

# Statistics of Natural Images and Models

Jinggang Huang  
Division of Applied Math  
Box F, Brown University  
Providence, RI02912  
jhuang@cfm.brown.edu

David Mumford  
Division of Applied Math  
Box F, Brown University  
Providence, RI02912  
David\_Mumford@brown.edu

## Abstract

*Large calibrated datasets of ‘random’ natural images have recently become available. These make possible precise and intensive statistical studies of the local nature of images. We report results ranging from the simplest single pixel intensity to joint distribution of 3 Haar wavelet responses. Some of these statistics shed light on old issues such as the near scale-invariance of image statistics and some are entirely new. We fit mathematical models to some of the statistics and explain others in terms of local image features.*

## 1 Introduction

There has been much attention recently to the statistics of natural images. For example, Ruderman [7] discusses the approximate scale invariance property of natural images and Field [4] linked the design of the biological vision system to the statistics of natural images. Zhu, Wu and Mumford [9] set up a general frame work for natural image modeling via exponential models. Simoncelli[1] uncovered significant dependencies of wavelet coefficients in natural image statistics. In most of these papers, simple statistics are calculated from which some properties are derived to prove some point. But little effort has been made to systematically investigate the exact statistics that underline natural images. Many of these papers base their calculation on a small set of images, casting doubt on how robust their results are. Also, because of the small sample sets, rare events (e.g. strong contrast edges) which are important visually may not show up frequently enough to stabilize the corresponding statistics. We tried to overcome these problems by using a very large calibrated image data base (about 4000  $1024 \times 1536$  images taken by digital camera), provided by J.H. van Hateren (for details, see [5]). Figure 1 shows some sample images from this data base. These images measure light in the world up to an unknown multiplicative constant in each of the image. We will only work on the log

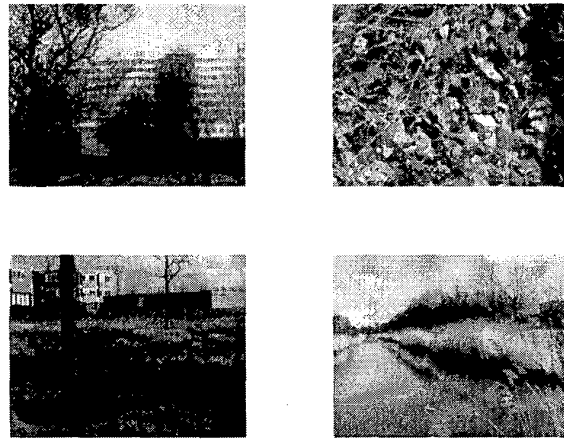


Figure 1: Four images from the data base

intensity, and use statistics which do not contain the constant (now an additive constant). We believe our work here can serve as a solid starting point for further image modeling and provide guidance in design of image processing and image compression systems.

We explain some symbols we will use in the paper: Assume  $X$  is a random variable on  $\mathbf{R}$ , we use  $\mu$  and  $\sigma^2$  to represent the mean and variance of  $X$ . We define:

$$\kappa = \frac{E(X - \mu)^4}{\sigma^4} \quad \mathcal{S} = \frac{E(X - \mu)^3}{\sigma^3}$$

where  $\kappa$  is the kurtosis,  $\mathcal{S}$  is the skewness. Assuming  $Y$  is another random variable on  $\mathbf{R}$ , we denote the differential entropy for  $X$  by  $\mathcal{H}(X)$ , and denote the mutual information between  $X$  and  $Y$  by  $\mathcal{I}(X, Y)$ , both in bits. We use differential entropy instead of discrete entropy, because the variables are real valued. For details, see [2]. All our pictures of probability distributions (or of normalized histograms) will be shown with

