On the Statistics of Natural Images

by

Ting-Li Chen

B.A., National Taiwan University, 1994
M.S., National Taiwan University, 1996

Thesis

Submitted in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy
in the Division of Applied Mathematics at Brown University

Ph.D. Advisor: Stuart Geman

PROVIDENCE, RHODE ISLAND

May 2005

This dissertation by Ting-Li Chen is accepted in its present form by the
Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of
Doctor of Philosophy

Date ....
Stuart Geman

Recommended to the Graduate Council

Date ....
Basilis Gidas

Date ....
Donald E. McClure

Approved by the Graduate Council

Date ....
Karen Newman, Ph.D
Dean of the Graduate School

The Vita of Ting-Li Chen

Ting-Li Chen was born in the City of Keelung, Taiwan on January 26^{th} 1972.

He graduated from Chieh-Kuo Senior High School in Taipei, Taiwan. In 1990, he began his undergraduate study in National Taiwan University, and received a Bachelor of Sciences degree in Mathematics in 1994. He completed his Master's degree in Mathematics from National Taiwan University in 1996, then worked as a research assistant in the Department of Mathematics, National Taiwan University.

Since the year of 1997, he has been attending the Ph.D. program in the Division of Applied Mathematics at Brown University. This dissertation was defended and completed in October 2004

Abstract of "On the Statistics of Natural Images," by Ting-Li Chen, Ph.D., Brown University, May 2005

This thesis studies the statistics of natural images. It contains two chapters. Chapter 1 is on the scale invariance, a mysterious property of natural images. Chapter 2 is on the compression related to statistics of natural images.

Chapter 1 explores evidence of scale invariance of natural images, and explains why natural images have this nice property.

Chapter 2 studies an image compression algorithm "LOCO", and establish some useful properties of residual entropy, which can provide a more efficient compression.

Acknowledgments

I want to thank all those who have helped me to complete this thesis.

Most of all, I want to express my deepest gratitude to my advisor, Professor Stuart Geman, for directing my research and also for supporting me throughout my time at Brown. He gave many stimulating ideas and valuable suggestions all the time. He also offered many additional help in preparing for the manuscript. This thesis would not be possible without his help.

I also want to thank my committee members, Professor Donald E. McClure and Professor Basilis Gidas, for being my thesis readers and giving valuable comments.

I want to give my best regards to Professor Chii-Ruey Hwang in Academia Sinica in Taiwan, who was my advisor for my Master's thesis. I want to thank him for introducing me to Brown University, for his continuous encourgaments and also for his help both in my research and my personal life.

My many thanks to Matthew Harrison, my officemate, and to all people in the Applied Mathematics for helping me in every aspect. I am also very grateful to my parents and my wife for their love and support.

Contents

List of Tables

List of Figures

Chapter 1

On the Scale Invariance of Natural Images

1.1 Introduction

Scale invariance is a mysterious property of natural images. It refers to the phenomenon that the distributions of many statistics of natural images are very close to those of scaled ones. At the first thought, scale invariance seems a trivial phenomenon. One might think that a scaled image can be produced by moving the camera backward. However, this is not correct. When an object is at distance d_1 to the camera and the camera is moved distance d_2 backward, the image of the object is down-scaled by a factor of $\frac{d_1}{d_1+d_2}$. This factor varies with respect to d_1. Therefor, unless all the objects are at the same distance to the camera, regardless of where we place the camera, we can not get an image in which all objects are scaled down with the same factor. In other words, a scaled image can not be viewed as another image taken from a difference distance. Therefore, scale invariance is not as trivial as one might think at the beginning.

When scale invariance of natural images is studied, a question quickly comes to one's mind: Why do natural images have such a property? There are models which explain scale invariance of natural images. For example, in the model given by Ruderman [2] [3] [4], objects with random shapes and sizes are randomly placed in an infinite image plane, and each object is independently painted with a gray tone chosen from a distribution. It was argued that if these objects have a power-law distribution of sizes, statistics of images of the plane also follow a power law, which is related to scale invariance.

In the model given by Mumford and Gidas [5], objects in images are also randomly placed and have random shapes and sizes. In contrast to the model given by Ruderman, Mumford and Gidas allowed for patterns of colors within each object. Images composed of all categories of patterns together have scale invariant property.

There is also a model given in Chi [1], in which the world is flat, and the objects are placed randomly not too far from the ground. Chi argued that the transformation of object distances into object sizes in the image plane (i.e. on the film, retina, or CCD

array) produces a size distribution following the $1/r^3$ rule. Since this rule of sizes is necessary for perfect scale invariance, this "flat-earth" model suggests an explanation for scale invariance. We continue this work with more detailed calculations, and we claim that the size of objects does approximately follow the $1/r^3$ rule in the flat-world model. However, we also find some evidence of images having a good scale invariance without the projection effect. Therefore, we propose a new model to explain scale invariance.

We mention here some related referrences. Balboa et al. [6] discussed occlusion effect to scaling. Lee et al. [7] developed "dead leaves model", which takes occlusions into account. Thomson [8] described a measure and showed that there are statistical consistencies in the phase spectra of natural scenes. Grenander [9] and Srivastava [10] introduced a Bessel Kernel form for modeling the marginal probabilities of the spectral components of images. Turiel et al. [11] [12] showed multiscaling properties of natural scenes. Field [13] [14] and Olshausen [15] showed the relation of the response properties to the structure of natural images. Burton et al. [16] discussed structures in natural scene and methods of optimum image coding. van Hateren et al. [17] developed a theory of maximizing sensory information.

In the following, we review the mathematical definition of scale invariance, and provide some examples in section 2. In section 3, we derive the $1/r^3$ rule of the size of objects, and explain it using the flat-world model. In section 4, we present evidence to show that the projection effect may not be the key factor in scale invariance. Finally, we propose a new model to explain scale invariance in local statistics of natural images.

1.2 Definition and Examples of Scale Invariance

1.2.1 Definitions

A digital image I of size $M \times N$ with L integer gray levels is a matrix with M rows and N columns, where $I(i,j) \in \{0, 1, 2, \ldots, L-1\}$ is the gray level at pixel (i,j).

To scale down an image by k, we partition an image into disjoint blocks of size $k \times k$, and take the average of gray level values of each block as its new value. More precisely, let $I^{(k)}$ be the down-scaled image by factor k, and block $B_{ij} = [(i-1)k + 1, ik] \times [(j-1)k + 1, jk]$ for $1 \le i \le \lfloor M/k \rfloor$ and $1 \le j \le \lfloor N/k \rfloor$. Then

$$I^{(k)}(i,j) = \frac{1}{k^2} \sum_{n=1}^{k} \sum_{m=1}^{k} I((i-1)k + n, (j-1)k + m).$$

We round off each average to an integer, so that the scaled image takes values in the same space as the original one does.

Real images are not truly scale invariant. But most, if not all, *local* statistics are found to be nearly scale invariant when tested on an ensemble of natural images. The phenomenon is surprisingly robust. Mathematically, this property is conveniently studied by pretending that images are truly scale invariant, meaning that the original image and the scaled image have the same distribution. However, since the size of the original image and that of the scaled one are different, the distributions of these two images can not be the same. In order to make a comparison between their distributions, we extend the image domain to $\mathbf{R^2}$ as follows.

Let $\phi(x,y)$ be a function defined on $\mathbf{R^2}$. We can think of ϕ as the underlying continuous image of I. The value of the pixel (i,j) of I is the average of $\phi(x,y)$ over its corresponding block on $\mathbf{R^2}$. That is,

$$I(i,j) = \frac{1}{d^2} \int_{[(i-1)d, id] \times [(j-1)d, jd]} \phi(x,y) dx dy,$$

where d is the unit length. From the definition of $I^{(k)}$,

$$\begin{aligned}
& I^{(k)}(i,j) \\
= & \frac{1}{k^2}\sum_{n=1}^{k}\sum_{m=1}^{k} I((i-1)k+n,(j-1)k+m) \\
= & \frac{1}{k^2}\sum_{n=1}^{k}\sum_{m=1}^{k}\frac{1}{d^2}\int_{[((i-1)k+n-1)d,((i-1)k+n)d]\times[((j-1)k+m-1)d,((j-1)k+m)d]}\phi(x,y)dxdy \\
= & \frac{1}{k^2d^2}\int_{[(i-1)kd,ikd]\times[(j-1)kd,jkd]}\phi(x,y)dxdy \\
= & \frac{1}{d^2}\int_{[(i-1)d,id]\times[(j-1)d,jd]}\phi(kx',ky')dx'dy'.
\end{aligned}$$

Therefore, we can view $\phi(kx,ky)$ as the underlying function of $I^{(k)}(i,j)$, while viewing $\phi(x,y)$ as the underlying function of $I(x,y)$. Now we can say that an image has scale invariance if

$$\phi(kx,ky) \stackrel{\mathcal{D}}{=} \phi(x,y).$$

In practice, it is impossible to check whether two infinite-dimensional random fields (such as $\phi(kx,ky)$ and $\phi(x,y)$) have the same distribution. However, the distribution of any filter should be the same for both fields. Furthermore, under the assumption of stationarity of images, it is shown in Chi[1] that the distribution of natural images itself is scale invariant if the marginal distribution of any filter is also scale invariant.

In the literature, the power spectrum of natural images is used to examine the property of scale invariance. Power spectrum $S(k)$ is defined as

$$S(k) = \frac{1}{2\pi}\int_0^{2\pi} d\theta \int_{\mathbf{R^2}} < \phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y}) > e^{-ik\mathbf{v}(\theta)\cdot\mathbf{y}}d\mathbf{y},$$

where for fixed $\mathbf{y}$, $< \phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y}) >$ represents the expected value (with respect to the random field ϕ) of the average of $\phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y})$ over all $\mathbf{x}$, $\mathbf{v}(\theta) = (\cos(\theta), \sin(\theta))$, and k is the magnitude of the spatial frequency. It is well-known that the power

spectrum of natural images takes the form of a power-law in the spatial frequency [4]:

$$S(k) = \frac{A}{k^{2-\eta}},$$

where η is the "anomalous exponent" (usually small), and A is a constant which determines the overall image contrast. Indeed, under the assumption of ergodicity of the distribution of images,

$$< \phi(\mathbf{x})\phi(\mathbf{x}+\mathbf{y}) >= \mathbf{E}(\phi(\mathbf{0})\phi(\mathbf{y})),$$

where E is over all ϕ. Then, with scale invariance property,

$$S(k) = \frac{S(1)}{k^2}.$$

This is exactly the case of $\eta = 0$. The proof can be found, for example, in [1].

1.2.2 Examples

The images that we use in this section are from "The Dutch Image Database", which was first used in [18]. These images were obtained with a Kodak DCS420 digital camera. The size of each image is 1536 horizontal by 1024 vertical pixels. Each pixel is a 2-byte unsigned integer. The pixels are linear in intensity. The intensity scaling is determined by the settings of the camera for each image. We randomly choose nine images from the database, and examine the scale invariance of each individual image as well as the ensemble of all nine images. Figure 1.1 shows the nine images we use.

We experiment with several statistics on these images. The first two statistics we use are horizontal derivatives and vertical derivatives. For digital images, we take the difference of gray level values between a pixel and its neighboring pixels as its

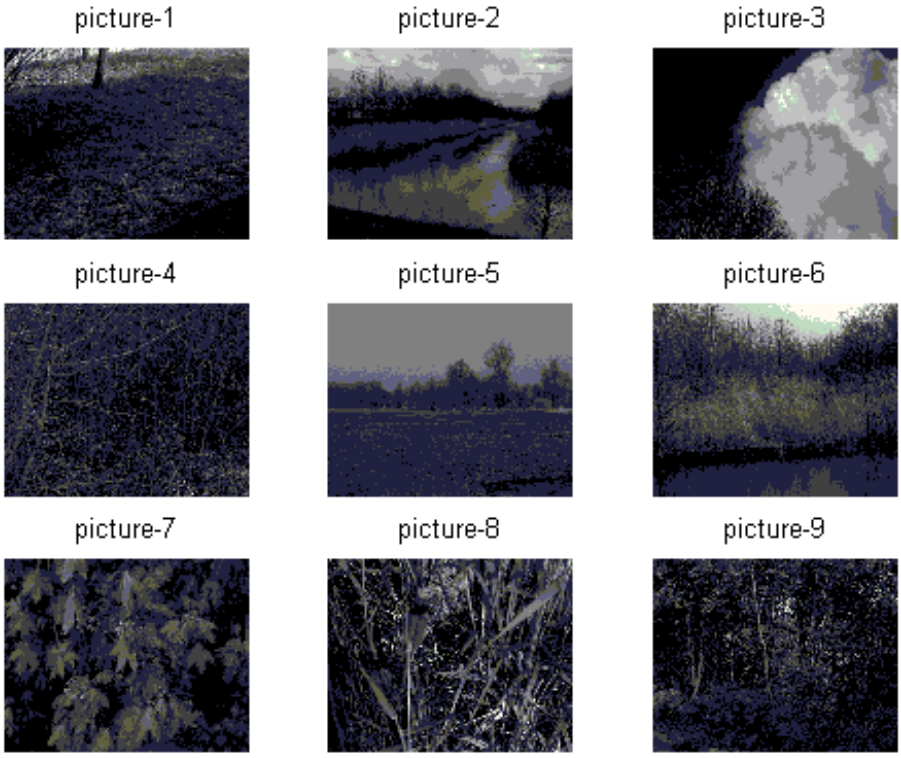

Figure 1.1: Nine sample images

derivatives. More precisely,

$$\begin{aligned} \nabla_x I(i,j) &= I(i,j+1) - I(i,j) \\ \nabla_y I(i,j) &= I(i+1,j) - I(i,j). \end{aligned}$$

Similarly, we also consider the filter

$$\nabla_x \nabla_y I(i,j) = I(i,j) - I(i,j+1) - I(i+1,j) + I(i+1,j+1).$$

We also experiment with ∇_x of the logarithm of each image. For each image I, we create a new image J by

$$J(i,j) = \log I(i,j).$$

Then we compute the marginal distributions of $\nabla_x J$, $\nabla_x J^{(2)}$ and $\nabla_x J^{(4)}$.

