,t)=y_m \sin\left(\frac{2\pi}{\lambda}(x-vt)\right). \tag{2.10}$$

Note that this has the form f(x-*v*t) required for a traveling wave, given by Equ. (2.3).

The *period* T of the wave is the time necessary for a point at any particular x coordinate to undergo one complete cycle of transverse motion. During this time T, the wave travels a distance *v*T that must correspond to one wavelength λ, so that

$$\lambda=vT. \tag{2.11}$$

The inverse of the period is called the *frequency* ν of the wave:

$$\nu=\frac{1}{T}. \tag{2.12}$$

Frequency has units of cycles per second, or *Hertz* (Hz). Substituting Eau. (2.12) into Equ. (2.10) gives another expression for the wave:

$$y(x,t)=y_m \sin\left(2\pi\left(\frac{x}{\lambda}-\frac{t}{T}\right)\right). \tag{2.13}$$

From this form it is clear that y, at any given time, has the same value at x, x+λ, x+2λ, and so on, and that y at any given position, has the same value at the times t, t+T, t+2T, and so on. To reduce Equ. (2.13) to a more compact form, we introduce two quantities, the *wave number* k and the *angular frequency* ω. These are defined as

$$\begin{aligned} k &\equiv \frac{2\pi}{\lambda}, \\ \omega &\equiv \frac{2\pi}{T} = 2\pi\nu. \end{aligned} \tag{2.14}$$

The wave number k is, like ω, and angular quantity, and units for both involve radians. Units for k might be, for instance, rad/m, and for ω, rad/s. In terms of these quantities, the equation of a sine wave traveling in the positive x direction is

$$y(x,t) = y_m \sin(kx - \omega t), \tag{2.15}$$

while the equation for a wave traveling in the negative x direction is

$$y(x,t) = y_m \sin(kx + \omega t). \tag{2.16}$$

Comparing Equs. (2.11) and (2.14), we see that the phase speed *v* of the wave is given by

$$v = \lambda\nu = \frac{\lambda}{T} = \frac{\omega}{k}. \tag{2.17}$$

Phase and Phase Constant

In the travelling waves of Equs. (2.15) and (2.16) we have assumed that the displacement of y is zero at the position x=0 and at time t=0. This, of course, need not be the case. The general expression for a sinusoidal wave traveling in the positive x direction is

$$y(x,t) = y_m \sin(kx - \omega t - \phi). \tag{2.18}$$

The quantity that appears in the argument of the sign, (kx-ωt-φ), is called the *phase* of the wave. Two waves with the same phase (or with phases differing by any integer multiple of 2π) are said to be "in phase"; they execute the same motion at the same time. The angle φ is sometimes called the *phase constant*. the phase constant does not affect the shape of the wave; it moves the wave forward or backward in space or time. To see this, we rewrite Equ. (2.18) in two equivalent forms:

$$\begin{aligned} y(x,t) &= y_m \sin\left[k\left(x - \frac{\phi}{k}\right) - \omega t\right], \\ y(x,t) &= y_m \sin\left[kx - \omega\left(t + \frac{\phi}{\omega}\right)\right]. \end{aligned} \tag{2.19}$$

The upper plot in **Figure 2-5** shows a "snapshot" at any time t of the two waves represented by Equ. (2.15) (in which φ=0) and Equ. (2.18). Note that any particular point on the wave described by the first of Equs. (2.19) (say a certain wave crest) is a distance φ/k *ahead of* the corresponding point in the wave described by Equ. (2.15). Equivalently, if we were to observe the displacement at a fixed position x resulting from each of the two waves represented by Equs. (2.15) and (2.18), we would obtain the result indicated by the lower plot in **Figure 2-5**. The wave described by the second of Equs. (2.19) is similarly *ahead of* the wave having φ=0, in this case by a time difference φ/ω.

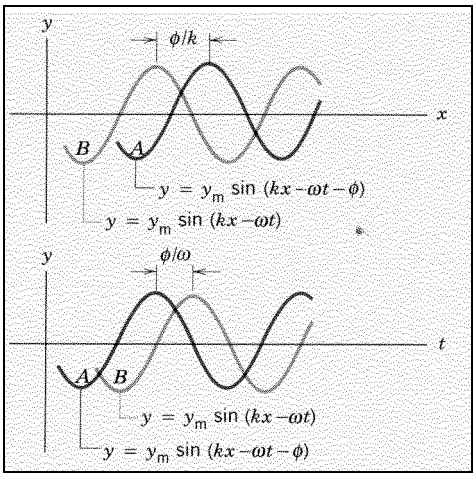
When the phase constant in Equ. (2.18) is positive, the corresponding wave is ahead of a wave described by a similar equation having φ=0. It is for this reason that we introduced the phase constant with a negative sign in Equ. (2.18). when one wave is ahead of another in time or space, it is said to "lead". On the other hand, putting a negative phase constant into Equ. (2.18) moves the corresponding wave behind the one with φ=0. such a wave is said to "lag". If we fix our attention on a particular point of the string, say x_1, the displacement y at that point can be written

$$y(t) = y_m \sin(\omega t + \phi'), \tag{2.20}$$

where we have substituted

$$\phi' = \phi - kx_1. \tag{2.21}$$

This is an expression of *simple harmonic motion*. Hence, any particular element of the string undergoes simple harmonic motion about its equilibrium position as this wavetrain travels along the string.

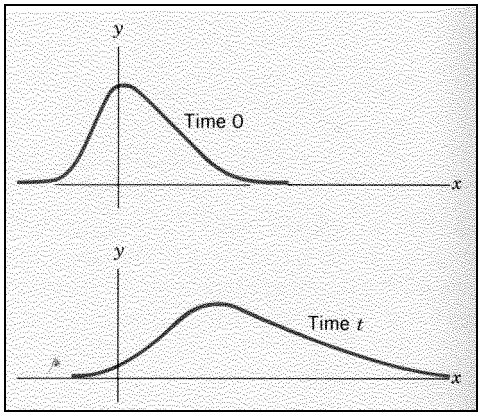


Two Sine Waves with Different Phase Constants Traveling in the Positive Direction

Figure 2-5

Group Speed and Dispersion

Pure sinusoidal waves are useful devices for helping us understand wave motion. In practice, we use other kinds of waves to transport energy and information. These waves may be periodic but non-sinusoidal, such as square waves or sawtooth waves, or they may be non-periodic pulses. We have used the phase speed to describe the motion of two kinds of waves: the pulse that preserves its shape as it travels and the pure sine wave. In other cases, we must use a different speed, called the *group speed*, which is the speed at which energy or information travels in a real wave. **Figure 2-6** shows a pulse traveling through a medium. The shape of the pulse



A Pulse Traveling Through a Dispersive Medium

Figure 2-6

changes as it travels; the pulse spreads out, or *disperses*. (Note that dispersion is not the same as energy dissipation. The energy content of the pulse in **Figure 2-6** remains constant as it travels, even though the pulse disperses. We are assuming, then, that the medium is dispersive, but not dissipative.) Any periodic wave can of course be regarded as the sum of a series of sinusoidal waves of different frequencies or wavelengths--the frequencies, amplitudes, and phases being prescribed by Fourier's Theorem. In most real media, the speed of propagation of these component waves (that is, the phase speed) depends on the frequency or wavelength of the particular component. Each component wave may travel with its own unique speed. Thus, as the wave travels, the phase relationships of the components may change, and the waveform of the sum of the components will correspondingly change as the wave travels. This is the origin of dispersion—the component waves travel

at different phase speeds. There is no simple relationship between the phase speeds of the components and the group speed of the wave; the relationship depends on the dispersion of the medium.

Some real media are approximately non-dispersive, in which case the wave keeps its shape, and all component waves travel with the same speed. Sound waves in air are one such an example. (If this were not the case, conservation would be impossible, because the waveform produced by one individual's vocal cords would be jumbled by the time it reached another's ear.) Light waves in vacuum are perfectly non-dispersive. Other media through which light can travel is, however, dispersive, giving rise to such effects as the spectrum of colors in rainbows. In a non-dispersive medium, all the component waves in a complex waveform travel at the same phase speed, and the group speed of the waveform is equal to that common value of the phase speed. Only in this case can we speak of the phase speed of the entire waveform.

Transverse Velocity of a Particle

The motion of a particle in a transverse wave such as that of **Figure 2-4** is in the y direction. the wave speed describes the motion of the wave along the direction of travel (the x direction). The wave speed does *not* characterize the transverse motion of the particles of the string. to find the transverse velocity of a particle of the string, we need the change in the y coordinate with time. We focus our attention on a single particle of the string, that is, on a certain coordinate x. We therefore need the derivative of y with respect to t at constant x, and we represent the particle velocity, which varies with x (the location of the particle) as well as with t, as u(x,t). Assuming we are dealing with a sinusoidal wave of the form of Equ. (2.18), we then have

$$\begin{aligned} u(x,t) &= \frac{\partial y}{\partial t} = \frac{\partial}{\partial t}\left[y_m \sin\left(k x - \omega t - \phi\right)\right], \\ &= -y_m \omega \cos\left(k x - \omega t - \phi\right). \end{aligned} \tag{2.22}$$

Differentiating again, the transverse acceleration a(x,t) of the particle at this location is

$$\begin{aligned} a(x,t) &= \frac{\partial^2 y}{\partial t^2} = \frac{\partial u}{\partial t} = -{y_m}^2 \sin(k x - \omega t - \phi), \\ &= -\omega^2 y. \end{aligned} \tag{2.23}$$

Power and Intensity in Wave Motion

If we were shaking the end of a string, a person at the other end could extract the resulting energy (which is transported along the string in the form of the potential and kinetic energy of its elements) and use the energy to do work on another system. Such transport of energy (and momentum) is in fact one of the purposes for which we produce waves. Here we consider the rate at which the string transports energy. **Figure 2-7** shows a snapshot of the wave at times t and t+dt. A point on the string at coordinate x has at time t a transverse velocity **u**, which has only a y component. This velocity, as we discussed in the previous section, is *not* related to the phase speed of the wave but instead has magnitude given by Equ. (2.22) with ϕ=0,

$$u(x,t) = -y_m \omega \cos\left(k x - \omega t\right). \tag{2.24}$$

for a sinusoidal wave of the form of Equ. (2.15). The force exerted on an element of the string by the element to its left is also shown in **Figure 2-7**. This force transmits energy at a rate given by

$$P = \mathbf{u} \cdot \mathbf{F} = u F_y. \tag{2.25}$$

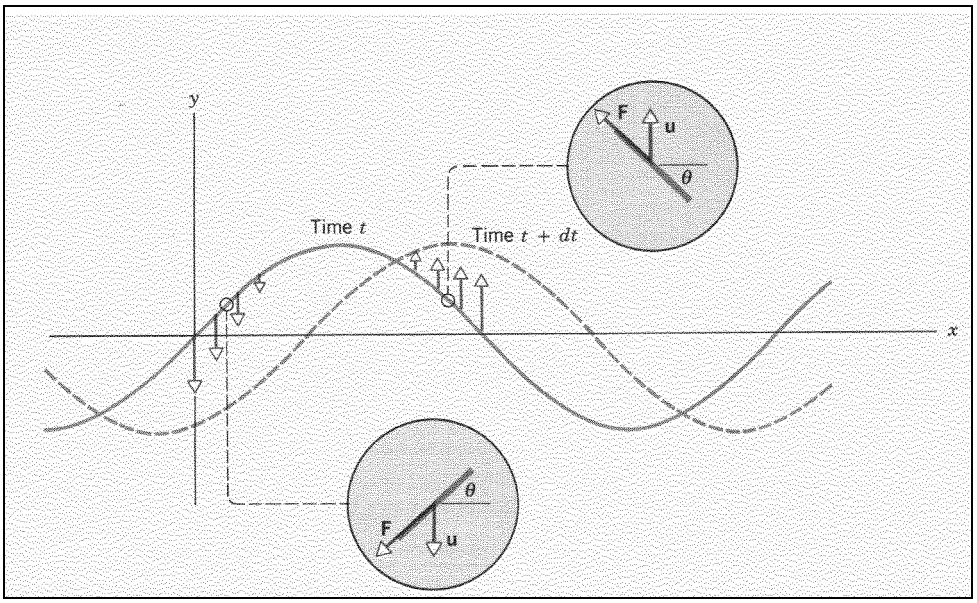
Only the component F_y of **F** along **u** contributes to the power; this component is F$\sin\theta$, which for small displacements can be approximated as

$$F \tan\theta = F \frac{\partial y}{\partial x}, \tag{2.26}$$

where $\partial y / \partial x$ is the slope of the string at the coordinate x.

Note that the y component of **F** is parallel to **u**, no matter whether that element of string happens to be moving up or down. Thus $u F_y \geq 0$, and so the power transmitted is never negative during the cycle of oscillation. *There is a continuous*

net flow of energy in the positive x direction (the direction of propagation of the wave). Substituting for the y component of the force, we obtain



Instantaneous Transverse Velocities of Different Points along a String

Figure 2-7

$$\begin{aligned} P &= u F_y = \left(\frac{\partial y}{\partial t}\right)\left(F\frac{\partial y}{\partial x}\right), \\ &= F\left[-\omega y_m \cos(kx-\omega t)\right]\left[-k y_m \cos(kx-\omega t)\right], \\ &= y_m^2 k\omega F \cos^2(kx-\omega t). \end{aligned} \tag{2.27}$$

Note also that the power or rate of flow of energy is not constant. This is because the input power oscillates: the work done by our hand as it moves the end of the string varies with the transverse displacement of that point. As energy is transported along the string, it is stored in each element of the string as a combination of kinetic and potential energy of deformation. The power input to the string is often taken to be the average over one period of motion. It is given by

$$\bar{P} = \frac{1}{T}\int_t^{t+T} P\,dt, \tag{2.28}$$

where T is the period. The average value of $\sin^2\theta$ or $\cos^2\theta$ over one cycle is $\frac{1}{2}$, so after substituting Equ. (2.27) into Equ. (2.28) we obtain

$$\bar{P} = \frac{1}{2} y_m^2 k\omega F. \tag{2.29}$$

If m is the mass per unit length of the string, we can show that

$$F = v^2 \mu. \tag{2.30}$$

Substituting this result and Equ. (2.17) into Equ (2.29) then gives

$$\begin{aligned}\overline{P} &= \frac{1}{2} y_m^2 \, k \omega v \nu \mu, \\ &= \frac{1}{2} y_m^2 \, k \omega \frac{\omega}{k} \nu \mu, \\ &= \frac{1}{2} y_m^2 \, \mu \nu \omega^2,\end{aligned} \tag{2.31}$$

a result that does not depend on x or t. The dependence of the rate of transfer of energy on the *square* of the wave amplitude and the *square* of the wave frequency is in general true for all types of waves, including light waves.

