2. Spatial Vision Models: Problems and Successes

Stanley A. Klein

2.1 Introduction

Computer vision has progressed slower than expected. Thus the existence of animal vision is comforting. At least we know that vision is not impossible. Human vision provides more than an existence proof since it is generally believed that insights from human vision are critical for progress in robot vision. A goal of this chapter is to point out aspects of human vision that may be of interest to workers in robotic vision. To that end, I will explore several classes of models of human suprathreshold visual discrimination. The models to be discussed are applicable to a wide range of tasks.

During the decade of the 70’s a successful model of visual detection was developed, where detection means that the test pattern is presented on a uniform field. This model, to be called the “standard” model, will be described in Section 2.2. The standard model assumes that there are spatial filters at many positions and orientations and with many sizes. In the early days of the field there had been some confusion regarding the bandwidths of the underlying mechanisms because some data pointed to very narrow bandwidth (one quarter octave) mechanisms (Sachs et al., 1971) and other data pointed to medium bandwidth (one octave wide) mechanisms (Blakemore and Campbell, 1969). Stromeyer and Klein (1975) showed how medium bandwidth mechanisms with tuning functions similar to those found in cortical receptive fields could account for all the detection data as long as one allowed spatial probability summation. Norma Graham’s excellent book (1989) reviews all aspects of the “channel story” that are connected with detection (see (Klein, 1992) for an alternative, somewhat biased version). The review of Graham’s book by Klein and Tyler (1992) points out that there are still a number of unanswered mysteries associated with detection. However, to a first approximation (accuracy up to factors of 2 or 3), the “channel” model proposed by Campbell and Robson (1968) does quite well. The story for suprathreshold discrimination is less clear, as will be discussed in the following...
sections.

Section 2.3 describes the bisection experiments of Klein and Levi (1985). The bisection thresholds are remarkably low and therefore provide a challenge to models of spatial vision. Two approaches for predicting these thresholds will be considered in Sections 2.4 and 2.6.

Section 2.4 presents the viewpoint model of Klein and Levi (1985). It is of special interest for people interested in machine vision because it uses a very interesting representation of an image. It is a representation that throws away local phase information. The intent of this representation is to keep high quality information about relative position while degrading absolute position information. The discussion of the viewpoint model is applicable to a wide class of filter models. The section ends with a discussion of the difficulties that confront filter models when applied to discrimination tasks. Work still needs to be done to learn how the local luminance distribution affects the mechanisms' gain control.

Section 2.5 goes back to the question of what is discrimination and introduces a model-free method that we call the Template Observer for calculating discrimination thresholds. Discrimination is similar to detection except that instead of presenting the test pattern on a uniform field it is presented on a suprathreshold pedestal. The Template Observer assumption is that as the pedestal strength is increased the discrimination threshold gradually becomes equal to the detection threshold. The detection threshold of simple stimuli such as sinusoidal gratings or lines and edges can be taken as the raw input data specifying the sensitivity of the visual system. This model-free approach doesn't always work. The presence of the background can make a big difference in some tasks and very little difference in others. Texture discrimination, shape from shading, and face recognition can all be framed as discrimination tasks. It will be of interest to see which of these tasks can be predicted from the Template Observer approach.

Section 2.6 presents a Template Observer calculation of the very low bisection thresholds presented in Section 2.3. This model-free calculation is a striking success of the Template Observer approach.

In Section 2.7 the relative merits and deficiencies of different modeling approaches will be considered. I will argue that intermixing the approaches may be needed to obtain good predictions over a wide range of conditions.

2.2 The Standard model

My first paper in vision (Stromeyer and Klein, 1971) is one of my favorites. In it we did a calculation that set the stage for all my subsequent modeling. The experiment involved detection of a 9 c/deg test grating with four pedestals: a) a 9 c/deg pedestal, b) a 3 c/deg pedestal (the test is the third harmonic), c) a 1.8 c/deg pedestal (the test is the fifth harmonic) and d) no pedestal. We found that the 9 and 3 c/deg pedestals strongly facilitated the visibility of the test pattern over the no pedestal condition. Nachmias and
Sansbury (1974) found a similar facilitation when the pedestal and test had the same spatial frequency. The finding that a grating could facilitate the visibility of the third harmonic seemed inconsistent with the narrowly tuned channels that were in the air in the early 1970's. Stromeyer and Klein (1974) first had to argue that the presumed very narrow tuning was an artifact of spatial probability summation. Then we had to argue that medium bandwidth (1.5 octave full bandwidth) channels could do the facilitation. We showed that a mechanism with the medium bandwidth tuning demonstrated in adaptation studies (Blakemore and Campbell, 1969) did an excellent job of quantitatively predicting the magnitude of the first plus third harmonic facilitation and the lack of facilitation of a first plus fifth harmonic. The optimal mechanism that was able to achieve the facilitation was found to be a mechanism whose peak sensitivity was near the second harmonic. The lower tail of this mechanism was stimulated by the suprathreshold fundamental (1 octave below the peak) and the upper tail was sensitive to the test pattern (1/2 octave above the peak). The threshold nonlinearity that was demonstrated by the 9 on 9 c/deg facilitation was used to also produce the 9 on 3 c/deg facilitation. The same calculation showed that the medium bandwidth mechanisms were narrow enough that no interactions between a first and a fifth harmonic should be found, in agreement with the data.

The model used to calculate the discrimination threshold is the "standard model" mentioned earlier. It was the basis for subsequent spatial vision models (Wilson and Bergen, 1979; Wilson and Gelb, 1984; Klein and Levi, 1985). The model assumes a number of underlying filters at multiple positions and with multiple peak spatial frequencies. The important additional feature that the model had over earlier models is the presence of a threshold nonlinearity relating the output and input of the filters. The next step in the model is to search through all the filter outputs to find the mechanism with the greatest differential response. The differential response is the response to the test plus pedestal minus the response to just the pedestal. The search across space was easy in our first plus third harmonic stimulus because we were able to argue that the mechanisms located at the zero crossings of the pedestal would be most sensitive for our particular task. So the computer only had to do a one-dimensional search over mechanisms with different peak spatial frequencies.