Theoretically, the range of ∇_x's is from -65535 to 65535. However, ∇_x's rarely fall outside [-2000,2000]. Therefore, when we compute the histogram of derivatives for each image, we use 201 bins which have equal width of 20 and are centered from -2000 to 2000. For the experiment $\nabla_x J$, we also use 201 bins but they are centered from -5 to 5. We normalize each histogram so that the sum is 1. Since most filter values are near 0, the histograms look very similar. We take logarithms of each histogram to explore more features.

Figure 1.2 to Figure 1.10 are the results for images from 1.1, and Figure 1.11 is that for the ensemble images. In each figure, (a) shows horizontal derivative $\nabla_x I(i,j)$, (b) shows vertical derivative $\nabla_y I(i,j)$, (c) shows $\nabla_x J(i,j)$, and (d) shows $\nabla_x \nabla_y I(i,j)$. In each sub-figure, the solid curve represents the empirical marginal distribution of the filter values of the original image I, the dashed curve represents that of the scaled image $I^{(2)}$, and the dotted curve represents that of the scaled image $I^{(4)}$.

For images with scale invariance property, we expect to see that the marginal distributions stay nearly the same after scaling. However, as shown in Figure 1.2 to

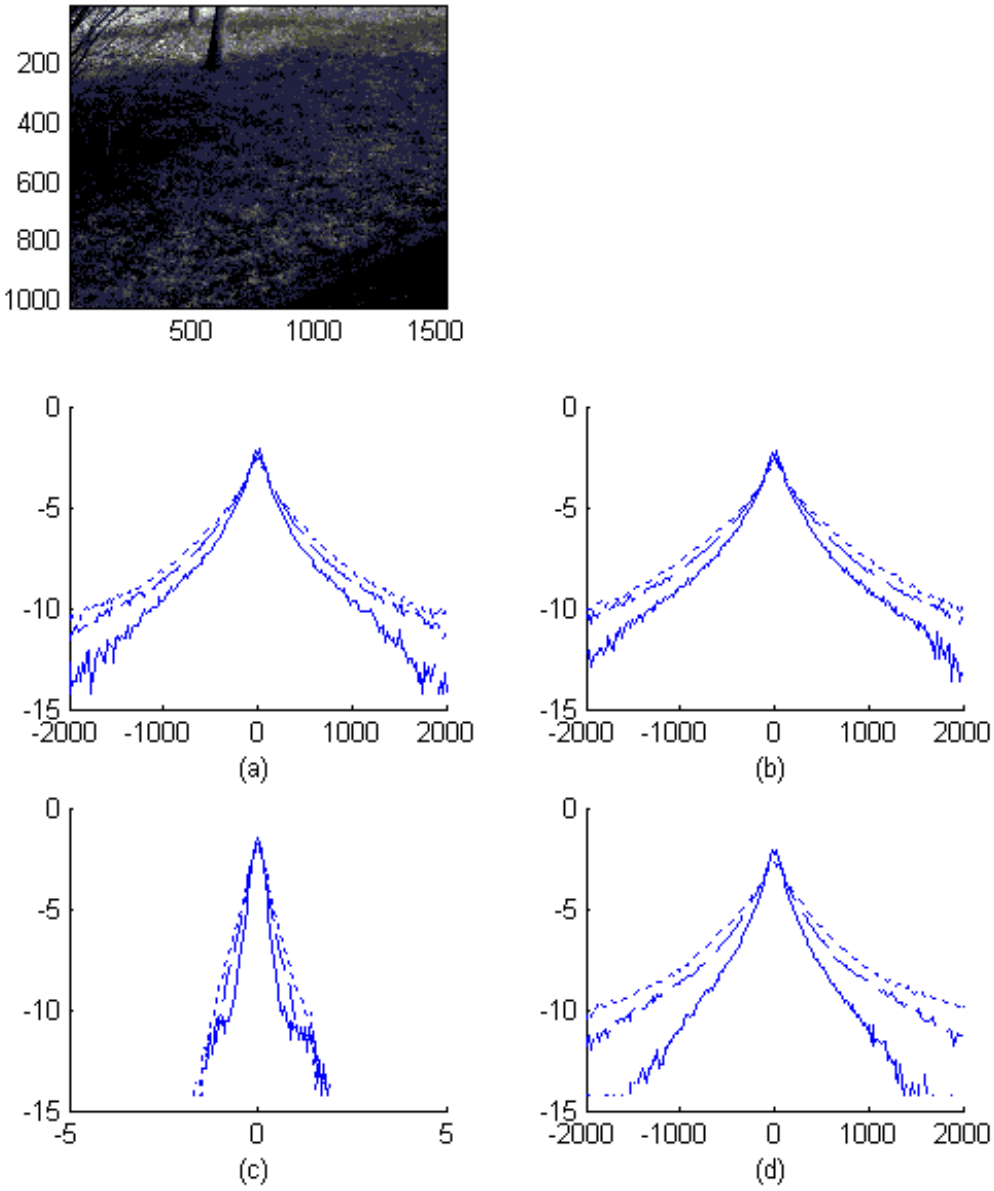

Figure 1.2: Logarithms of marginal distributions of different filters for picture-1 from Figure 1. Solid curve:I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

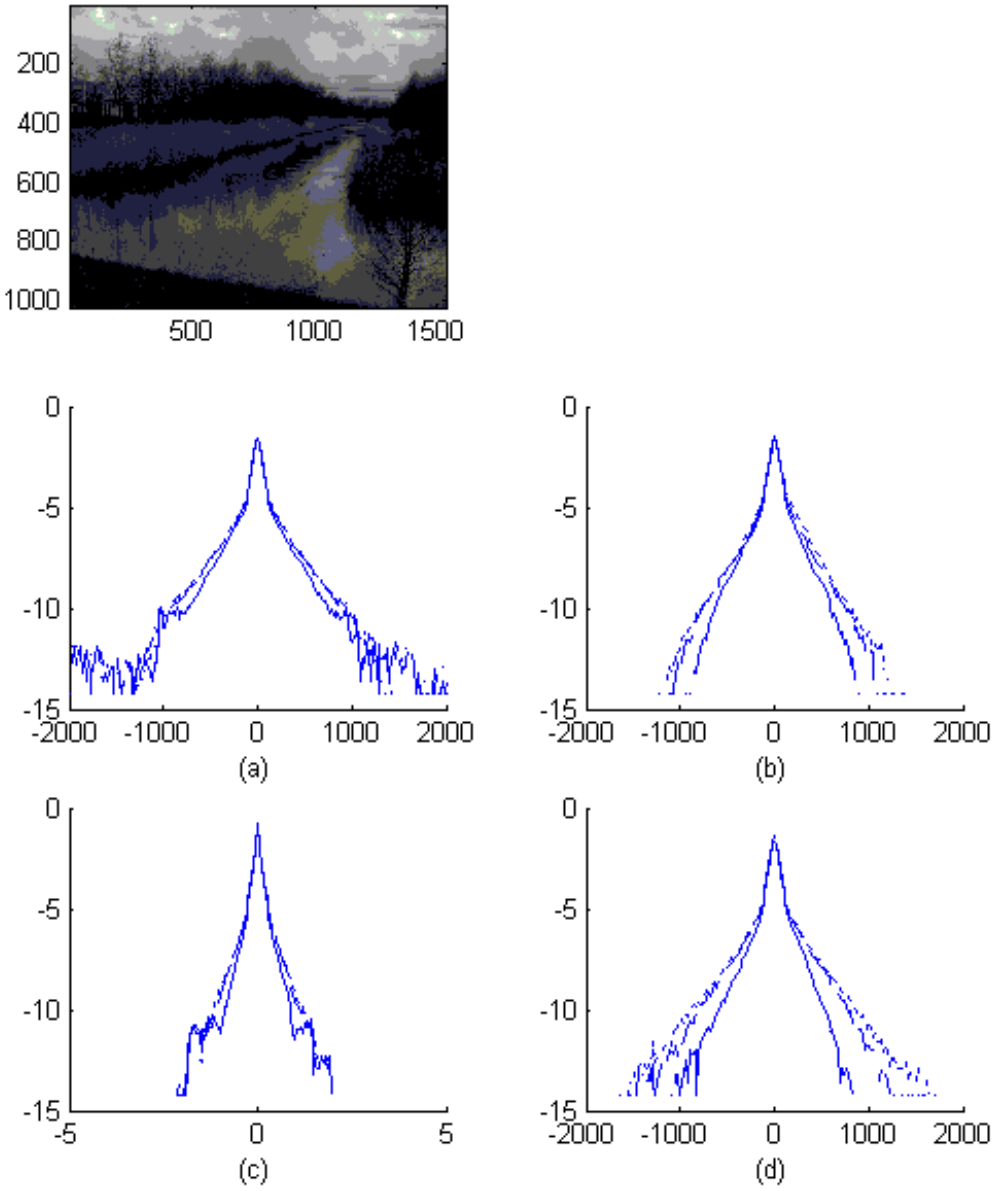

Figure 1.3: Logarithms of marginal distributions of different filters for picture-2 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

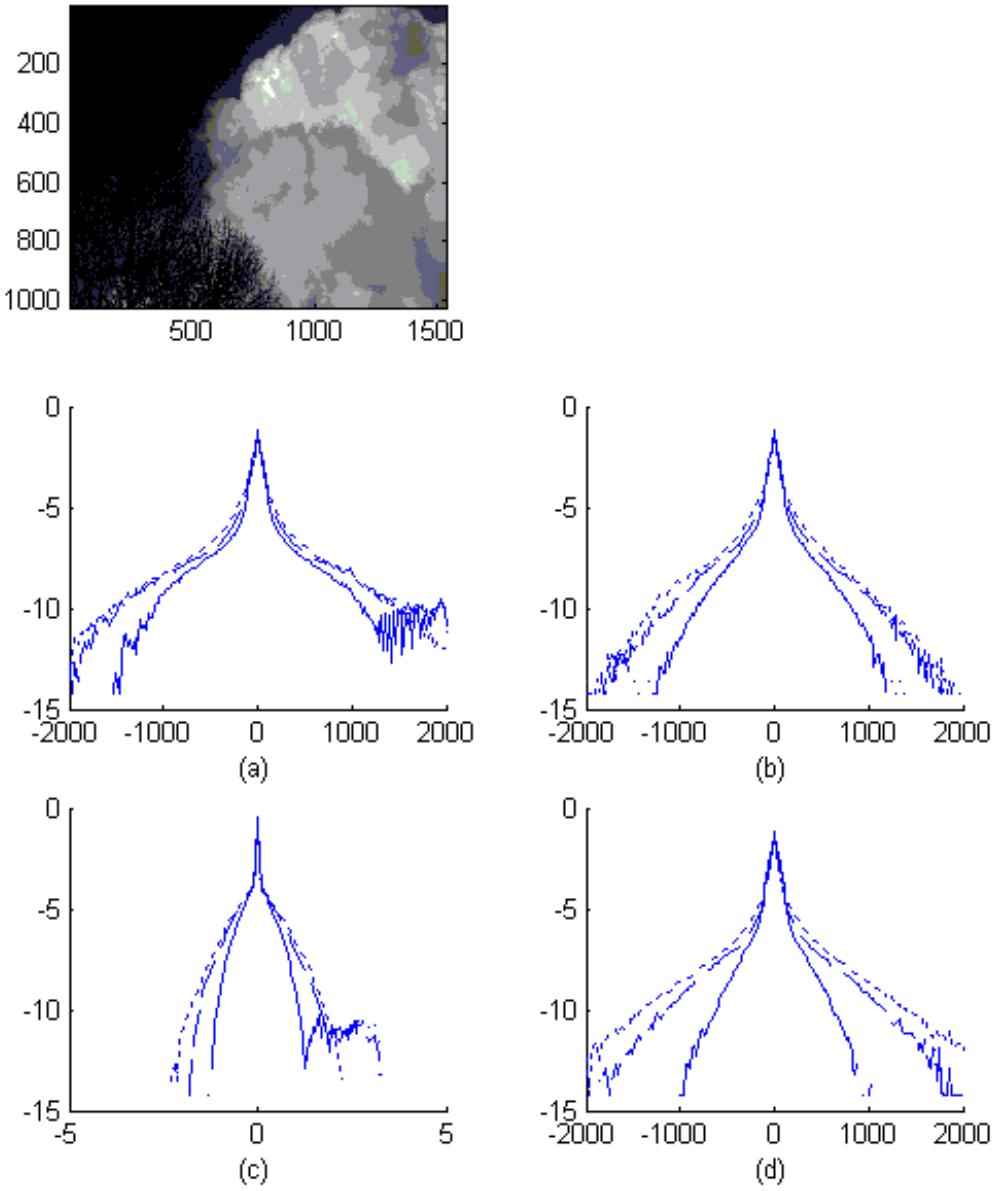

Figure 1.4: Logarithms of marginal distributions of different filters for picture-3 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

