Metamers of the Early Visual System

Weihan Kong

May 2014

Submitted to
New York University
Courant Institute of Mathematical Sciences
Department of Computer Science
in Partial Fulfillment of the Requirements for the Degree of
Master of Science

Thesis supervisor: Eero P. Simoncelli

____

Eero P. Simoncelli

____

Yann LeCun

Acknowledgements

I would like to thank my advisor Eero Simoncelli for offering me this opportunity to do neuroscience research and an elegant lab space to study. His wisdom, passion and decency influenced me profoundly.

Abstract

The visual information processing of human starts at retina. We developed a population model for retinal ganglion cells(RGC) whose receptive field sizes grow with eccentricity. We generated visual metamers, stimuli that are perceptually the same but physically different, to test the model. We developed a spatial-variant filtering method for approximation of linear shift-variant systems based on an idea of warped Fourier transform. Detailed procedure and results are discussed. The model provides a basis for building metamers of cascaded models of the visual pathway.

Contents

1 Introduction

The visual system is one of the most important systems in human brain. As the initial part of the visual pathway, the properties of retina and V1 has been extensively explored [3, 7] for decades but the computation performed by later stages such as inferotemporal cortex remain a mystery. A recent effort [2] that provided a functional account of neurons in V2 using visual metamers examplified a quantitative framework for assessing models of visual systems. A cascaded model involving all the well known areas and a corresponding metamer generating method are desired to test the accuracy of our understanding to the visual system through psychophysics. We chose ganglion cells as the starting point and built a linear model to represent the early visual system as a basis for later stages.

The main idea of this framework is to explore what information is lost in each stage of the visual system. Previous phyiological studies are helpful in directing this research. It has been shown the sizes of receptive fields of ganglion cells and neurons in later stages grow with eccentricity [2]. We hypothesized that this property introduce loss of information in the ensamble response to the visual input. Stimuli whose responses coincides because of this loss of information should look the same to human observers. They are refered to as metamers.

We developed a spatial-variant filtering method to generate metamers for this model of early visual system and did casual psychophysics to estimate the parameters that turned out to match the physiological data. The model and metamer generating method are ready to be generalized to models with adapation and non-linearity and provide a first building block for the cascaded model.

2 Metamers

2.1 What Are Metamers

Perceptual metamers are stimuli that are physically different, but perceptually the same. In vision science, the stimuli are usually images observed by human subjects.

This term is believed to appear firstly in the study of *Trichromacy.* Back in the middle of 17th century, Issac Newton found out that light can be decomposed into rays of different wavelength and characterized by its spectrum. About two hundred years later, a series of psychophysics experiments, called *color-matching* experiments, led to the conclusion that only three different colors are needed to represent all the colors that are distinguishable by human [7].

In these experiments, subjects looked at a bipartite field where the left side was illuminated by a test light of arbitrary spectrum and the right side was illuminated by a combination of three primary lights of fixed spectra. The subjects were asked to adjust the intensity of the three primary lights until the mixed lights on the right side appear the same with the test light on the left side.

It turned out that any test light can be matched with some combination of three well-chosen primary lights. Except for a few special cases(e.g. colorblind subjects), every subject gave the same combination for the same test light and the result was stable over time. Different test lights that result in the same combination of primary lights are indistinguishable to human observers. They are metamers.

The fact that the degrees of freedom of human perception for color is 3 not only is the foundation of pervasively used colorful display technologies, but also has important implications for how the human visual system works: it reduces the dimension of incoming color information from very high (basically infinite since the spectrum is continuous) to 3 and opens the possibilities for metamers.

2.2 Metamers of Linear Systems

The study of Trichromacy can also make an excellent example of how the theory of linear systems can be used to describe the mechanism of the brain and how metamers can be derived from it.

In the color-matching experiment, we can see the spectrum of the test light as input, the combination of the primary lights that the subject adjusts to match the test light as output, and the whole process, from the subject's eyes receiving the light to his/her brain processing the information to adjusting the combination, as a system. The input is a vector of N elements, denoted as l, if we divide the spectrum into N bins. The output is a vector with 3 numbers, denoted as r. The system is denoted as M. The whole process can be denoted as

$$r = Ml$$

Investigators in the 19th century had discovered that within a certain range this system obeys the rule of superposition: (1) if the intensity of the test light is increased by a factor, the result combination would also be increased by the same factor; (2) if two test lights are added together, the result combination would be the sum of the combination of the two individual test lights. Specifically,

$$M(\alpha l) = \alpha(Ml)$$

$$M(l_1 + l_2) = M(l_1) + M(l_2)$$

This means M is a linear system and can be described as a 3 by N matrix. From linear algebra we know such a matrix has at most rank of 3 and must have a *nullspace* [6]. If the difference of the two lights, $l_1 - l_2$

lies in the nullspace of M, we must have $Ml_1 = Ml_2$. Then l_1 and l_2 are metamers.

2.3 Why Are Metamers Useful

If we want to verify whether a model correctly describes the behavior of a real system, ideally we would like to test every possible input on both the model and the real system to see whether their outputs coincide. However this is impossible in the research of the visual system for two reasons.

First of all, the number of possible inputs is infinite. The input of the visual system at a given location and time is described by the plenoptic function [1]

$$P(x, y, \lambda)$$

where λ is wavelength, (x, y) is the coordinate of the light projected onto the picture plane of the eye. Even though we can limit the range and resolution of these parameters and quantize the energy of incoming light, a reasonable approximation would still result in a huge number of possibilities that is not feasible to test one by one. This is refered to as the *Curse of Dimensionality.*

Secondly, in the case of Trichromacy the output of the system are three numbers produced by the subject. It is well defined and easy to measure. But the output of the visual system, including retina, is the activities of neurons that presumably project onto the next stage of the visual pathway, which is not accurately defined and very hard to measure, especially in human.

Although we cannot measure how information is represented in the response of the early visual system directly, we can measure what information is thrown away by it. If we assume the brain cannot access the information lost in the early visual system, no matter how complicated the later stage of the processing is, such information cannot be recovered and the loss can be confirmed by the behavioral response of the subject in psychophysics experiments.

An issue here is the information could be lost in the later stage of the visual pathway or anywhere during the complicated computation of consciousness and dicision making. We can reasonably rule this out for two reasons. (1) The psychophysics task would be low-level image comparison, so the complexity of the later stages should not affect the behavioral response very much. (2) The model is built based on the physiological characteristics of the early visual system so it is unlikely the later stages also throw away the information in the same structural manner predicted by the model. Therefore by doing psychophysics experiment, we can verify what information is thrown away by the early visual system.

The confirmation of information loss does not verify the model completely, but is very informative. For linear systems, the lost information represents the nullspace of the system. This approach therefore verifies the partition of the nullspace and the complement of the nullspace. For nonlinear systems, this procedure confirms the implicit manifold structure within the input space. Each manifold corresponds to a set of inputs whose responses are the same. Their differences can be seen as lying in the generalized "nullspace" of the system given the specific input. These are strong evidence that the model closely approximates the real system given the enormous number of possibilities for a high-dimensional system.

We can confirm such nullspaces and manifolds of the model by randomly drawing pairs of points in these spaces and check that they are metamers. To test the model more strictly, we would like to randomly draw points that lie on the boundary of theses spaces, the most "dangerous" points, where the model is the most "vulnerable" to be falsified.

2.4 Psychophysics Paradigm

The ABX task is used in this study. The subject is asked to maintain his/her focus on the center of the screen. In a single session, the subject views a sequence of two images where one of them is the original image(a regular gray-color photo) and the other is a hypothesized metamer of it. Then the subject views the original image again and indicate which one of the first two images is identical to the thrid image.

In this study, only casual psychophysics experiments were done by the author, no recruited subject or behavioral data analysis was involved.

3 Physiology of the Early Visual System

3.1 Retina

The basic structure of retina are three layers of cells and two layers of synaptic interconnections between them. From outer to inner, there are photoreceptors, bipolar, horizontal and amacrine cells and ganglion cells [7]. Optic nerve is the pathway that transmit visual information from retina to other area of the brain.