This model is called a peak detection model since only the optimal mechanism contributes to the discrimination decision. Wilson's version of the standard model (Wilson and Bergen, 1979; Wilson and Gelb, 1984; Wilson, 1986) differs in two respects from the Stromeyer and Klein (1974) model. First, rather than assuming a continuum of mechanisms, Wilson's mechanisms are sparsely distributed in space and spatial frequency. They are separated by about an octave in spatial frequency and by about a quarter cycle in space. This sparse sampling has the advantage that calculations can be done more rapidly (a big deal a decade ago when personal computers were slow). Unfortunately, the sparse mechanism sampling has the disadvantage that it predicts a variety of cues that are not compatible with the data (Wilson, 1986; Mayer and Kim, 1986). The second respect in
which Wilson’s version differs is that rather than looking at just the optimal mechanism he does a “line element” probability summation of the information contributed by all the mechanisms. As we argued (Stromeyer and Klein, 1974; Stromeyer and Klein, 1975) spatial probability summation is indeed needed to explain data when spatial beats are present. For most tasks, however, probability summation doesn’t have a big effect on the results. Wilson has applied the standard model to a number of diverse discrimination tasks with fairly good success. I will, however, offer a critique (and a defense) of this model in Section 2.6. First, I would like to show how the standard model (with modifications) can explain the data that holds the Guinness record for best acuity.

2.3 The Guinness experiment

The main data to be discussed in this chapter comes from two bisection experiments done by Klein and Levi (1985). In the first experiment the stimulus was three bright lines on a computer monitor. A diagram of the stimulus is shown next to the “with overlap” legend in Figure 2.1. The observer’s task was to indicate whether the central test line appeared above or below the midpoint of two outside reference lines. The angular separation between the test and reference lines is the independent variable. The solid triangles in Figure 2.1 show that when the separation is larger than 1.4 min, bisection thresholds are about 1/60 of the separation. For smaller separations the thresholds rise rapidly so that the best threshold is a displacement of 1.4 sec. In subsequent years we have come to realize that the bisection data shown by triangles in Figure 2.1, can not be explained by a unitary model. As will be shown in Figure 2.5 there are two regimes (Levi and Klein, 1990). Below separations of about 10 min is the filter regime, to be discussed in Section 2.4. At larger separations is the local sign regime, where thresholds are relatively independent of the characteristics of the lines such as their spatial frequency content (Levi and Klein, 1993) or their contrast. The proportionality between threshold and separation in the local sign regime is explained by the increasing position uncertainty in peripheral vision. The data in Figure 2.1 represented by filled circles are for the case in which the test line does not overlap the pair of reference lines. For this nonoverlapping stimulus, filters are not able to operate effectively and so thresholds sharply increase in the filter regime.

In the second experiment, a pair of flanks was added 1.3 min outside the reference lines producing a 5-line bisection task. Thresholds decreased to .85 ± .04 sec (at 75% correct). This very small value was found consistently over about 10 runs. The angular threshold of .85 sec is remarkably small, corresponding to a linear distance of 1/4 inch at a distance of a mile, or to an eighth of the wavelength of yellow light when imaged on the retina. This bisection threshold presently holds the Guinness record for best position acuity (Mcfarlan, 1991). As a challenge for others to collect similar data, I will share my confidence that the record can be reduced to .6 sec without much trouble. I hope readers of this chapter will
Figure 2.1: Bisection thresholds for subject DL as a function of the separation between the lines are shown by the filled triangles. The dashed line has a slope of 1, showing that threshold is a more-or-less constant fraction of the separation down to about 1.2 min. For still closer separations, thresholds increase rapidly. The filled circles show thresholds for bisection with the luminance cue removed by placing the test line adjacent to (see inset) the reference lines. At small separations this becomes a vernier task.
Figure 2.2: Schematic representation of the filter sampling for the viewpoint calculation. Shown here are pairs of symmetric and antisymmetric receptive fields of different sizes and at different locations. In the standard models of discrimination only the symmetric mechanisms are usually considered. In the viewpoint model, a Pythagorean sum is taken of the outputs of the pair of matched even and odd symmetric mechanisms to eliminate local phase information.

be inspired to run to the lab to try to claim the record for themselves.

2.4 The Viewpoint model

In modeling the bisection data described in Section 2.3, we decided to add one new ingredient to the standard model. To be more exact, we decided to subtract an ingredient, the local phase, which leads to a novel representation for spatial information. Since this representation will be of interest to researchers in computer vision we will first review the motivation for seeking a new image representation. The standard model looks at the output of a bank of filters such as shown in Figure 2.2. The filters can be assumed to be located at a continuum of positions and with a continuum of sizes. Figure 2.2 shows that at each position there is both a symmetric filter and an antisymmetric filter. Many versions of the standard model only have symmetric filters (Wilson and Bergen, 1979; Wilson and Gelb, 1984; Malik and Perona, 1990; Bennett and Banks, 1991). There is a prob-
Figure 2.3: This staff representation of music is presented as a vivid reminder of what is accomplished by the viewprint representation. In the case of music the abscissa is coarse time and the ordinate is coarse temporal frequency. A musician looking at this representation will be able to “hear” the sound of the music. This representation has lost the information about the absolute phase of each note. For the viewprint, time is replaced by space and temporal frequency is replaced by spatial frequency. Our hypothesis is that this representation of an image is less degraded by eye movements and by some uncertainty in the locations of the underlying mechanisms within a hypercolumn.