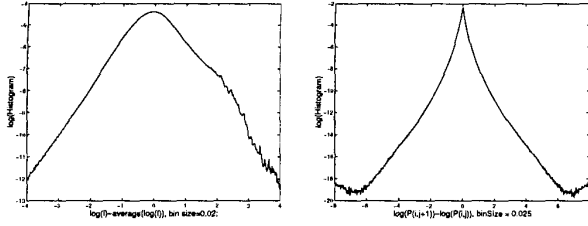


Figure 2: left: log histogram of  $\ln(I(i, j)) - \text{average}(\ln(I))$ , right: log histogram of  $\ln(I(i, j)) - \ln(I(i, j + 1))$

the vertical scale not probability but *log of probability*: this is very important as it shows the non-Gaussian nature of these probability distributions more clearly and shows especially the nature of the tails. We will regard each image  $I$  in the data set as a  $1024 \times 1536$  matrix, and  $I(i, j)$  represents the intensity at position  $(i, j)$ .

## 2 Single Pixel Statistics

The left image of Figure 2 shows  $\ln(\text{Histogram})$  of the random variable  $\ln(I(i, j)) - \text{average}(\ln(I))$ . Constants associated to this statistics are  $\mu = 0$ ,  $\sigma = 0.79$ ,  $\mathcal{S} = 0.22$ ,  $\kappa = 4.56$ ,  $\mathcal{H} = 1.66$ . From the log plot and the skewness  $\mathcal{S}$ , we can see that this statistic is not symmetric. One important reason is the presence of a portion of sky in many images, which is quite different from other parts of images, always with a high intensity value. Another interesting feature is the linear tail in the left half. Obviously, this statistic is non Gaussian, although the center part of the log plot does show a parabola shape. The kurtosis is bigger than 3 (the value for Gaussians) but not very large.

## 3 Derivative Statistics

We now look at the marginal distribution of horizontal derivatives, which in discrete case, is simply the difference between two adjacent pixels in a row, i.e.  $D = \ln(I(i, j)) - \ln(I(i, j + 1))$ . The right image of Figure 2 shows the  $\ln(\text{Histogram})$  of  $D$ , and here are some constants associated to it:  $\mu = 0$ ,  $\sigma = 0.260$ ,  $\mathcal{S} = 0.085$ ,  $\kappa = 17.43$ ,  $\mathcal{H} = -0.24$ . The value of  $\sigma$  is quite interesting: random horizontally adjacent pixels have roughly 25% more or less energy. The kurtosis is now very large *but* one should be aware that this statistic is very sensitive to outliers and so the number cannot be taken very seriously.

We wish to model this statistic: notice that the shape of the histogram has a distinct peak at 0, and

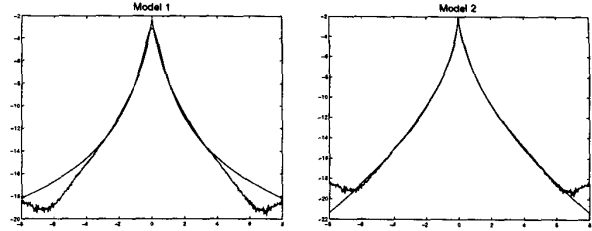


Figure 3: The two models with best fitting constants

a concave tail. Writing the density function for  $D$  as  $f(x)$ , we consider the following models:

$$\begin{aligned} \text{Model 1 } f(x) &= \frac{1}{Z} \cdot \frac{1}{(1 + x^2/s^2)^t} \\ \text{Model 2 } f(x) &= \frac{1}{Z} \cdot e^{-|x/s|^\alpha} \end{aligned}$$

$Z$  is fixed since the integral of  $f(x)$  is 1, so both models have 2 free parameters. *Model 1* is the  $t$ -distribution and *Model 2* is the generalized Laplace distribution, whose parameters  $s, \alpha$  are directly related to the variance and kurtosis by:

$$\sigma^2 = \frac{s^2 \Gamma(\frac{3}{\alpha})}{\Gamma(\frac{1}{\alpha})} \quad \text{and} \quad \kappa = \frac{\Gamma(\frac{1}{\alpha}) \Gamma(\frac{5}{\alpha})}{\Gamma^2(\frac{3}{\alpha})} \quad (3.1)$$

Figure 3 shows the best fit of both models to  $\ln(f(x))$ . We choose parameters such that the mean square error is minimal, obtaining  $t = 2.60$  for model 1 and  $\alpha = 0.55$  for model 2. It's obvious that model 2 is better in the tails. Note that in Model 2, the parameters of the model can be calculated (numerically) directly from the variance and kurtosis using (3.1), we find such calculated model is very close to the best fitting model.