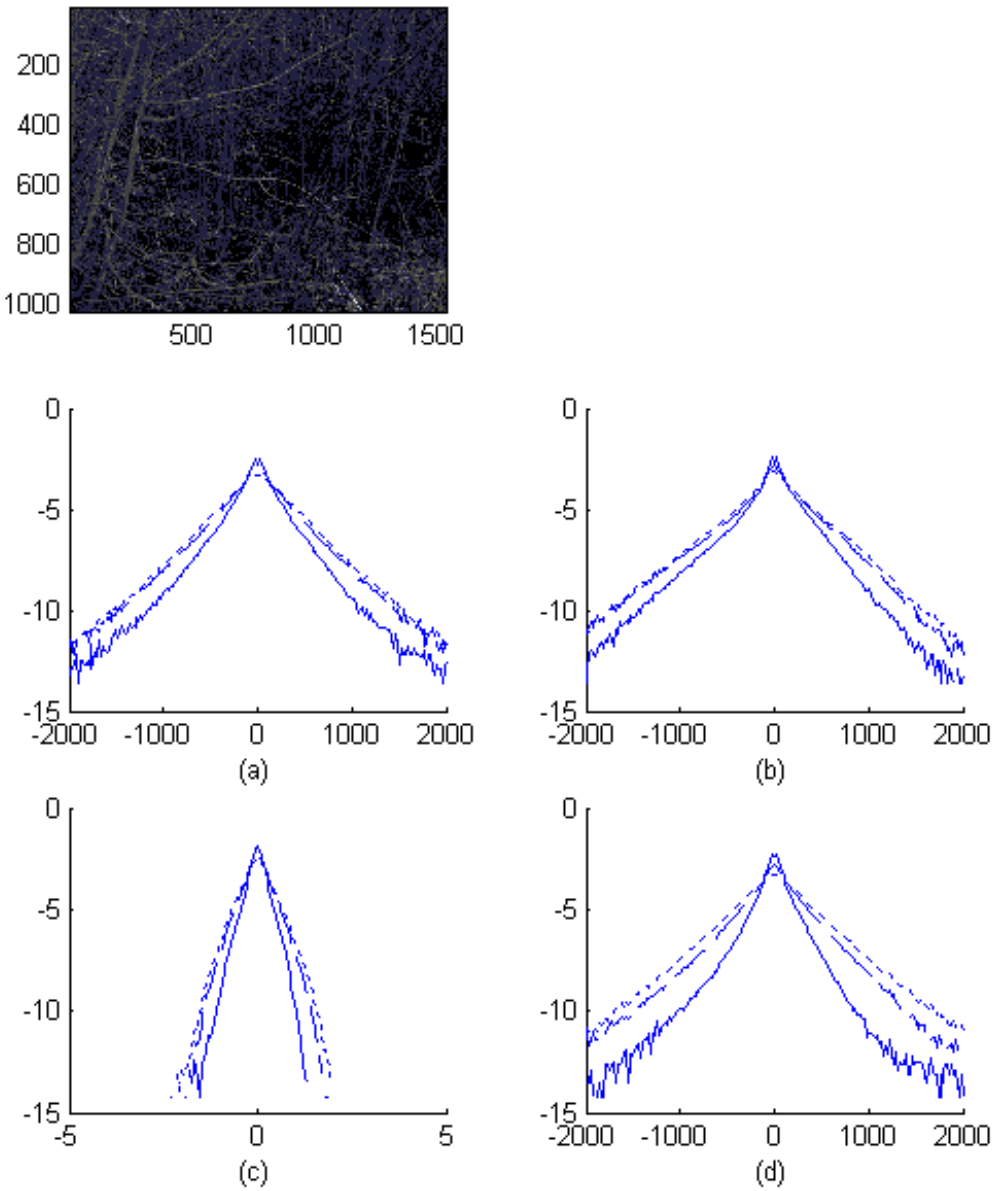

Figure 1.5: Logarithms of marginal distributions of different filters for picture-4 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

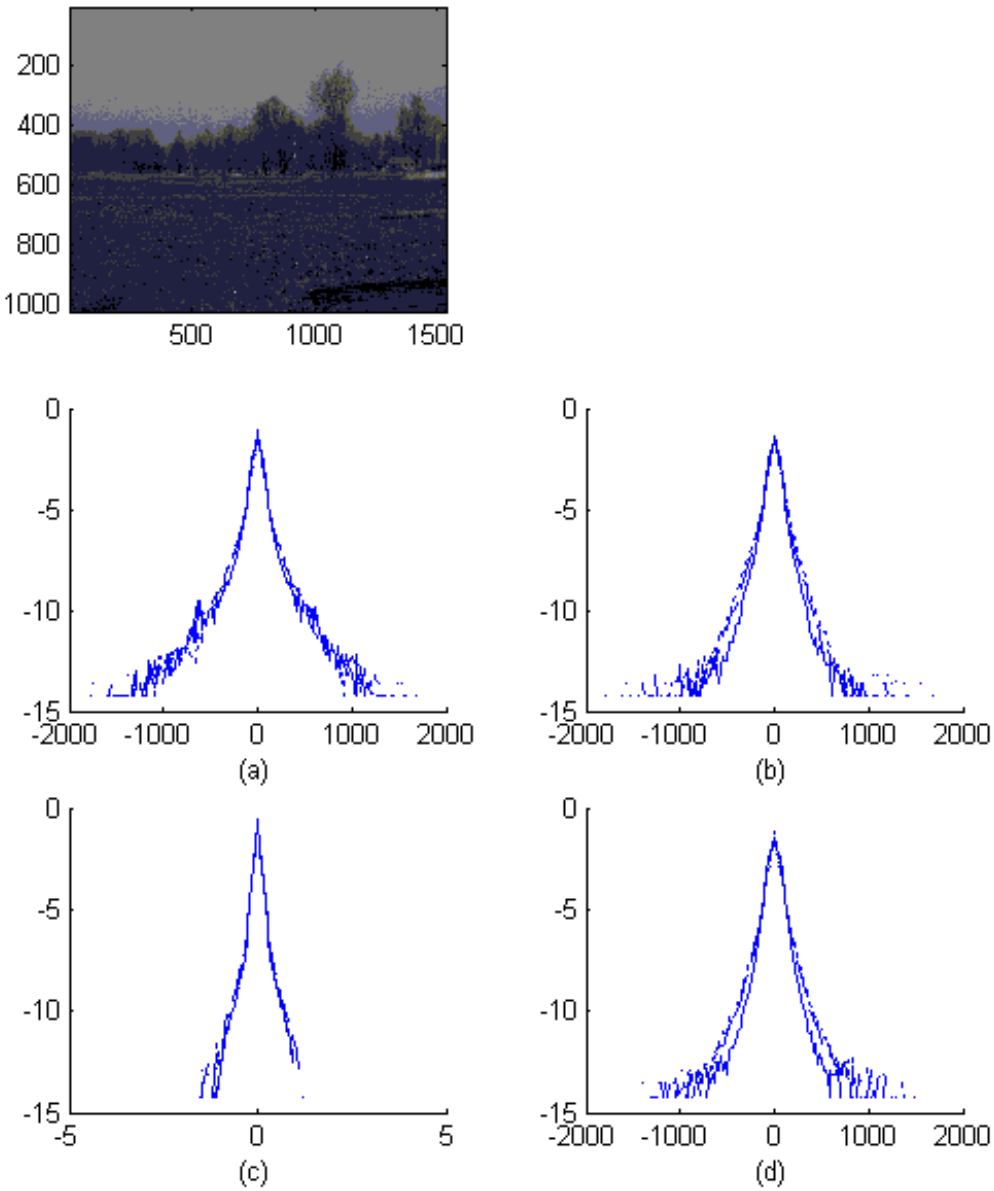

Figure 1.6: Logarithms of marginal distributions of different filters for picture-5 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

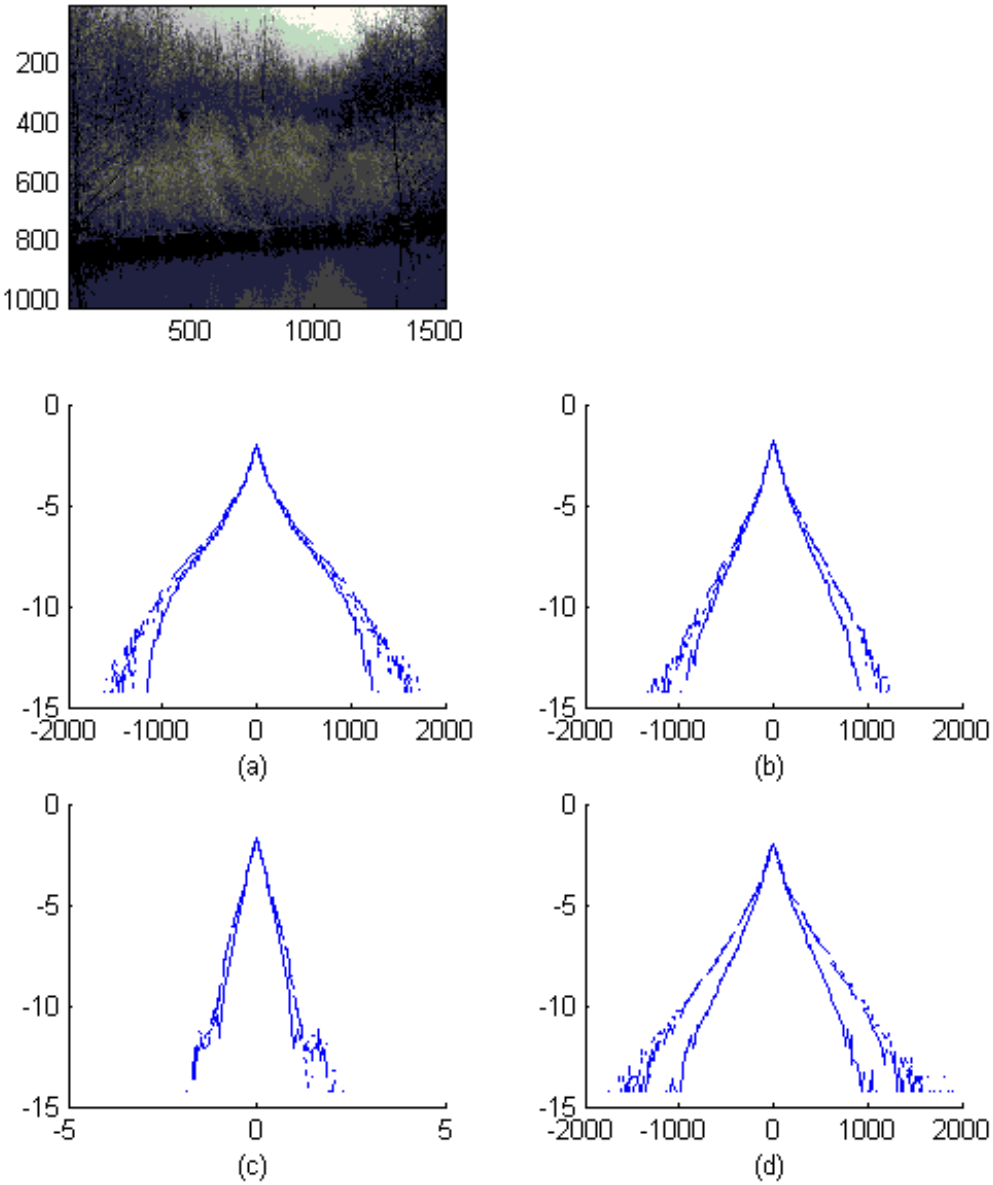

Figure 1.7: Logarithms of marginal distributions of different filters for picture-6 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

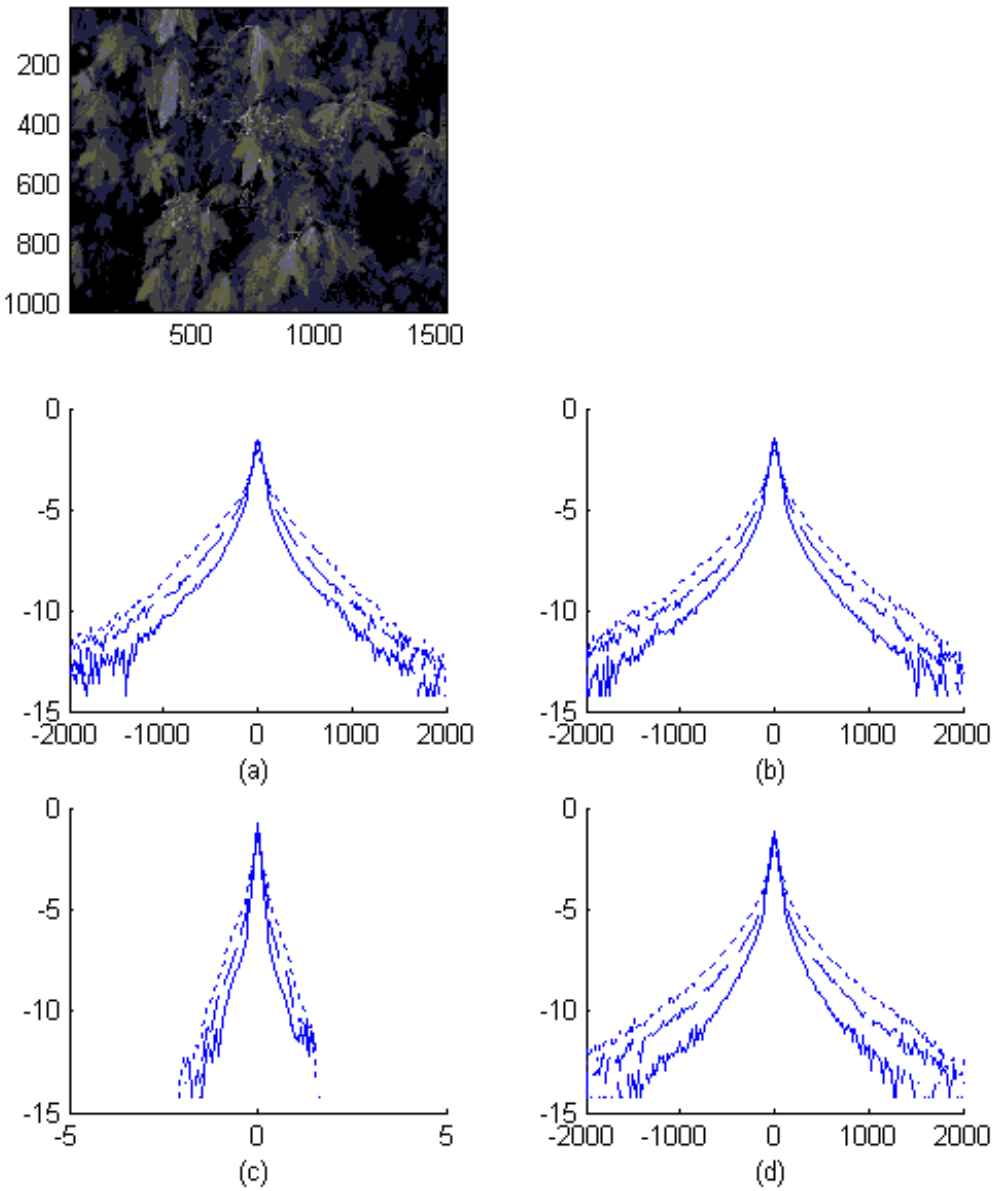

Figure 1.8: Logarithms of marginal distributions of different filters for picture-7 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