Problem with representing the image in terms of the output of symmetric mechanisms. The representation becomes too sensitive to the absolute position of the mechanism. A small uncertainty in position of the mechanism (e.g. due to an eye movement or simply due to a neural miscalibration) would drastically change the mechanism’s expected output. Thus in the standard model the position of each mechanism must be known with great accuracy in order to make the filter outputs meaningful.

In order to develop a representation of the image that is less sensitive to the exact location of each mechanism we decided to look at a representation that is commonly used in audition, as shown in Figure 2.3. In the auditory version the horizontal axis is time with time increasing to the right and the vertical axis is temporal frequency with frequency increasing to the top. What is special about this representation is that local phase has been eliminated. Consider the sound represented by the upper-left most note in Figure
2.3. It is called an "E" note. Since it is a fifth above the standard "A" its frequency is 440 Hz x 3/2 = 660 Hz. This representation tells us the precise frequency of the waveform, corresponding to the distance between peaks of the wave, but it doesn't tell us anything about the phase of the wave. That is a good thing for the pianist, because if accurate phase had been needed the piano key would have had to be hit with a timing accuracy much better than a millisecond; well beyond human capability. A similar representation (a voiceprint) is used for analyzing human speech into formants and also for describing bird songs. We believe that this music representation is also useful for vision. When applied to vision we call this representation a "viewprint" because of its parallel to the voiceprint of audition.

Stromeyer and Klein (1975) first made use of this representation in order to show how probability summation would act differently on amplitude modulated and frequency modulated gratings. In order to produce a viewprint one must first assume a tuning function for the underlying mechanisms. Stromeyer and Klein assumed mechanisms with the spatial frequency tuning found by Blakemore and Campbell (1969) in adaptation studies. It is quite easy to produce a viewprint for a general tuning function when the stimulus is specified in the spatial frequency domain. See the Appendix of Stromeyer and Klein (1975) for details. The viewprint representation was used by Klein and Levi (1985) to account for the bisection data discussed in Section 2.3. In order to simplify the viewprint calculation for the spatial domain a new class of basis functions, called Cauchy functions, were introduced. The symmetric and antisymmetric Cauchy functions can be written as:

\[ S_n(x) = \text{Re}(1 + ix/\sigma)^{-n} = \cos^n(\theta) \cos(n\theta) \]

\[ A_n(x) = \text{Im}(1 + ix/\sigma)^{-n} = \cos^n(\theta) \sin(n\theta) \]

with

\[ \tan(\theta) = x/\sigma \]

where \( x \) is position, \( \sigma \) sets the scale or size of the mechanism and \( n \) sets the bandwidth or number of oscillations of the mechanism (the receptive field has about \( 5\sqrt{n} \) cycles).

The Cauchy functions are special because the symmetric and antisymmetric functions are Hilbert pairs. That means that they have the same Fourier amplitude spectrum, which is given by:

\[ C_n(f) = f^n \exp(-f\sigma) \]

The peak of the tuning function is at \( f = n/\sigma \), its mean is at \( f = (n+1)/\sigma \) and its variance is \( \text{var} = (n+1)/\sigma^2 \). One of the special attractions of the Cauchy functions is that because of the Cauchy contour integration theorem, it is easy to find analytic expression for the convolution of a Cauchy function with many standard functions.

The functions \( S_3(x) \) and \( A_3(x) \) corresponding to \( n = 3 \) are shown in Figure 2.2. These are broad bandwidth mechanisms with full bandwidths of about 2 octaves. Klein and Levi (1985) used the \( n = 3 \) and \( n = 5 \) Cauchy functions for predicting the bisection data.
Spatial Vision Models: Problems and Successes

When designing a filter model one must decide on the shape of the filter. One might decide to use filters with pretty orthogonality properties such as the wavelet basis functions. One might decide to use filters that have a pretty analytic structure in both space and spatial frequency such as the Cauchy functions. Or for good reasons one might want to use Hermite functions that Gabor (1946) thought minimized the joint space-spatial frequency uncertainty. Beware of the Hermite! Klein and Beutter (1992) showed that for the class of functions produced by an nth order polynomial times a Gaussian, the nth order Hermite function maximized rather than minimized the joint space-spatial frequency uncertainty.

In order to calculate a viewpoint of an image, for each size, $\sigma$, one convolves the image with $S_n(x)$ and $A_n(x)$. For the bisection stimulus this convolution consists of a sum of three terms, one for each line of the stimulus.

\[ S(x_m, \sigma) = S_n(x_1 - x_m) + S_n(x_2 - x_m) + S_n(x_3 - x_m) \]  \hspace{1cm} (2.5)

\[ A(x_m, \sigma) = A_n(x_1 - x_m) + A_n(x_2 - x_m) + A_n(x_3 - x_m) \]  \hspace{1cm} (2.6)

where $x_m$ is the location of the center of the mechanism and $x_i$ are the locations of the three lines of the bisection stimulus. We are using the notation $S(x_m, \sigma)$ as the amount of stimulation of a symmetric mechanism centered at $x_m$ and with a size $\sigma$.

So far everything is as it would have been in the standard model. Now comes the new step that eliminates the local phase. Phase is eliminated by simply taking the Pythagorean sum of the responses of the symmetric and the antisymmetric Cauchy functions.

\[ P(x_m, \sigma) = \left[ S(x_m, \sigma)^2 + A(x_m, \sigma)^2 \right]^{\frac{1}{2}} \]  \hspace{1cm} (2.7)

In order to connect the filter output to psychophysical thresholds, the filter sensitivity must be normalized according to the contrast sensitivity function and one must introduce the nonlinear transducer function that produces facilitation at low contrasts (discussed in Section 2.2) and saturation at high contrasts. The intent of these nonlinearities is to convert the output of Eq. 2.7 into a measure of the signal to noise of the visual system, $d'$. Klein and Levi (1985) show that if the filters are normalized by the contrast sensitivity function, the signal to noise ratio of a mechanism is given by:

\[ d' = 2.5\ln[1 + .5P(x_m, \sigma)^2] \]  \hspace{1cm} (2.8)