## 4 Joint Distribution of Two Adjacent Pixels

Figure 4, left, shows the joint distribution of the intensities  $p_1$  and  $p_2$  at two horizontally adjacent pixels, where  $p_1 = \ln(I(i, j)) - \text{average}(\ln(I))$ ,  $p_2 = \ln(I(i, j+1)) - \text{average}(\ln(I))$ . The constants we found are:  $\mathcal{H}(p_1, p_2) = 1.51$ ,  $\mathcal{I}(p_1, p_2) = 1.80$ . Notice that the mutual information  $\mathcal{H}$  between adjacent pixels is a large number, indicating that adjacent pixels are highly correlated. On the other hand, we can see from the contour plot, that there is some symmetry along  $p_1 = p_2$ , and a rough symmetry along  $p_1 = -p_2$ , we may guess that the sum and the difference of two adjacent pixels are more likely independent. Figure 4,

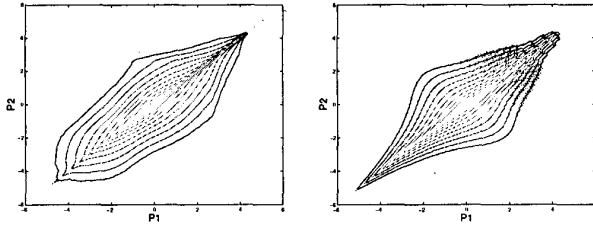


Figure 4: Left figure: Joint Histogram of  $p_1$  and  $p_2$ , Right figure: The product density function of  $p_1 - p_2$  and  $p_1 + p_2$

right, shows the product distribution of the marginals of  $p_1 + p_2$  and  $p_1 - p_2$ . Comparing the two contour plots, we can see that at the center part (where the density is much higher than other places) the product distribution and the original distribution are very similar, but the shape of the level curves away from  $(0, 0)$  becomes quite different. The mutual information between  $p_1 + p_2$  and  $p_1 - p_2$  is 0.0255. Compared to that of  $p_1$  and  $p_2$ , it's very small, indicating a rough independence between  $p_1 + p_2$  and  $p_1 - p_2$  from information theory point of view.

## 5 Joint Statistics in the Wavelet Domain

### 5.1 Definitions

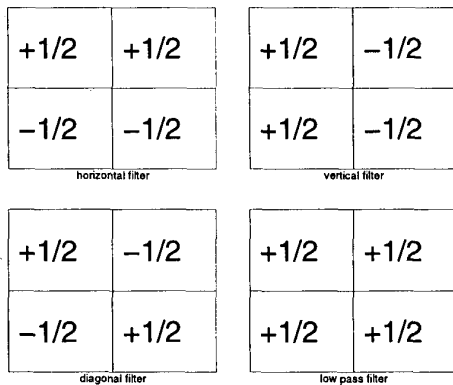


Figure 5: Haar Filters

The idea of applying wavelet pyramids to image processing has proven to be very successful. Here we use the simplest wavelets: Haar wavelets. Figure 5 shows the four basic Haar filters in 2D. Assuming we have a  $8 \times 8$  image, we can apply the horizontal filter