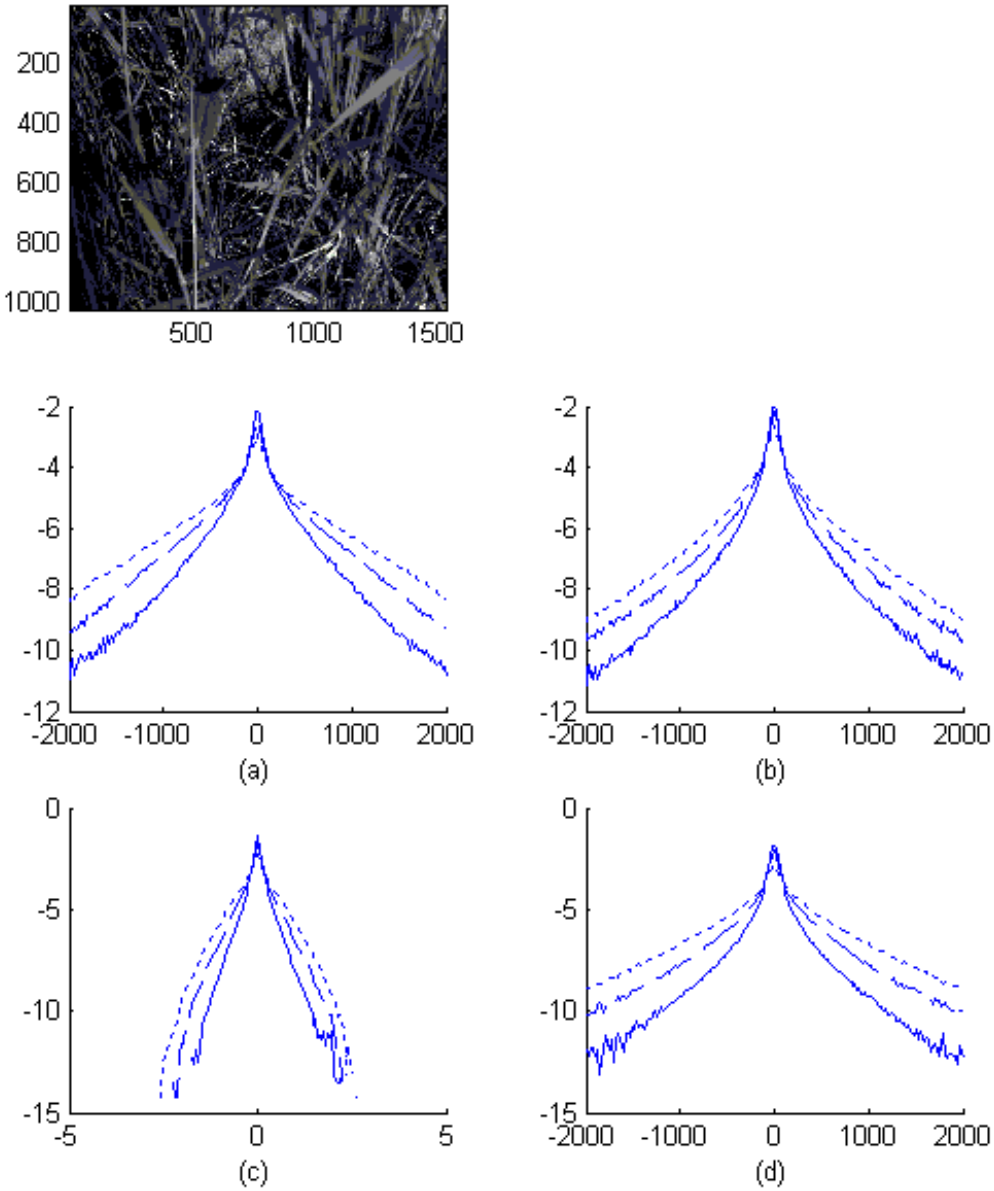

Figure 1.9: Logarithms of marginal distributions of different filters for picture-8 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

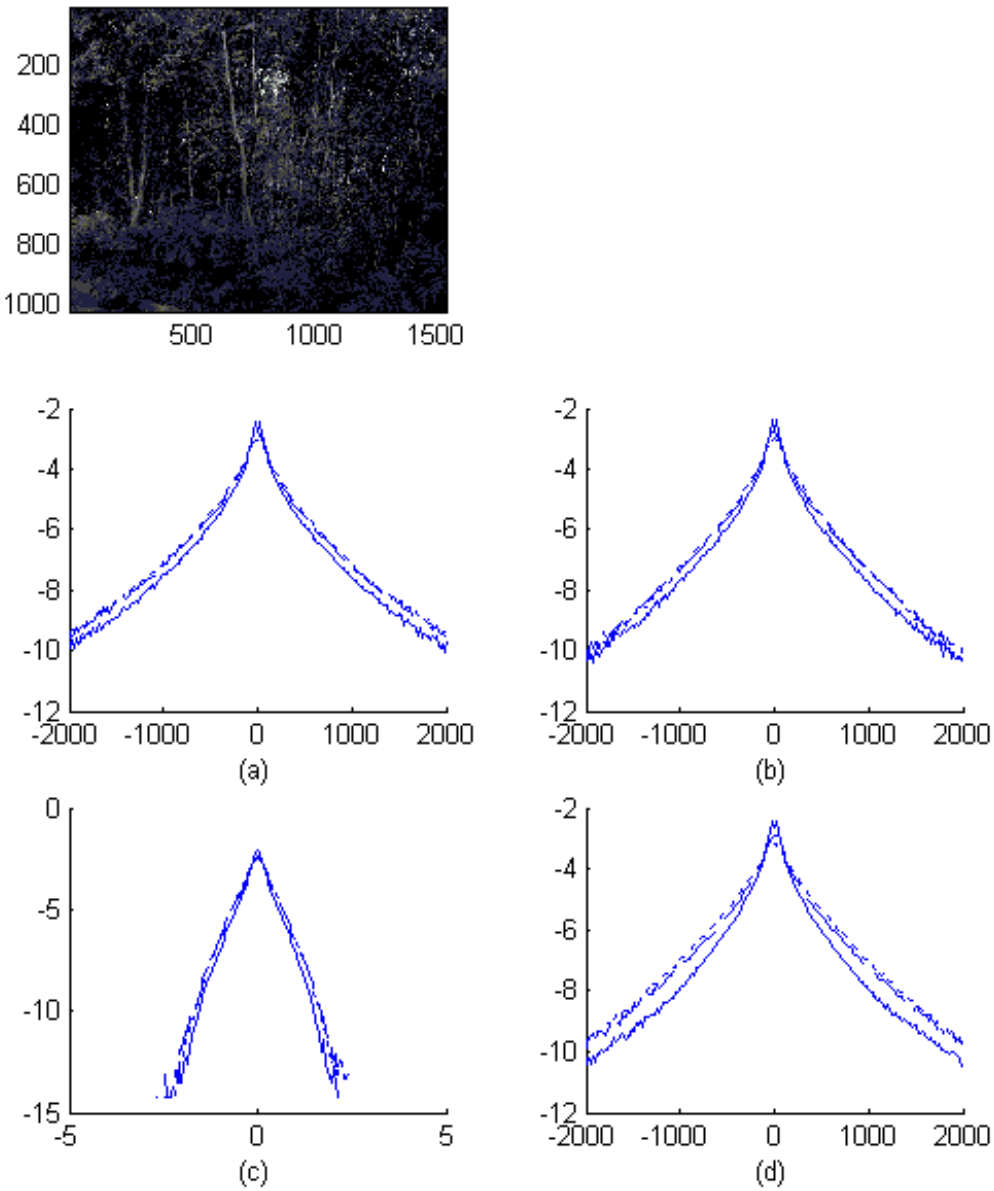

Figure 1.10: Logarithms of marginal distributions of different filters for picture-9 from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

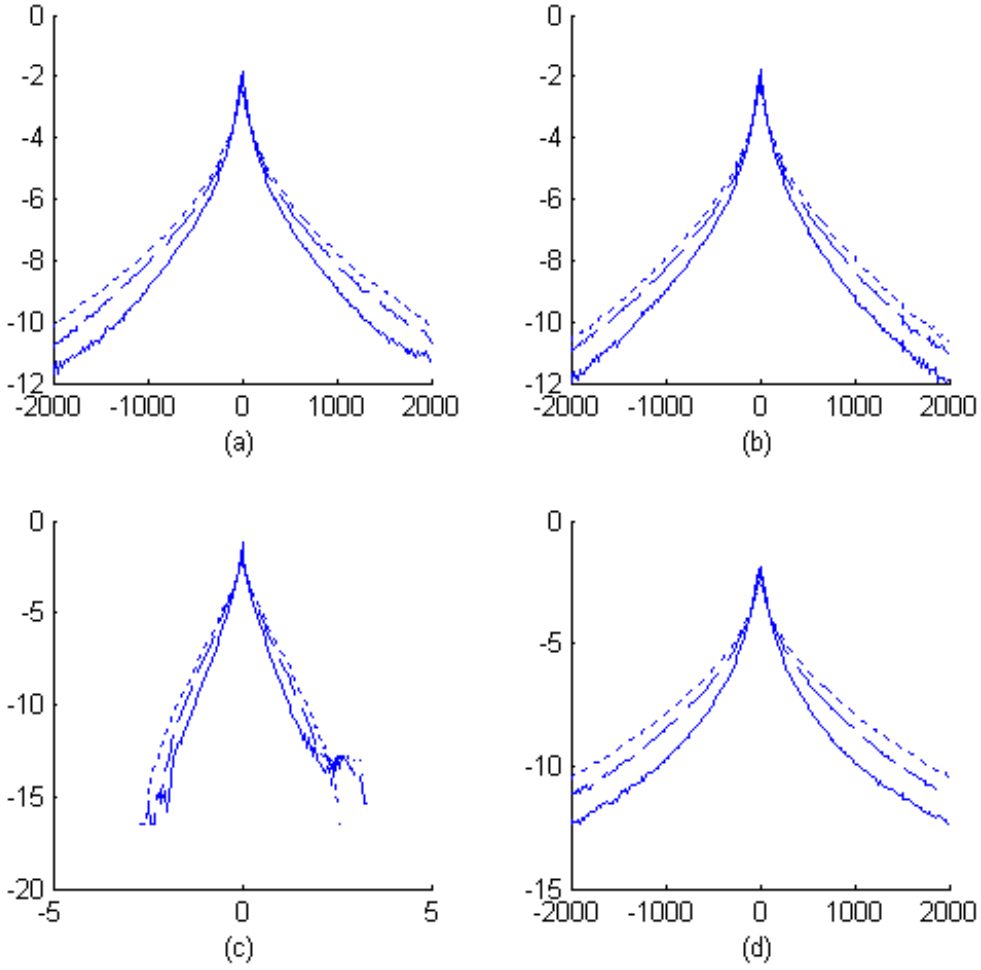

Figure 1.11: Logarithms of marginal distributions of different filters for the ensemble images of all pictures from Figure 1. Solid curve: I, dashed curve: $I^{(2)}$, and dotted curve: $I^{(4)}$. (a) $\nabla_x I$, (b) $\nabla_y I$, (c) $\nabla_x J$, and (d) $\nabla_x \nabla_y I$.

Figure 1.11, not all of the marginal distributions stay nearly the same. We observed that, among nine images, image 2, 5, 6, 9 demonstrate better scale invariance property. We also observed that these four images cover broader views and show richer features than the other five images do. In Section 4, we propose a model to explain why images with broader views and richer features have better scale invariance.

In fact, figure 1.2 to figure 1.11 show that all of the marginal distributions, including the four images with better scale invariance, have heavier tails after scaling. However, there was no such phenomenon observed by Chi [1] that the tails are heavier after scaling. Our explanation for this discrepancy is that the images we used were less noisy. In Section 4, we will discuss the noise effect on scale invariance. In general, the noisier an image is, the less heavy tail the marginal distribution has after scaling.

From Figure 1.2 to Figure 1.11, we also found that the behavior of the marginal distributions is very consistent with respect to different filters for each image. In other words, an image either has good scale invariance for every filter, or does not have good scale invariance for any filter.

1.3 The size of objects in the natural images

In this section, following an argument by Mumford (personal communication), we first derive the distribution that the size of objects has to obey in images with perfect scale invariance. Then we examine, empirically, the distribution of object sizes in natural images. Last, we build a model, which is an elaboration of the "flat-earth" model, to explain the distribution of sizes in natural images.

1.3.1 Mathematical considerations: the $1/r^3$ law

By perfect scale invariance with respect to the size of objects, we mean that the distribution of the size of objects in the image stays unchanged after the image gets scaled. In the following, we derive the distribution that the size of object has to follow

for perfect scale invariance.