After an initial blurring caused by the lens, the photons are received by the photoreceptors which lead to hyperpolarization of the cell. The signal is transmitted and processed along different pathways in the two plexiform layers and then reach the ganglion cells. There are different types of ganglion cells but for simplisity we do not distinguish them in this study. The computation before and of the ganglion cells results in the formation their *receptive fields*, which usually have a center-surround structure.

The receptive field of the ganglion cell can be seen as a vector used to predict the response of the neuron for any input light pattern. The response is computed by taking the dot product of the receptive field and the input image. If the receptive fields of ganglion cells are all the same and orderly

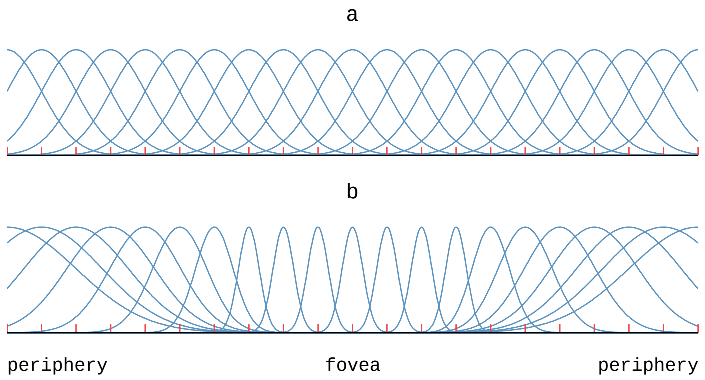

Figure 1: Illustrition of Gaussian-shape receptive field of ganglion cells layed out together. (a) Uniform size. Works like a convolution and filtering. (b) Varying size. If the spacing of two input samples is about the same size of the receptive field of ganglion cell(like the onse at fovea), the receptive field can be seen as a discrete delta function.

arranged, the ensamble response of a population of ganglion cells can be seen as the input image filtered by that receptive field. A Gaussian-shape receptive field would behave like a low-pass filter(Figure 1a). A receptive field having the form of a single spike(a delta function) would completely preserve the input because its Fourier transform is a constant function of value 1.

3.2 Eccentricity-Dependency of Receptive Fields

The property of ganglion cells we would like to exploit here is the eccentricity-dependency of the sizes of receptive fields. Physiological studies have shown that the sizes of receptive fields of ganglion cells(also V1, V2 and later stages) grows proportionally with its distance from fovea.

This proportional size-to-eccentricity relationship is described by a hinge function with a slope factor. Near fovea the size of receptive field is constant 1 so the information at fovea is(almost) perfectly preserved. At periphery the receptive field grows bigger and may produce a blurring effect on the input and throws high-frequency information away. In the following sections

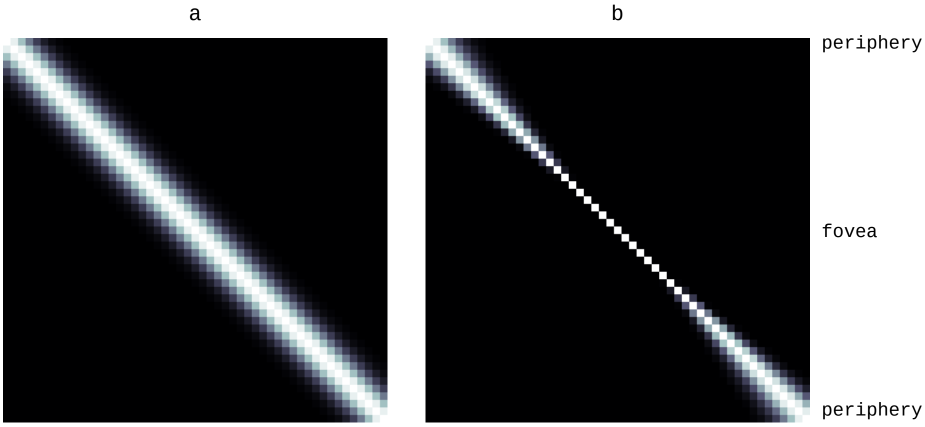

Figure 2: Illustrition of transform matrix consists of Gaussian-shape receptive fields. (a) Uniform size. Works like a convolution and filtering. (b) Varying size. At fovea the receptive fields are delta functions.

I carefully describe the consequences of this property and how the model is biult based on it.

4 Linear Models

4.1 Shift-Invariant Systems

If the ganglion cells form a uniform, densely packed mosaic and their receptive fields are the same, the ensamble response of this population to an image input could be described as a linear shift-invariant system. Without loss of generality, we can conceive such a system for 1-D signal input and generalize to 2-D image input. The system is described by a circulant matrix(except at boundary) with shifted versions of the same row, which is the receptive field of the ganglion cell(Figure 2a).

The eigenvectors of such a system are complex exponentials(essentially sinusoidal functions) and the behavior of the system is characterized by the frequency response of the receptive field(its discrete Fourier transform).

If we simply assume the receptive field is a low-pass filter, the ensamble response of the ganglion cells is then the low-pass filtered version of the input signal. The nullspace of the system is spanned by the high frequency sinusoidal eigenvectors and the information that the system throws away is the high frequency details of the input signal.

For this simplest model, the metamers can be generated by adding noise in the high frequency band of the original image. The prediction is that if high frequency information is thrwon away by ganglion cells, thses noise should be invisible to human observer.

4.2 Shift-Variant Systems

However, the system of ganglion cells is not shift-invariant. The size of receptive field grows with eccentricity. In 1-D scenario, the rows of the transform matrix are not only shifted but also stretched versions of each other(Figure 2b). The receptive fields at fovea are basically delta functions while the receptive fields at periphery are low-pass filters of growing sizes. The more distant the ganglion cell is from the fovea, the bigger the size of its receptive field.

Since there is no simple description of the behavior of such a system as in the case of shift-invariant systems, the most straightforward way to see what the system does is to actually build the transform matrix and perform Singular Value Decomposition(SVD). Since it is still a linear system, the nullspace can be read out by finding the basis vectors in the input space with corresponding singular values that are zeros.

The problem of this approch is that the transform matrix is too big. For reasonable images of resolution 1000 by 1000, the dimension of the matrix is 1 million by 1 million. For 8 bytes double datatype, the matrix takes up 8 TB storage space, which cannot fit in most computing devices. So we have to find alternative ways to characterize the behavior of the system.

4.3 Warpping The Input Space

The system described above cannot be understood in frequency domain and analyzed using Fourier transform because it is not a convolutional operation. The shape of the filter changes as it “slides” through the input signal. Fortunately, not every aspect of the shape changes, only the sizes. For continuous signals, the result of convolution at each location point is the integral of the signal multiplied by the shifted filter. Changing the size of the filter is equivalent to changing the size of the signal. Thus, even though it is the filter that is stretched, we can instead view it as the signal beging stretched. The mapping between the original input space and the warped space can be described by a warping function, derived as follows.

The receptive field size $s(x)$ is defined relative to the input sample spacing. Eccentricity x is defined as number of pixels from the center of the image, $x = 0$ corresponds to the center of the image. Denote the minimal receptive field size as $s_0 = 1$, the size-to-eccentricity ratio(the slope) as p. The eccentricity range within which the receptive field size is 1 is denoted $x_0 = s_0/p$. The receptive field size eccentricity-dependency function is then defined as

$$s(x) = \max(px, s_0)$$

Eccentricity in the conceived warped space is denoted as y. Eccentricity range within which the receptive field size is 1 in the warped space is $y_0 = x_0/s_0 = 1/p$. Since moving dx in the original input space corresponds to moving $dy = \frac{1}{px}dx$ in the warped space, we have

$$\begin{aligned} dy &= \frac{1}{px}dx \\ y &= \frac{1}{p}\ln x + C \end{aligned}$$

when $x \geq x_0$. Since $y = y_0$ when $x = x_0$,

$$\begin{aligned} y_0 &= \frac{1}{p} \ln x_0 + C \\ C &= \frac{1}{p} \ln \frac{p}{s_0} + \frac{1}{p} \end{aligned}$$

Thus the warping function is

$$y(x) = \begin{cases} \frac{1}{p}(\ln \frac{x}{x_0} + 1) & \text{if } x \geq x_0 \\ x/r_0 & \text{if } \leq x_0 \end{cases}$$

Figure 3 illustrates the eccentricity-dependency function and the warping function. The warped receptive fields all have the same size in the warped space. The spatial-variant operation becomes convolution. Fourier analysis is applicable in the warped space.