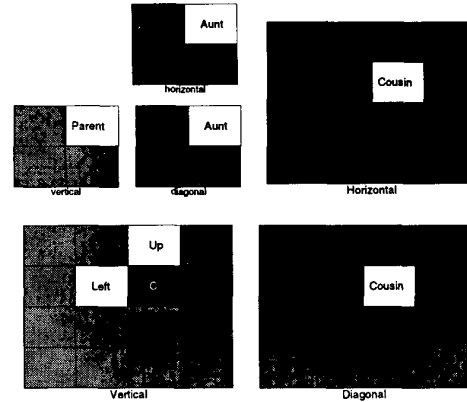


Figure 6: Relations Between Wavelet Coefficients

on non overlapping  $2 \times 2$  blocks of the image, to get a  $4 \times 4$  matrix of responses, called the first level horizontal sub band. Similarly, we have the first vertical and diagonal sub bands. Next, apply the low pass filter on the image, getting a  $4 \times 4$  image, and repeat the above procedure on this smaller image to get the second level horizontal, vertical and diagonal sub bands, all of them with dimension  $2 \times 2$ . This procedure can go on till we get a  $1 \times 1$  image. This way we get the wavelet pyramid of sub bands whose statistics we wish to study. Figure 6 shows the sub bands of the first two levels. In order to describe the relative positions of wavelet coefficients in sub bands, we borrow some of the definitions given in [1]: we call the coefficients at adjacent spatial locations in the same sub band *brothers*, (left, right, upper, lower brothers according to the relative positions), call the coefficients in same level, same position, but different orientations *cousins*. And call the relationship of coarse-scale coefficients and finer-scale coefficients *parent* and *child*. Figure 6 shows a coefficient  $C$  in the vertical sub band, and its relatives.

### 5.2 2D Joint Statistics for Haar Wavelet

In this subsection, we look at the joint statistics of different coefficient pairs. Where we have used the Haar wavelet for simplicity, the general model we find here also fits to other wavelet bases. Figure 7 shows the mesh plot of the joint statistics of the pair horizontal component( $hc$ ), and its vertical cousin( $vc$ ). We found that: for any angle  $0 \leq \theta \leq \pi$ , the cross section along  $hc = \tan(\theta)vc$  is similar to the derivative statistics we got in section 3, and the 'generalized Laplacian'

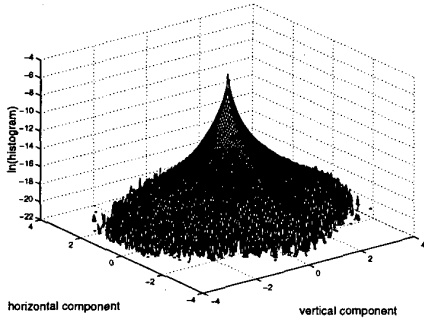


Figure 7: Mesh Plot of the  $\ln(\text{Joint Histogram})$  of Horizontal Component and its Vertical Cousin

model can be easily used to fit the cross sections here. This suggests, as a first step, fitting a model

$$f(x, y) = e^{C_1(\theta) + (\frac{r}{\sigma(\theta)})^\alpha(\theta)} \quad (5.1)$$

The fittings are pretty good, Figure 9 shows the *worst* fitting case (with largest mean square error) among all cross sections, even which fits very well.

Using this method, we successfully fit other pairs of wavelet coefficients. Figure 8 shows the contour plot of the joint density functions of different pairs. We see that 2D joint density functions have variable and complex structures. But for a specific pair, we may come up with some specific simple model. For example, notice that the level curves of the horizontal component( $hc$ ) and its vertical cousin( $hv$ ) pair show a shape similar to the curve defined by:  $|x| + |y| = 1$ , hence we may model the density function as:  $f(ch, cv) = e^{C_1 + C_2(|ch| + |cv|)^\alpha}$  for some constants  $C_1, C_2, \alpha$ . All these shapes reflect typical local features in the images, such as the fact that horizontal and vertical edges are more common than diagonal ones. We leave to a later paper the detailed exploration of these features.