In a three-dimensional wave such as a light, it is often more useful to specify the *intensity* of the wave. The intensity I is defined as the *average power per unit area transmitted across an area A normal to the direction in which the wave is traveling*, or

$$I = \frac{\overline{P}}{A}. \tag{2.32}$$

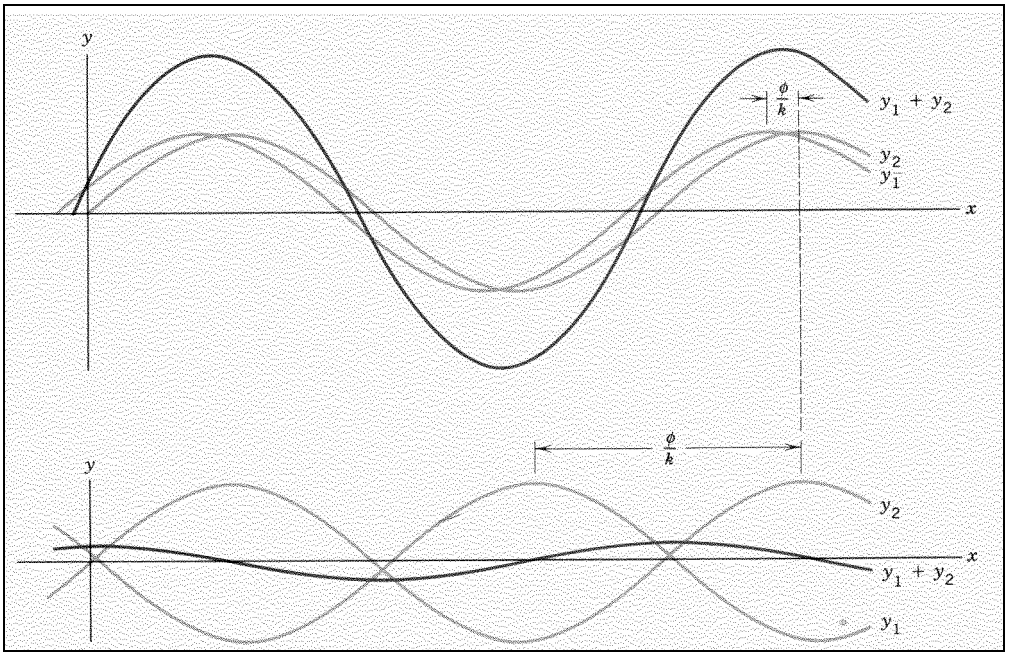
Just as with power in the wave traveling along the string, the intensity of any wave Is always proportional to the square of the amplitude. For circular or spherical waves, the amplitude is not constant as the wavefront travels. For example, in the case or a spherical wavefront, the intensity at any point a distance r from its point source is, by Equ. (2.32),

$$I = \frac{\overline{P}}{4 \pi r^2}. \tag{2.33}$$

The wave intensity thus varies inversely as the square of its distance from the source. Since the intensity is proportional to the square of the wave's amplitude, the amplitude must also vary, in this case inversely with the distance from the source. Thus, in doubling the distance form the source, the amplitude of a spherical wave decreases by half, and its intensity by one quarter.

Interference of Waves

When two or more waves combine at a particular point, they are said to *interfere*, and the phenomenon is called *interference*. As we shall see, the resultant waveform is strongly dependent on the relative phases of the interfering waves. Let us first consider two transverse sinusoidal waves of equal amplitude and wavelength, which travel in the x direction with the same speed. We take the phase constant of one wave to be ϕ, while the other has $\phi=0$. **Figure 2-8** shows the combined waveform at a particular time for the two



Interference of Two Waves

Figure 2-8

cases of ϕ nearly zero (the waves are nearly in phase) and ϕ nearly 180^o (the waves are nearly out of phase). We can see by merely adding the individual displacements at each x that in the first case there is nearly complete *reinforcement* of the two waves and the resultant has nearly double the amplitude of the individual components, while in the second case there is nearly complete *cancellation* at every point and the resultant amplitude is close to zero. These cases are known, respectively, as *constructive* interference and *destructive* interference.

Let us see how interference arises from the equations for the waves. We consider a general case in which the two have phase constants ϕ_1 and ϕ_2, respectively. The equations for them are then, from Equ. (2.18),

$$\begin{aligned} y_1(x,t) &= y_m \sin(k\,x - \omega t - \phi_1), \\ y_2(x,t) &= y_m \sin(k\,x - \omega t - \phi_2). \end{aligned} \tag{2.34}$$

Because this is a linear system, we can use superposition to find the resultant wave. Summing the two waves together, then, gives

$$\begin{aligned} y(x,t) &= y_1(x,t) + y_2(x,t), \\ &= y_m \left[\sin(k\,x - \omega t - \phi_1 + \sin(k\,x - \omega t - \phi_2)\right]. \end{aligned} \tag{2.35}$$

Applying the trigonometric identity

$$\sin A + \sin B = 2\sin\left[\frac{1}{2}(A+B)\right]\cos\left[\frac{1}{2}(A-B)\right], \tag{2.36}$$

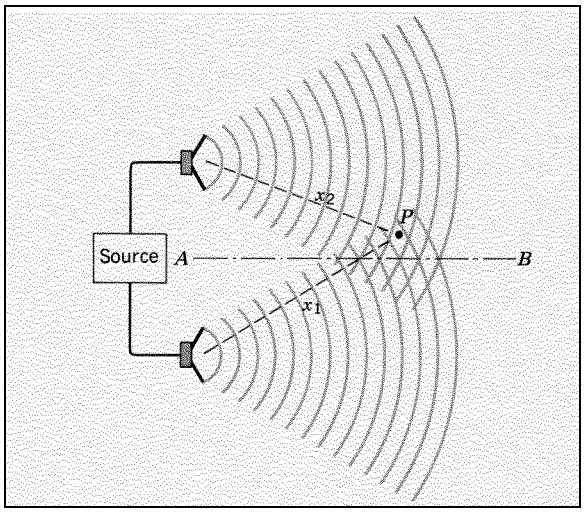
we obtain, after some rearrangement,

$$y(x,t) = 2\,y_m \cos\left(\frac{\phi_2 - \phi_1}{2}\right)\sin\left(k\,x - \omega t - \frac{\phi_1 + \phi_2}{2}\right). \tag{2.37}$$

The quantity $\Delta\phi = \phi_2 - \phi_1$ is called the *phase difference* between the two waves.

The resultant wave corresponds to a new wave having the same frequency but with an amplitude $2\,y_m\left|\cos\left(\Delta\phi/2\right)\right|$. If $\Delta\phi$ is very small (compared to 180^o) the resultant amplitude is nearly $2y_m$ (the upper plot in **Figure 2-8**). When $\Delta\phi$ is close to 180^o, on the other hand, the resultant amplitude is nearly zero (the lower plot in **Figure 2-8**). Notice that Equ. (2.37) has the form of a sinusoidal wave. Thus adding two sine waves of the same wavelength and amplitude always gives a sine wave of the identical wavelength. We can also add components that have the same wavelength but different amplitudes. In this case the resultant again is a sine wave with the identical wavelength, but the resultant amplitude does not have the simple form given by Equ. (2.37). If the individual amplitudes are y_{1m} and y_{2m}, then if the waves are in phase ($\Delta\phi=0$) the resultant amplitude is $y_{1m} + y_{2m}$, while if they are out of phase ($\Delta\phi=180^o$) the resultant amplitude is $\left|y_{1m} - y_{2m}\right|$. There can be no complete destructive interference in this case, although there is partial destructive interference.

Figure 2-9 shows an example of the occurrence of interference effects. Here two loudspeakers are driven from the same source.



Two Loudspeakers Driven by a Common Source
Figure 2-9

At points equidistant from the speakers (on the line AB, which represents the entire midplane), there is complete constructive interference if the speakers are driven in phase ($\Delta\phi=0$). There are also other points P where the waves arrive in phase and interfere constructively. That is, we can shift one of the waves in **Figure 2-9** by a phase constant of any integer multiple of 2π (or by a distance of any whole number of wavelengths), and the combined waveform is unchanged. these other points of constructive interference are located wherever the difference in distance to the two speakers is a whole number of wavelengths:

$$\left|x_1 - x_2\right| = \lambda,\ 2\lambda,\ 3\lambda,\ \ldots\ . \tag{2.38}$$

At other points P, the differing distances x_1 and x_2 result in the waves possible arriving at P out of phase, even if they started out in phase at the speakers. The listening environment might therefore have "dead spots" where partial or complete destructive interference occurs for a particular wavelength λ. Maximal destructive interference occurs at points where

$$\left|x_1 - x_2\right| = \frac{\lambda}{2},\ 3\frac{\lambda}{2},\ 5\frac{\lambda}{2},\ \ldots\ , \tag{2.39}$$

corresponding to a phase difference of 180^o, 540^o, 900^o, and so on.

Of course, if the speakers emit a mixture of many different wavelengths, some points P might show destructive interference for one wavelength and constructive interference for another. The critical factor in determining the locations of the maxima and minima of sound intensity is the *path difference* $\left|x_1 - x_2\right|$. At points not on the midplane represented by the

line AB, the tow components arrive with different amplitudes (because the distances from the speakers are not the same). There will thus be no complete destructive interference. (In certain geometries, it is possible for the sound radiated from the *back* of the speaker to interfere with the sound radiated from the front. These two waves are 180^o out of phase, and their interference can reduce the sound intensity at locations in front of the speaker. loudspeaker enclosures are designed to eliminate this effect.)

B. The Electromagnetic Field

The state of excitation which is established in space by the presence of electric charges is said to constitute an *electromagnetic field*. It is represented by two vectors, **E** and **B**, which we'll refer to as the *Electric and Magnetic Field Vectors*, respectively. The Electric Field is the space around a *non-moving*, charged particle that exerts a force of electric origin on other *non-moving*, charged particles. We define the electric field **E** associated with a certain collection of charges in terms of the force **F** exerted on a positive test charge q_0 at a particular point,

$$\mathbf{E} = \lim_{q_0 \to 0} \frac{\mathbf{F}}{q_0} \frac{\text{Newtons}}{\text{Coulomb}}, \tag{2.40}$$

where the limit is included to remind us that the test charge must be sufficiently small so as not to disturb the distribution of charges whose electric field we are trying to measure. Since q_0 is a positive scalar, the direction of **E** is the same as the direction of **F**. We are all familiar with electric field diagrams such as that shown in **Figure 2-10**. The lines here are *lines of force*. Such a diagram contains three important pieces of information regarding the field:

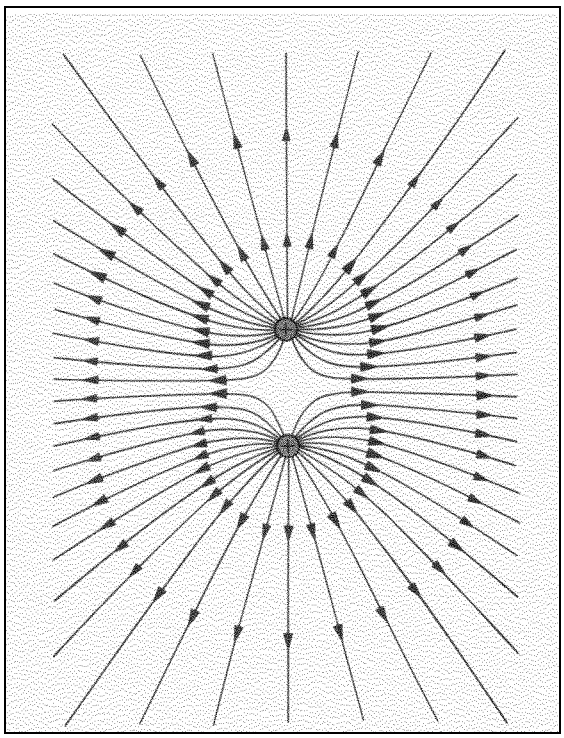
1. The lines of force give the direction of the electric field at any point, and hence the direction of the force acting on a positively charged particle at that point.
2. The lines of force originate on positive charges and terminate on negative charges. (In the figure, the negative charges are assumed to be at infinity).
3. The lines of force are drawn to that the number of lines per unit cross-sectional area (perpendicular to the lines) is proportional to the magnitude of the electric field. (The closer the lines are together, then, the stronger the field.)

Magnetic Fields are a little more complicated, due to the fact that *magnetic monopoles* do not exist (or are so exceedingly rare that relationships involving them are of no practical value). We thus define a Magnetic Field as the space around a *moving*, charged particle (or permanent magnet) that exerts a magnetic force on other *moving* charges (or magnets). We define the magnetic field **B** by the relation

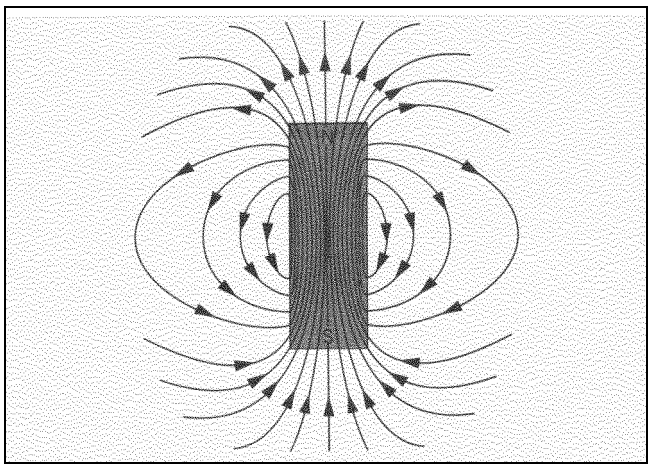
$$\mathbf{F} = \lim_{q_0 \to 0} q_0 \, \mathbf{v} \times \mathbf{B}, \tag{2.41}$$

where **v** x **B** denotes the vector product. **Figure 2-11** shows a magnetic field diagram, The essential properties of this diagram, though similar to those of the electric field diagram, also differ in important ways:

1. The lines of force give the direction of the magnetic field at any point. The direction of the magnetic force applied to a moving charge, as specified by the vector product in Equ. (2.41), is however *perpendicular* to the lines of B, as well as the direction of travel of the charge.
2. The lines of force form a closed loop, as indicated in the figure.
3. The lines of force are drawn so that the number of lines per unit cross-sectional area (perpendicular to the lines) is proportional to the magnitude of the magnetic field. (The closer the lines are together, then, the stronger the field.)