Let N be the total number of objects in an image $[0, l_1] \times [0, l_2]$, and $N(a, b)$ be the expected number of objects of size between a and b. When we scale down this image by s, the scaled image has size $[0, \frac{l_1}{s}] \times [0, \frac{l_2}{s}]$, and objects of size between a and b in the original image become those of size between $\frac{a}{s}$ and $\frac{b}{s}$ in the scaled image. Then we create a new image of size $[0, l_1] \times [0, l_2]$, which consists of $s \times s$ identical scaled images mentioned above. For perfect scale invariance, the statistics of this new image should be the same as those of the original one. Then we have

$$N(\frac{a}{s}, \frac{b}{s}) = N^{(s)}(\frac{a}{s}, \frac{b}{s}) = s^2 N(a, b),$$

where $N^{(s)}(a, b)$ is the expected number of objects of size between a and b in the scaled image. Note that the total number of objects in the new image is s^2 times as many as that in the original one. This will be addressed shortly, by putting a lower bound on object size. Let $f(r)$ be the density of objects of size r. The equality above turns into

$$N \cdot \int_{\frac{a}{s}}^{\frac{b}{s}} f(r)dr = N \cdot s^2 \int_a^b f(r)dr.$$

Taking derivatives on both sides with respect to b, we obtain

$$\begin{aligned} & \tfrac{1}{s} f(\tfrac{b}{s}) = s^2 f(b) \\ \Rightarrow \quad & f(\tfrac{b}{s}) = s^3 f(b) \\ \Rightarrow \quad & f(r) \propto \tfrac{1}{r^3} \end{aligned}$$

The density has a non-integrable singularity at zero, so we put a lower bound on object size, say m. (This will also insure that the total number of objects is preserved under scaling, and shown below.) After normalizing, we get

$$f(r) = \frac{2m^2}{r^3}, \quad r \in [m, \infty).$$

Now we examine whether the total number of objects is preserved under scaling. With the assumption of a lower bound m on object size, objects of size between m and $s \cdot m$ in the original image become invisible after we scale down the image by s. The number of objects of sizes between m and $s \cdot m$ in the original image is

$$N \cdot \int_m^{s \cdot m} f(r)dr = N \cdot \frac{s^2 - 1}{s^2}.$$

In other words, there are only $N \cdot \frac{1}{s^2}$ objects visible after scaling. Along with the fact that the total number of objects in the new image is s^2 times as many as that in the original one, there are $s^2 \cdot N \cdot \frac{1}{s^2} = N$ objects in the new image. Therefore, the total number of objects is preserved under scaling.

1.3.2 The area law

In the real world, objects have different shapes. Instead of examining the distribution of the size of objects in the image, it is easier to examine the area of objects. In this section, we will derive a rule with respect to the area of objects. This rule is equivalent to the rule that the size of objects follows $1/r^3$. Then we find evidences that the area of objects in natural images actually follows the rule.

Equivalence of size and area scaling rules

An equivalent statement to "the size of objects follows $1/r^3$ rule" is "the area of objects in the image follows $1/A^2$ rule".

Let N be the total number of objects in an image $[0, l_1] \times [0, l_2]$, and $N_A(a, b)$ be the expected number of objects of area between a and b in an image $[0, l_1] \times [0, l_2]$. After we scale down this image by s, the scaled image has size $[0, \frac{l_1}{s}] \times [0, \frac{l_2}{s}]$, and objects of area between a and b in the original image become those of area between $\frac{a}{s^2}$ and $\frac{b}{s^2}$ in the scaled image. Then we create a new image of size $[0, l_1] \times [0, l_2]$, which consists of $s \times s$ identical scaled images as mentioned above. For perfect scale

invariance, the statistics of this new image should be the same as those of the original one. By similar calculation in previous section, we have

$$N(\frac{a}{s^2}, \frac{b}{s^2}) = s^2 N(a, b).$$

Let $g(A)$ be the density of objects of area A. The equality above turns into

$$N \cdot \int_{\frac{a}{s^2}}^{\frac{b}{s^2}} g(A)dA = N \cdot s^2 \int_a^b g(A)dA.$$

Taking derivatives on both sides with respect to b, we obtain

$$\begin{aligned} & \tfrac{1}{s^2} g(\tfrac{b}{s^2}) = s^2 g(b) \\ \Rightarrow \quad & g(\tfrac{b}{s}) = s^2 g(b) \\ \Rightarrow \quad & g(A) \propto \tfrac{1}{A^2}. \end{aligned}$$

Similarly, we put a lower bound on object area, say m_A. After normalizing, we get

$$g(A) = \frac{m_A}{A^2}, \quad A \in [m_A, \infty).$$

Evidence that the area of objects in the image follows $1/A^2$ rule

Experiments in Alvarez-Morel's paper [19]:

Consider a digital image I of size $H \times L$, with G integer gray levels, and write $I(i, j)$ for the gray level at pixel (i, j). Let k be an integer less than G. Let N_1 be the first integer such that more than $\frac{HL}{k}$ pixels have a gray level less than N_1, and similarly, N_l be the first integer such that more than $l\frac{HL}{k}$ pixels have a gray level less than N_l. Define $I_l(i, j)$ by

$$I_l(i, j) = \begin{cases} 1 & \text{if } I(i, j) \in [N_{l-1}, N_l) \\ 0 & \text{otherwise.} \end{cases}$$

Alvarez and Morel call these images "k-bilevels of I." This method classifies all pixels in image I into k categories, and each category has approximately the same number of pixels.

In any of the k-bilevels of I, each connected component of 1's is viewed as an object. Alvarez and Morel then calculate $f(s)$ as the number of connected components with area s in all k-bilevels of I. They consider both 4-connexity and 8-connexity.

For fixed k, consider the set of points

$$S = \{(\log(s), \log(f(s))), 0 \leq s \leq T_{\max}\},$$

where

$$T_{\max} = \inf\{s : f(s) = 0\} - 1.$$

Alvarez and Morel perform linear regressions on the set S to find the straight line (in the log-log coordinates) $(\log(f(s))) = A - \alpha \log(s)$, which is the best fit to S in the least squares sense.

They obtained the value of α between 1.5 and 3 for images of natural scenes, and between 2.5 and 3.5 for texture images. In the ideal case $f(A) = \frac{C}{A^2}$, the value of α is expected to be 2.

Our experiments:

In the experiment as described above [19], connected components are viewed as objects. However, these connected components may be far from real objects in real images. In order for connected components to look more like real objects, we consider two neighbor pixels as connected if their absolute difference in gray level values is smaller than k. We partition the whole image into disjoint connected components. Each such connected component is viewed as an object, and the area of each object is calculated. The last step of our experiment is to perform similar linear regressions.

We experiment with the nine images in Figure-1.1. Figure-1.12(a) shows that there are much more pixels of low gray level values in these nine images. Figure-

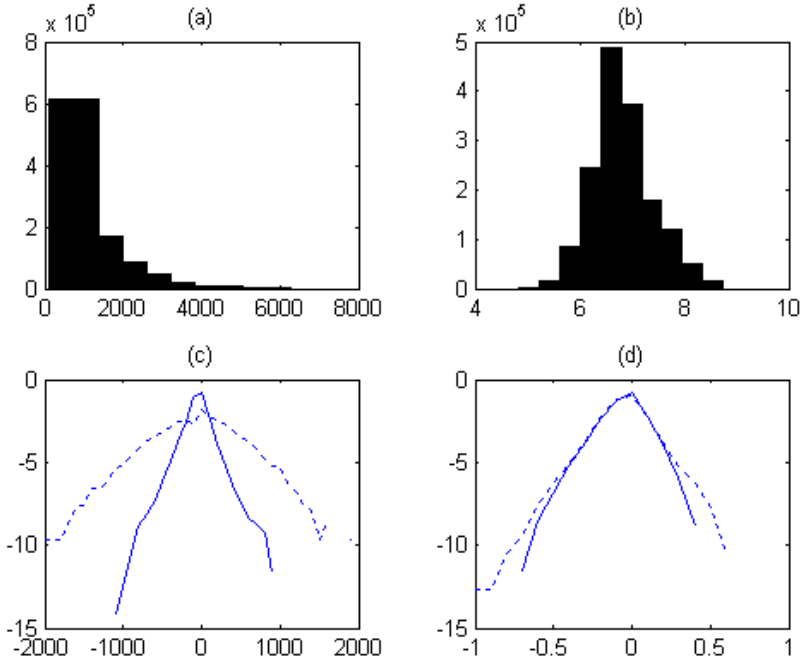

Figure 1.12: (a) Histogram of gray level values. (b) Histogram of the logarithm of gray level values. (c) Logarithm of histogram of $\nabla_x I$. Solid curve: gray level values < 2000, dotted curve: gray level values > 4000. (d) Logarithm of histogram of $\nabla_x \log I$. Solid curve: gray level values < 6.4, dotted curve: gray level values > 7.2.

1.12(b) shows that the histogram of the logarithm of gray level values is bell-shaped. In our experiment, we define two neighboring pixels as connected if their absolute difference is smaller than a fixed constant k. Under this setup, we hope to see that the percentage of two neighboring pixels being connected stays roughly the same regardless of the gray level values. Therefore, we compare the distribution of ∇_x on pixels of low gray level values with that on pixels of high gray level values. We calculate ∇_x when both the gray level values of neighbor pixels are less than 2000, and when both are greater than 6000, respectively. The result presented in Figure-1.12(c) shows that ∇_x is much larger when both neighbor pixels are of high gray level values. We do the same calculations to the logarithm of the image, and the result presented in Figure-1.12(d) shows that, ∇_x from different gray level values are close

to each other. Therefore, throughout our experiment, we use the logarithm of images instead of original images.

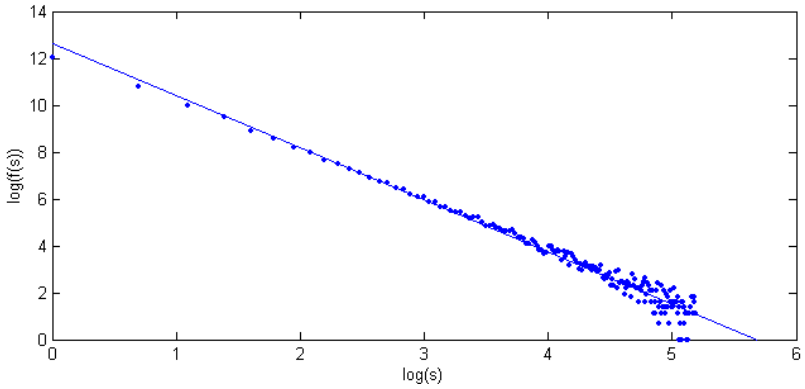

Figure 1.13: The distribution of area of image-1 and $k = 5$. Both x-axis and y-axis are in logarithm scale.

We take the logarithm of each image I, then adjust the logarithms of gray level values to integers between 0 and 255. Explicitly, the new image J is

$$J = [\log(I+1) \cdot \frac{255}{\max \log(I+1) - \min \log(I+1)}].$$

Figure-1.13 shows the result of images presented in Figure-1.2, and using $k = 5$. We can see that points of $(\log(s), \log(f(s)))$ distribute very close to a line. A linear regression model $\log(f(s)) = A - \alpha \log(s)$ is performed, and it gives that $\hat{A} = 12.62$ and $\hat{\alpha} = -2.22$.

From Figure-1.13, we also observe the following. First, our linear model fits better when $\log(s)$ is smaller than when $\log(s)$ is larger. Second, most of $(\log(s), \log(f(s)))$ lie under the fitted regression line when $\log(s)$ is smaller, while $(\log(s), \log(f(s)))$ distribute both above and below the fitted regression line when $\log(s)$ is larger. These observations can be explained by the fact that we have less data of larger $\log(s)$. It is possible that, once the data of larger values of $\log(s)$'s are removed, we can

	regression on S				regression on S'			
k	5	6	7	8	5	6	7	8
Picture-1	2.22	2.17	2.13	2.21	2.03	2.04	2.07	2.17
Picture-2	2.32	2.35	2.27	2.20	2.15	2.13	2.07	2.02
Picture-3	2.26	2.24	2.14	2.11	2.42	2.31	2.22	2.19
Picture-4	2.30	2.12	2.07	2.02	2.23	2.14	2.12	2.11
Picture-5	2.12	2.09	2.33	2.41	1.96	1.93	2.07	2.17
Picture-6	2.58	2.30	2.19	2.08	2.09	1.97	1.91	1.91
Picture-7	2.32	2.46	2.57	2.59	2.40	2.43	2.53	2.48
Picture-8	2.05	1.97	1.99	1.93	2.38	2.37	2.37	2.38
Picture-9	2.43	2.32	2.30	2.26	2.42	2.32	2.25	2.19

Table 1.1: regression coefficients α of images from Figure 1.1 for different k's

improve the over-estimation problem of our linear model for smaller values of $\log(s)$'s. Therefore, in addition to a linear regression on

$$S = \{(\log(s), \log(f(s))), 0 \leq s \leq T_{\max}\}$$

we also perform a linear regression on

$$S' = \{(\log(s), \log(f(s))), 0 \leq s \leq 15\}.$$

Table-1.1 presents the regression coefficients α from the regression model $(\log(f(s))) = A - \alpha \log(s)$ on S and S', which are calculated from nine images shown in Figure-1.1, with k =5, 6, 7, 8, respectively. It show that α is indeed close to 2.

1.3.3 The flat-earth explanation for object scaling

Scaling of physical objects

If the size of physical objects in the 3-D world follows the $1/r^3$ rule, then the size of objects in the image will also follow this rule. This is a strong assumption, and there is no evidence to support this assumption.

The effect of distance to the lens

We claim that the distance of objects is the key factor that the size of objects in a image follows the $1/r^3$ rule. In the following, we start with building up a "flat-earth" model (proposed by Chi [1]), then we derive the distribution of the distance of objects based on this model. With this distribution of distance, then we explain the property of scale invariance of object size.

Derivation of the flat-earth model:

Assume that the world is flat and objects are placed randomly not too far from the ground. Explicitly, all objects are placed in the region $V = \mathbf{R}^+ \times \mathbf{R} \times [-H_1, H_2]$, where $\{z = -H_1\}$ is the ground, and $\{z = H_2\}$ is the upper bound plane. Assume that $H_2 > H_1$, which means the distance from the lens to the upper bound plane is larger than that to the ground. An ideal pinhole lens is placed on the origin, and an image film is placed on $-1 \times [-l_1, l_1] \times [-l_2, l_2]$. The centers of objects are distributed by a homogeneous Poisson process with density parameter μ, where μ is the expected number of occurrences in any region of volume 1. For each location (x_i, y_i, z_i) generated from the Poisson process, an object is randomly sampled from a fixed distribution $\mathcal{O}$. All objects in $\mathcal{O}$ are planar templates with different sizes and shapes. The object sampled is placed on the plane $x = x_i$ and centered at (x_i, y_i, z_i). If an object o is not entirely in V, we only consider the part of object inside V, which is $o \bigcap V$.

The previous paragraph describes the setup of our flat-earth model. Our goal is to derive the density function of the distances of objects to the lens. More explicitly, each pixel in the image film represents one part of an object, the ground, or the sky. We want to calculate the distribution of the distance from the lens to the corresponding object. The distance here means the distance in x-direction, or the projection of the distance to the x-axis.