### 5.3 3D joint statistics for haar wavelet coefficients

Here we just present the joint statistics for the triple: horizontal component, its vertical cousin and diagonal cousin. Figure 10 shows a level surface(counterpart of a single level curve in Figure 8) viewed from different angles. We can see more structures here. There are four corners on the ‘horizontal’ and ‘vertical’ axes, which imply that the probability density heavily concentrates on these two axis.

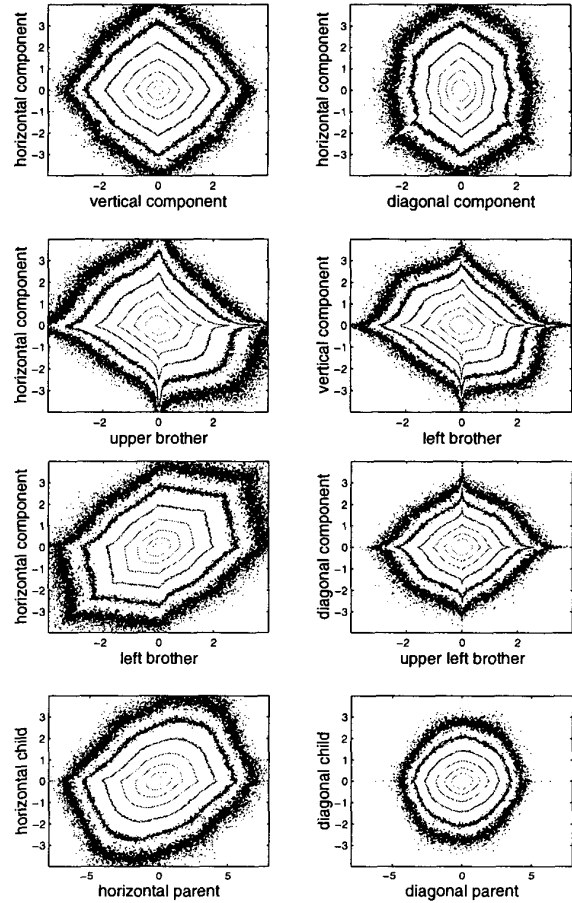


Figure 8: Contour Plot of the  $\log(\text{histogram})$  of several wavelet coefficient pairs

### 5.4 An interesting phenomena and our explanation

In [1], the authors observed a interesting phenomena in the joint histograms from images. Let  $hp$  represents horizontal coefficient at parent level, and  $hc$  represents horizontal coefficient at child level, they observed that the conditional histogram  $H(\ln|hc| | \ln|hp|)$  has a shape shown in Figure 11 (we will explain how we get this figure next). The filters they use are separable Quadrature Mirror Filters(QMF). We repeated their calculation of 2D histograms using QMF and found that the shape of level curves of joint histogram  $H(hc, hp)$  is very close to a circle, similar to that of the diagonal parent & diagonal child pair, shown in figure 8. So our general model (5.1) reduces

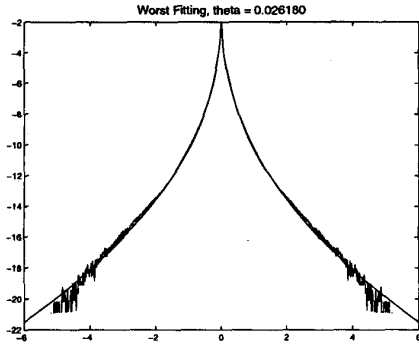


Figure 9: The cross section of Figure 6, that has the worst fitting curve

to:

$$f(hc, hp) = C_1 e^{-C_2 r^\alpha} \quad (5.2)$$

where  $r = \sqrt{hc^2 + hp^2}$ , and  $C_1, C_2, \alpha$  are just constants, independent of  $\theta$ . We found that this model fits fairly well to the data for  $C_2 = 1$  and  $\alpha = 0.5$ . From the phenomena they observed, the authors concluded in [1] that, the conditional expectation,  $E(|hc| || hp)$  is approximately proportional to  $P$ , and derive a very simple linear predictor from it. For the left part, they suspected these low-amplitude coefficients are dominated by quantization and other noise sources.