Lines of Force Surrounding Two Equal Positive Charges
Figure 2-10



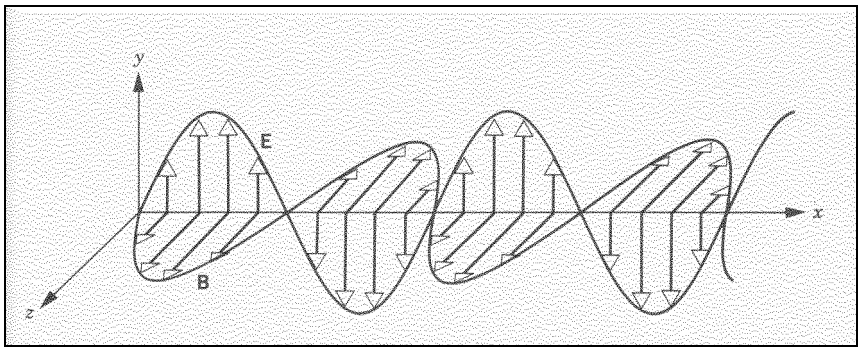
The Magnetic Lines of Force Produced by a Bar Magnet
Figure 2-11

The magnetic field vector **B** has the units $\frac{\text{Newtons}}{\text{Coulomb}\cdot\frac{\text{Meter}}{\text{Second}}}$ or Tesla .

If all field quantities are independent of time, and if, moreover, there are no moving charges present, the field is said to be *static*. If all the field quantities are time independent, but moving charges are present, we denote the field as a *stationary field*. In optical fields the field vectors are very rapidly varying functions of time, but the sources of the field are usually such

that, when averages over any macroscopic time interval are considered rather than the instantaneous values, the properties of the field are found to be independent of the instant of time at which the average is taken. The work *stationary* is often used in a wider sense to describe a field of this type.

Combining the ideas of the previous two sections, we show in **Figure 2-12** a sinusoidally varying electromagnetic traveling wave, propagating in the x direction. All such waves have four important properties, which follow from an analysis of Maxwell's Equations,



A Sinusoidally Varying Electromagnetic Traveling Wave
Figure 2-12

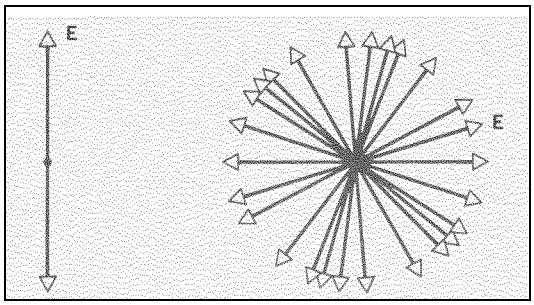
the governing equations of electromagnetism:

1. The electric and magnetic field vectors **E** and **B** are always perpendicular to one another, as sell as the direction of propagation of the wave.
2. **E** and **B** are always in phase.
3. The ratio of the amplitudes Em and B_m of **E** and **B** is the wave speed *v*, i.e.,

$$\frac{E_m}{B_m} = v \,. \tag{2.42}$$

4. The speed of an electromagnetic wave in a vacuum is the speed of light, i.e., $v = c$.

The wave illustrated in **Figure 2-12** is said to be *linearly polarized* (or *plane polarized*). This means that the **E** field remains in a fixed direction (here the y direction) as the wave propagates. By convention, we define the *direction of polarization* of the wave to be the direction of the **E** vector. The plane determined by the **E** vector and the direction of propagation of the wave (the xy plane in **Figure 2-12**) is called the *plane of polarization* of the wave. Ordinary sources of light, such as incandescent bulbs or the sun emit waves whose planes of polarization are randomly oriented about the direction of propagation. **Figure 2-13a** shows polarized light, as viewed from the direction of propagation. (Here only the **E** vector is shown). Unpolarized light, then, would be as shown in **Figure 2-13b**. En either case the light continues to be a transverse traveling wave.



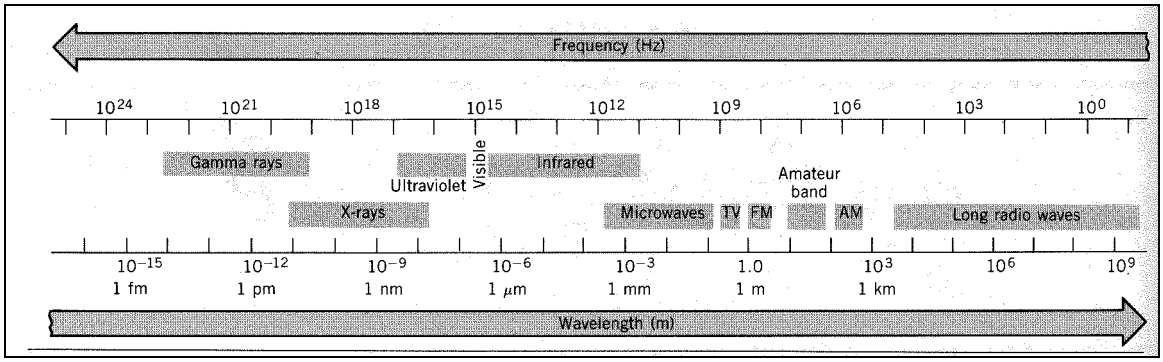
a **b**

Unpolarized Light, Viewed From Along the Direction of Propagation
Figure 2-13

C. Light and the Electromagnetic Spectrum

The electromagnetic spectrum, shown in **Figure 2-14**, includes a broad range of different kinds of radiation from a variety of sources. They differ greatly in their properties, in the means of their production, and in the ways we observe them. They share other features in common, however: They can all be described in terms of electric and magnetic fields, and they all travel through vacuum with the same speed (the speed of light). In fact, from the fundamental point of view, they differ *only* in wavelength (or frequency). The names given to the various regions of the spectrum have to do only with the way the different types of waves are produced or observed.; they have nothing to do with any fundamental property of the waves. Other than the difference in their wavelengths, there is no experimental way to distinguish a wave in the visible region from one in the infrared region; the waves have identical mathematical descriptions. There are no gaps in the spectrum, nor are there sharp boundaries between the various categories.

What we call light comprises only a very small part of the overall electromagnetic radiation spectrum. **Table 2-1** gives a finer breakdown of the visible portion of the spectrum. As indicated, our visual system distinguishes wavelength/frequency through the perception of color. Note from **Figure 2-14** and **Table 2-1** that light waves have extremely high frequencies (and correspondingly, extremely short wavelengths). The *instantaneous electromagnetic field* is way beyond the capabilities of our eyes to detect. Instead, our visual systems detect its *intensity*, which, as we showed in the previous section, is proportional to the square of the amplitude and



The Electromagnetic Spectrum

Figure 2-14

the square of the frequency. In general, only the electric field is important in optics. This is because the interaction of electromagnetic radiation with matter involves interaction of the radiation with the electrons in the material, and whereas the electric field interactswith all electrons, the magnetic field only interacts with fast moving electrons. The electrons in optical materials such as our eye move sufficiently slowly that their speed can be neglected and therefore they are only affected by the electric field component of the radiation.

The sources of visible light depend ultimately on the motion of electrons. Electrons in atoms can be raised from their lowest energy state to higher states by various means, such as heating the substance or by passing an electric current through it. When the electrons eventually drop back to their lowest levels, the atoms emit radiation that may be in the visible region of the spectrum. Emission of visible light is particularly likely when the outer valence electrons are the ones making the transitions. The color of the light tells us something about the atoms or the object from which it was emitted. The most familiar source of light is the sun. Its surface emits radiation across the entire electromagnetic spectrum, but its most intense radiation is in the region we define as visible, and the sun's radiant intensity peaks at a wavelength of about 550nm, which corresponds precisely to the peak in the sensitivity of a "standard" human observer. (This suggests that, through natural selection, our eyes evolved in such a way that their sensitivity matched the sun's spectrum.)

Color	Vacuum Wavelength (10^{-9} Meter)	Frequency (10^{12} Cycles per Second)
Red	780-622	384-482
Orange	622-597	482-503
Yellow	597-577	503-520
Green	577-492	520-610
Blue	492-455	610-659
Violet	455-390	659-769

The Visible Spectrum

Table 2-1

D. Geometrical and Physical Optics

In our description of wave motion, we used the *ray* as a convenient way to represent the motion of a train of waves; the ray is perpendicular to the wavefronts and indicates the direction of travel of the wave. A ray is a convenient geometrical construction that is often helpful in studying the optical behavior of a system such as a lens. A ray is not a physical entity, however, and it is not possible to isolate one. Consider a train of plane light waves of wavelength λ incident on a barrier in which there is a slit of width a. As suggested by **Figure 2-15a**, if $\lambda \gg a$, the waves pass through the slit, and the barrier forms a sharp "shadow". As we make the slit smaller, we find that the light flares out into what was formerly the shadow of the barrier, as shown in **Figure 2-15b**. This phenomenon, known as *diffraction*, occurs when the size of the slit (or other obstacle) in the path of the wave is comparable to the wavelength. (We consider diffraction in detail later.) Note (**Figure 2-15c**) that diffraction becomes more pronounced as the slit width becomes smaller; thus an attempt to isolate a single ray will be futile.

If a is a measure of the smallest transverse dimension of a slit or obstacle, then the effects of diffraction can be ignored if the ratio a/λ is small large enough. In this case, the light appears to travel in straight-line paths that we can represent as rays. This is the condition for *geometrical optics*, also known as *ray optics*. When a light beam encounters such obstacles as mirrors, lenses, or prisms whose lateral size is much greater than the wavelength of light, we are safe in using the equations of geometrical optics. Perhaps the two most important result derived from geometrical optics are the Law of Reflection and the Law of Refraction. Both involve light passing from one medium to another, as in **Figure 2-16**. (Notice here how rays are used to represent the paths of the incident, reflected, and refracted light waves.) The *Law of Reflection* states that the reflected ray lies in the same plane as the incident ray, and that

$$\theta_1' = \theta_1 . \tag{2.43}$$

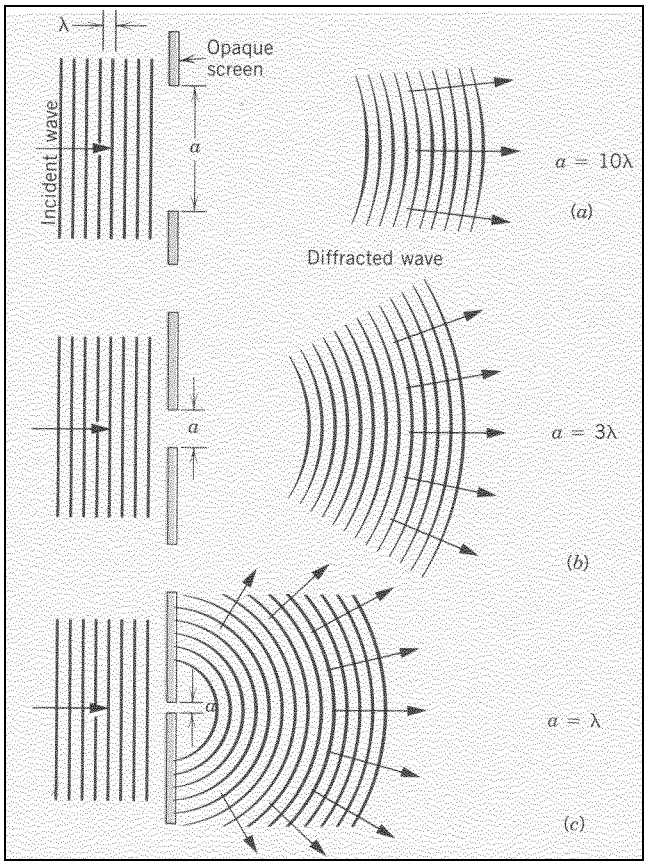
The *Law of Refraction* states that the refracted ray lies in the same plane as the incident ray, and that

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 , \tag{2.44}$$

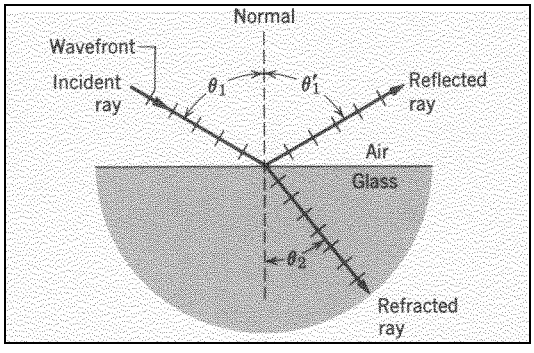
where the n_i are the indices of refraction, defined as the ration of the speed of light c in vacuum to the speed of light *v* in the corresponding medium,

$$n_i = \frac{c}{v_i}. \tag{2.45}$$

If the condition for geometrical optics is not met, we cannot describe the behavior of light by rays but must take its wave nature specifically into account. In this case we are in the realm of *physical optics* or *wave optics*, which includes geometrical optics as a limiting case. Physical Optics, on the other hand, describes the phenomena of Interference and Diffraction, which we take up shortly.