Mathematically, we define a random variable X as follows: Randomly choose a pixel from the image film. Let a ray start from this pixel, go through the origin, and

go on until it hits something. Let X be the x-distance from the lens to the object or the ground which the ray hits. $X = \infty$ if the ray does not hit anything. Our goal is to calculate the density function $f(x)$ for $X = x < \infty$, and $\Pr(X = \infty)$. Let $d(i, j)$ be the x-distance from the lens to the corresponding object of pixel (i, j), and $\tilde{d}(y, z)$, the continuous version of d, be that of $(-1, y, z)$ in the image film. Then

$$\begin{aligned} f(x)dx &= \Pr(x \le X < x + dx) \\ &= \frac{\text{expected number of}\{(i,j)|x \le d(i,j) < x + dx\}}{\text{total pixels in the image film}} \\ &= \frac{\int_{-l_2}^{l_2} \int_{-l_1}^{l_1} 1_{(x \le \tilde{d}(y,z) < x+dx)} dydz}{4l_1 l_2} \end{aligned}$$

Therefore, it suffices to calculate the expected area in the image film where $x \le \tilde{d}(y, z) < x + dx$.

When $x \le \tilde{d}(y, z) < x+dx$ for some (y, z) in the image film, the ray from $(-1, y, z)$ through the lens (the origin) hits either an object or the ground at x-distance between x and $x + dx$. If the ray hits an object, there is an object at $(s, -sy, -sz)$ for some $s \in [x, x + dx)$, and there is no object at $(s, -sy, -sz)$ for all $s < x$. If the ray hits the ground, the location where the ground is hit is $(\frac{H_1}{z}, \frac{-H_1 \cdot y}{z}, -H_1)$. Moreover, there is no object at $(s, -sy, -sz)$ for all $s < \frac{H_1}{z}$, and $x \le \frac{H_1}{z} < x + dx$. In both scenarios as mentioned above, we need the probability that there is no object along the ray from the origin to any fixed point. This probability clearly depends on the density of objects.

For any $(x, y, z) \in V$, let r be the ray from (x, y, z) to $(x + 1, \frac{x+1}{x}y, \frac{x+1}{x}z)$, and let λ be the expected number of objects which the ray r intersects. Suppose that all objects are of the same size and shape as some fixed $o \in \mathcal{O}$. Since o is planar, we can describe it as $o \subset \mathbf{R^2}$, and place its center at the origin. On any x-plane, an object centered at (x, u, v) covers the point (x, y, z), if $(y - u, z - v) \in o$. Define

$$-o = \{(-u, -v)|(u, v) \in o\}.$$

Then, for any $(x, y, z) \in V$, an object will cover (x, y, z) if its center falls in

$$-o + (x, y, z) \equiv \{(x, y - u, z - v) | (u, v) \in o\}.$$

Therefore, the ray r intersects an object if the center of the object falls in

$$r - o \equiv \{(s, \frac{s}{x}y - u, \frac{s}{x}z - v) | x \leq s < x + 1 \text{ and } (u, v) \in o\}.$$

The volume of $r - o$ is $|o| \times 1$. Along with the assumption that centers of objects are distributed by a homogeneous Poisson process with density parameter μ, where μ is the expected number of occurrences in any region of volume 1, the expected number of centers in $r - o$ is $\mu \times |o|$. This is the expected number of objects which r intersects. We can run the same argument to any object $o \in \mathcal{O}$, then

$$\begin{aligned} \lambda &= \int \mu \cdot |o| \cdot dm(o) \\ &= \mu \cdot \int |o| \cdot dm(o), \end{aligned}$$

where m is the probability measure on $\mathcal{O}$. $\int |o| \cdot dm(o)$ is the mean areas of objects.

Let $G(x, y, z)$ be the probability that there is no object at (sx, sy, sz) for all $0 < s < 1$. Let v be a vector (x_1, y_1, z_1), and $G_v(x) \equiv G(x, x \cdot \frac{y_1}{x_1}, x \cdot \frac{z_1}{x_1})$. Then, $-dG_v(x) = G_v(x) - G_v(x + dx)$ is the probability that there is no object in the direction v before $(x, x \cdot \frac{y_1}{x_1}, x \cdot \frac{z_1}{x_1})$, and there are some objects between $(x, x \cdot \frac{y_1}{x_1}, x \cdot \frac{z_1}{x_1})$ and $(x + dx, (x + dx) \cdot \frac{y_1}{x_1}, (x + dx) \cdot \frac{z_1}{x_1})$. On the other hand, consider r as the ray from $(x, x \cdot \frac{y_1}{x_1}, x \cdot \frac{z_1}{x_1})$ to $(x + dx, (x + dx) \cdot \frac{y_1}{x_1}, (x + dx) \cdot \frac{z_1}{x_1})$, the expected number of objects which r intersects is $\frac{\lambda}{x_1} dx$. r is a short ray with length $|v| \cdot dx$. Since objects are distributed by a homogeneous Poisson process, the probability that r intersects more than one object is $o(\cdot dx)$. Therefore, the probability that r intersects some objects is $\lambda \cdot dx$. We have

$$-dG_v(x) = G_v(x) \cdot \lambda \cdot dx.$$

Along with the fact $G_v(0) = 1$, the solution of the differential equation above is $G_v(x) = e^{-\lambda x}$. Let $x = x_1$, we have

$$\begin{aligned} G(x_1, y_1, z_1) &= G_v(x_1) \\ &= e^{-\lambda x_1}. \end{aligned}$$

As discussed earlier, if $\tilde{d} = x$ for some point in the image film, the ray from that point through the lens will hit either an object or the ground at x-distance x. We decompose $f(x)$ into two parts. One is f_1 when the ray hits an object, and the other is $f_2(x)$ is when it hits the ground. In the following, we will calculate $f_1(x)$ and $f_2(x)$ separately. Then we have $f(x)$ as $f_1(x) + f_2(x)$.

Consider the $y - z$ plane at distance x. The region of this plane that is viewable in the image film is

$$x \times [-l_1 x, l_1 x] \times [\max(-H_1, -l_2 x), \min(H_2, l_2 x)].$$

The area of this region is

$$\begin{aligned} & (l_1 x - (-l_1 x))(\min(H_2, l_2 x) - \max(-H_1, -l_2 x)) \\ = \; & 2l_1 x(\min(H_1, l_2 x) + \min(H_2, l_2 x)) \\ = \; & 2l_1 l_2 x(\min(\frac{H_1}{l_2}, x) + \min(\frac{H_2}{l_2}, x)). \end{aligned}$$

The expected number of objects in

$$[x, x + dx) \times [-l_1 x, l_1 x] \times [\max(-H_1, -l_2 x), \min(H_2, l_2 x)]$$

is

$$2l_1 l_2 x(\min(\frac{H_1}{l_2}, x) + \min(\frac{H_2}{l_2}, x)) \times \mu dx.$$

Since the projection of a unit square at distance x onto the image is a square with

area $\frac{1}{x^2}$, the expected area of these objects in the image film ignoring the occlusion effect and the boundary effect[1] is

$$\begin{aligned} & 2l_1l_2x(\min(\frac{H_1}{l_2},x)+\min(\frac{H_2}{l_2},x))\times\lambda dx/x^2 \\ = \quad & 2l_1l_2\lambda(\min(\frac{H_1}{xl_2},1)+\min(\frac{H_2}{xl_2},1))dx. \end{aligned}$$

Then $f_1(x)dx$ is the amount above times the occlusion effect $G(x,y,z)=e^{-\lambda x}$, divided by the total area of image film $4l_1l_2$. Recall that $H_1 < H_2$. We have

$$f_1(x)=\begin{cases} \lambda e^{-\lambda x} & x < \frac{H_1}{l_2} \\ \frac{1}{2}(\frac{H_1}{xl_2}+1)\lambda e^{-\lambda x} & \frac{H_1}{l_2} \le x < \frac{H_2}{l_2} \\ \frac{1}{2}\frac{H_1+H_2}{xl_2}\lambda e^{-\lambda x} & \frac{H_2}{l_2} \le x \end{cases}.$$

Next, we will derive $f_2(x)$. While objects from x-distance between x and $x+dx$ can appear at any location in the image film, the ground from x-distance between x and $x+dx$ has to appear in the region

$$-1\times[-l_1,l_1]\times[\frac{H_1}{x+dx},\frac{H_1}{x}]$$

of the image film. The area of this region is

$$\begin{aligned} & 2l_1\times(\frac{1}{x}-\frac{1}{x+dx})H_1 \\ = \quad & 2l_1\frac{H_1}{x(x+dx)}dx. \end{aligned}$$

[1]Some objects may exceed the boundary. Therefore, the expected area is smaller than the product of the expected number of objects and the mean area of objects. However, if we also distribute objects outside the boundary, some objects centered outside the boundary will have parts falling inside the boundary. Furthermore, if we distribute the outside objects with the same density as the inside objects, the expected outside area from objects centered inside the boundary is equal to the expected inside area from objects centered outside the boundary. Therefore, we have the expected area is exactly the same as the product of the expected number of objects and the mean area of objects.

Then $f_2(x)dx$ is this area times the occlusion effect $G(x,y,z) = e^{-\lambda x}$, divided by he total area of image film $4l_1l_2$. Therefore,

$$f_2(x) = \frac{H_1}{2x^2 l_2} e^{-\lambda x}.$$

The above is true only when $x > \frac{H_1}{l_2}$, since the projection of the ground at $x-$distance smaller than $\frac{H_1}{l_2}$ can not appear in the image film.

Combine both results of $f_1(x)$ and $f_2(x)$, we have

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x < \frac{H_1}{l_2} \\ \frac{1}{2}(\lambda + \frac{H_1\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x} & \frac{H_1}{l_2} \leq x < \frac{H_2}{l_2} \\ \frac{1}{2}(\frac{H_1\lambda}{xl_2} + \frac{H_2\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x} & \frac{H_2}{l_2} \leq x \end{cases}.$$

Last, we calculate $\Pr(X = \infty)$. When $X = \infty$, the ray from a pixel through the origin does not hit anything. We can pretend that there is a ceiling $z = H_2$, and every ray which does not hit anything will hit somewhere in the ceiling. Let $f_3(x)$ be the density function that $\tilde{d}(y,z) = \infty$ and the ray hits the ceiling at x-distance x. Using the same calculation as the derivation of $f_2(x)$, we have

$$f_3(x) = \frac{H_2}{2x^2 l_2} e^{-\lambda x}.$$

for $x > \frac{H_2}{l_2}$. Therefore,

$$\begin{aligned} \Pr(X = \infty) &= \int_{\frac{H_2}{l_2}}^{\infty} f_3(x)dx \\ &= \frac{H_2}{2l_2} \int_{\frac{H_2}{l_2}}^{\infty} \frac{1}{x^2} e^{-\lambda x} dx. \end{aligned}$$

To summarize, under the flat-world model, the distribution of distance is

$$
\begin{aligned}
f(x) &= \begin{cases} \lambda e^{-\lambda x} & x < \frac{H_1}{l_2} \\ \frac{1}{2}(\lambda + \frac{H_1\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x} & \frac{H_1}{l_2} \le x < \frac{H_2}{l_2} \\ \frac{1}{2}(\frac{H_1\lambda}{xl_2} + \frac{H_2\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x} & \frac{H_2}{l_2} \le x \end{cases} \\
\Pr(X=\infty) &= \frac{H_2}{2l_2}\int_{\frac{H_2}{l_2}}^{\infty} \frac{1}{x^2} e^{-\lambda x} dx.
\end{aligned}
$$

We can double check that $\int_0^\infty f(x)dx = 1 - \Pr(X=\infty)$:

$$
\begin{aligned}
& \int_0^\infty f(x)dx \\
= & \int_0^{\frac{H_1}{l_2}} \lambda e^{-\lambda x}dx + \int_{\frac{H_1}{l_2}}^{\frac{H_2}{l_2}} \frac{1}{2}(\lambda + \frac{H_1\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x}dx \\
& + \int_{\frac{H_2}{l_2}}^{\infty} \frac{1}{2}(\frac{H_1\lambda}{xl_2} + \frac{H_2\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x}dx \\
= & \frac{1}{2}\int_0^{\frac{H_1}{l_2}} \lambda e^{-\lambda x}dx + \frac{1}{2}\int_{\frac{H_1}{l_2}}^{\infty} (\frac{H_1\lambda}{xl_2} + \frac{H_1}{x^2 l_2})e^{-\lambda x}dx \\
& + \frac{1}{2}\int_0^{\frac{H_2}{l_2}} \lambda e^{-\lambda x}dx + \frac{1}{2}\int_{\frac{H_2}{l_2}}^{\infty} \frac{H_2\lambda}{xl_2}\lambda e^{-\lambda x}dx \\
= & -\frac{1}{2}e^{-\lambda x}|_{x=0}^{\frac{H_1}{l2}} - \frac{H_2}{2xl_2}e^{-\lambda x}|_{x=\frac{H_1}{l2}}^{\infty} + \int_{\frac{H_1}{2l_2}}^{\infty} e^{-\lambda x}(-\frac{H_1}{2x^2 l_2})dx + \int_{\frac{H_1}{2l_2}}^{\infty} \frac{H_1}{2x^2 l_2}e^{-\lambda x}dx \\
& -\frac{1}{2}e^{-\lambda x}|_{x=0}^{\frac{H_2}{l2}} - \frac{H_2}{2xl_2}e^{-\lambda x}|_{x=\frac{H_2}{2l_2}}^{\infty} + \int_{\frac{H_2}{2l_2}}^{\infty} e^{-\lambda x}(-\frac{H_2}{2x^2 l_2})dx \\
= & -\frac{1}{2}(e^{-\lambda\frac{H_1}{l_2}} - 1) + \frac{1}{2}(e^{-\lambda\frac{H_1}{l_2}} - 0) \\
& -\frac{1}{2}(e^{-\lambda\frac{H_2}{l_2}} - 1) + \frac{1}{2}(e^{-\lambda\frac{H_2}{l_2}} - 0) - \int_{\frac{H_1}{2l_2}}^{\infty} \frac{H_2}{2x^2 l_2}e^{-\lambda x}dx \\
= & 1 - \int_{\frac{H_1}{2l_2}}^{\infty} \frac{H_2}{2x^2 l_2}e^{-\lambda x}dx \\
= & 1 - \Pr(X=\infty)
\end{aligned}
$$