This phenomena can be actually explained using (5.2). Since  $f(hc, hp)$  is symmetric, the density function for  $(|hc|, |hp|)$  should have the same expression, only that  $C_1$  is different. So we just assume  $hc > 0, hp > 0$ . Let  $x = \ln(hc)$  and  $y = \ln(hp)$ , then the joint density function for  $x, y$  is:

$$g(x, y) = C_1 e^x e^y e^{-C_2 r^\alpha}$$

where  $r = \sqrt{e^{2x} + e^{2y}}$ . The conditional density function will be:

$$g(x|y) = C(y) e^x e^{-C_2 r^\alpha}$$

Figure 11 shows the numerically calculated  $g(x | y)$ . Next, we explain analytically this phenomenon. For fixed  $y$ , let's find  $x$  which maximize  $g(x | y)$ . Set the derivatives of  $g(x | y)$  to zero, and substitute  $C_2$  by 1, we get the equation:

$$e^{2y} = e^{2x} (\alpha^{-\frac{1}{1-\alpha/2}} e^{2x \frac{\alpha}{2-\alpha}} - 1) \quad (5.3)$$

From this equation, it is easy to see that:

$$x \approx (1 - \frac{\alpha}{2})y - \frac{1}{2} \ln \alpha, \quad \text{if } y \rightarrow \infty$$

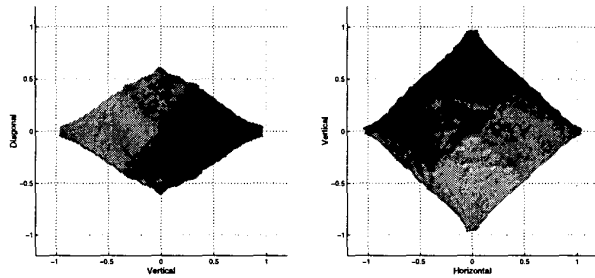
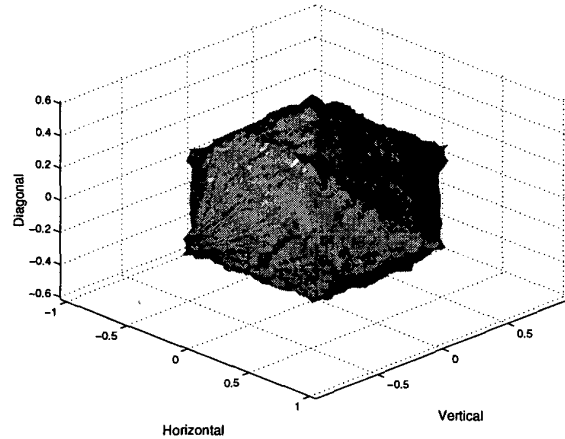


Figure 10: An equi-surface of 3d joint histogram of horizontal component, its vertical cousin and diagonal cousin, viewed from three different angles

$$x \rightarrow -\frac{1}{\alpha} \ln \alpha, \quad \text{if } y \rightarrow -\infty.$$

## 6 Long Range Covariances

All the statistics we talked above are small scale statistics, i.e. they are about single or a few nearby pixels. The most important long range statistic is the covariance of two pixel values:

$$C(x, y) = \langle I(x, y) I(0, 0) \rangle$$

Here  $\langle \rangle$  is the expectation, taking over all the images. However, our images are samples of a distribution which is only well-defined up to an additive constant, so we replace this statistic by the 'difference function':

$$D(x, y) = \langle |I(x, y) - I(0, 0)|^2 \rangle$$

which is related to the covariance by

$$D(x, y) + 2C(x, y) = \text{constant}$$

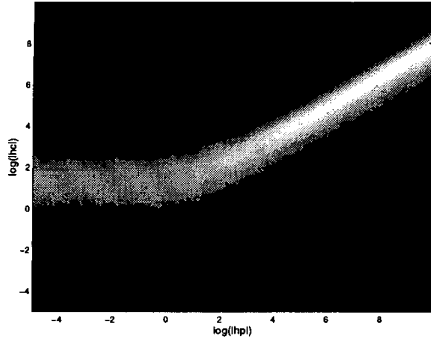


Figure 11: The conditional histogram  $H(\ln|hc| | \ln|hp|)$  calculated from our model. Bright parts mean high values

when both are well defined.