A Train of Plain Light Waves Incident Upon Slits of Varying Widths
Figure 2-15

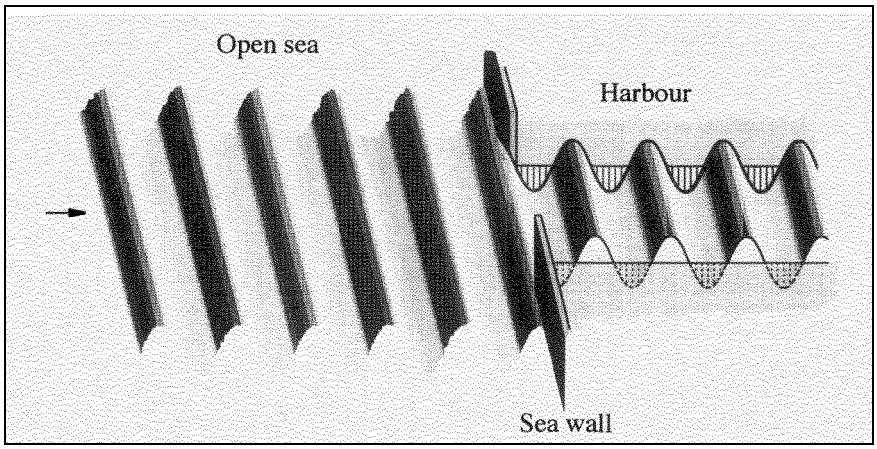


Geometrical Optics and the Laws of Reflection and Refraction
Figure 2-16

E. The Theory of Diffraction[1]

Huygen's Principle

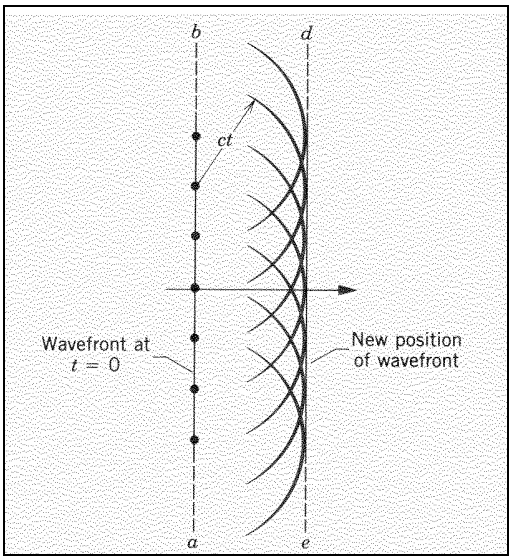
In a number of situations, it can be observed that when a wave motion passes through an aperture (opening), the waves bend around the edge of the aperture. Sea waves entering an enclosed harbor, for example, will bend around the seawall. The effect is called *diffraction*. To understand it better, let us consider first what would happen if diffraction did not take place. **Figure 2-17** shows parallel waves entering a harbor in the absence of diffraction. Upon passing through the seawall, the waves would remain parallel



Sea Waves Entering an Enclosed Harbor in the Absence of Diffraction

Figure 2-17

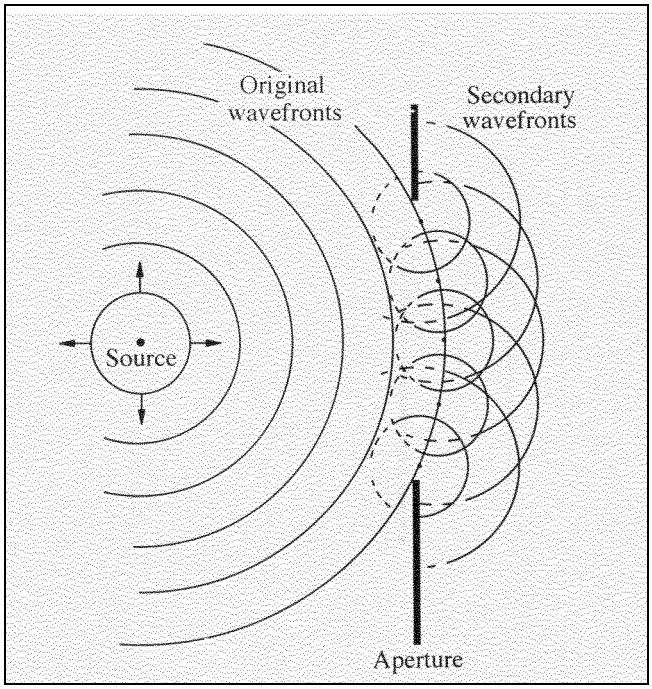
with a sharp edge at their boundary, producing a vertical wall of water with a sinusoidal profile at the edge, as shown in the figure. This, of course, is not possible. Huygens (1690) suggested that every point on a wave motion becomes a new and secondary spherical source with an amplitude equal to the amplitude of the original wave at that point. This is known as Huygen's Principle. For an unbounded wave, the effect of these secondary waves cancels out except in the original direction of the wave, and in that direction their cumulative effect is identical to the original wave motion, as shown if **Figure 2-18**. These secondary wave fronts only become manifest near the



Huygens' Wavelets Applied to an Unbounded Wave

Figure 2-18

edge of a bounded wave, for example, when a wave passes through an aperture, as in **Figure 2-19**. Returning now to our water wave example in **Figure 2-17**, and applying these ideas to just one of the edges of the seawall, we have the construction shown in **Figure 2-20**. Summing the amplitudes of these secondary wave fronts, taking care to account for phase differences, gives the result of

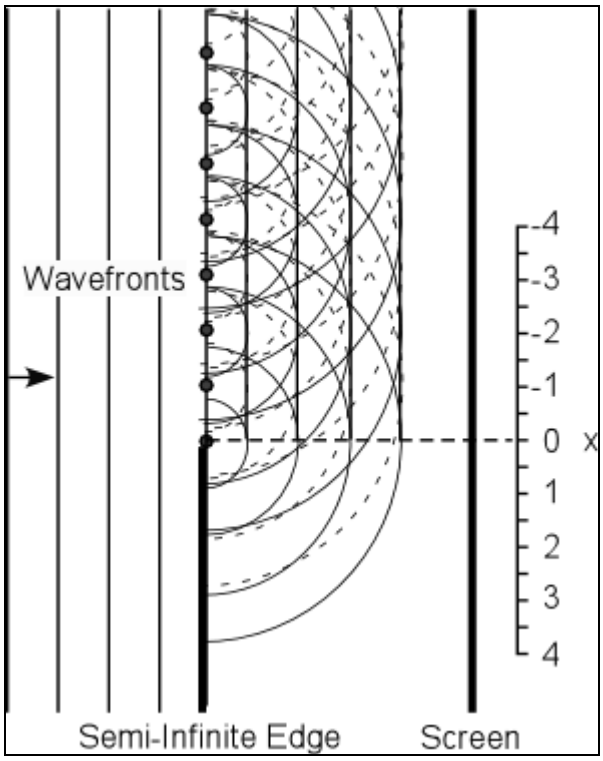


Huygens' Principle and Diffraction by an Aperture

Figure 2-19

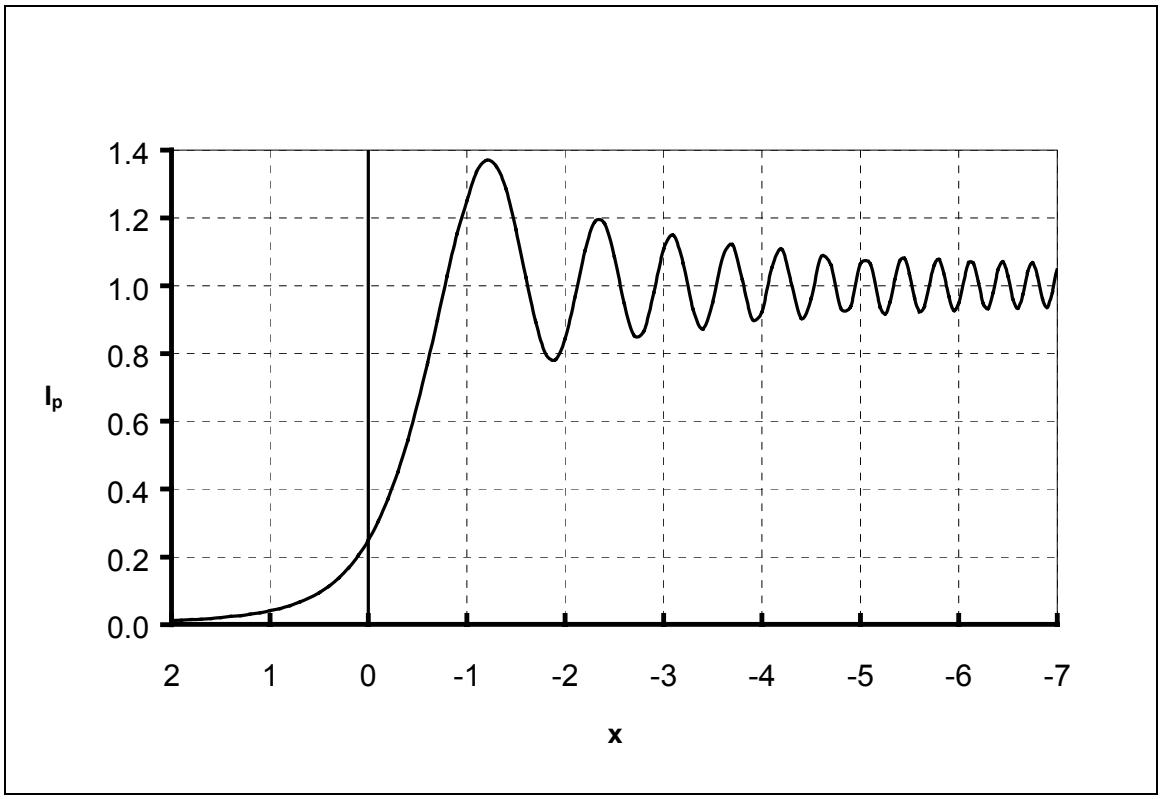
Figure 2-21. The shape of a water wave after entering the harbor, then, will be a superposition of two sinusoidal shapes at right angles to one another. The first is the "pure sinusoid" of **Figure 2-17** (evident when viewed at right angles to the direction of propagation), and the second is the "damped sinusoid" of **Figure 2-21** (evident when viewed along the direction of propagation**)**. The wave's overall amplitude at any particular point along it will be a decreasing function of that point's distance from the seawall opening. If the seawall opening is "wide" (many times the wavelength of the water waves), then the edge effect of **Figure 2-21** will occur over a relatively insignificant length and in the middle the wave will have just the sinusoidal shape of **Figure 2-17**. This is the case of geometrical optics, in which diffraction is ignored, which we described previously. Note that the edge effects are still there, but when applying the assumptions of geometrical optics we are more interested in conditions in the central region of the opening, far from the edge. When viewed from a distance under these circumstances, the edges of the wave appear "sharp" and the seawall appears to cast a "sharp shadow". If the opening is relatively narrow, the effects of the two edges overlap all along the (transverse) length of the wave and hence must be taken into account—the condition of wave optics discussed previously. These ideas carry directly over to the three dimensional case of planar or spherical light waves passing through a small aperture. We are interested in understanding the effect of light passing through the pupil of the eye to form an image on the retina. As we'll see, diffraction must be taken into account here.

Fresnel explained diffraction in terms of this principle and the interference between all the secondary sources. This development is known as the Huygens-Fresnel theory and has enabled the mathematical development of the theory of diffraction and the calculation of the light level at any point beyond an aperture due to diffraction at the aperture. The Huygens-Fresnel theory was further developed by Kirchoff.[2]



Huygens' Wavelets Applied to an Infinite Edge

Figure 2-20



Diffraction Around an Infinite Edge

Figure 2-21

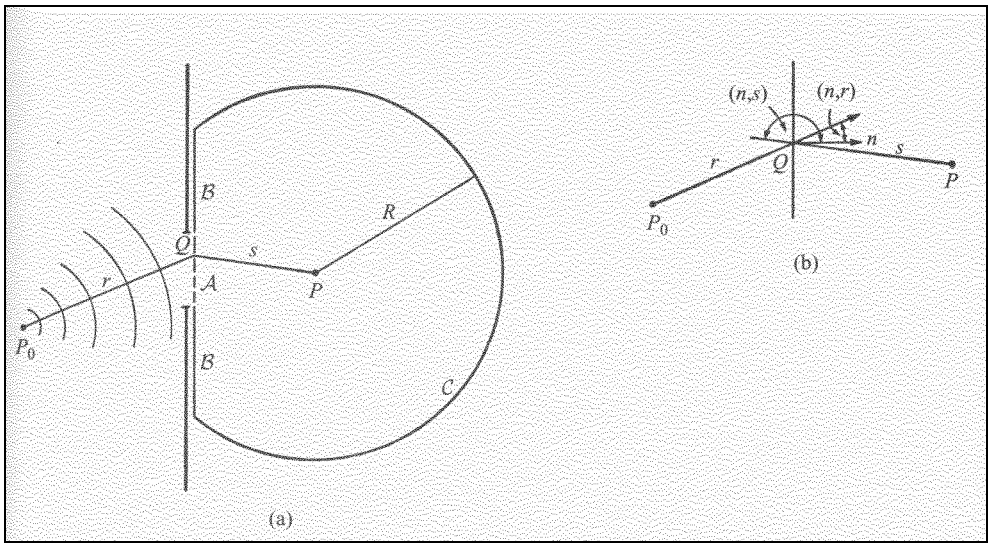
Fresnel and Fraunhofer Diffraction

Consider a monochromatic wave emanating from a point P_0, that has propagated through an opening A in a plane opaque screen, as in **Figure 2-22**. If the linear dimensions of the opening are small compared to the distances P_0 and P (though large compared to the wavelength λ of the beam), then the disturbance U at point P is given by the *Fresnel—Kirkhoff Diffraction Formula*[3],

$$U(P) = -\frac{iA}{2\lambda}\iint_A \frac{e^{ik(r+s)}}{rs}\left[\cos(n,r) - \cos(n,s)\right] dS\,. \tag{3.46}$$

Here A is a constant, the angles (n,r) and (n,s) are as defined in the figure, and

$$k = \frac{\omega}{c} = \frac{2\pi}{\lambda}. \tag{3.47}$$



Fresnel--Kirchoff Diffraction

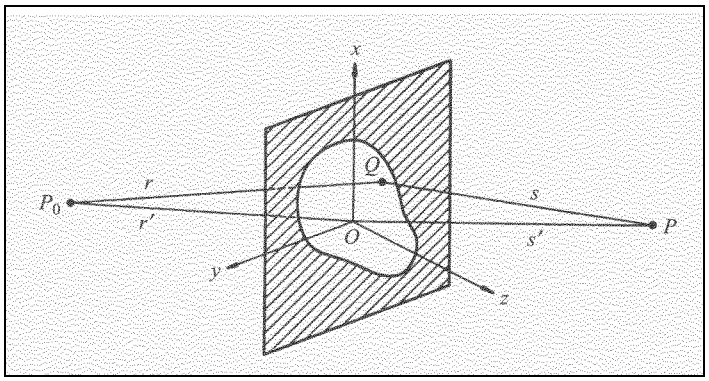
Figure 2-22

As the element dS varies over the domain of integration A, r+s will in general change by very many wavelengths, so that the factor $e^{ik(r+s)}$ will oscillate rapidly. On the other hand, since the distances to the points P_0 and P from the screen are assumed large compared to the linear dimensions of the aperture, the factor [cos(n,r) – cos(n,s)] will not vary appreciably over the aperture and may therefore be taken outside the integral. Further, if O is any point in the aperture, the angles which the lines P_0O and OP make with P_0P will be small. We may then replace this factor by 2 cos δ, where δ is the angle between P_0P and the normal to the screen. Finally, the factor 1/rs may be approximated by 1/r's', where r' and s' are the distances from P_0 and P to O. With these, then, Equ. (3.46) reduces to

$$U(P) \approx -\frac{A\,i\cos\delta}{\lambda\,r's'}\iint_A e^{ik(r+s)}\,dS\,. \tag{3.48}$$