Evidence for flat-earth model:

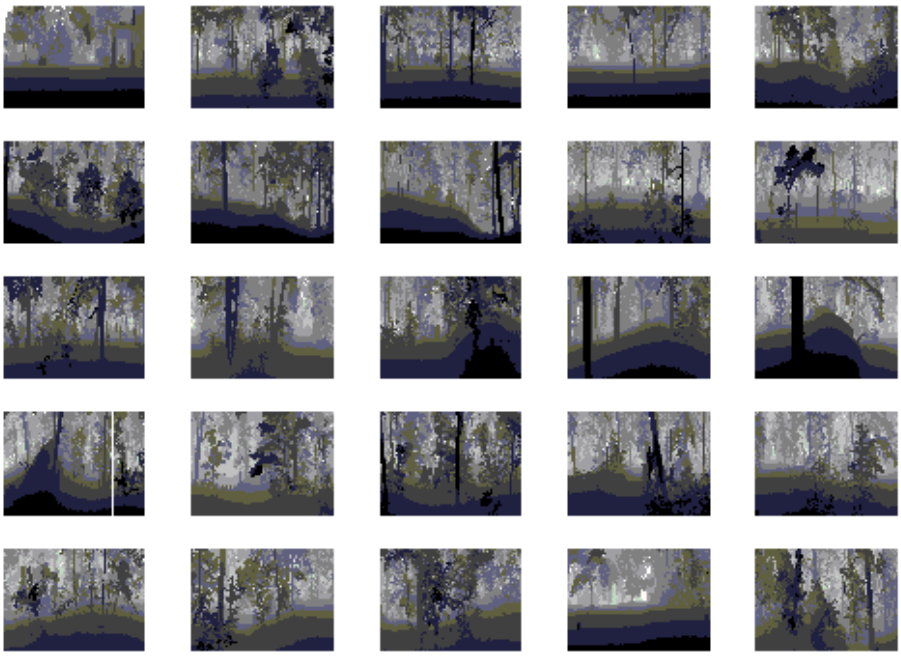

Figure 1.14: log(range+1) for images from "Brown Range Image Database"

We use "Brown Range Image Database" to test whether the flat-earth model is a better fit to the real world. This database contains 197 range images collected by Ann Lee and Jinggang Huang. Some preliminary analysis on these images were presented in [20]. The images have been collected with a laser range-finder with a rotating mirror (3D imaging sensor LMS-Z210 by Riegl). Each image contains 444x1440 measurements with an angular separation of 0.18 deg. The field of view is thus 80 degrees vertically and 259 degrees horizontally. Each measurement is calculated from the time of flight of the laser beam. The operational range of the sensor is typically 2-200m. The laser wavelength of the range-finder is 0.9 mu m, which is in the near infra-red region. The data set consists of images which can be categorized as "forest", "residential", and "interior" scenes. We use all twenty-five images categorized as forest. Fig-1.14 shows these twenty-five images.

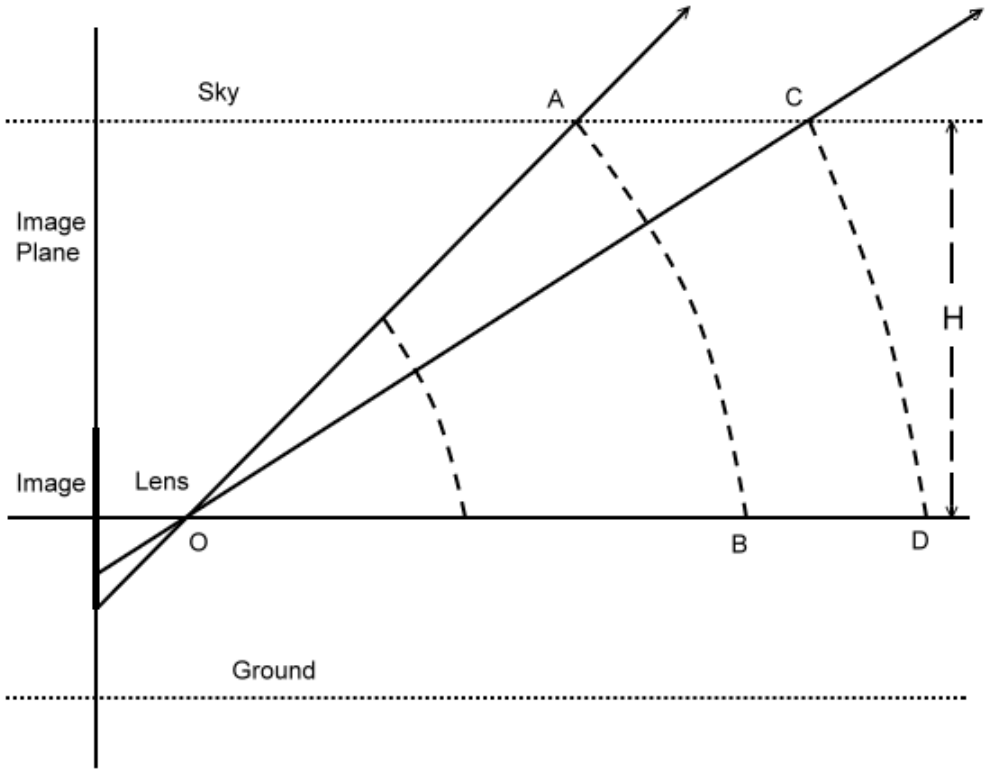

Figure 1.15:

The image matrices in this database were created uniformly in angle, and the range is the actual distance ρ, instead of the x-distance. In order to test the flat-earth model with this data, we need to re-derive the distribution, this time for ρ rather than x. Here, we only consider the upper half of these images, so that there is no ground effect. Since the field of view of the whole image is 80 degrees vertically, that of the upper half is 40 degrees. Then, in Fig-1.15, the angle of AOB is 40°. Let ρ_1 = the length of $\overline{OA}$ and ρ_2 be that of $\overline{OC}$. Ignoring the occlusion effect, the ratio of the expected area in the image film coming from distance $/rho_2$ to that from distance $/rho_1$ is the ratio of the angle of COD to that of AOB. Since the angle of COD is $\arcsin(\frac{H}{\overline{0C}})$, the ratio is $\frac{\arcsin(\frac{H}{\overline{0C}})}{40°}$. When $\rho \leq \rho_1 = \frac{H}{\sin(40°)}$, the expected area coming from distance ρ is a constant. In other words, the ratio to that from distance

ρ_1 is 1. Along with the occlusion effect $e^{-\rho\lambda}$, we have the new density function f as

$$\begin{aligned} f(\rho) &\propto \begin{cases} e^{-\rho\lambda} & \text{if } \rho \leq \frac{H}{\sin(40^\circ)} \\ \frac{\arcsin(\frac{H}{\rho})}{40^\circ} e^{-\rho\lambda} & \text{if } \rho > \frac{H}{\sin(40^\circ)} \end{cases} \\ &\propto \arcsin(\min(\sin(40^\circ), \frac{H}{\rho})) \cdot e^{-\rho\lambda}. \end{aligned}$$

In the 3D-world model, the expected area coming from any distance ρ should be a constant if the occlusion effect is ignored. With the occlusion effect, we have $f_{3D}(\rho) = \lambda e^{-\rho\lambda}$ for any ρ. Note that, when ρ is large, $\arcsin \frac{H}{\rho} \approx \frac{H}{\rho}$. Therefore, $f(\rho) \propto \frac{e^{-\rho\lambda}}{\rho}$ for large ρ.

We calculate the empirical distribution of the distance of objects from these range images. The result is presented in Figure-1.16. There are very few objects placed near the camera. When $\rho < 3.2$ (meters), the number of objects decreases as ρ approaches to zero. This contradicts to both the density function derived from the flat-world model and that from the 3D-world model. However, this observation is not surprising since objects near the camera are very likely avoided when pictures are taken. In addition, we also found that the curve has two shapes with a break point at around $\rho = 20$. The curve decreases with a faster rate for $\rho > 20$. This observation is expected for the flat-world model, as the density function has different forms for $r \leq \frac{H}{\sin(40^\circ)}$ and $r > \frac{H}{\sin(40^\circ)}$. However, this observation is not expected for the 3D-world model.

In the following, we fit the curve in Figure-1.16 using the 3D-world model and the flat-world model, respectively. As objects near the camera are very likely avoided when pictures are taken, we only use the part of the data where $\rho > 10$. As shown in Figure-1.17, the dotted line is the fitted curve using the 3D-world model, and the dashed line is the one using the flat-world model. For the 3D-world model, the conditional density function is

$$f_{3D}(\rho) = \lambda e^{-(\rho-10)\lambda},$$

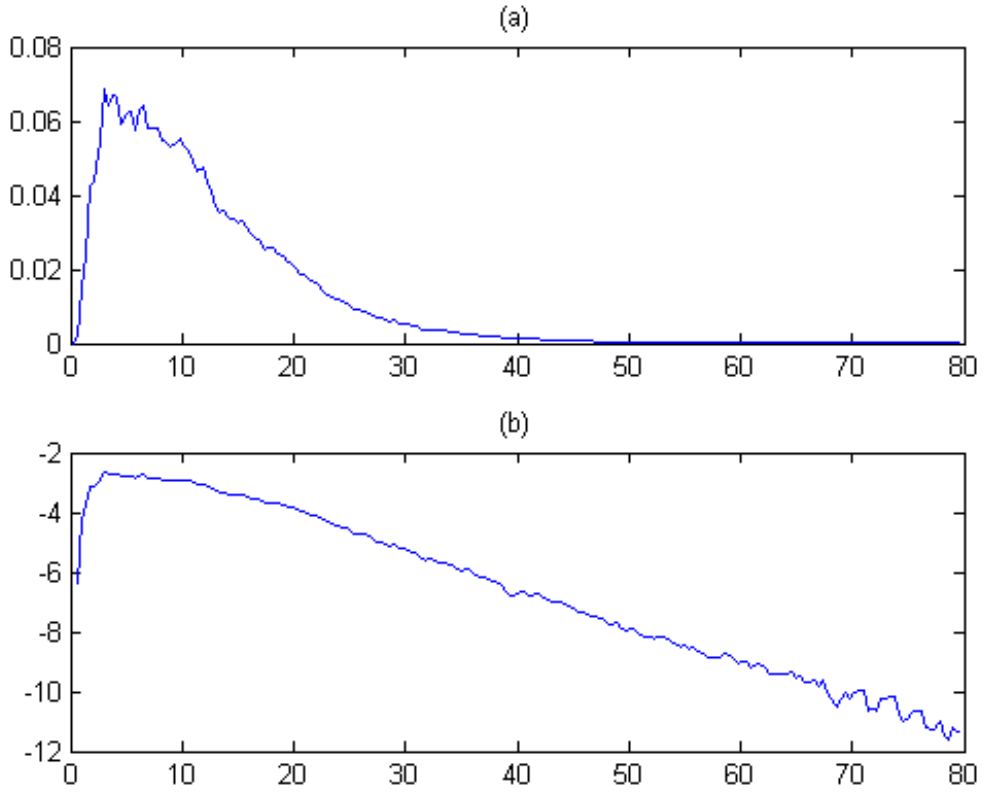

Figure 1.16: (a) the distribution of the distance of objects. (b) the log distribution of the distance of objects

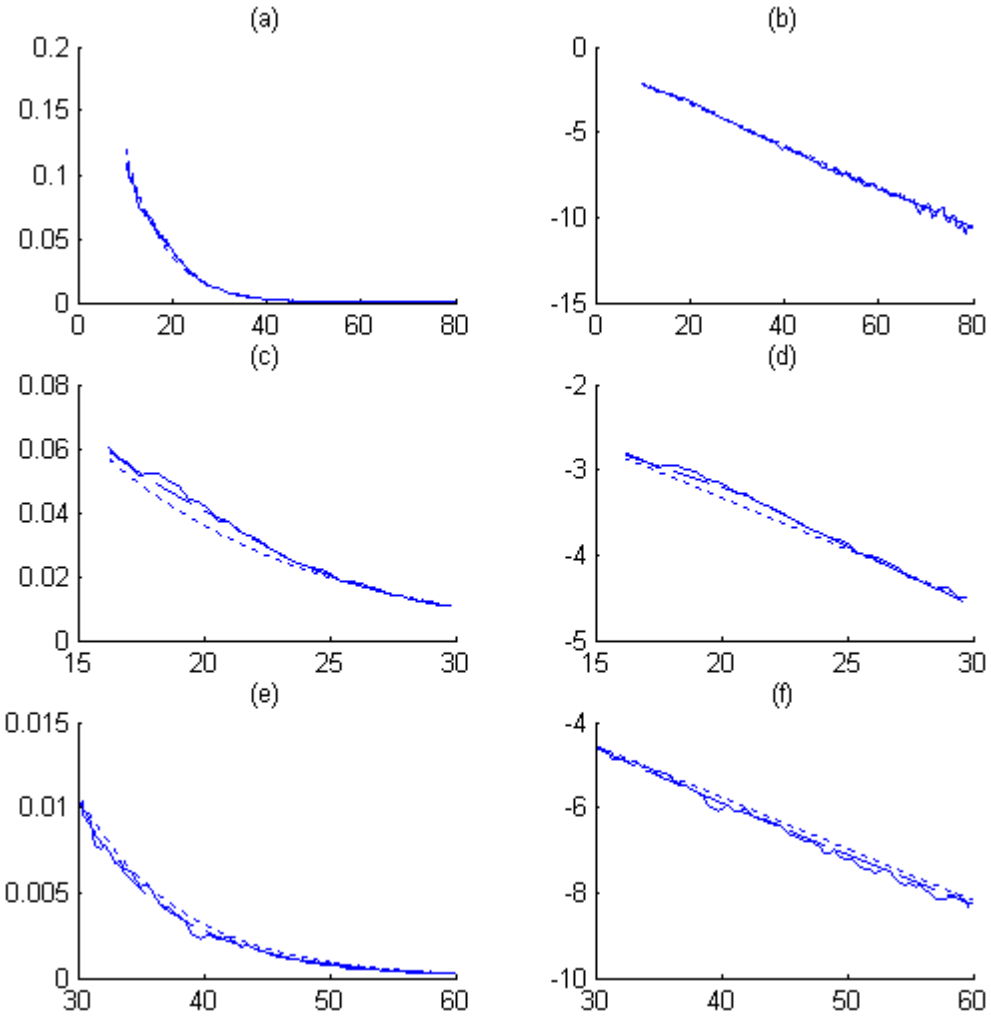

Figure 1.17: Solid curve: the empirical conditional density function. Dotted curve: the fitted curve to the 3D-world model. Dashed curve: the fitted curve to the flat-world model. (a) original scale (b) logarithm scale in y-axes (c) original scale restricted to the domain $16 < \rho < 30$ (d) logarithm scale in y-axis, and the domain restricted to $16 < \rho < 30$ (e) original scale restricted to the domain $30 < \rho < 60$ (f) logarithm scale in y-axis, and the domain restricted to $30 < \rho < 60$.

and the maximum likelihood estimate is $\hat{\lambda}$ is 0.112. For the flat-world model, the conditional density function is

$$f(\rho) = c \cdot \arcsin(\min(\sin(40^\circ), \frac{H}{\rho})) \cdot e^{-\rho\lambda},$$

and the maximum likelihood estimates are $\hat{H}$=13.50 and $\hat{\lambda} = 0.099$. Figure 1.17 shows that, the dashed curve fitted by the flat-world model not only is closer to the empirical curve, but also capture its shape (a break point around $\frac{h}{sin(40^\circ)} \approx 21$). Note: The flat-earth model should have a better fit since it has one more parameter than the 3D-world model does.