In [7], Ruderman calculated the ‘one-dimensional’ difference function, i.e., he took the average of  $D(x, y)$  over all directions, and got a one dimensional function  $D_1(x)$  to which he fit a scaling model:

$$D_1(x) = C_1 + C_2|x|^{-\eta} \quad (6.1)$$

These covariance models correspond in frequency domain to the power spectrum  $\sim \frac{1}{f^{2-\eta}}$ . If  $\eta$  goes to 0, note that  $1 - r^{-\eta} = 1 - e^{-\eta \log r} \approx \eta \log r$  giving us the model

$$D_1(x) = C_1 + C_2 \log(|x|) \quad (6.2)$$

which is the model implied by the assumption that  $2 \times 2$  block averages of the image  $I$  have the same distribution as  $I$  [8]. The best fitting constants Ruderman found from his image dataset are:  $C_1 = 0.79$ ,  $C_2 = -0.64$  and  $\eta = 0.19$ .

We calculated the two dimensional  $D(x, y)$  from our data set. Using a Fourier transformation technique, we actually took into our calculation *all* the possible pixel pairs within distance of 500 pixels. The statistics we got are very stable, and we can look more closely at the tail of the statistics, and even take delicate operations like derivatives on them. The upper two images in Figure 12 show the contour and mesh plot of  $D(x, y)$  we got. The lower two show the two cross sections along horizontal and vertical direction. We can see, the cross section along vertical direction grows faster than that along the horizontal direction. We believe the main reason is that, in many images, there is a portion of sky at top, and ground at bottom and the large difference between them will contribute a lot to the difference function along their vertical direction.

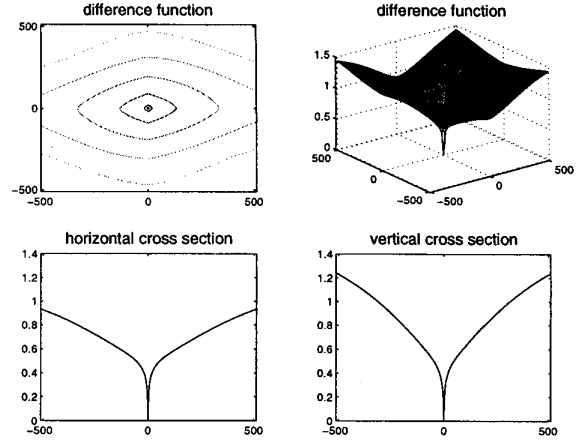


Figure 12: Difference Function

The upper left image in Figure 13 shows the log-log plot of the derivative of the positive part of horizontal cross section. The base we used when we took the log operation is 2. We see that, between 2 and 5 (corresponding to distances of 4 and 32 pixels), the derivative is close to a straight line, with a slope -1.19. If we use model (6.1), then  $\eta = -(-1.19 + 1) = 0.19$ , which is exactly what Ruderman got. But notice how the log-log plot begins to turn and becomes almost a horizontal line around 8. This clearly indicates that there exists a linear term, i.e. we can model it as:

$$D_1(x) = C_1 + C_2|x|^{-\eta} + C_3|x|$$

Generalizing it to  $D(x, y)$ , we seek a model:

$$D(x, y) = C_1(\theta) + C_2(\theta)r^{-\eta} + C_3(\theta)r$$

where,  $r = \sqrt{x^2 + y^2}$ , and  $\theta = \tan^{-1}(\frac{y}{x})$ . The best fitting  $\eta$  we got is 0.32, and the best fitting  $C_1(\theta)$ ,  $C_2(\theta)$  and  $C_3(\theta)$  are shown in figure 13. The maximum fitting error we got is 0.0035, which is very small, considering the range of  $D(x, y)$  is between 0 and 0.8, and the large area on which we fit the model (a annulus with  $4 < r < 200$ ).