Although the idea of simply warping the input space sounds appealing, it is not feasible for discrete signals because there is simple method to reconstruct the continuous input signal and resample it non-uniformly without losing information.

Another idea is that since the sinusoidal basis in the warped space corresponds to warped sinusoids in the original space, maybe we should build a basis consising of a set of warped sinusoids whose frequency grow with eccentricity. However, it turns out that the discrete version of these warped sinusoids are not orthogonal and there is no straightforward way to construct the complemental set of basis for the high frequencies at periphery.

4.4 Spatial-Variant Filtering

The method we end up using in this study is based on the thoughts about warping the space but only manipulate the signal in the original input space without actually performing the warping. The "ground truth" of what we want to achieve is described in the warped space by the frequency response

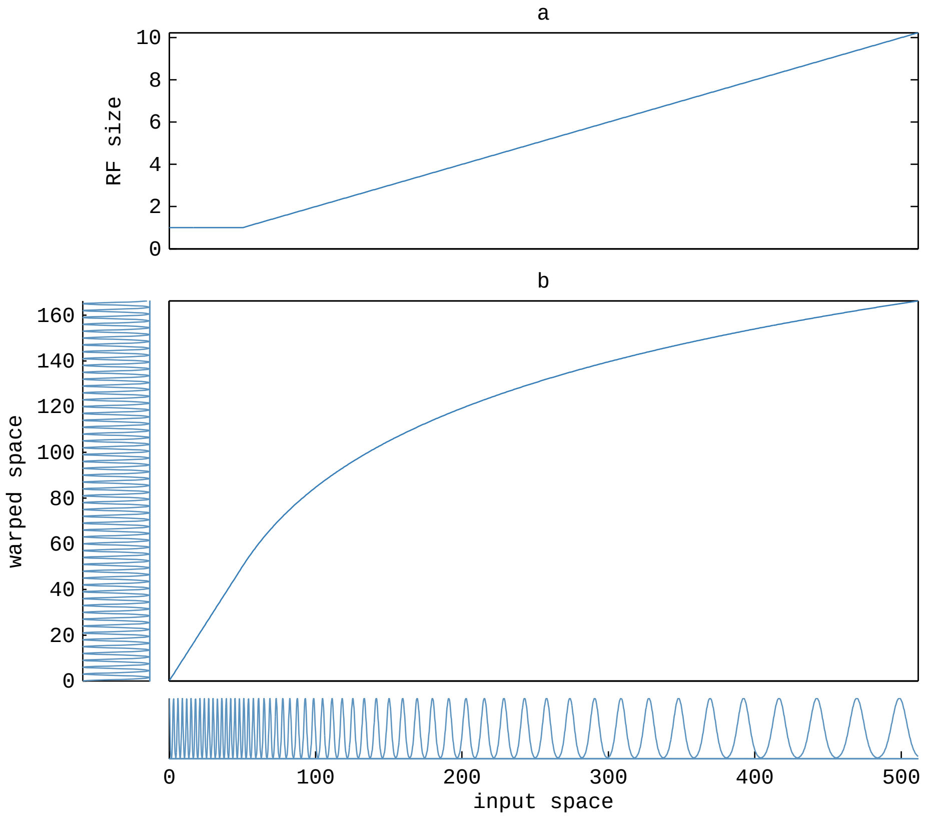

Figure 3: (a) Eccentricity-dependency function. (b) Warping function. The size of fovea is chosen so that the warping function is smooth. The warped receptive fields all have the same size. Notice not all receptive fields are drawn here for clearity.

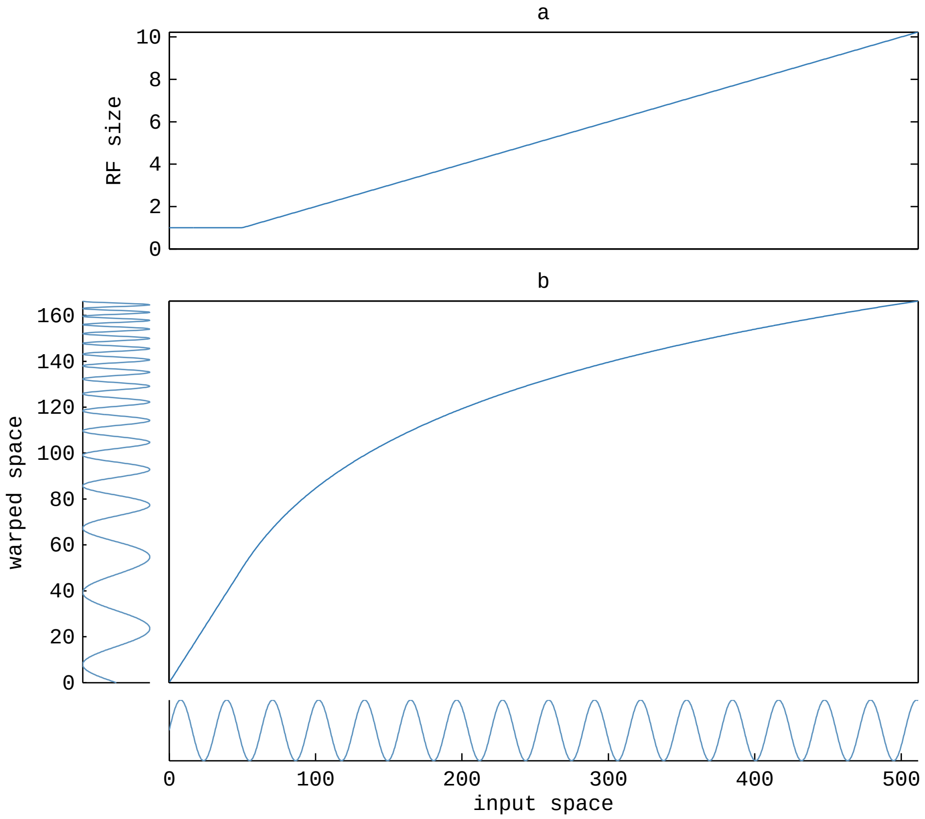

Figure 4: (a) Eccentricity-dependency function. (b) Warping function. The size of fovea is chosen so that the warping function is smooth. A warped sinusoidal basis in the warped space has higher frequency at periphery.

of the filter. We noticed that a sinusoidal basis function of some frequency in the original input space corresponds to a warped sinusoids in the warped space with higher frequency at periphery(Figure 4). For a given sinsoidal basis, every location in the original space corresponds to a different frequency in the warped domain. Therefore we should modify this frequency differently at different location according to its corresponding frequency in the warped space.

Specifically, the procedure of spatial-variant filtering is

1. Divide the frequency domain the of input signal into 20 octave sub-

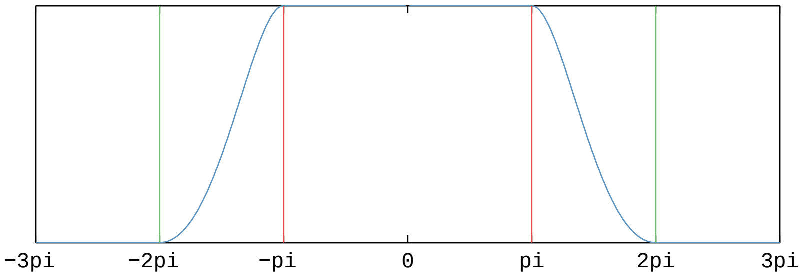

Figure 5: Frequency response of the receptive field in the warped space.

bands. No subsampling is needed for simplicity because the Nyquist sampling rate varies with location.

2. For each subband, multiply the whole signal by a spatial mask function whose value at each location is determined by how much the center frequency of this subband should be modulated in the warped space.