We'll now take a Cartesian reference system with origin O in the aperature and with the x- and y- axes in the plane of the aperture. We'll choose the positive z direction to point into the half-space that contains the point P of observation, as shown in **Figure 2-23**. If (x_0, y_0, z_0) and (x, y, z) are the coordinates of P_0 and P, respectively, and (ξ, η) the coordinates of a point Q in the aperture, we then have

$$\begin{aligned} r^2 &= (x_0 - \xi)^2 + (y_0 - \eta)^2 + z_0^2\,, \\ s^2 &= (x - \xi)^2 + (y - \eta)^2 + z^2\,, \\ r'^2 &= x_0^2 + y_0^2 + z_0^2\,, \\ S'^2 &= x^2 + y^2 + z^2\,. \end{aligned} \tag{3.49}$$



Diffraction at an Aperture in a Plane Screen

Figure 2-23

Hence,

$$\begin{aligned} r &= r'^2 - 2(x_0\xi + y_0\eta) + \xi^2 + \eta^2, \\ s &= s'^2 - 2(x\xi + y\eta) + \xi^2 + \eta^2. \end{aligned} \tag{3.50}$$

Since we assumed that the linear dimensions of the aperture are small compared to both r' and s', we may expand r and s as power series in ξ/r', η/r', ξ/s', and η/s'. We then obtain

$$\begin{aligned} r &\approx r'^2 - \frac{x_0\xi + y_0\eta}{r'} + \frac{\xi^2+\eta^2}{2r'} - \frac{(x_0\xi + y_0\eta)^2}{2r'^3} - \cdots, \\ s &\approx s'^2 - \frac{x\xi + y\eta}{s'} + \frac{\xi^2+\eta^2}{2s'} - \frac{(x\xi + y\eta)^2}{2s'^3} - \cdots. \end{aligned} \tag{3.51}$$

Substituting these results into Equ. (3.48) then gives

$$U(P) = -\frac{i\cos\delta}{\lambda}\frac{A e^{ik(r'+s')}}{r's'}\iint_A e^{ikf(\xi,\eta)}\, d\xi\, d\eta, \tag{3.52}$$

where

$$f(\xi,\eta) = \frac{x_0\xi + y_0\eta}{r'} - \frac{x\xi + y\eta}{s'} - \frac{\xi^2+\eta^2}{2r'} + \frac{\xi^2+\eta^2}{2s'} - \frac{(x_0\xi + y_0\eta)^2}{2r'^3} - \frac{(x\xi + y\eta)^2}{2s'^3} \cdots. \tag{3.53}$$

If we denote (l_0, m_0) and (l, m) as the first two direction cosines, i.e.,

$$\begin{aligned} l_0 &= -\frac{x_0}{r'}, \qquad l = \frac{x}{s'}, \\ m_0 &= -\frac{y_0}{r'}, \qquad m = \frac{y}{s'}, \end{aligned} \tag{3.54}$$

then Equ. (3.53) may be written as

$$f(\xi,\eta) = (l_0 - l)\xi + (m_0 - m)\eta + \frac{1}{2}\left[\left(\frac{1}{r'} + \frac{1}{s'}\right)(\xi^2+\eta^2) - \frac{(l_0\xi + m_0\eta)^2}{r'} - \frac{(l\xi + m\eta)^2}{s'}\right]\cdots. \tag{3.55}$$

We have thus reduced the problem of determining the light disturbance at P to the evaluation of the integral in Equ. (3.55). Naturally, the evaluation is simpler to carry out when the quadratic and higher-order terms in ξ and η may be neglected in f. In this case we have *Fraunhofer Diffraction*. When the quadratic terms cannot be neglected we have *Fresnel Diffraction*. Fortunately, the simpler case of Fraunhofer Diffraction is of much greater importance in optics, and the focus of our interest here as well.

Strictly speaking, the second and higher-order terms disappear only in the limiting case $r' \to \infty$, $s' \to \infty$, that is, when both the source and the point of observation are at infinity. [The factor A in Equ. (3.52) must then be assumed to tend to inifinty in the same fashion as r's', in order to keep U(P) bounded.] It is, however, evident that the second-order terms will not appreciably contribute to the integral if

$$\frac{1}{2}k\left|\left(\frac{1}{r'}+\frac{1}{s'}\right)\left(\xi^2+\eta^2\right)-\frac{\left(l_0\,\xi+m_0\,\eta\right)^2}{r'}-\frac{\left(l\xi+m\eta\right)^2}{s'}\right| \leq 2\pi. \tag{3.56}$$

With this in mind we can immediately recognize certain conditions under which Equ. (3.56) is satisfied. If we make use of inequalities of the form $\left(l_0\,\xi+m_0\,\eta\right)^2 \leq \left(l_0^{\,2}+m_0^{\,2}\right)\left(\xi^2+\eta^2\right)$ and remember that l_0^2, m_0^2, l^2, and m^2 cannot exceed unity, we find that Equ. (3.56) will be satisfied if

$$r' \geq \frac{\left(\xi^2+\eta^2\right)_{max}}{\lambda}, \text{ and} \qquad s' \geq \frac{\left(\xi^2+\eta^2\right)_{max}}{\lambda}, \tag{3.57}$$

or if

$$\frac{1}{r'}+\frac{1}{s'}=0, \qquad \text{and} \qquad l_0^2, m_0^2, l^2, m^2 \geq \frac{|r'|\lambda}{\left(\xi^2+\eta^2\right)_{max}}. \tag{3.58}$$

Equs. (3.57) give an estimate of the distances r' and s' for which the Fraunhofer representation may be used. Equs. (3.58) imply that Fraunhofer Diffraction also occurs when the point of observation is situated in a plane parallel to that of the aperture, provided that both the point of observation and the source are sufficiently close to the z-axis[4]. Here two cases may be distinguished: when r' is negative, the wave-fronts incident upon the aperture are concave to the direction of propagation. That is, P_0 is a center of convergence and not of divergence of the incident wave. This case is of great practical importance, as it arises in the image space of a well-corrected, centered system that images a point source which is not far from the axis. The resulting Fraunhofer pattern may be considered as arising from the diffraction of the image--forming wave on the exit pupil. In other words, Fraunhofer phenomena occur in the focal plane of a well corrected lens, in the region near the r's' axis. The second case implied by Equ. (3.58) in which Fraunhofer Diffraction occurs corresponds to positive r'. Here the wave-fronts are convex to the direction of propagation, and the diffraction phenomena are virtual, being apparently formed on a screen through the source P_0. This case arises, for example, when a aperture is held in front of the eye, or the object glass of a telescope adjusted for distant vision of the light source. In this analysis we'll be concerned with the first of these cases. Assuming the conditions of the first case, we can neglect the quadratic and higher-order terms in ξ and η in Equ. (3.55) and write Equ. (3.52) as

$$U(P)=-\frac{i\cos\delta}{\lambda}\frac{A\,e^{ik(r'+s')}}{r's'}\iint_A e^{-ik(p\xi+q\eta)}\,d\xi\,d\eta, \tag{3.59}$$

where

$$p=l-l_0,\ q=m-m_0. \tag{3.60}$$

The constant appearing in front of the integral in Equ. (3.59) is unwieldy, but it can be simplified into a more convenient form. Let E be the total energy incident upon the aperture. By the law of conservation of energy the total energy that reaches the plane of observation must also be equal to E, so that we have

$$\iint\left|U(p,q)\right|^2 dp\,dq=\frac{E}{R^2}, \tag{3.61}$$

where R is the distance from O to the point at which the line P_0Q intersects the plane of observation, as shown in **Figure 2-23**. In Equ. (3.61) the integration extends over all possible values of p and q. Equ. (3.59) may be rewritten as a Fourier Integral

$$U(p,q)=\iint G(\xi,\eta)\,e^{-\frac{2\pi i}{\lambda}(p\xi+q\eta)}\,d\xi\,d\eta, \tag{3.62}$$

where the *Pupil Function* $G(\xi, \eta)$ is given by

$$G(\xi,\eta)=\begin{cases}-\dfrac{i\cos\delta}{\lambda}\dfrac{A\,e^{ik(r'+s')}}{r's'} & \text{for points inside the aperture}\\ 0 & \text{for points outside the aperture}\end{cases}. \tag{3.63}$$

Here the integration extends over the entire ξ, η plane. Applying Parseval's Theorem[5] to Equ. (3.62) gives

$$\begin{aligned}\frac{1}{\lambda^2}\iint\left|U(p,q)\right|^2 dp\,dq &= \iint\left|G(\xi,\eta)\,e^{-\frac{2\pi i}{\lambda}(p\xi+q\eta)}\right|^2 d\xi\,d\eta\,,\\ &= \iint\left|G(\xi,\eta)\right|^2\left|e^{-\frac{2\pi i}{\lambda}(p\xi+q\eta)}\right|^2 d\xi\,d\eta\,,\\ &= \iint\left|G(\xi,\eta)\right|^2|1|^2\,d\xi\,d\eta\,,\\ &= \iint\left|G(\xi,\eta)\right|^2 d\xi\,d\eta\,,\\ &= |C|^2\iint_A d\xi\,d\eta\,.\end{aligned} \tag{3.64}$$

Substituting Equ. (3.61) for the left-hand side integral and noting that the right-hand side integral is just the area of the aperture, which we'll denote as D, Equ. (3.64) reduces to

$$\frac{1}{\lambda^2}\frac{E}{R^2}=|C|^2 D\,. \tag{3.65}$$

Thus,

$$C=\frac{1}{\lambda R}\left(\frac{E}{D}\right)^{\frac{1}{2}}, \tag{3.66}$$

and the basic Fraunhofer Diffraction Integral [Equ. (3.59)] becomes

$$U(p,q)=\frac{1}{\lambda R}\left(\frac{E}{D}\right)^{\frac{1}{2}}\iint_A e^{-ik(p\xi+q\eta)}\,d\xi\,d\eta\,. \tag{3.67}$$

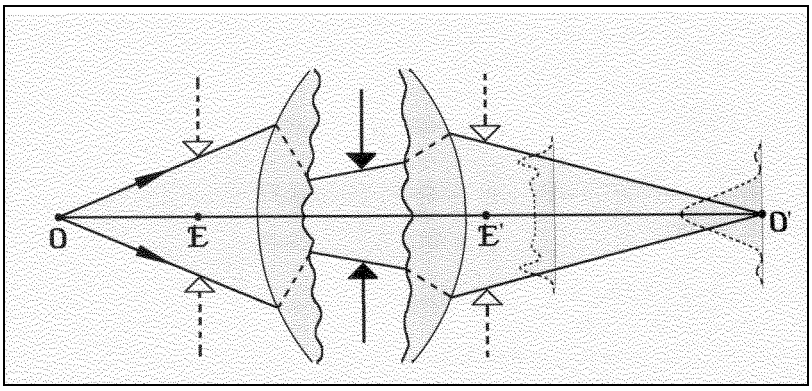
From this, we can see that the intensity $I_0=|U(0,0)|^2$ at the center of the pattern p=q=0 is given by

$$I_0=\left(\frac{1}{\lambda R}\right)^2\frac{E}{D}\left(\iint_A d\xi\,d\eta\right)^2=\frac{E\,D}{\lambda^2 R^2}=C^2 D^2\,. \tag{3.68}$$

To see in a more physical way how Fresnel and Fraunhofer Diffraction are related, consider **Figure 2-24**, which shows a focussed, monochromatic and aberration-free beam, arising from an axial monochromatic point source. The continuous lines are the geometrical optics boundary and the dotted lines indicate the actual form of the light distributions across the beam at two representative positions along the beam. For the position closer to the optical system, a Fresnel Diffraction Pattern exists. Here the light distribution at either extreme is similar to the light distribution at the edge of the beam, similar to **Figure 2-21**. In the center of the beam there is no fringing and the light distribution is as predicted by geometrical optics. If we now move this plane towards the focus at O', the edges of the beam become closer, the flat central portion of the light distribution narrows, and finally the fringe systems at the edges merge and at O' the light level finally takes the form of a Fraunhofer Diffraction pattern. Because the Fresnel diffraction pattern varies in going from the optical system to the focus at O', there is no unique pattern in this region. In the plane of the focus, however, the diffraction pattern is unique. Fraunhofer Diffraction, then, is a special case of Fresnel Diffraction and is in fact defined as the diffraction effect in the focal plane of a beam. Here we are interested in Fraunhofer Diffraction because it (along with any aberrations present) predicts the light distribution in the image plane. This particular case--the light distribution in the image of a point source for an aberraton-free beam--is also known as the *Diffraction-Limited Point Spread Function*.

Ideally, we would measure retinal image quality by first imaging a point source of light onto the retina, and then recording that image using light reflected back out of the eye. (Mathematical manipulation is required to account for the fact that the recorded light rays have passed through the eye twice.) Until the advent of lasers and very sensitive detectors, though, this

approach was not possible because a conventional "point" source of light contained too little light to be detected after it had been reflected from the retina. A suitable and simple alternative was to use a "line" source in place of a point source. Such a source emits perhaps a hundred times for luminous flux than a point source of the same width. (Ideally, of course, both sources would have zero width.) Such



Diffraction Patterns Along an Aberration-Free, Focused Beam

Figure 2-24

a light distribution is called a *Line Spread Function*. Because this part of the visual system is linear, such a function can be obtained by summing (or in the limit, integrating) point spread functions. Thus, the line spread function is the sum of a line of point spread functions. Continuing with this reasoning, we could also construct an *Edge Spread Function* (a sum or integral of line spread functions). We could also construct the light distribution resulting from an image whose luminance varied as a square wave or a sine wave. We could also go the other way, if we wished, and recover the point spread function upon which a particular line spread function was based (and similarly for edge, square wave, and sinusoidal patterns).

Students of Systems and Control Theory may by now have come to suspect that the point spread function is the *impulse response* of the optical system comprising the eye (with the independent variables being the three spatial dimensions, as opposed to time). This is indeed the case, and we define the *Optical Transfer Function* $G(\sigma)$ as the Fourier Transform of the point spread function. (In systems theory we usually deal with Laplace Transforms, but since here we are interested in sinusoidal inputs it is more convenient to work with Fourier Transforms, which can be used in a completely analogous manner.) In words, the Optical Transfer Function is a complex-valued function of the spatial frequency σ that measures the reduction in amplitude and change in phase of sinusoidal patterns imaged by an optical system. Because it is derived from an impulse function, the optical transfer function (in the case of the human visual system) can provide information on the eye's response (in terms of amplitude attenuation and phase shift) to any input. The optical transfer function is traditionally written as the product of two other transfer functions, the *Modulation Transfer Function* $G_m(\sigma)$, which gives the attenuation of the input amplitude, and the *Phase Transfer Function* $G_p(\sigma)$, which gives the phase shift. Thus,

$$G(\sigma)=G_m(\sigma)G_p(\sigma). \tag{3.69}$$

We'll next derive the Point Spread Function for a defect-free eye, and from this develop the corresponding Optical Transfer Function. (This diffraction-limited case represents the upper limit of the eye's capabilities.) Finally, once this foundation is laid we will introduce the *Contrast Sensitivity Function*, a measure of the eye's actual performance in this regard.