One interesting observation we make is that  $C_1(\theta) + C_2(\theta)$  is almost zero, hence we may fit our model with one less parameter:

$$D(x, y) = C_2(\theta)(1 - r^{-\eta}) + C_3(\theta)r$$

Since  $C_3(\theta)$  is very small, the linear term can be omitted when  $r$  is small, we get (for  $r$  small):

$$D(x, y) \approx C(\theta)\ln(r)$$

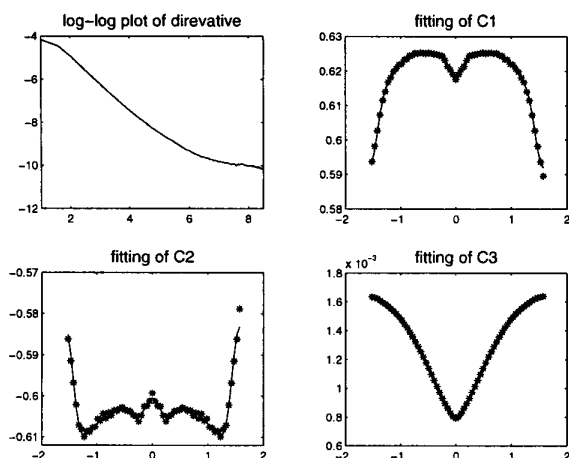


Figure 13: log-log plot of the horizontal cross section and some fitting constants, see text

This shows that while random images seem very close to samples from a scale-invariant process, there are also systematic deviations from scale invariance on a large scale.

## 7 Discussion

We have presented several statistics from natural images. From the statistics about derivatives and wavelet coefficients, our data suggests that statistics involving linear filters (with mean 0) can be modeled by 'generalized Laplace' distributions. Another new observation we make is that there is a linear tail in the Difference Functions, which indicates that, strictly speaking, natural images may not have the full scale invariance property. However we found that this property holds almost exactly *locally* (i.e. for filters with small supports).

## 8 Acknowledgments

We would like to thank Hans van Hateren for allowing us to use his image data set and clarifying questions we had about it.

## References

- [1] R. W. Buccigrossi and E. P. Simoncelli, "Image Compression via Joint Statistical Characterization in the Wavelet Domain". *GRASP Laboratory Technical Report #414, Univ. of Pennsylvania*, available at: <http://www.cns.nyu.edu/~eero/publications.html>
- [2] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., 1985

- [3] D.J. Field "Relations between the statistics of natural images and the response properties of cortical cells." *Journal of the Optical Society of America A*, 4:2379-2394, Dec 1987
- [4] B.A. Olshausen and D.J. Field "Natural image statistics and efficient coding". *Network: Computation in Neural Systems* 7:333-339,1996
- [5] J.H. van Hateren and A. van der Schaaf. "Independent Component Filters of Natural Images Compared with Simple Cells in Primary Visual Cortex" *Proc.R.Soc.Lond. B* 265:359-366,1998
- [6] D.Mumford, Song Chun Zhu and Basilis Gidas. "Stochastic Models for Generic Images", in preparation.
- [7] D.L. Ruderman "Origins of Scaling in Natural Images". *Vision Research* vol 37, No23, 3385-3395, 1997
- [8] Ya. G. Sinai. "Self-Similar Probability Distributions". *Theory of Probability and Its Applications* Volume XXI, 64-80, 1976
- [9] S Zhu, Y Wu and D Mumford. "Filters, random fields and maximum entropy(FRAME) -towards the unified theory for texture modeling". *IEEE Conf. Computer Vision and Patt Rec* June 1996