3. Add the modulated subband back together to get the filtered signal.

Figure 5 shows the frequency response of the filter we used in this study in the conceived warped space. The green lines indicates the range of $[-\pi, \pi]$ and the red lines indicates $[-2\pi, 2\pi]$. Since all the information is preserved at fovea, the frequency response is constant 1 within $[-\pi, \pi]$. The response goes down to zero as the frequency goes to 2π.

We can see what the spatial-variant filtering does in two ways.

From the spatial perspective, at fovea, the whole set of sinusoidal basis in the original space look the same with their counterparts in the warped space because the warping function is straight line with slope$= 1$ near fovea. Thus the signal at fovea is not modified because they are not modified in the warped space within frequency range $[-\pi, \pi]$. At periphery, however, the set of sinusoidal basis are warped to be higher frequency sinusoids in the warped space and have a frequency response smaller than 1 (even 0 for very high frequencies). Thus these frequencies at periphery would be scaled down.

From the frequential perspective, the low frequency sinusoidal basis in the original space is not modified because even the peripheral part of it does not exceed π in the warped space. For high frequency sinusoidal basis, the foveal part stays unmodified but the perpheral part is scaled down because it is warped into a frequency higher than π in the warped space.

Figure 6 demonstrates the result of spatial-variant filtering on 1-D white noise. The lowest frequency band is defined by the range of frequencies that remain unmodified. The rest of the frequencies are divided into 20 subbands so that the center frequency of the each band is bigger than the previous subband by a constant factor. The 20 subbands and the lowest frequency band completely cover the whole range from $-\pi$ to π. The spatial maks function for each subband is plotted. The most "inner" mask function corresponds to the subband of highest frequencies.

Figure 6(c) shows the white noise as the original input signal and Figure 6(d) shows the spatial-variant filtered signal whose high frequencies at periphery is filtered out.

Figure 7 shows the spatial-frequential modification of the filtering operation. The low frequencies (rows at center) are preserved. The high frequencies (rows at top and bottom) are scaled down at periphery. At fovea(columns at center), all frequencies are preserved. At periphery(columns at left and right), high frequencies are scaled down.

Figure 8 shows at different locations how well the spatial-variant filtering(blue line) approximates the desired filtering at that location(red dash line). The approximation has a lumpy shape due to the limited number of subbands but is not seriously undermining the effectiveness of the operation.

To verify whether the spatial-variant filtering approximates the effect of the transform matrix, we constructed the 1-dimensional transform matrix. The filter on each row is computed by inverse Fourier transforming the frequency response defined above and stretched accordingly. We performed

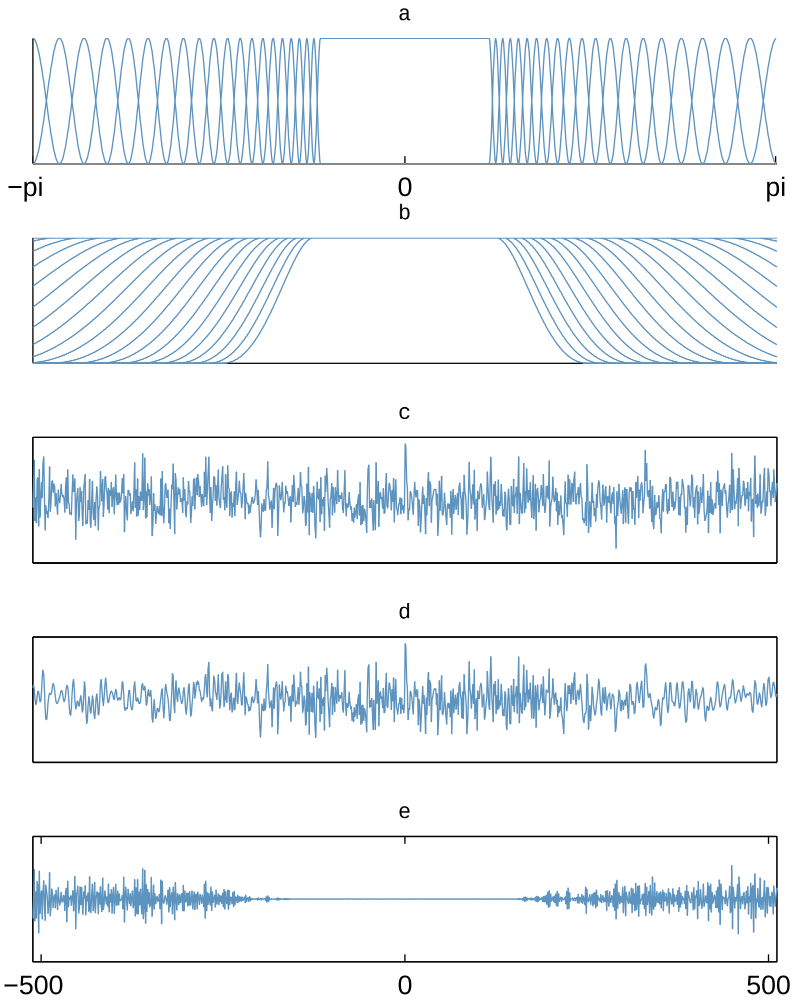

Figure 6: (a) Partition of frequency domain. (b) Spatial mask functions for each band. (c) A white noise signal as input signal. (d) Spatial-variant filtered signal. (e) Difference between the input signal and the filtered signal.

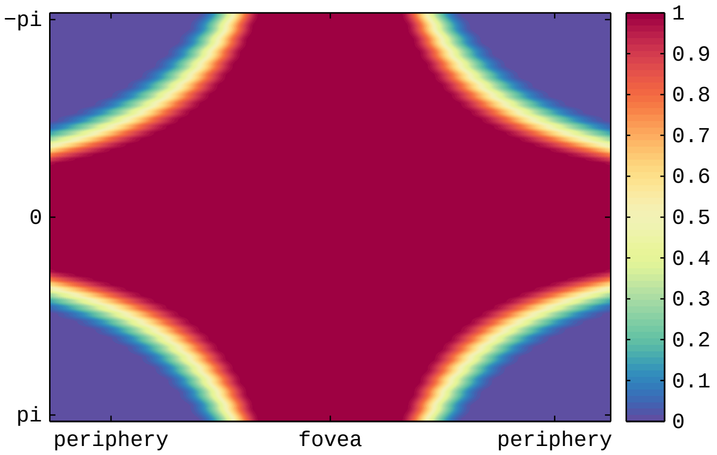

Figure 7: Illustrition of the effect of spatial-variant filtering operation in spatial-freuency domain.

SVD on it

$$svd(M) = USV^T$$

and projected the filtered signal onto the basis functions(columns of V) we get from SVD. The directions of $v's$ with small corresponding singular values in S are the components that are actually scaled down or thrown away if we apply the matrix to the input signal. If the filtering operation and the transform matrix is equivalent, the result signal should contain little contents in these thrown-away directions.

Figure 9(a) shows the actual columns of V of the transform matrxi for 1-dimensional signal. Figure 9(b) shows the corresponding singular values for each column of V. The basis functions look messy but qualitatively what we would expect. The directions with constant singular values are near the fovea and directions with small singualr values have non-zero values at periphery(see very carefully). Strangely, the directions with singular values bigger than 1 also have contents in periphery but it is more or less low frequencies. It is not well understood why there are singular values that are bigger than 1 since the frequency response of the filters are at most 1 in frequency domain.