Fraunhofer Diffraction Due to a Circular Aperture

In analyzing the diffraction pattern due to a circular aperture it is more convenient to use polar instead of rectangular coordinates. Let (ρ, θ) be the polar coordinates of a typical point in the aperture, so that

$$\rho\cos\theta=\xi, \qquad \rho\sin\theta=\eta; \tag{3.70}$$

and let (w, ψ) be the coordinates of a point P in the diffraction pattern, referred to the geometrical image of the source. Then

$$w\cos\psi=p, \qquad w\sin\psi=q\ . \tag{3.71}$$

From the definition of p and q [Equ. (3.60)] it follows that $w=\sqrt{p^2+q^2}$ is the sine of the angle which the direction (p, q) makes with the central direction p=q=0. The diffraction integral given by Equ. (3.59) then becomes

$$U(P)=C\int_0^a\int_0^{2\pi} e^{-ik\rho w\cos(\theta-\psi)}\,\rho\,d\rho\,d\theta\,. \tag{3.72}$$

Now we have the well-known integral representation of the Bessel Function $J_n(z)$:

$$\frac{i^{-n}}{2\pi}\int_0^{2\pi} e^{ix\cos\alpha}\,e^{in\alpha}\,d\alpha=J_n(x)\,. \tag{3.73}$$

Equ. (3.72) therefore reduces to

$$U(P)=2\pi C\int_0^a J_0(k\rho w)\,d\rho\,. \tag{3.74}$$

We also have the well-known recurrence relation

$$\frac{d}{dx}\left[x^{n+1}J_{n+1}(x)\right]=x^{n+1}J_n(x)\,. \tag{3.75}$$

Integrating Equ. (3.75) for n=0, then gives

$$x\,J_1(x)=\int_0^x x'J_0(x')\,dx'. \tag{3.76}$$

From Equs. (3.74) and (3.76) it follows that

$$U(P)=CD\left[\frac{2J_1(kaw)}{kaw}\right], \tag{3.77}$$

and the intensity is given by

$$I(P)=|U(P)|^2=I_0\left[\frac{2J_1(kaw)}{kaw}\right]^2, \tag{3.78}$$

or

$$\frac{I(P)}{I_0}=\left[\frac{2J_1(kaw)}{kaw}\right]^2, \tag{3.79}$$

where, per Equ. (3.68),

$$I_0=C^2D^2=\frac{ED}{\lambda^2R^2}\,. \tag{3.80}$$

Equ. (3.78) is the celebrated formula first derived in a somewhat different form by Airy[6].

Equ. (3.79) gives the normalized intensity distribution in the neighborhood of the geometrical image, and is plotted in **Figure 2-25**. It has a principle maximum $I(P)/I_0=1$ at $(kaw)=0$, and with increasing (kaw) it oscillates with gradually diminishing amplitude. The intensity is zero (minimum) for values for (kaw) given by $J_1(kaw)=0$. The minima are no longer equidistant, as shown in **Table 2-2**. The positions of the secondary maxima are given by the values of (kaw) that satisfy the equation

$$\frac{d}{d(kaw)}\left[\frac{J_1(kaw)}{kaw}\right]=0, \tag{3.81}$$

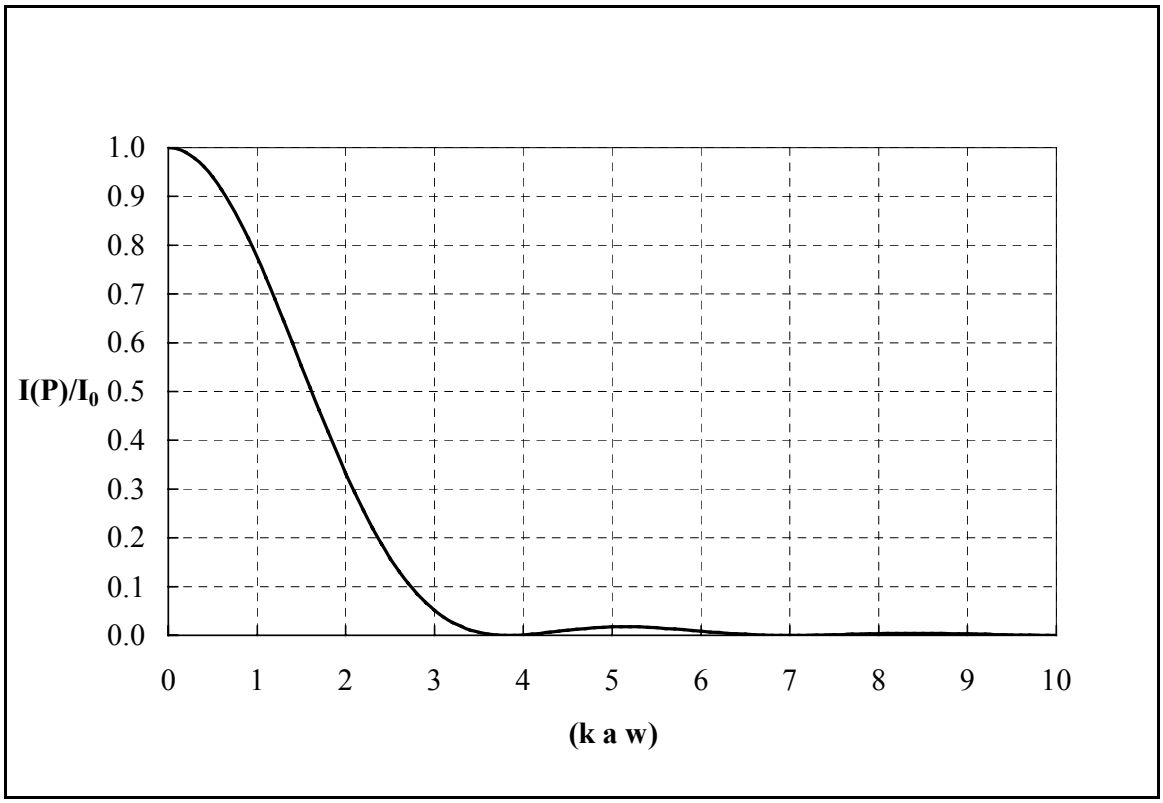
or, using the relation [analogous to Equ. (3.75)]

$$\frac{d}{d(k\,a\,w)}\left[(k\,a\,w)^n\,J_n(k\,a\,w)\right]=-(k\,a\,w)^{-n}\,J_{n+1}(k\,a\,w)\,, \tag{3.82}$$

by the roots of the equation

$$J_2(k\,a\,w)=0\,. \tag{3.83}$$

With increasing $(k\,a\,w)$ the separation between two successive minima or two successive maxima approaches the value π. The result is a pattern (the *Airy Pattern*) consisting of a bright disc (the *Airy Disc*), centered on the geometrical image p=q=0 of the source, surrounded by concentric bright and dark rings, shown in **Figure 2-26**. `The intensity of the bright rings decreases rapidly with their radius and normally only the first one or two rings are bright enough to be visible to the naked eye.



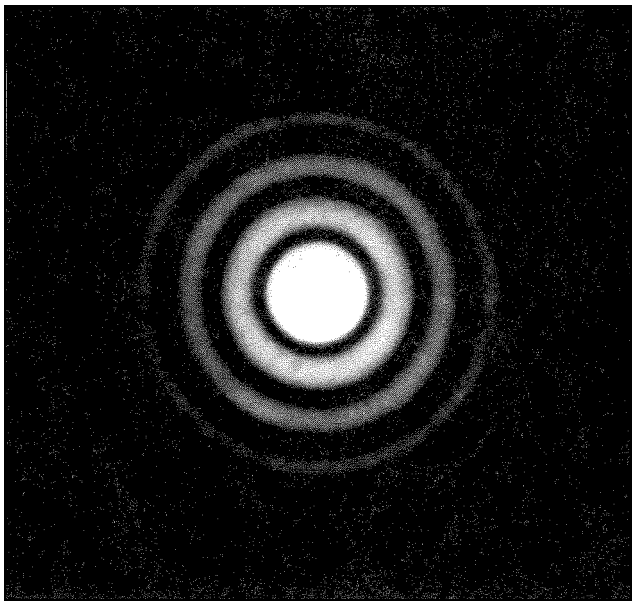
Fraunhofer Diffraction at a Circular Aperture: The Function $\frac{I(P)}{I_0}=\left[\frac{2J_1(k\,a\,w)}{k\,a\,w}\right]^2$

Figure 2-25

(k a w)	$\frac{I(P)}{I_0}$	Maximum/Minimum
0	1	Maximum
1.220π=3.833	0	Minimum
1.635π=5.136	0.0175	Maximum
2.233π=7.016	0	Minimum
2.679π=8.417	0.0042	Maximum
3.238π=10.174	0	Minimum
3.699π=11.620	0.0016	Maximum

The First Few Maxima and Minima of $\frac{I(P)}{I_0}$

Table 2-2



Fraunhofer Diffraction Pattern of a Circular Aperture (The Airy Pattern)
Figure 2-26

F. The Optical Transfer Function

The *Fourier Transform* $F(\sigma)$ of the complex-valued function f(x) is defined as

$$F(\sigma)=\int_{-\infty}^{\infty} f(x)\,e^{-2\pi i\sigma x}\,dx. \tag{3.1}$$

Here Equ. (3.78) corresponds to f(x), with $x=(k\,a\,w)$. The Optical Transfer Function, then, is given by

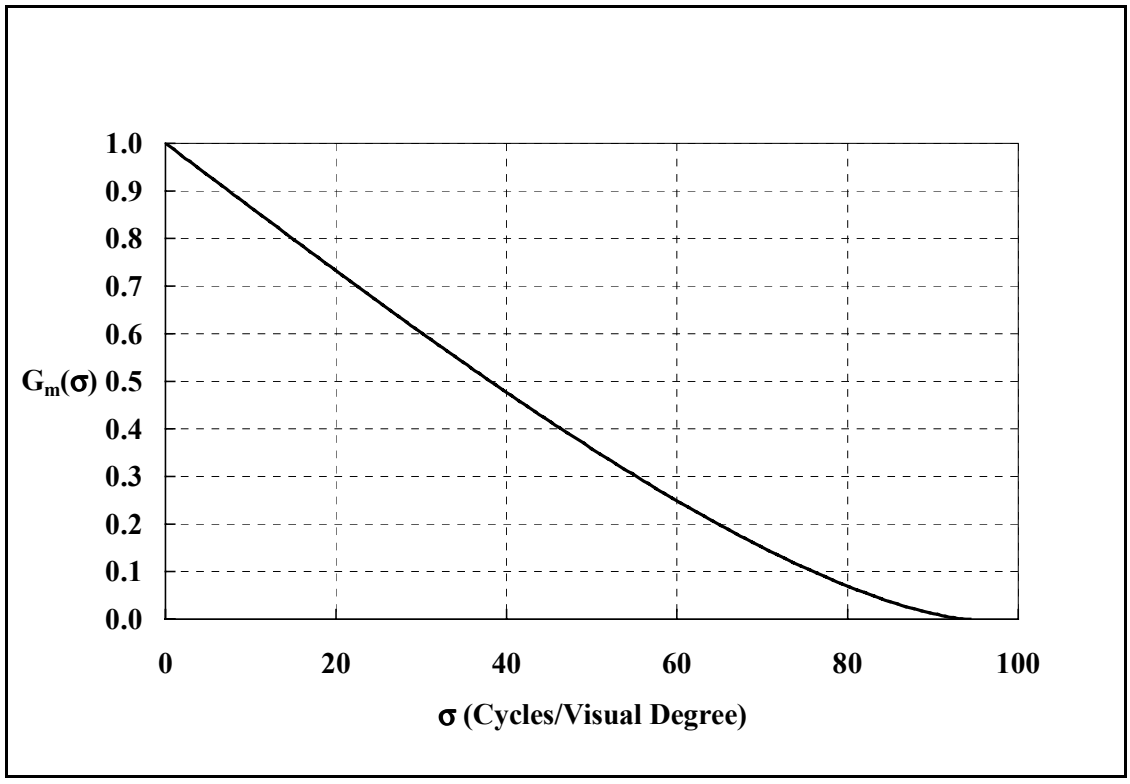
$$G(\sigma)=\int_{-\infty}^{\infty}\left[\frac{2J_1(k\,a\,w)}{k\,a\,w}\right]^2 e^{-2\pi i\sigma(k\,a\,w)}\,d(k\,a\,w), \tag{3.84}$$

Since in this case the diffraction-limited point spread function is even-symmetric, $G(\sigma)$ will have no phase shift associated with it, and the Optical and Modulation Transfer Functions will be one and the same. Equ. (3.84) has been evaluated numerically, and the result plotted in **Figure 2-27**. σ here is the spatial frequency (in cycles per degree of visual angle) of the harmonic input signal.