Figure 2.4 shows what the viewpoint of a three-line bisection stimulus looks like. The stimulus is three lines separated by 4 min. In the lower panel the middle line is given an offset of .08 min corresponding to a Weber fraction of 1/50. This value is chosen because it corresponds to the threshold for seeing the offset, as shown in Figure 2.1. In the upper panel the offset of the middle line is half-threshold (.04 min). The curved lines labelled 1, 6, 10, 14, and 18 are iso-$d'$ contours. Thus a value of 10 means that that particular
Figure 2.A: Viewprints for the three-line bisection stimulus using third order Cauchy functions as the underlying mechanisms. The interline separation was 4.0 arc min. The two panels correspond to two offsets of the test line. Lower panel: Threshold offset (0.03 min, corresponding to a 1/50 Weber fraction). Upper panel: A half-threshold (0.04 min) offset. The shading represents one bit of absolute phase (the sign bit). The contour lines are the iso-δ' lines of the contrast-response function as indicated by the numbers by the lines. The clusters of slightly smaller numbers are the differential response and indicate the discriminability of the offset. Negative differential-response values are indicated by the reversed contrast numbers.
mechanism is 10 jnds (just noticeable differences) above threshold. The dotted contour lines are circling a region of the viewprint near a depression. The number 14 at the bottom of each plot indicates that the associated dotted contour line is at a level of \( d' = 14 \).

The two solid dots may be the most important features of the viewprint. They are a pair of null points. A null point corresponds to a mechanism whose spatial position, \( x_m \), and size, \( \sigma \), are such that neither the symmetric nor the antisymmetric mechanisms are stimulated. For a two-line stimulus the null point has a location, \( x_m \), that is halfway between the two lines, so that the antisymmetric mechanism doesn’t contribute. The size of the mechanism is determined by the condition that the stimulus lines fall on the zero crossings of the symmetric mechanism. The location of the zero crossing is given by Eq. 2.1:

\[
\cos(n\theta) = 0
\]

which implies

\[
\theta = \pi/2n
\]

(2.9)

The angle \( \theta \) is given by

\[
\theta = x/\sigma = \text{sep}/2\sigma
\]

(2.10)

since the location of each line is a distance \( \text{sep}/2 \) from the center of the mechanism, where \( \text{sep} \) is the separation between the pair of stimulus lines. Combining Eqs 2.10 and 2.11 shows that the mechanism size corresponding to the null point is:

\[
\sigma = \text{sep}/(2\tan(\pi/2n))
\]

(2.12)

For \( n = 3 \), \( \sigma = \text{sep} \sqrt{3}/2 = .866\text{sep} \). For the three-line stimulus the location of the null points is close to where they would be for the two-line case.

The reason that the null points are so important is that they are the main features that significantly change when the middle line is shifted by a threshold amount. Notice that in the upper panel of Figure 2.4 the two null points are at about the same height (same \( \sigma \)). In the lower panel, the two null points are at visibly different values of \( \sigma \). Other aspects of the viewprint are less sensitive to the shift of the test line. This shift in the vertical height of the nullpoints leads to a visible differential response of the mechanisms that are used to detect the offset.

The numbers printed near the null points are the differential responses in \( d' \) units. The differential response is simply the difference between the \( d' \) value of each mechanism after the middle line is given its offset and the \( d' \) value before the offset (when the middle line is exactly at the bisection point). A value of 1 means that that mechanism has a unity signal to noise ratio (\( \Delta d' = 1 \)). The reversed contrast values in Figure 2.4 indicate that the mechanism activity with the displacement is less than the activity with no displacement. If one would enlarge the region close to the null point one would find mechanisms whose differential response is larger than \( \Delta d' = 2 \). This is because the slope of the viewprint
activity becomes extremely steep. Thus we have the unexpected problem that rather than having difficulty in predicting the very small bisection offsets, the model predicts offsets that are even smaller than what the human can see. Klein and Levi (1985) introduced a .5 min "final" blur to smooth out the null point. This final blur is a free parameter that can change the magnitude of the predicted threshold. Thresholds predicted by the viewprint model are shown in Figure 2.5. These predictions were originally shown by Klein and Levi (1987). For separations of less than 10 min the agreement with the data of Figure 2.1 is excellent (but remember that there was a free parameter that controls the final blurring). For separations greater than 10 min the viewprint predicts a constant threshold since thresholds are based on measuring the positions of each line separately. This is the local sign regime where as discussed by Klein and Levi (1987) and Levi and Klein
(1990, 1992) thresholds are degraded because of poorer spatial localization in peripheral vision. It would be nice to develop an approach for predicting the bisection threshold that is less sensitive to the many assumptions connected with the viewprint model. That is the topic of the next section.

A most serious question for computer vision is whether the viewprint representation can be inverted to reconstruct the original image. From what we have said so far the answer is clearly, NO! The process of taking the Pythagorean sum erased all the local phase information, so that it is not possible to tell whether a three line stimulus consists of light lines or dark lines. In order to recapture some of the phase information, we appended to the viewprint one bit of phase information. This phase information is indicated by the shading that is seen in Figure 2.4. The unshaded regions indicate the locations where the symmetric mechanisms receive positive stimulation. The shaded regions correspond to negative stimulation. Thus the viewprint representation is really a dual representation. The one bit of phase information is treated much the same as the color of the stimulus. There are clearly many tasks that depend on knowing the “color” or local phase. However, the “color” information might be coarser than the intensity information, in both spatial resolution and in intensity resolution. It is quite possible that the finest spatial discriminations, such as needed for the bisection task, are based on the Pythagorean sum where the local phase is eliminated. The advantage of using the Pythagorean sum is that it has greater spatial smoothness than does a representation that maintains local phase. Thus it is less degraded by eye movements and other factors contributing to spatial uncertainty, such as any uncertainty in the location of the mechanism.