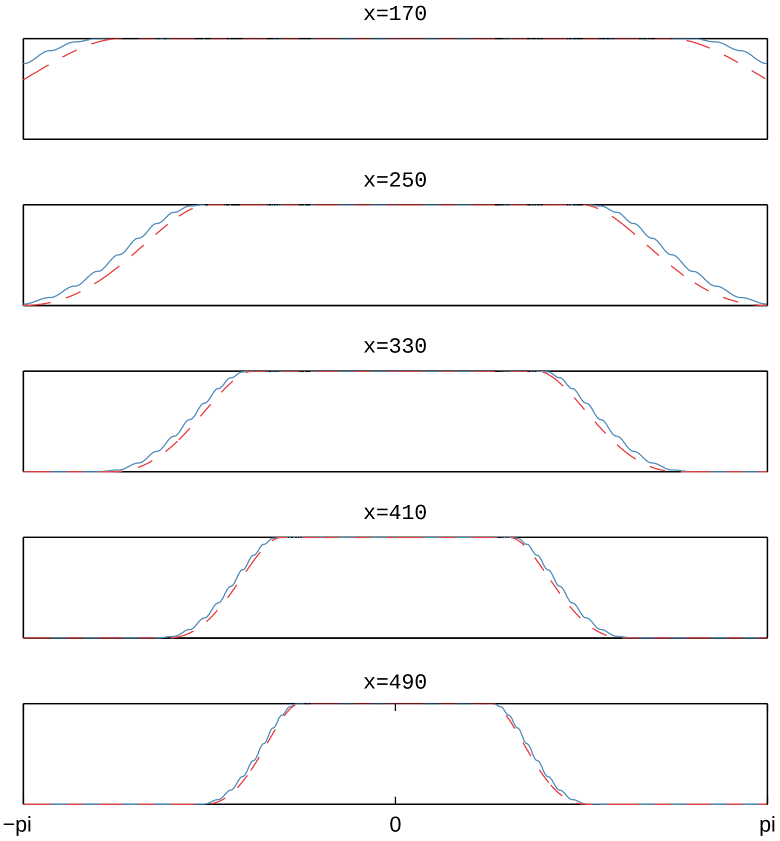

Figure 8: Approximation of the spatial-variant filtering at different location.

The projection of spatial-variant filtered signal and its residual onto the columns of V is shown in Figure 9(c) and (d). The red lines in every plot indicate the threshold for “small” singular values which is 10% of the maximal singular value. The columns of V with singular values smaller than the threshold are considered the basis for “nullspace” of the matrix while the columns of V with singular values bigger than the threshold are considered the basis for the complement of the “nullspace”. The distribution of the energy of the filtered signal and the residual on the “nullspace” and the complement of the “nullspace” is listed below. We can see the spatial-variant filtering successfully separates the signal into a part in the “nullspace” and another part in the complement of the “nullspace” and thus well approximates what the transform matrix actually does.

	“nullspace”	complement of “nullspace”
filtered signal	0.3%	99.7%
residual	87.6%	12.4%

4.5 Internal Noise

The linear model discussed above is problematic in three aspects. (1) The system actually does not have a nullspace if the receptive field we use is Gaussian or any filter whose Fourier transform is non-zero in all frequencies. The singular values of the system could be very small, but never zero. No two inputs would result in the same response and metamers are impossible. (2) This model says foveal vision is perfect because the receptive field is basically delta function. However in thereal early visual system, even the fovea does not preserve all the information. It is always possible to add a little bit of noise at fovea without being noticed by the subject. (3) This “hard” notion of metamers would result in difficulty in explaining the psychophysics data because the subject will only respond to the same pair

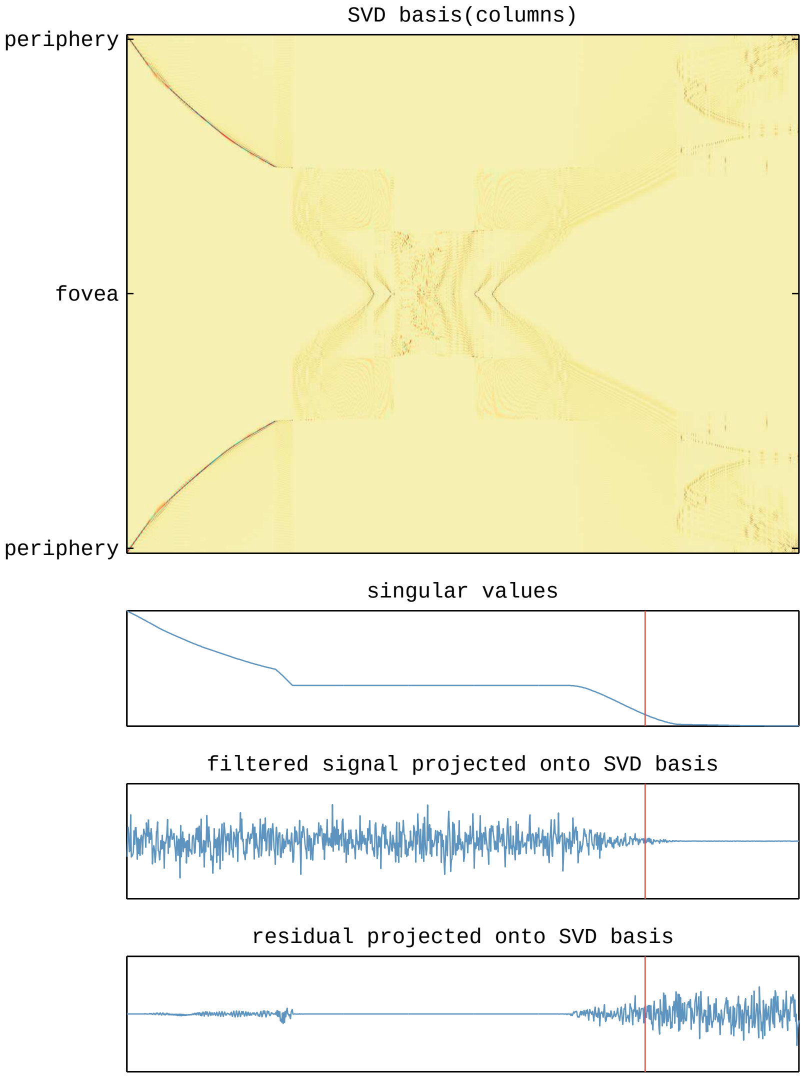

Figure 9: SVD analysis of the transform matrix. The spatial-variant filtered signal lies mostly in the directions with substantial singular values. The operation well approximates the behavior of the transform matrix.

of hypothesized metamers consistently with a certain probability depending on how metameric they are.

To solve these problems we take into account the fact that the internal response of the subject's brain is noisy. Specifically, noise is added to the output of the linear system. Two inputs are considered metamers if their responses are “close” enough and invisible noise at fovea is made possible in the generated metamers because of this generalized notion of nullspace.

From signal detection theory [4], the distributions of internal response of two images with additive Gaussian noise are two overlapping Gaussian distributions if their true response is close to each other. The variance of the distribution represents the magnitude of the noise. During the experiment, the subject views two images in sequence and internally draws two samples, one from each of these two distributions. When the subject views the third image, he/her draws another sample from one of these two distributions and must decide which distribution this sample comes from. The decision is made by determining which one of the first two samples is closer to the third sample.

If the two images produce almost the same true response, it is very hard for the subject to tell the difference between them and the probability that he/she gives the right answer is around 50%(Figure 10a). If the true responses of the two images are far away from each other, the subject would almost always give the right answer because of the separation of the distributions of the noisy internal responses. If the subject performs slightly better than chance but not quite good, the distributions are heavily overlapped(Figure 10b). The percentage correct of the subject's performance depends on the distance between the mean of the two distributions and the variance of the distributions.

We can set a threshold on the percentage correct of the subject's performance on this task and use the corresponding distance value between the

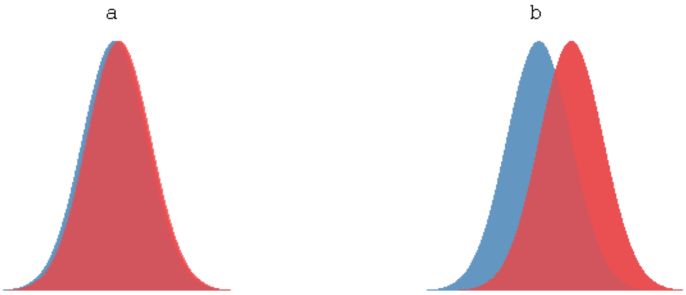

Figure 10: 1-dimensional signal detection theory.

mean to define metamerism, i.e., two images are metamers if their responses are within such distance. A typical value to use is simply the standard deviation of the distribution of the noise and the corresponding percentage correct is about 58%(Figure 10b) for 1-dimensional signal.