The optical and modulation transfer functions can be determined based on experimental data. Campbell and Gubisch[7], for example, calculated the modulation transfer functions from their measured white light line spread functions. Their mean result for a 3mm pupil size is shown in **Figure 2-28**. The diffraction –limited modulation transfer function of **Figure 2-27** is also plotted here for comparison. We can also measure the modulation transfer function of the eye by subjective methods; Campbell and Green[8] accomplished this by measuring two somewhat related functions:

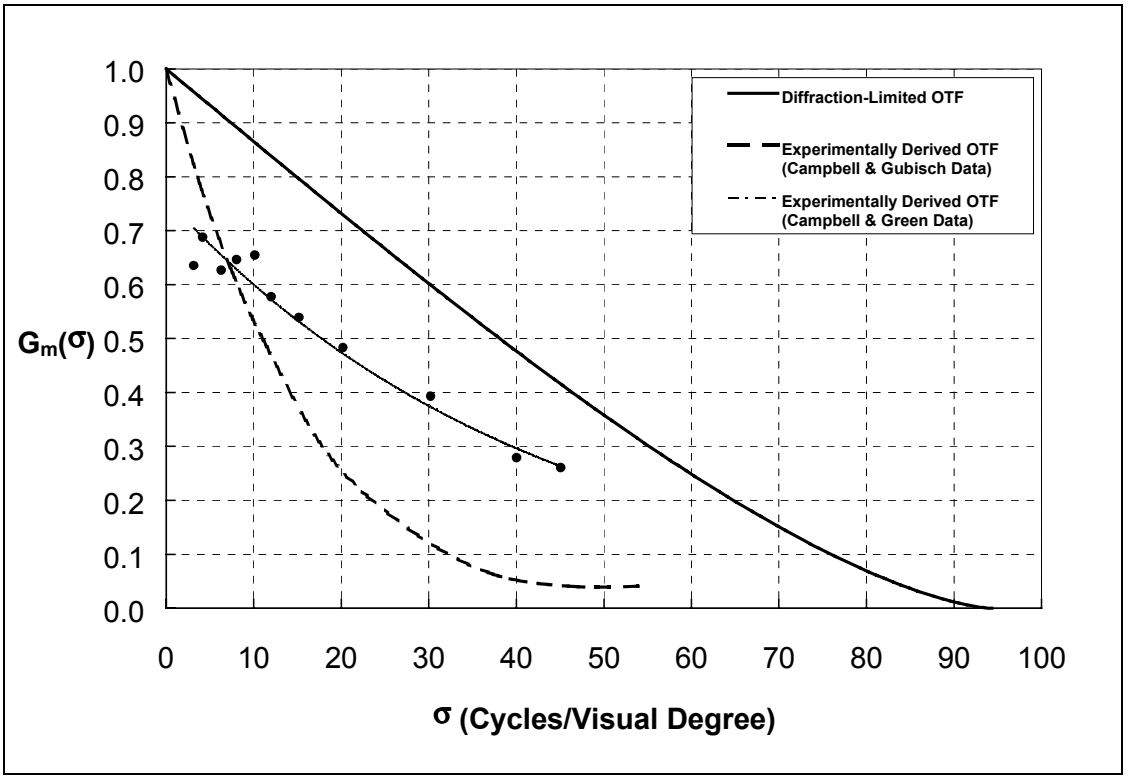
1. The *Total Contrast Threshold Function*, which is obtained by presenting sinusoidal patterns on an oscilloscope screen to an observor and then by some suitable psychophysical procedure varying the contrast until the subject decides that it is just visible. In this case, the final result is affected by the aberrations and diffraction properties of the eye as well as by neural factors.
2. The *Contrast Threshold Function*, which is obtained by using an interferometer to project sinusoidal patterns directly onto the retina. In this case the results are not affected by aberrations or diffraction, only neural factors.

Note that these measures of contrast are not modulation transfer functions, because they are influenced by the neural system. The modulation transfer function does emerge, however, if we take the ratio of the Contrast Threshold Function to the Total Contrast Threshold Function. (This in effect "divides out" the neural factors.) These results are also plotted in **Figure 2-28**. They are not strictly comparable with the two other curves in the figure because a pupil diameter of 2 mm (vs. 3 mm for the other curves) was used, and two different wavelengths of light (530 nm and 632.8 nm vs, 550 nm for the other curves) were used. If parameters and two different wavelengths of light (530 nm and 632.8 nm vs, 550 nm for the other curves) were used. If parameters consistent with our other data had been used this curve would be pushed down towards the one reflecting the Green and Gubisch results. Even with this, though, we see that the experimental data agrees well with the theory, and we get an idea of the degree to which the visual system is affected by the imperfections associated with the eye's optics.



Optical (Modulation) Transfer Function for an Aberration-Free Eye (3mm Pupil Diameter)

Figure 2-27



Experimentally Determined Optical (Modulation) Transfer Functions

Figure 2-28

G. Summary

We have now accomplished the objective of these two appendices—to understand one of the central elements of what we think of as "seeing": contrast sensitivity. We have in Appendix I investigated the pertinent physiological elements involved in this task and the manner in which they're organized to accomplish it. In Appendix II we investigated the nature of light, and seen how the human visual system has tuned itself to accommodate its properties. More concretely, we have seen that the visual system acts as a filter to detect luminance patterns, and that those exhibiting the greatest contrast are the ones that attract our notice. Tying the two appendices together, we find that this filtering function is carried out for light/dark (black and white) patterns, red/green patterns, and blue/yellow patterns. Thus the scene is sampled in three ways to detect its salient features. We will use these conclusions in developing our model.

Perhaps the most profound conclusion, however, that the engineer can take from this is that human vision is eminently understandable, and that it lends itself to the same kinds of analysis as any other physical system. Further, if we can analyze it, we can optimize its role in larger systems (such as automobile control) that it is an integral part of.

---

[1] Taken from Smith and Atchison, *'The Eye and Visual Optical Instruments'*, Cambridge University Press, Cambridge, England, 1997, Chapter 26.

[2] Here and in what follows we assume that the incoming waves are monochromatic and *coherent*. We say that the waves emitted by separate sources are coherent if they all have the same phase difference, which remains constant with respect to time. In reality, light waves are never strictly monochromatic, nor are they strictly coherent. This, though, is a complex subject which, fortunately, we need not take up in order to achieve the insights and results we seek. In what follows, then, we will retain the assumption of monochromatic, coherent waves. See Garbuny, *'Phycical Optics'*, Academic Press, New York, 1965, Chapters 1-3, and 6 for a discussion of these ideas.

[3] Born and Wolf, *'Principles of Optics'*, Cambridge University Press, Cambridge, England, 1999, pp. 421-425.

[4] This is the so-called *paraxial approximation* (also referred to as *Gaussian Optics*). It assumes that all rays (actually, just the rays of interest to us) passing through the aperture are almost parallel to the optial axis, here the z-axis. This in turn allows us to make the usual small angle approximations, $\sin\theta \approx \tan\theta \approx \theta$ and $\cos\theta \approx 1$, as well as the others described above.

[5] I.N. Sneddon, *Fourier Transforms*, New York, McGraw-Hill, 1951, pp. 25 and 44.

[6] G.B. Airy, *Transactions of the Cambridge Philosophical Society*, **5** (1835), 283.

[7] Campbell and Gubisch, *The Optical Quality of the Human Eye*, Journal of Physiology, Cambridge University Press, Cambridge, UK, October, 1966 (Volume 186, No. 3, pp. 576-593).

[88] Campbell and Green, *Optical and Retinal Factors Affecting Visual Resolution*, Journal of Physiology, Cambridge University Press, Cambridge, UK, December, 1965 (Volume 181, No. 3, pp. 558-578).

Appendix 3

The Matlab M-Files CSF01 and CSF02

Appendix 3: CSF01

Appendix 3: CSF01

```
home ; clear all ; close all ;
format compact
pause(.01)

%*************************************************************************
% Set Directories for Input and Output Files
%*************************************************************************
addpath('C:\Conspicuity\Matlab Files\','-end');
InPutPath='C:\Conspicuity\Input\';
OutPutPath='C:\Conspicuity\Output\';

%*************************************************************************
% Initialize Parameters
%*************************************************************************
InputFileName=strcat(InPutPath,'Img0014-02.bmp');
OutputFileName=strcat(OutPutPath,'RoadSideScene');
InitialPlots=1    ;
PlotFinalResult=0 ;
PlotScoringGrid=0 ;

MonCSF=[0.7692;0.9155;0.9993;1.000;0.9169;0.7713;0.5952;0.4212;0.2734;0.1626;0.0888
;0.0440];
RGCSF=[1.0000;0.9171;0.8114;0.6823;0.5341;0.3778;0.2315;0.1158;0.0435;0.0109;0.0015
;0.0001];
BYCSF=[1.0000;0.9185;0.8145;0.6872;0.5403;0.3846;0.2378;0.1205;0.0461;0.0118;0.0017
;0.0001];

FilterFreq = 2.^[-1:.5:4.5]    ; FreqEnd = length(FilterFreq) ;
Ang        = pi/6.*[0:1:6]     ; AngEnd  = length(Ang)        ;
Pool       = 4                 ; Rho     = 3*sqrt(log(2))/pi  ;
Count      = 0                 ; dtsum   = 0                  ;

set(0,'units','inches')
ScreenSizeInch=get(0,'screensize');
set(0,'units','pixels')
ScreenSizePix=get(0,'screensize') ; SW=ScreenSizePix(3) ; SH=ScreenSizePix(4);
PixperInch=ScreenSizePix(1,3:4)./ScreenSizeInch(1,3:4);

%*************************************************************************
% Load Scene
%*************************************************************************
Scene=imread(InputFileName,'bmp');
Scene=double(Scene);
minmum=min(min(min(Scene)));
Scene=Scene-minmum;
maxmum=max(max(max(Scene)));
Scene=Scene/maxmum;

[HPix,WPix,z]=size(Scene);
D=25 ; W=6; H=W*HPix/WPix ;
Alpha=2*atan((W/2)/D)*180/pi;
dW=floor(WPix/Alpha) ; dH=dW ;

%*************************************************************************
% Create and Plot (If Desired) Black/White, Red/Green, and Blue/Yellow Sub-Images
%*************************************************************************
MonMatrix=(Scene(:,:,1)+Scene(:,:,2)+Scene(:,:,3))/3.;
MonScene(:,:,1)=MonMatrix ; MonScene(:,:,2)=MonMatrix ; MonScene(:,:,3)=MonMatrix ;
```