I will end this section on filter models by pointing out some problems. The main trouble with filter models is that too many assumptions are needed. The typical filter model (Wilson, 1986) determines the sensitivity and nonlinearities of the mechanisms by using masking data involving small patches of sinusoids. One then assumes that these same mechanism properties are still valid when applied to a very different stimulus, such as three-line bisection on a dark background. There are so many ingredients that go into a model: the mechanism bandwidth, sensitivity, spatial sampling, spatial frequency sampling, nonlinearities, phase sampling, position uncertainty, orientation tuning, aspect ratio (length to width), type of probability summation. With so many ingredients it is difficult to know which aspects of the model are critical and which are irrelevant.

A second problem with present filter models, as stated earlier, is the lack of a good understanding of luminance gain control. This problem directly affects the viewprint prediction of the three-line and five-line bisection data. One needs to know the sensitivity of the mechanisms near the middle line. We showed (Klein and Levi, 1985) that bisection thresholds are not strongly affected by the line luminance. Further research on the spatial properties of the luminance gain control is needed.
2.5 The Template Observer

In this section we show how Geisler's photon based, "stimulus known exactly" Ideal Observer Model (1984, 1985, 1989) can be transformed to a contrast based model that has greater applicability to photopic vision. It works as follows. Suppose the observer's task is to discriminate between patterns A and B. Let \( M_i(1 + E_{A_i}) \) and \( M_i(1 + E_{B_i}) \) be the expected quantum catches in the ith cone for the two patterns, where \( M_i \) is the quantum catch due to the mean luminance background and \( E_{A_i} \) and \( E_{B_i} \) are fractional deviations from background. We have separated out the "mean luminance" \( M_i \) from the "contrast" \( E_{A_i} \) in order to develop an ideal observer model that only depends on contrast. In this "stimulus known exactly" case the ideal observer calculates the log of the likelihood ratio, which, from binomial statistics is simply given by

\[
\text{LogLik} = \sum_i O_i \ln((1 + E_{A_i})/(1 + E_{B_i}))
\]

where \( O_i \) is the observed number of photons absorbed in the ith cone and the summation is over all the cones. If the likelihood ratio exceeds a criterion level then one chooses pattern A. Geisler and Davila (1985) pointed out that for reasonably low contrasts, \( E_i \ll 1 \), the log likelihood becomes:

\[
\text{LogLik} \approx \sum_i O_i (E_{A_i} - E_{B_i})
\]

This is a template match between the observed quantum catch and the expected catch of the difference in patterns to be discriminated. We will call this decision rule "The Template Observer". If one calls pattern \( E_{B_i} \) the "pedestal" pattern, then \( E_{A_i} \) becomes the "test plus pedestal" where the test pattern, the template, is \( E_{Ti} = E_{A_i} - E_{B_i} \), and therefore:

\[
\text{LogLik} \approx \sum_i O_i E_{Ti}
\]

This equation is remarkably simple. It says that the most efficient way to detect a pattern is to overlap the stimulus with a matched template whose profile has the shape of the pattern to be detected. In this section we show how this matched filter approach can be used for predicting vernier acuity. In the next section the method is applied to the three line bisection task discussed in Sections 2.3 and 2.4. As will be seen, this simple idea goes a long way towards explaining many features of spatial vision.

Geisler tested the ideal observer prediction with experiments on 2-dot contrast discrimination, 2-dot resolution and 2-dot separation discrimination. For all three tasks the human observer had an efficiency of between 10% and 20% of the ideal. That is, in order to make the discriminations the human needed to absorb about 5 to 10 times the number of photons than did the ideal observer. Averaged over three observers the efficiency for resolution and separation was 11% and 12% respectively. This result is exciting since it implies that hyperacuity is not "better" than resolution acuity. They are the same when measured with the proper metric! However, the ideal observer's predictions being about
3 times too small (square root of efficiency) implies that the noise limiting performance comes mainly from sources other than photon statistics. Thus one can not take the ideal observer predictions too seriously. One of the goals of my research in the last few years is to show how a modified ideal observer approach (based on the Template Observer) can be applied to the high luminance regime where factors due to photon noise can be replaced with the contrast based factors actually limiting performance.

Many experiments in spatial vision can be understood in terms of the Template Observer framework defined above (in our earlier publications we called it the “test-pedestal” framework). Theories similar to the Template Observer have been proposed before for vernier acuity. Watt’s orthoaxial mechanism (Watt et al., 1983; Watt, 1984) calculates the amount of “error energy” produced by the vernier offset. Morgan and Aiba (1986) developed a theory to assess the “cone” activation produced by the offset. However, these researchers did not have a reliable way to calculate the absolute thresholds. Our new element is that we advocate measuring the visibility of the template in a detection task. This and the next section show how hyperacuity thresholds can be predicted from the individual’s sensitivity to contrast.

Consider vernier acuity of an edge or a line. The edge offset is produced by adding a thin line to one half of the edge pedestal as shown in the left panel of Figure 6. The line offset is produced by adding a thin dipole to one half of the line pedestal as shown in the right panel. A dipole is a pair of equal strength, opposite polarity abutting lines. Klein, Casson and Carney (1990) proposed the shockingly simple idea that vernier thresholds could be understood in terms of the visibility of the test line (for edge vernier acuity) or the test dipole (for line vernier). Since the suprathreshold pedestal masks the test this prediction is not expected to hold at high pedestal contrasts. However, for low contrast pedestals, the masking should be reduced, so our hypothesis was that low contrast vernier acuity should be predicted by the detection threshold of a thin line or dipole. As shown by the data in Figure 2.7 for the case of line vernier acuity for two observers this hypothesis was correct! In the left panel the threshold units are minutes and in the right panel the data replotted with thresholds in dipole strength units of %min² (see Klein, 1989 for details on multipole units). The dipole strength is given by the strength of the pedestal line (in units of %min, which is the % contrast of the line times its width in min) times the vernier offset (in min). Thus, the ordinate of Figure 2.7b is obtained by multiplying the ordinate of Figure 2.7a by the abscissa. The advantage of replotothe data as in Figure 2.7b is that the dipole detection thresholds can be directly plotted on the same graph as shown by the leftmost points connected by a dotted line. This new way of plotting the vernier data can be called a tvp curve (threshold vs. intensity). One might quibble about whether it should be called tvp (threshold vs. pedestal strength) rather than tvi, but tvi has a nice ring to it.