In general, for N-dimensional signals, the distance between two responses is the L2-norm of the difference of the responses. For a given percentage correct, a different value of difference of responses of two metamers would be used rather than simply the standard deviation of the noise. For example, for 100-dimensional signals with white noise of standard deviation 1, a distance of 3 would give percent correct 60%. This distance is estimated through experiments.

5 Generation of Metamers

5.1 Inverting the Difference of Responses

Figure 11 shows the linear model that was conceptually used in this study. Specifically, denote the transform matrix as M, the original image input as x_0, and the metamer we would like to synthesize as $x_0 + \Delta x$ and their responses to be r_0 and $r_0 + \Delta r$ respectively. Since M is just a linear transformation,

$$\begin{aligned}\Delta r &= (r_0 + \Delta r) - r_0 \\ &= M(x_0 + \Delta x) - Mx_0 \\ &= (Mx_0 + M\Delta x) - Mx_0 \\ &= M\Delta x\end{aligned}$$

Conceptually, Δx can be found by

$$\Delta x = M^{-1}\Delta r$$

5.2 Constraints on Difference of Responses

According to the basic idea of the model, the first requirement on Δr for $x_0 + \Delta x$ to be a metamer of x_0 is

$$||\Delta r|| \leq d$$

where d is the distance threshold between the responses of the two images for some given percent correct. Thus, any Δx whose response has a norm smaller than d can be added directly to the original image to get a metamer for it. On the other hand, metamers can be generated by taking any N-dimensional vector whose norm is smaller than d, inverse transforming it back into the input space and adding it to the original image.

Furthermore, in order to expect the same psychophysics performance for all the metamers we generate,

$$||\Delta r|| = d$$

is required. This makes sure all the metamers are barely visible and fits the goal of testing the "most dangerous" points in the nullspace of the model.

However, the norm constraint should not be the only constraint on $r\Delta r$. Its content also needs to be uniformly distributed spatially. If Δr has no

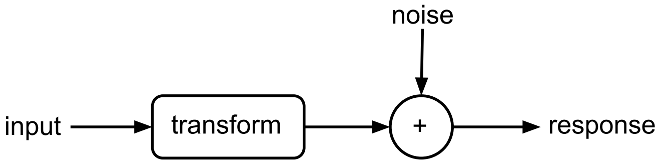

Figure 11: The linear model.

content somewhere, it must end up having higher magnitude content somewhere else to keep the norm unchanged. But this excessive magnitude would result in a strong distortion in the original space that would pop out perceptually and make the distortion no longer just barely visible as a whole. The reason is that perceptual distortion is inherently local and almost only depends on the response of a local group of ganglion cells.

Finally, the response Δr should also be generated by a random process because we would like to randomly test different points in the nullspace of the model by generating multiple metamers for a single image.

A good and simple choice that meet all these three requirements is Gaussian white noise. It can be easily generated, its norm is reasonable stable (because it follows χ^2 distribution) and its content is uniformly distributed spatially.

5.3 Inverse Spatial-Variant Filtering

Once Δr is determined, inverse transform can be performed to generate distortion in the input space. Since the forward transform is well approximated by the spatial-variant filtering operation, the inverse transform is just the inverse operation:

1. Divide the frequency domain of the signal into subbands
2. Spatially mask each subband by how much it should be inversely scaled up in the warped space

3. Add the processed subband together to get the distortion in the input domain

The problem here is that the frequency response of the receptive field is defined so that it scales the high frequency down to zero, so the inverse transform should scale it up to infinity, which does not make sense because there is no way to display that image. Furthermore, such absurd prediction comes from the over-simplicity of linear models that totally ignore the adaptability and nonlinearity of the ganglion cells.

For the linear model, an *ad hoc* solution is to contrain the inverse spatial-variant filtering operation so that the scale factor cannot exceed $1/\epsilon$ and the generated metamers can be normally displayed. This can also be seen as redefining the frequency response of the receptive field so that the high frequencies are not thrown away but scaled down by at most ϵ.

For example metamers generated using this method, see Figure 15, Figure 16, Figure 17, Figure 18.

6 Measurements

6.1 Parameters Estimation

There are three parameters for the linear model, σ, p and ϵ. σ represents the magnitude of the internal noise and is estimated through casual experiments by adding white noise to a uniform gray background and see whether it is visible.

p represents the slope of the eccentricity-dependency function and ϵ represents the minimal scaling factor of the frequency response of the receptive field. Multiple values for both parameters were tried and the pair that generated the biggest but barely visible distortion were chosen.

Figure 12 shows conceptually the expected result for different choices of (p, ϵ) pairs. The blue lines are frequency response of the receptive field. The

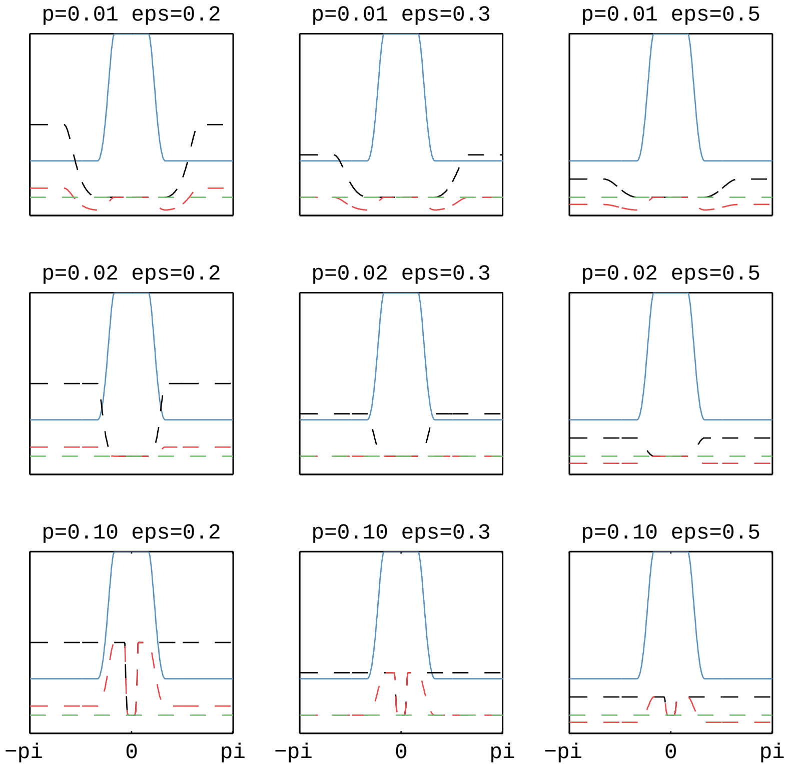

Figure 12: Expected result for different choices of (p, ϵ) pair.

black dashed lines are spectra of distrotions that are generated using the corresponding parameters printed above each plot. The green lines are the defined threshold responses for barely visible distortion. The red lines are the result of filtering the noise with the actual receptive field. Obviously, either a too-big p or a too-small ϵ could cause the response to be above the threshold and the distortion becomes visible. But a too-small p or a too-big ϵ could cause the distortion not visible enough to reach the required psychophysics performances. Only when p and ϵ are both at the right value can the distortion be barely visible.

Casual psychophysics experiments were done by the author and the estimated p coincides with the physiological data of ganglion cells. [5]

6.2 Display Constraint

The synthesized metamers are designed for displays with square pixel lattice, but the field of view of human eye is measured by angles. Under the assumption that the display would not cover a big visual angle, the sampling rate of photoreceptors could be matched with the pixel density of the display by choosing the right viewing distance from the display.

For example, using the common sampling rate of L- and M-cones, 120 samples per degree, and an ordinary Apple 23" display of pixel density 98.4 pixel per inch, the viewing distance for the photoreceptors to match the pixels one-by-one is 1.77 meters. If one views at a closer distance, the resolution of the display is too low. The highest acuity of the eye is not exploited. If one views at a farther distance, the resolution of the display is too high, the finest detail of the image cannot possibily be grasped by the eye. The ability of the display is wasted. Therefore it is best to view at the critical distance where the sampling rate of photoreceptors matches the resolution of the display.