Appendix 3: CSF01

```
RGMatrix=Scene(:,:,1)-Scene(:,:,2);
if min(min(RGMatrix))<0
  RGMatrix=RGMatrix-min(min(RGMatrix));
end
if max(max(RGMatrix))>1
  RGMatrix=RGMatrix/max(max(RGMatrix));
end
RGScene(:,:,1)=RGMatrix ; RGScene(:,:,2)=RGMatrix ; RGScene(:,:,3)=RGMatrix ;

BYMatrix=(Scene(:,:,1)+Scene(:,:,2)-2*Scene(:,:,3));
if min(min(BYMatrix))<0
  BYMatrix=BYMatrix-min(min(BYMatrix));
end
if max(max(BYMatrix))>1
  BYMatrix=BYMatrix/max(max(BYMatrix));
end
BYScene(:,:,1)=BYMatrix ; BYScene(:,:,2)=BYMatrix ; BYScene(:,:,3)=BYMatrix ;

if InitialPlots==1
  figure(1)
    image(Scene)
    Position=[10 (SH-HPix-75) WPix HPix];
    set(gcf,'Position',Position);
    title('Figure 1: Original Scene','fontsize',12,'fontweight','bold')
  figure(2)
    Position=[SW-WPix-10 (SH-HPix-75) WPix HPix];
    set(gcf,'Position',Position);
    image(MonScene)
    title('Figure 2: Monochrome Scene','fontsize',12,'fontweight','bold')
  figure(3)
    Position=[10 40 WPix HPix];
    set(gcf,'Position',Position);
    image(RGScene)
    title('Figure 3: Red-Green Scene','fontsize',12,'fontweight','bold')
  figure(4)
    Position=[SW-WPix-10 40 WPix HPix];
    set(gcf,'Position',Position);
    image(BYScene)
    title('Figure 4: Blue-Yellow Scene','fontsize',12,'fontweight','bold')
    pause(.01)
end

save(OutputFileName,'Scene','MonScene','RGScene','BYScene');
imwrite(Scene,strcat(OutPutPath,'OriginalScene.bmp'),'bmp');
imwrite(MonScene,strcat(OutPutPath,'BWSubImage.bmp'),'bmp');
imwrite(RGScene,strcat(OutPutPath,'RGSubImage.bmp'),'bmp');
imwrite(BYScene,strcat(OutPutPath,'BYSubImage.bmp'),'bmp');

%*******************************************************************************
% Create Filtered Sub-Images
%*******************************************************************************
Filter=1*dW;
FilterGridx=[round(-Filter/2):round(Filter/2)]'/round(Filter/2);
FilterGridy=FilterGridx;
[FilterX, FilterY] = meshgrid(FilterGridx,FilterGridy);

HPix=HPix-length(FilterGridx)+1 ; WPix=WPix-length(FilterGridy)+1  ;
```

```
MonPoolImage = zeros(HPix,WPix) ;
RGPoolImage  = zeros(HPix,WPix) ;
BYPoolImage  = zeros(HPix,WPix) ;

disp(sprintf('Beginning First of %3i Convolutions at
%s',AngEnd*FreqEnd,datestr(Now,14)));
for iang = 1:AngEnd
  Xrot =  FilterX*cos(Ang(iang)) + FilterY*sin(Ang(iang));
  Yrot = -FilterX*sin(Ang(iang)) + FilterY*cos(Ang(iang));
  for ifreq = 1:FreqEnd
    t0      = clock;
    Count   = Count+1;
    Lambda  = Rho/(FilterFreq(ifreq)/dW);
    Filter  = zeros(length(FilterGridy),length(FilterGridx));
    Filter  = exp(-(Xrot.^2 + Yrot.^2)/(Lambda.^2));
    Filter  = Filter.*cos(2*pi*(FilterFreq(ifreq)/dW)*Xrot);
    Filter  = round(Filter);

    MonConvResult = conv2(MonMatrix,Filter,'valid');
    RGConvResult  = conv2( RGMatrix,Filter,'valid');
    BYConvResult  = conv2( BYMatrix,Filter,'valid');

    MonPoolImage  = MonPoolImage+MonCSF(ifreq)*(MonConvResult.^Pool);
    RGPoolImage   = RGPoolImage +RGCSF(ifreq) *(RGConvResult .^Pool);
    BYPoolImage   = BYPoolImage +BYCSF(ifreq) *(BYConvResult .^Pool);

    dt(Count)=etime(clock,t0);
    dtsum=dtsum+dt(Count);
    DT=dtsum/Count;
    ToGo=(AngEnd*FreqEnd-Count)*DT;
    Now=now;
    Line1=sprintf('Convolution %3i%3i: Iteration %3i of %3i; dt=%3i:%5.2f min:sec;
',...

iang,ifreq,Count,AngEnd*FreqEnd,floor(dt(Count)/60),rem(dt(Count),60));
    Line2=sprintf('DT=%3i:%5.2f; Completion in %3i:%5.2f or at %s',...
                  floor(DT/60),rem(DT,60),floor(ToGo/60),rem(ToGo,60),...
                  datestr(Now+ToGo/(24*3600),14));
    disp([Line1 Line2]);
  end
end
disp(sprintf('Convolutions completed at %s',datestr(now,14)))

%*************************************************************************
% Scale Pooled Images (If Necessary) Such That all Pixels Lie Between 0 and 1
%*************************************************************************
MonPoolImage=MonPoolImage-min(min(MonPoolImage));
MonPoolImage=MonPoolImage/max(max(MonPoolImage));
MonFinalResult=MonPoolImage.^(1/Pool);
MonOverallAvgInt=sum(sum(MonFinalResult))/(WPix*HPix);

RGPoolImage=RGPoolImage-min(min(RGPoolImage));
RGPoolImage=RGPoolImage/max(max(RGPoolImage));
RGFinalResult=RGPoolImage.^(1/Pool);
RGOverallAvgInt=sum(sum(RGFinalResult))/(WPix*HPix);

BYPoolImage=BYPoolImage-min(min(BYPoolImage));
```

```
BYPoolImage=BYPoolImage/max(max(BYPoolImage));
BYFinalResult=BYPoolImage.^(1/Pool);
BYOverallAvgInt=sum(sum(BYFinalResult))/(WPix*HPix);

save(OutputFileName,'dt','MonPoolImage','MonFinalResult',...
                         'RGPoolImage' ,'RGFinalResult' ,...
                         'BYPoolImage' ,'BYFinalResult' ,
                                                  '-APPEND');

Image(:,:,1)=MonFinalResult; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'ReconBW.bmp'),'bmp');

Image(:,:,1)=RGFinalResult; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'ReconRG.bmp'),'bmp');

Image(:,:,1)=BYFinalResult; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'ReconBY.bmp'),'bmp');

%*************************************************************************
% Construct Scoring Grid
%*************************************************************************
Width=dW/2                  ; Height=dH/2                ;
nx=round((WPix-1)/Width)    ; ny=round((HPix-1)/Height)  ;
Width=(WPix-1)/(nx-1)       ; Height=(HPix-1)/(ny-1)     ;
xn(1,1)=1                   ; yn(1,1)=1                  ;
for i=2:nx
  xn(i,1)=xn(i-1,1)+Width ;
end
for i=2:ny
  yn(i,1)=yn(i-1,1)+Height;
end
xn=round(xn)                ; yn=round(yn)               ;

for i=1:nx-1
  for j=1:ny-1
    Area=(xn(i+1)-xn(i)) * (yn(j+1)-yn(j));
    MonAvgInt(i,j)=sum(sum(MonFinalResult(yn(j):yn(j+1),xn(i):xn(i+1))))/Area;
    MonScore(i,j)=MonAvgInt(i,j)/MonOverallAvgInt;
    MonScoreGrid(yn(j):yn(j+1),xn(i):xn(i+1))=MonScore(i,j);

    RGAvgInt(i,j)=sum(sum(RGFinalResult(yn(j):yn(j+1),xn(i):xn(i+1))))/Area;
    RGScore(i,j)=RGAvgInt(i,j)/RGOverallAvgInt;
    RGScoreGrid(yn(j):yn(j+1),xn(i):xn(i+1))=RGScore(i,j);

    BYAvgInt(i,j)=sum(sum(BYFinalResult(yn(j):yn(j+1),xn(i):xn(i+1))))/Area;
    BYScore(i,j)=BYAvgInt(i,j)/BYOverallAvgInt;
    BYScoreGrid(yn(j):yn(j+1),xn(i):xn(i+1))=BYScore(i,j);
    end
end

MonScoreGrid=abs(MonScoreGrid-1);
RGScoreGrid=abs(RGScoreGrid-1);
BYScoreGrid=abs(BYScoreGrid-1);

MinScoreGrid=min([min(min(MonScoreGrid)) min(min(RGScoreGrid))
min(min(BYScoreGrid))]);
MaxScoreGrid=max([max(max(MonScoreGrid)) max(max(RGScoreGrid))
max(max(BYScoreGrid))]);
```

```
MonScore=(MonScoreGrid-MinScoreGrid)/MaxScoreGrid;
RGScore=(RGScoreGrid-MinScoreGrid)/MaxScoreGrid;
BYScore=(BYScoreGrid-MinScoreGrid)/MaxScoreGrid;
OverallScoreGrid=(MonScoreGrid+RGScoreGrid+BYScoreGrid)/3;
OverallScoreGrid=OverallScoreGrid-min(min(OverallScoreGrid));
OverallScoreGrid=OverallScoreGrid/max(max(OverallScoreGrid));

MonScoreGrid=MonScoreGrid-min(min(MonScoreGrid));
MonScoreGrid=MonScoreGrid/max(max(MonScoreGrid));
RGScoreGrid=RGScoreGrid-min(min(RGScoreGrid));
RGScoreGrid=RGScoreGrid/max(max(RGScoreGrid));
BYScoreGrid=BYScoreGrid-min(min(BYScoreGrid));
BYScoreGrid=BYScoreGrid/max(max(BYScoreGrid));

save(OutputFileName,'nx','ny','xn','yn',...
                    'MonOverallAvgInt'  ,'RGOverallAvgInt','BYOverallAvgInt',...
                    'MonScoreGrid'      ,'RGScoreGrid'    ,'BYScoreGrid'    ,...
                    'OverallScoreGrid'  ,                                '-APPEND');

Image(:,:,1)=OverallScoreGrid; Image(:,:,2)=Image(:,:,1);
Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'OverallScore.bmp'),'bmp');

Image(:,:,1)=MonScoreGrid; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'BWScore.bmp'),'bmp');

Image(:,:,1)=RGScoreGrid; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'RGScore.bmp'),'bmp');

Image(:,:,1)=BYScoreGrid; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
imwrite(Image,strcat(OutPutPath,'BYScore.bmp'),'bmp');

%*************************************************************************
% Plot Reconstructed Sub-Images (If Desired)
%*************************************************************************
if PlotFinalResult==1
  figure(5)
    Position=[10 (SH-HPix-75) WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=MonFinalResult; Image(:,:,2)=Image(:,:,1);
Image(:,:,3)=Image(:,:,1);
    image(Image)
    title('Figure 5: Final Result: Monochrome
Scene','fontsize',12,'fontweight','bold')
  figure(6)
    Position=[SW-WPix-10 (SH-HPix-75) WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=RGFinalResult; Image(:,:,2)=Image(:,:,1);
Image(:,:,3)=Image(:,:,1);
    image(Image)
    title('Figure 6: Final Result: Red-Green
Scene','fontsize',12,'fontweight','bold')
  figure(7)
    Position=[10 40 WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=BYFinalResult; Image(:,:,2)=Image(:,:,1);
Image(:,:,3)=Image(:,:,1);
    image(Image)
```

```
    title('Figure 7: Final Result: Blue-Yellow
Scene','fontsize',12,'fontweight','bold')
    pause(.01)
end

%*************************************************************************
% Plot Scoring Grid (If Desired)
%*************************************************************************
if PlotScoringGrid==1
figure(8)
    Position=[10 (SH-HPix-75) WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=MonScoreGrid; Image(:,:,2)=Image(:,:,1);
Image(:,:,3)=Image(:,:,1);
    image(Image)
    title('Figure 8: Scoring Grid: Monochrome
Scene','fontsize',12,'fontweight','bold')
  figure(9)
    Position=[SW-WPix-10 (SH-HPix-75) WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=RGScoreGrid; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
    image(Image)
    title('Figure 9: Scoring Grid: Red-Green
Scene','fontsize',12,'fontweight','bold')
  figure(10)
    Position=[10 40 WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=BYScoreGrid; Image(:,:,2)=Image(:,:,1); Image(:,:,3)=Image(:,:,1);
    image(Image)
    title('Figure 10: Scoring Grid: Blue-Yellow
Scene','fontsize',12,'fontweight','bold')
  figure(11)
    Position=[SW-WPix-10 40 WPix HPix];
    set(gcf,'Position',Position);
    Image(:,:,1)=OverallScoreGrid; Image(:,:,2)=Image(:,:,1);
Image(:,:,3)=Image(:,:,1);
    image(Image)
    title('Figure 11: Overall Scoring Grid','fontsize',12,'fontweight','bold')
    pause(.01)
end
return
```

Appendix 3: CSF01

Table 1

Main Variables Used by CSF01

#	Variable	Description
1	Alpha	Visual Angle (in Degrees) Subtended by Scene
2	Ang	Aray of Rotational Angles That Will be Applied to Original Scene
3	BYCSF	Array of Blue/Yellow Contrast Sensitivity Function Values Corresponding to Spatial Filter Frequencies Contained in FilterFreq
4	BYFinalResult	Final Reconstructed Blue/Yellow Subimage
5	BYScene	Blue/Yellow Subimage of Scene
6	BYScoreGrid	Blue/Yellow Scoring Grid Array
7	D	Distance (in Meters) from Observer to Actual Scene
8	Filter	Two-Dimensional Array of Filter Values for a Particular Spatial Frequency and Rotational Angle
9	FilterFreq	Array of Spatial Filter Frequencies That Will be Applied to Original Scene
10	FilterX	Unrotated X-Coordinate of Filters
11	FilterY	Unrotated Y-Coordinate of Filters
12	FreqEnd	Size of FilterFreq Array
13	H	Height (in Meters) of Area Encompassed by Scene
14	InputFileName	Full Path and Name of File Containing Scene to be Analyzed
15	InputPath	Directory Containing Scene to be Analyzed
16	Lambda	Parameter Used in Construction of Filters
17	MonCSF	Array of Monochrome Contrast Sensitivity Function Values Corresponding to Spatial Filter Frequencies Contained in FilterFreq
18	MonFinalResult	Final Reconstructed Monochrome Subimage
19	MonScene	Monochrome Subimage of Scene
20	MonScoreGrid	Monochrome Scoring Grid Array
21	OutputFileName	Full Path and Name of File to Which Results Will be Saved
22	OutPutPath	Directory to Which Results Will be Saved
23	OverallScoreGrid	Sum of Monochrome, Red/Green, and Blue/Yellow Scoring Grid Arrays
24	Pool	Pooling Factor
25	RGCSF	Array of Red/Green Contrast Sensitivity Function Values Corresponding to Spatial Filter Frequencies Contained in FilterFreq
26	RGFinalResult	Final Reconstructed Red/Green Subimage
27	RGScene	Red/Green Subimage of Scene
28	RGScoreGrid	Red/Green Scoring Grid Array
29	Rho	Parameter Used in Construction of Filters
30	Scene	The Original Scene to be Analyzed (Assumed to be a Bitmapped Scene)
31	W	Width (in Meters) of Area Encompassed by Scene
32	Xrot	Rotated X-Coordinate Used in Construction of Filters
33	Yrot	Rotated Y-Coordinate Used in Construction of Filters

Table 2

Intermediate and Temporary Variables Used by CSF01

#	Variable	Description
1	AngEnd	Size of Ang Array
2	Area	Area of a Scoring Grid Square in Pixels
3	BYAvgInt	Array Giving the Average Intensity of Each Blue/Yellow Scoring Grid Square
4	BYConvResult	Result of the convolution of the Current Filter with the Blue/Green Subimage
5	BYMatrix	Intermediate Array Used to Construct Blue/Yellow Sub-Image
6	BYOverallAvgIint	Average Intensity of the Entire Blue/Green Reconstructed Subimage
7	BYPoolImage	Intermediate Array Used in Calculating Final Reconstructed Blue/Yellow Sub-Image
8	BYScore	Array Giving the Ratio of Average Intensity to Overall Intensity for Each Blue/Yellow Scoring Grid Square
9	dH	Number of Vertical Pixels per Degree of Visual Angle
10	dW	Number of Horizontal Pixels per Degree of Visual Angle
11	FilterGridX	Intermediate Array Used in Constructing Filters
12	FilterGridY	Intermediate Array Used in Constructing Filters
13	Height	Height (in Pixels) of a Square One Degree of Visual Angle to a Side (=Width)
14	HPix	Width (in Pixels) of Scene
15	iang	For-Loop Variable
16	ifreq	For-Loop Variable
17	InitialPlots	=1 if Original Scene, Monochrome, RG and BY Subimages are to be drawn =0 if not
18	MaxScoreGrid	Intermediate Variable Used to Scale OverallScore Grid Between 0 and 1
19	MinScoreGrid	Intermediate Variable Used to Scale OverallScore Grid Between 0 and 1
20	MonAvgInt	Array Giving the Average Intensity of Each Monochrome Scoring Grid Square
21	MonConvRes	Result of the convolution of the Current Filter with the Monochrome Subimage
22	MonMatrix	Intermediate Array Used to Construct Monochrome Sub-Image
23	MonOverallAvgInt	Average Intensity of the Entire Monochrome Reconstructed Subimage
24	MonPoolImage	Intermediate Array Used in Calculating Final Reconstructed Monochrome Sub-image
25	MonScore	Array Giving the Ratio of Average Intensity to Overall Intensity for Each Monochrome Scoring Grid Square
26	nx	Width in Scoring Grid Squares of Scene
27	ny	Height in Scoring Grid Squares of Scene
28	PixperInch	Array Giving the Horizontal and Vertical Resolution (in Pixels per Inch) of the Monitor Screen
29	PlotFinalResult	=1 if Reconstructed Monochrome, RG and BY Subimages are to be drawn =0 if not
30	PlotScoringGrid	=1 if Monochrome, RG and BY Scoring Grids are to be drawn, =0 if not
31	Position	Position of Lower Left Corner (in Pixels) of Figure to be Drawn
32	RGAvgInt	Array Giving the Average Intensity of Each Red/Green Scoring Grid Square
33	RGConvResult	Result of the convolution of the Current Filter with the Red/Green Subimage
34	RGMatrix	Intermediate Array Used to Construct Red/Green Sub-Image
35	RGOverallAvgInt	Average Intensity of the Entire Red/Green Reconstructed Subimage
36	RGPoolImage	Intermediate Array Used in Calculating Final Reconstructed Red/Green Sub-image
37	RGScore	Array Giving the Ratio of Average Intensity to Overall Intensity for Each Red/Green Scoring Grid Square
38	ScreenSizeInch	Array Giving the Width and Height (in Inches) of the Monitor Screen
39	ScreenSizePix	Array Giving the Width and Height (in Pixels) of the Monitor Screen
40	SH	Monitor Screen Height in Pixels (Used to Position Figures)
41	SW	Monitor Screen Width in Pixels (Used to Position Figures)
42	Width	Width (in Pixels) of a Square One Degree of Visual Angle to a Side
43	Wpix	Height (in Pixels) of Scene
44	xn	Array Containing the Left-Most X-Coordinate of Each Scoring Grid Square
45	yn	Array Containing the Bottom-Most Y-Coordinate of Each Scoring Grid Square
46	z	Number of “Layers” in Scene (For Bitmapped Scenes, z=3)

Appendix 3: CSF01

Table 3

Progress Variables Used by CSF01

#	Variable	Description
1	Count	Progress Variable Giving the Number of Convolutions Completed Thus Far
3	dt	Progress Variable Giving the Time to Complete the Current convolution
5	DT	Progress Variable Giving the Average Time to Complete a Convolution, Based on the Convolutions Completed Thus Far
4	dtsum	Progress Variable Giving the Time Taken to Complete the Number of Convolutions Specified by Count
7	Line1	String Variables Giving Estimate of Time Remaining to Complete Convolutions
8	Line2	String Variables Giving Estimate of Time Remaining to Complete Convolutions
7	Now	Progress Variable Giving the Current Time
2	t0	Progress Variable Giving the Time at the Start of the Current Convolution
6	ToGo	Progress Variable Giving an Estimate of the Time Required to Complete the Remaining Convolutions