As discussed by Klein, Casson and Carney (1990) the dipole detection threshold can also be plotted on the left panel by dividing the dipole threshold (in %min²) by the line
Figure 2.6: Illustration of how the Template Observer accounts for vernier acuity. The vernier offset of an edge is produced by adding a line to one half of the edge. Vernier offset of a line is produced by adding a dipole to one half of the line. When the line in (a) has the same contrast as the edge the edge will be shifted by the thickness of the line. Similarly for the dipole on a line pedestal in (b). If the test pattern had lower contrast then the centroid of the pedestal would shift by a smaller amount.
Figure 2.7: Panel (a) shows the vernier thresholds (in min) for a line for two observers as a function of line strength. Line strength has units of %min which is the % contrast of the line times its thickness in min. The solid line with a slope of -0.7 matches the slope of the data. The vertical location of the line has been shifted for ease of viewing. The error bars are 1 SE. Panel (b) shows the same data replotted using the same abscissa and with the vertical axis being the strength of the test dipole in units of %min^2 rather than vernier offset. This ordinate is the ordinate of panel (a) times the abscissa. The slope of the line is +0.3 since the slope is that of panel (a) plus one. The left-most data, connected by dotted lines, are the detection thresholds of a dipole on a uniform field background (zero pedestal strength). The ordinate for vernier acuity at low pedestal strength is close to the detection thresholds. The detection thresholds are also plotted in panel (a) by dividing the ordinate by the abscissa of panel (b). The resulting value corresponds to Ricco's diameter for detecting a dipole. These results provide a model-free prediction of vernier acuity of a line.
detection threshold (in %min) which is the abscissa value at which the detection datum is plotted. This quotient of the dipole and line detection thresholds are the counterparts of Ricco’s summation area for dipole detection (see (Klein et al., 1990) for details).

Figure 2.7b shows that when the dipole thresholds in the vernier task are extrapolated to the line detection threshold, the predicted detection threshold is in good agreement with the data. For observer SK Figure 2.7b shows that at the lowest two pedestal contrasts (below three times the line detection threshold) the vernier thresholds are elevated. This is not surprising since the target is barely visible. Geisler and Davila (1985) find that the ideal observer prediction of a two-dot hyperacuity task shows a similar degradation when the pedestal strength is close to its detection threshold and when the background luminance is not zero.

What have we accomplished so far? Sometimes when I look at Figure 2.7b I am impressed that an important result is shown. It looks like a model-free prediction of vernier acuity. No assumptions were needed about the linear and nonlinear properties of the underlying mechanisms. The only thing not explained is the shallow slope of the tvi curve whereby thresholds increase with masking contrast. However, sometimes when I look at Figure 2.7 it looks trivial. I say to myself, how could it be any other way. If the early stage of vision involves a linear filter then of course the visibility of the test pattern will not be masked at low pedestal contrasts, it seems like a tautology without content since Figure 2.7 shows that “of course” line vernier acuity must be based on the visibility of the dipole. However, in peripheral vision and for strabismic (but not anisometropic) amblyopes, vernier thresholds are 2-3 times worse than what the Template Observer would predict (Levi and Klein, 1992). Thus, the extra 2-3 fold loss in peripheral vernier acuity shown in Figure 2.7b can not be explained by retinal factors such as poor quantum catch and early saturation of the peripheral cones since all retinal factors are balanced in our Template Observer (low contrast) approach. Levi, Klein and Aitsebaomo (1985) proposed that there might be a number of extra factors that are likely to be of cortical origin that account for the extra degradation found in peripheral vision and in strabismic amblyopes. We hypothesized that in peripheral vision undersampling or neural scrambling will affect a position task such as vernier acuity but have minimal effect on a detection task.

As was stated in the preceding paragraph one missing ingredient is an explanation of the tvi slope whereby a high contrast line pedestal masks the dipole detection. In order to account for the tvi slope in a separate experiment we measured dipole contrast discrimination (Carney and Klein, 1989). The observer’s task is to measure the visibility of a dipole test pattern, just as before, except now the masking pedestal is a suprathreshold dipole rather than a suprathreshold line. The data differ in two ways from that shown in Figure 2.7b. First, the tvi slope of the contrast discrimination is steeper than the vernier data, .6 vs. .3. Slopes of about .6 are commonly found in contrast discrimination studies (Legge and Foley, 1980). Second, when the abscissa is transformed to contrast threshold units (to be able to compare the line and dipole pedestals) the dipole detection
thresholds are lower in the contrast discrimination task. This is not surprising since a pedestal facilitation effect is found in contrast discrimination but not in vernier acuity.

A detailed spatial vision model is needed to account for these discrepancies. I think it is clear why the tvı slope should be shallower for the vernier task. In a model with orientation tuned mechanisms, as the pedestal contrast increases the optimal mechanism for detecting the vernier offset changes. At higher contrasts one could use a mechanism tilted at a larger angle to avoid the masking. The consequence is that the masking effect would be smaller than if the same mechanism had been used at all contrasts. Exactly the same effect was found by Stromeyer and Klein (1974) in the task of detecting a 9 c/deg grating in the presence of a 3 c/deg pedestal as compared to a 9 c/deg pedestal, as a function of pedestal contrast.