However, the problem of viewing at this distance is that the visual angle that the display covers is too small and the peripheral vision is not tested. The trade-off between measuring "more periphery" and "higher frequencies at fovea" is shown in Figure 13.

red dashed line pixel size(in visual angle) as a function of the size of display(in visual angle). The closer the viewing distance, the wider the display(in visual angle), the bigger the pixel size(in visual angle).

magnet line cones size(in visual angle), assumed constant everywhere, from physiological data

blue line ganglion cell size(in visual angle), eccentricity dependent, slope from phyiological data

above green line and between black lines range of display, ganglion cells below the green line are not fully used because the resolution of display is lower than the sampling rate of the foveal ganglion cell

7 Conclusions

Based on one important property of ganglion cells that the receptive fields sizes grow with eccentricity, we have built a simple linear model with additive Gaussian internal noise. In order to test the model using visual metamers, we developed a spatial-variant filtering operation that approximates the behavior of linear shift-variant systems. Such systems cannot be explained using the theory of linear shift-invariant system and Fourier analysis. The accuracy of the approximation was verified in 1-D case and the method was generalized to 2-D to generate metamers. The model has three parameters that were estimated by casual psychophysics experiments. The estimated slope of eccentricity-dependency function coincides with the physiological data of ganglion cells, suggesting that the models successfully captures the main property of the ganglion cell population.

The *ad hoc* solution to the problem of non-invertability of the filtering operation needs to be revised in the future by taking into account the adaptability and non-linearity of the ganglion cells. Luminance and contrast gain control was not discussed here but are very important component for a robust model of the early visual system and needs to be carefully established before we can build cascaded models for the visual pathway.

Shortening viewing distance ends up testing more peripheral ganglion cells by giving up testing the acuity of foveal ganglion cells. Fortunately, choice of viewing distance does not affect how the metamers should be generated on the square lattice of the display because of the proportionality between the width of the display(in visual angle) and the pixel size(in visual angle, see Figure 14).

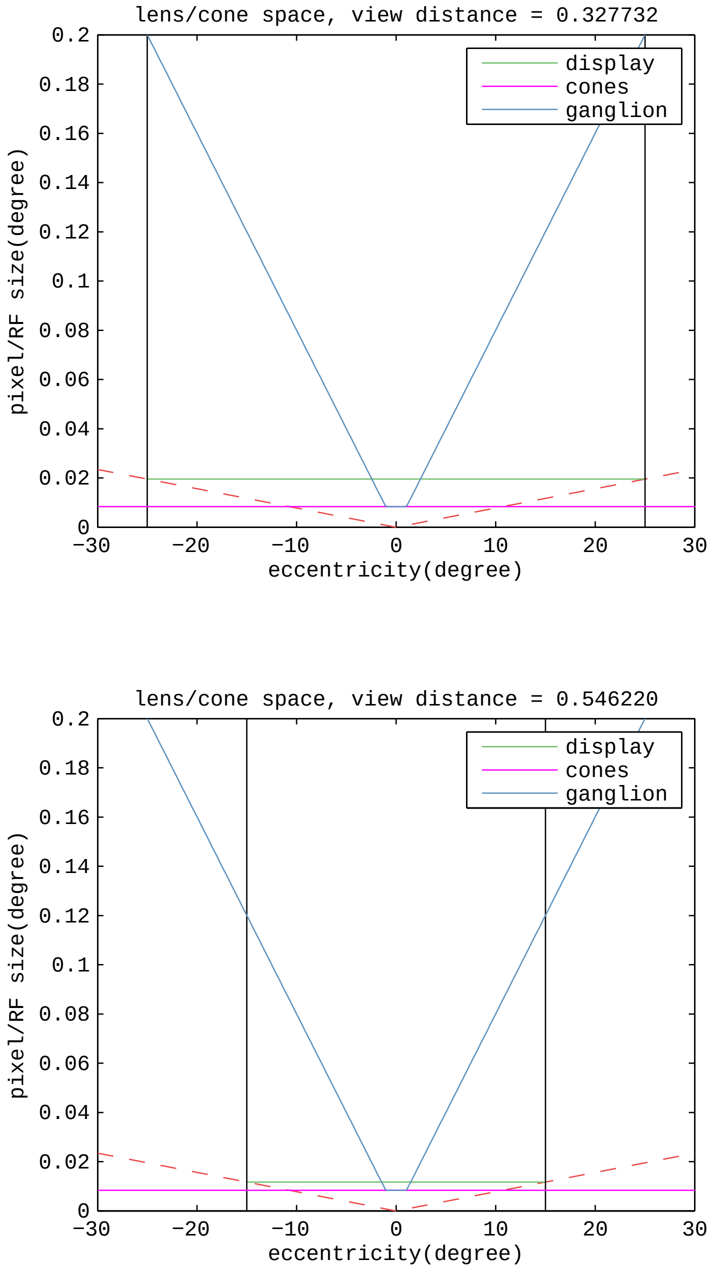

Figure 13: Relationship between viewing distance and the resolution of the display relative to the sampling rate of ganglion cells. Different viewing distances test different parts and abilities of the ganglion cell population.

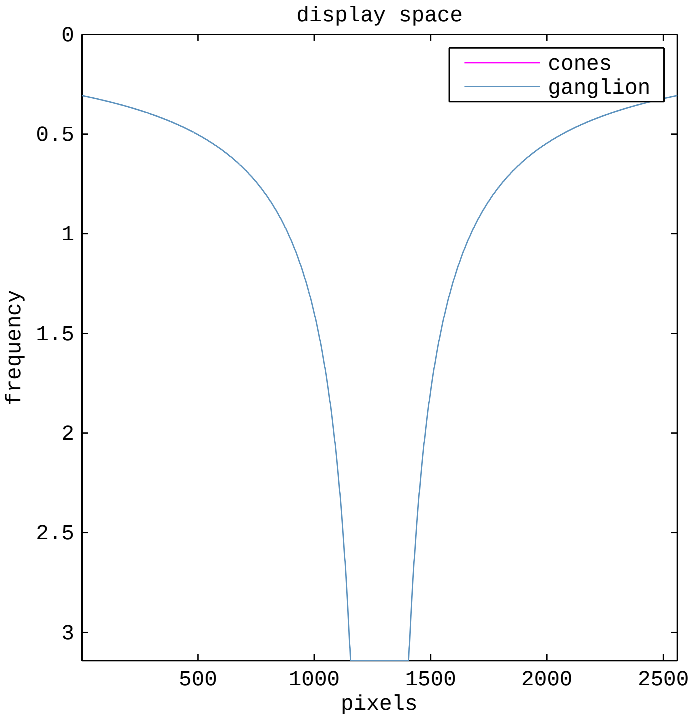

Figure 14: The area under the blue lines are contents that would be thrown away by the early visual system by prediction, and it is independent of the viewing distance. We should always generate metamers according to this configuration but different viewing distances would test different parts of the ganglion cell population.