Recently, we have applied the Template Observer to vernier acuity and contrast discrimination (jnd) of sinusoidal gratings (Hu et al., 1992). In the jnd stimulus the sinusoidal test grating was added in-phase to half of the pedestal whereas in the vernier stimulus the same test grating was added with a 90 deg phase shift to the pedestal (the phase was actually slightly greater than 90 deg in order to keep the contrast constant). Our use of sinusoids has an advantage over multipoles (edges, lines and dipoles) since both the test and pedestal have the same spatial frequency content for both the vernier task and the jnd task. This will simplify any future filter modeling.

Some of the results using sinusoids are shown in Figure 2.8: a) At low contrasts and for a mid-range of spatial frequencies (5 - 10 c/deg) the vernier and jnd thresholds are about equal. b) Vernier thresholds rise slower with increasing pedestal contrast than jnd thresholds (power exponents of .4 vs. .6). c) Vernier thresholds degrade with increasing spatial frequency whereas jnd thresholds are relatively unaffected. The Template Observer accounts for the success of the first-order effect of point a). The different tvı slopes of point b) are not bothersome because oriented mechanisms are able to avoid some of the masking at high contrasts and the facilitation at low contrasts as discussed earlier. What is somewhat bothersome is the sharp degradation of vernier thresholds at high spatial frequencies. In order to better understand this loss, further experiments were done varying the length of the gratings and the gap between the two halves. It was surprisingly found that at high spatial frequencies vernier thresholds were reduced when the grating length was shortened (holding the number of cycles constant). A small gap of about 2 min was used. Under these conditions the jnd thresholds were elevated. For short gratings as the gap increased the jnd thresholds were reduced but the vernier thresholds were elevated. The point of these results is that we found that one should compare the jnd and vernier data each under their optimal conditions. When this was done, the differences in spatial frequency tuning of the vernier and jnd configurations was much reduced. The Template Observer is not the final word, it is only the beginning. The Template Observer is used to predict discrimination thresholds under optimal conditions. It is clear that thresholds can be degraded by many stimulus manipulations. For example, separating the features
Figure 2.8: Data comparing vernier acuity and contrast discrimination (jnd) of sinusoidal gratings for two observers. Both thresholds are measured in contrast units. For the vernier task the shift of one half of the grating is thought of as being produced by adding a low contrast test grating, 90 deg phase shift. The data show that to first order, vernier acuity is well predicted by contrast jnd. There are small differences in the dependence of the vernier and jnd data on pedestal contrast and on spatial frequency. The difference in spatial frequency dependence is lessened if an optimal stimulus length and optimal gap is used. The difference in contrast dependence can be understood if the vernier data is modelled using orientation tuned mechanisms that do not have full overlap with the pedestal grating.
in a vernier task will sharply degrade vernier thresholds without affecting detection or jnd thresholds. Insight about the underlying mechanisms and their interactions is gained by looking at departures of the data from what is expected of the Template Observer.

2.6 Applying the Template Observer to the Guinnes data

In order to develop one's intuition about how the Template Observer works it is instructive to apply the formalism to make a "model-free" calculation of the bisection data discussed in Section 2.3. This section will be the first time that this calculation has been published.

The stimulus that gave the best thresholds had 5 lines that were almost equally spaced. The task was to detect a shift of the middle line. As seen in Figure 2.6b a shift of a thin line is equivalent to adding a dipole to the line. The Template Observer model would say that under optimal conditions, the bisection threshold is equal to the dipole detection threshold. The dipole detection threshold is based on the visibility of a dipole on a uniform field. For our stimulus a uniform field corresponds to the case when the 5 lines are so close together that they become blurred, thus producing a uniform field. We will discuss later what happens when the separation between lines increases.

A dipole consists of a pair of opposite polarity, closely spaced thin lines. The strength of a dipole is equal to the strength of one of the constituent lines times the distance between the pair of lines. The strength of the line, in turn, equals the thickness of the line times its contrast. For the present case of closely spaced thin lines, the thickness of a line equals the line spacing and the contrast equals 100%. This method for calculating line strength can be justified by blurring each line so that the grating of thin lines becomes a uniform field. As discussed by Klein (1989) a blurring operation does not change the strength of a multipole (including a line). In that case the thickness of each line equals the line separation. The contrast produced by shifting the line can now be calculated as

\[ c = \Delta L/L = -L/L = -100\% . \]

We are now ready to proceed with the calculation. Suppose the line spacing in the bisection stimulus is \( s \) min, so the line strength is 100\% min, and suppose the bisection threshold is \( \delta \). The dipole strength, \( D \), is then given by:

\[ D = s6100\% \text{min}^2. \tag{2.15} \]

The leftmost datum in Figure 2.6b shows each observer's dipole detection threshold. For one observer the threshold is slightly less than 1 \%min\(^2\) and for the other observer it is about 1.7 \%min\(^2\). The average gives a threshold of about \( D = 1.3 \% \text{min}^2 \). Eq. 2.15 can be used to get an expression for the bisection threshold in terms of the line spacing:

\[ \delta = D/100\%s = .013/s. \tag{2.16} \]
For a separation of 1.3 min which produced the best bisection thresholds (Klein, 1985) Eq. 2.16 leads to a predicted threshold of:

\[ \delta = \frac{0.13}{1.3} = 0.01 \text{ min} = 0.6 \text{ sec} \]  \hspace{1cm} (2.17)

This predicted value of 0.6 sec is very close to the experimental result of about 1.3 sec at \( d' = 1 \) (84% correct) or \( d = 0.85 \) sec at \( d' = 0.68 \) (75% correct). In more recent experiments (Carney and Klein, 1989) we have found bisection thresholds as low as 0.85 min (\( d' = 1 \)).

These experimental values are slightly higher than the Template Observer prediction. This may not be surprising since in actuality the stimulus was not a uniform field, but rather it was a very narrow stimulus field consisting of only 5 lines. Thus the edges of the stimulus could have masking effects. Also the background wasn't a uniform luminance. At a separation of 1.3 min there was a clear gap between lines. This could easily cause some elevation of threshold.