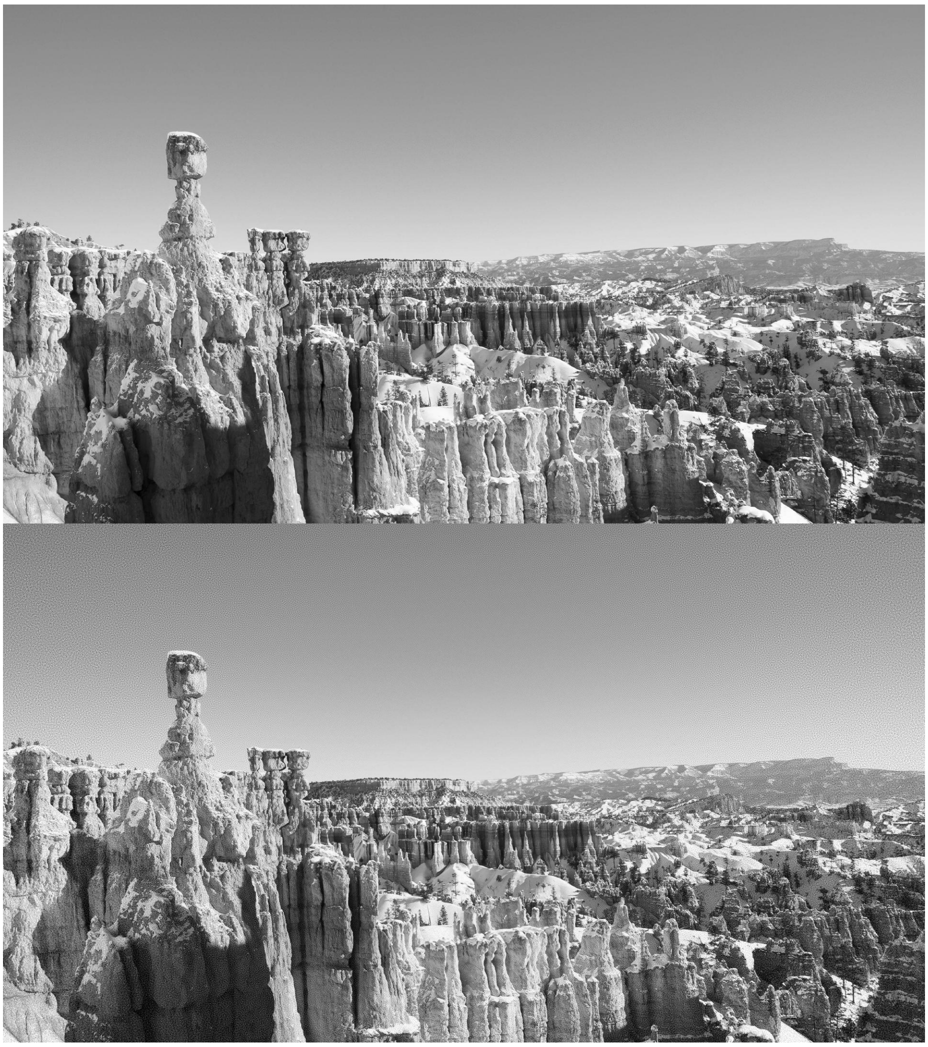

Figure 15: Top: original image; Bottom: metamer; Center is clean, periphery is filled with noise. MSE: 477. See Figure 16 for partial magnified.

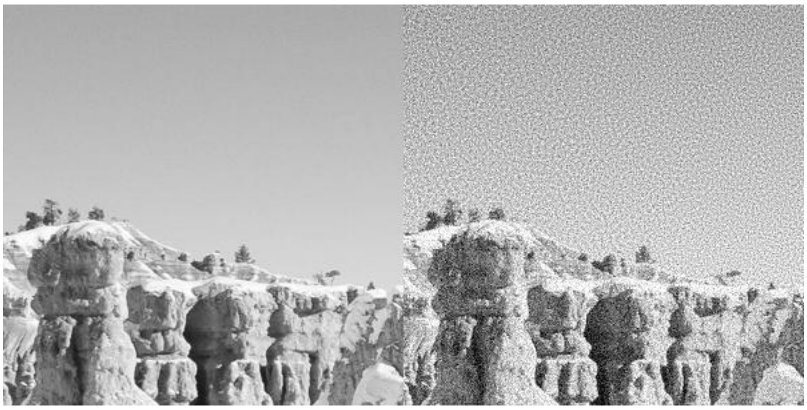

Figure 16: Partial magnified. Left: original image; Right: metamer

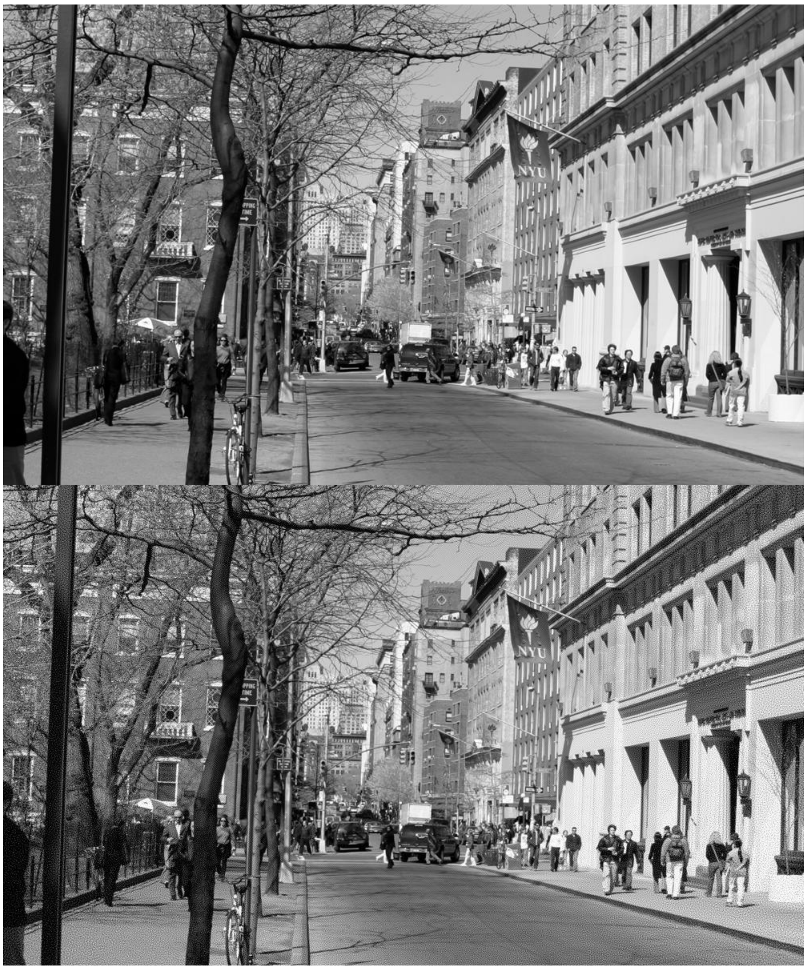

Figure 17: Top: original image; Bottom: metamer; Center is clean, periphery is filled with noise. MSE: 374. See Figure 18 for partial magnified.

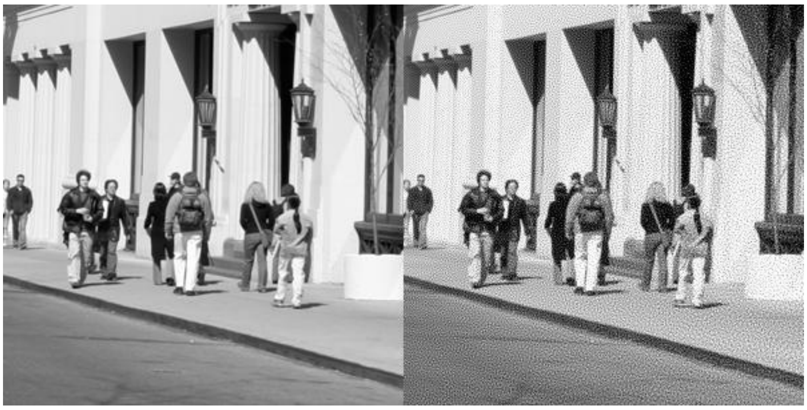

Figure 18: Partial magnified. Left: original image; Right: metamer

References

[1] Edward H Adelson and James R Bergen. The plenoptic function and the elements of early vision.

[2] Jeremy Freeman and Eero P Simoncelli. Metamers of the ventral stream. *Nat Neurosci*, 14(9):1195–1201, 09 2011.

[3] David H. Hubel. Exploration of the primary visual cortex, 1955-78. *Nature*, 299(5883):515–524, 10 1982.

[4] N.A. Macmillan and C.D. Creelman. *Detection Theory: A User's Guide.* Lawrence Erlbaum Associates, 2005.

[5] V H Perry, R Oehler, and A Cowey. Retinal ganglion cells that project to the dorsal lateral geniculate nucleus in the macaque monkey. *Neuroscience*, 12(4):1101–1123, Aug 1984.

[6] G. Strang. *Introduction to Linear Algebra.* Wellesley-Cambridge Press, 2003.

[7] Brian A Wandell. Foundations of vision. 1995.