As seen in Eq. 2.16 if the separation between lines is halved, the bisection threshold should be doubled. Figure 2.1 shows that the bisection thresholds actually degrade faster than is predicted by Eq. 2.16. This extra degradation is undoubtedly due to the very small extent of the stimulus. When the line separation is 1 min the two reference lines are only 2 min apart. This is too small to provide the uniform background that is needed for optimal visibility of the dipole. The introduction of the two extra flanking lines in the 5 line bisection stimulus provided enough of a uniform field to allow thresholds close to that predicted by Eq. 2.17.

I consider the agreement of this crude calculation with the actual data to be most amazing. I would be very surprised if any filter model could come as close. As discussed in Section 2.4, the viewprint calculation involved a number of assumptions such as the mechanism bandwidth and a final blurring stage. Wilson (1986) also needed an assumption about the effective contrast of the stimulus to obtain the proper vertical scaling of the predicted bisection threshold. The calculation done in this section did not need any assumptions at all!

### 2.7 A comparison of three approaches

This last section compares three approaches for modeling spatial vision: 1) the "general framework" approach for texture discrimination, 2) the "multiparameter" detailed modeling approach, and 3) the model-free Template Observer approach discussed in Sections 2.5 and 2.6.

#### 2.7.1 General framework, coarse models

The general framework approaches for texture discrimination of Malik and Perona (1990) and of Klein and Tyler (1986) will be discussed here. Malik and Perona (1990) seek to
explain the data of the sort that Julesz used for developing his higher order statistics and later his texton model (Julesz, 1981). The model that they use is quite similar to the standard model discussed in Section 2.2. They assume symmetric filters of multiple sizes, orientations and positions. Since they are applying the filters to suprathreshold stimuli rather than to detection tasks they added some contrast gain control to the model in order to normalize the filter outputs. *Their model is able to robustly predict the relative difficulty of texture discriminations for a wide range of different textures.*

Klein and Tyler (1986) seek to predict texture discrimination of repetitive textures. Their formalism based on higher order correlation functions is an outgrowth of the Julesz “statistics” approach (Julesz, 1962). The goal of this approach is to classify the different types of phase discrimination. The model produces an ordering of which stimuli should produce easier phase discriminations.

In these “general framework” models no attempt is made to carefully measure the sensitivity or the tuning of the underlying mechanisms since the absolute magnitude of the discriminations are not predicted. Only the relative discriminability of different patterns are predicted. These models do not have sufficient structure to predict details such as the cusp shown in the bisection data of Figure 2.1 for separations of 2 min. The advantage of coarse “general framework” models is that they focus attention on what aspects of the model are essential and what are irrelevant. Thus for the texture discriminations considered by Malik and Perona the detailed assumptions going into the Wilson model are not relevant.

### 2.7.2 Multiparameter “exact” modeling

The filter models of Wilson and co-workers (Wilson and Bergen, 1979; Wilson and Gelb, 1984 ; Wilson, 1986; Wilson, 1991a) and of Klein and Levi (1985) have a quite different goal than the coarse models described in the preceding paragraphs. The goal here is to make precise predictions of the data. Wilson’s multiparameter models are often misunderstood. It is thought by some that with so many parameters one could fit anything. That is unfair to Wilson. He actually has amazingly few free parameters. The shapes and sensitivities of his underlying filters are fixed by masking studies that were done years ago. Often one or two new parameters are needed for each different data set. For example, in amblyopic vision (Wilson, 1991a) some parameters are needed to characterize extra spatial losses that would not have shown up in the masking studies that set the filter characteristics. The hope is that these one or two extra parameters would be able to fit dozens of data points under a variety of different conditions.

If the structure of the underlying model is correct then the rich structure of the data should be able to be captured without detailed curve fitting. For example, let us consider the bisection data shown in Figure 2.1. Notice the dramatic cusp that appears at a separation of 2 min. As discussed in our article (Klein and Levi, 1985) this cusp has
been replicated in independent experiments. A good filter model should produce the cusp naturally without needing a delicate cancellation of terms. As seen in Figure 2.5 the simple viewprint model fails in this regard. Wilson's model is able to produce a cusp (Wilson, 1986) however, it produces too many cusps. The Wilson model cusps are produced by sparse sampling in space and spatial frequency.

It is easy to criticize a detailed model such as Wilson's. For example, the limited number of spatial frequency channels has been questioned by the smooth spatial frequency discrimination data of May and Kim (1986). However, one must remember that it is exactly because detailed models are easy to criticize that they are valuable. Detailed models make very detailed predictions. Much can be learned from accurate predictions. Remember that Newton's theory of gravity is accurate to many, many decimal places. Extremely accurate measurements are needed to show that Newton's predictions are wrong and that Einstein's are correct. Coarse models such as those of Malik and Perona (1990) or Klein and Tyler (1986) might provide an excellent general structures for modeling, but since they don't make accurate predictions they may be hard to disprove. The subtle nuances of the data, such as the bisection cusp at 2 min separations, provide insight into the inner workings of the visual system. Detailed quantitative are needed to capture these details.

2.7.3 Template Observer

The job of the Template Observer model discussed in Section 2.5 and 2.6 is to determine the overall sensitivity of discrimination tasks. This is exactly what filter models do poorly. It is difficult for filter models to predict the same hyperacuity thresholds for a wide variety of stimuli (Westheimer and McKee, 1977). In order to get the overall sensitivity correct, filter models typically must adjust one of the parameters of the model. Thus the Template Observer model is complementary to the models based on underlying mechanisms discussed previously. It is likely that successful predictions of discrimination tasks will require a combination of these approaches.

Acknowledgements

This research was supported by an Air Force Office of Sponsored Research grant AFOSR 89-0238. I thank Charles Stromeyer, Dennis Levi and Thom Carney for their collaborations over the past 20 years that made my research possible.
9. References