From the flat-earth model to the $1/r^3$ rule:

$1/r^3$ rule says that, the density function of the size of objects in an invariant image is proportional to $1/r^3$. Instead of calculating the expected numbers of objects with size r, it is easier to calculate the expected total area of objects with size r. Let $g(r)$ be the density function of objects with size r. For simplicity, we define the object size r as the square root of the object area. If

$$g(r) \propto 1/r^3,$$

we have

$$\int_a^b r^2 g(r) dr = \int_{2a}^{2b} r^2 g(r) dr,$$

for any $0 < a < b$. Furthermore, the converse is also ture. That is, statements in the following are equivalent.

1. The size of objects follows $1/r^3$ rule.
2. The total area of objects with sizes between a and b is the same as that between $2a$ and $2b$

In the following, we will show that (2) is asymptotically true under our flat-earth

model.

First, we assume that all physical objects in the 3-D world are of the same size, say, 1. (This will be generalized shortly.) Recall that we define object size as the square root of object area. Therefore, all physical objects of size 1 means all their area in 3-D world is 1. An object from distance x has area $\frac{1}{x^2}$ in the image. Then its size r in the image is $\frac{1}{x}$. Recall that the density function of the distance to objects is

$$f_1(x) \propto \begin{cases} 4l_1 l_2 \lambda e^{-\lambda x} & x < \frac{H_1}{l_2} \\ 2l_1(\frac{H_1}{x} + l_2)\lambda e^{-\lambda x} & \frac{H_1}{l_2} \leq x < \frac{H_2}{l_2} \\ 2l_1 \frac{H_1+H_2}{x} \lambda e^{-\lambda x} & \frac{H_2}{l_2} \leq x \end{cases}.$$

Note that $x \geq \frac{H_2}{l_2}$ is equivalent to $r \leq \frac{l_2}{H_2}$. Therefore, if $a < b \leq \frac{l_2}{H_2}$, the total area of objects with sizes between a and b is

$$N(a, b, \lambda) = \int_a^b f_1(x)dx = C \cdot \int_a^b \frac{1}{x} \lambda e^{-\lambda x} dx.$$

Similarly, if $2a < 2b \leq \frac{l_2}{H_2}$,

$$N(2a, 2b, \lambda) = \int_{2a}^{2b} f_1(x)dx = C \cdot \int_{2a}^{2b} \frac{1}{x} \lambda e^{-\lambda x} dx.$$

Since $e^{-\lambda x} \approx 1$ when λ is small enough, we expect that $N(a, b, \lambda) \approx N(2a, 2b, \lambda)$ when λ is small enough. Precisely, when $0 < a < b \leq \frac{l_2}{2H_2}$, for any $\epsilon > 0$, there exists $\lambda_0 > 0$, such that $e^{-2\lambda b} > 1 - \epsilon$ for all $\lambda < \lambda_0$. Therefore, we have

$$\begin{aligned} (1-\epsilon) \cdot C \cdot \int_a^b \frac{\lambda}{x} dx &< N(a, b, \lambda) < C \cdot \int_a^b \frac{\lambda}{x} dx \\ (1-\epsilon) \cdot C \cdot \int_{2a}^{2b} \frac{\lambda}{x} dx &< N(2a, 2b, \lambda) < C \cdot \int_{2a}^{2b} \frac{\lambda}{x} dx. \end{aligned}$$

Along with the fact

$$\int_a^b \frac{\lambda}{x} dx = \int_{2a}^{2b} \frac{\lambda}{x} dx,$$

we have

$$1-\epsilon < \frac{N(a,b,\lambda)}{N(2a,2b,\lambda)} < \frac{1}{1-\epsilon}.$$

That is,

$$\lim_{\lambda\to 0} \frac{N(a,b,\lambda)}{N(2a,2b,\lambda)} = 1.$$

Now, instead of assuming that all physical objects in 3-D are of size 1, we assume all are of size m. In this case, an object from distance x is of size $r = \frac{m}{x}$ in the image. Hence, the condition $x \geq \frac{H_2}{l_2}$ is now equivalent to $r \leq \frac{l_2 \cdot m}{H_2}$. We can run all the arguments again, and obtain a similar result as follows. For any $0 < a < b \leq \frac{l_2 \cdot m}{2H_2}$,

$$\lim_{\lambda\to 0} \frac{N(a,b,\lambda)}{N(2a,2b,\lambda)} = 1.$$

In general, assume that physical objects have sizes from a fixed distribution, and the minimal size is m_0. Then we prove the following Theorem.

Theorem 1.1 *For any* $0 < a < b \leq \frac{l_2 \cdot m_0}{2H_2}$,

$$\lim_{\lambda\to 0} \frac{N(a,b,\lambda)}{N(2a,2b,\lambda)} = 1$$

Proof:

For each size m, recall that

$$\begin{aligned}(1-\epsilon)\cdot C\cdot \int_a^b \frac{\lambda}{x}dx \quad &< N_m(a,b,\lambda) < \quad C\cdot \int_a^b \frac{\lambda}{x}dx \\ (1-\epsilon)\cdot C\cdot \int_{2a}^{2b} \frac{\lambda}{x}dx \quad &< N_m(2a,2b,\lambda) < \quad C\cdot \int_{2a}^{2b} \frac{\lambda}{x}dx.\end{aligned}$$

Since these inequalities are true for all m, we can integrate out m, and have

$$\begin{aligned}(1-\epsilon)\cdot C\cdot \int_a^b \frac{\lambda}{x}dx \quad &< N(a,b,\lambda) < \quad C\cdot \int_a^b \frac{\lambda}{x}dx \\ (1-\epsilon)\cdot C\cdot \int_{2a}^{2b} \frac{\lambda}{x}dx \quad &< N(2a,2b,\lambda) < \quad C\cdot \int_{2a}^{2b} \frac{\lambda}{x}dx.\end{aligned}$$

Therefore,

$$1-\epsilon < \frac{N(a,b,\lambda)}{N(2a,2b,\lambda)} < \frac{1}{1-\epsilon}$$

when $0 < a < b \leq \frac{l_2 \cdot m}{2H_2}$ for all m. And this condition is equivalent to $0 < a < b \leq \frac{l_2 \cdot m_0}{2H_2}$

□

Remark: This limit is $\frac{1}{2}$ for 3D-world model.

1.4 The scale invariance of local statistics

In this section, we explain the scale invariance of local statistics in natural images. We first explain it using $1/r^3$ rule, which was established based on the projection effect in the previous section. Then we find evidence showing that the projection effect might not be the key reason for natural images to scale well. Therefore, we proceed with proposing a new model. This new model provides three effects to explain the property of scale invariance. We also claim that, even though the projection effect may not be a major factor, it still plays a role in improving the property of scale invariance.

1.4.1 Explanation by the projection effect

In the previous section, we built up a model, and explained that $1/r^3$ rule is asymptotically true as λ approaches zero. In this model, the distance distribution based on the projection effect is the key that $1/r^3$ rule holds. Naturally, we want to investigate whether this model can also explain the scale invariance of local statistics in natural images.

In Ruderman [4], objects with random shapes and sizes are randomly placed in an infinite image plane, and each object is independently painted with a gray tone chosen from a distribution. It was argued that if these objects have a power-law distribution of sizes, images of the plane also follow a power law, which is related to scale invariance. Therefore, in the point of view of power-spectrum, the projection

effect can also be applied to explain scale invariance property.

In the following, we explain scale invariance in the point of view of another definition: marginal distribution stays the same after scaling. Suppose that S is a local statistic (or filter). Let $h(s)$ be the density function of S of the original image I, $h^{(k)}(s)$ be that of the scaled image $I^{(k)}$, $f(x)$ be the density function of the x-distance described in previous section, and $h(s|x)$ be the density function of S conditional on the x-distance x. Then

$$\begin{aligned} h^{(k)}(s) &= \int h^{(k)}(s|x)f(x)dx \\ &= \int h(s|x')f(\frac{x'}{k})\frac{dx'}{k} \qquad (x' = kx) \\ &\approx \int h(s|x')kf(x')\frac{dx'}{k} \qquad (f(x) \approx \frac{c}{x}) \\ &= \int h(s|x')f(x')dx' \\ &= h(s). \end{aligned}$$

Therefore, based on $f(x) \approx \frac{c}{x}$ derived from the projection effect, $h^{(k)}(s) \approx h(s)$, which represents scale invariance.

1.4.2 Evidence

In the previous subsection, we explained scale invariance of local statistics based on the distance distribution derived from the projection effect. However, we do not know whether this is the key factor to the scale invariance of local statistics. In the following, we will present some kinds of images which do not involve distances still have good scale invariance property. Therefore, the distance effect may not be the main reason that images have good scale invariance.

Texture images

To question whether the distance effect is the key factor to cause scale invariance, we explore features in images without the distance effect, and observe whether they

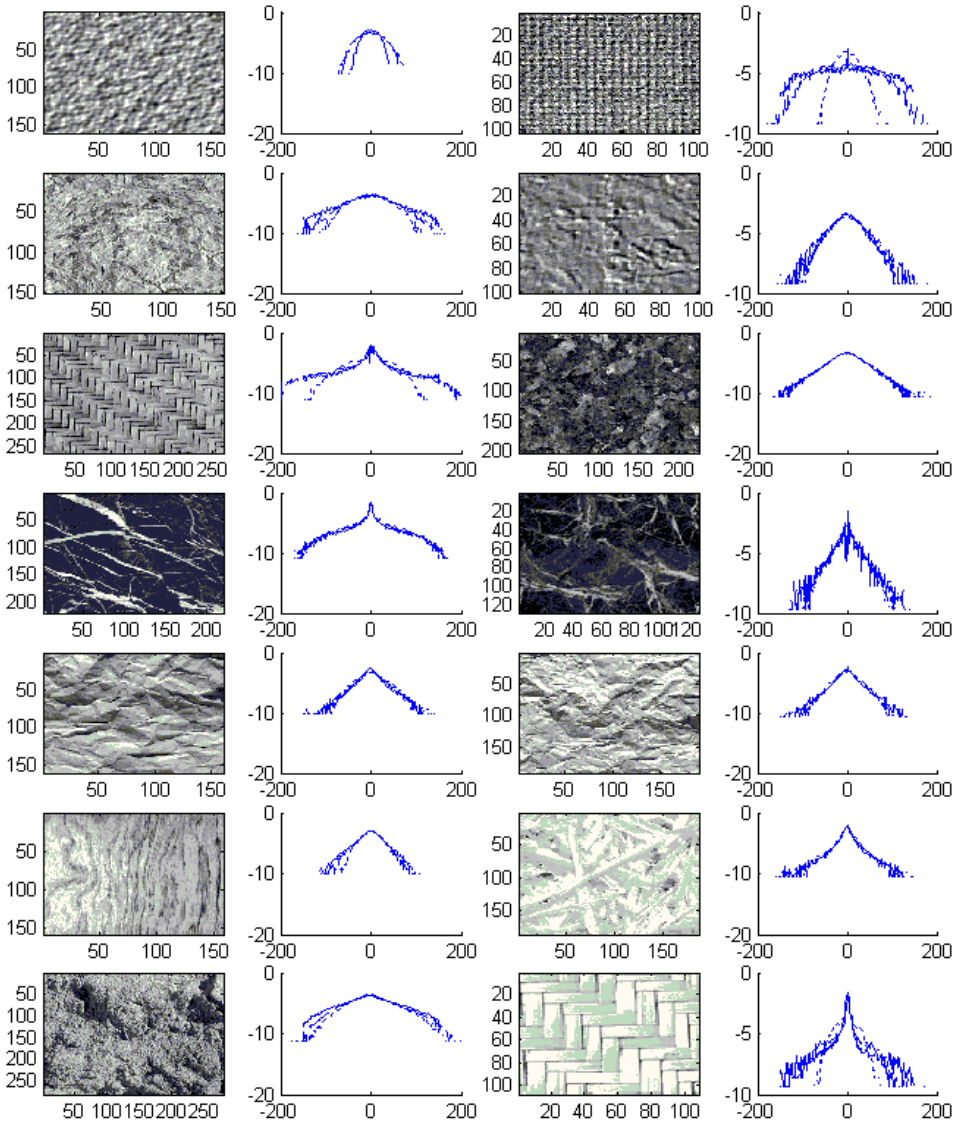

Figure 1.18: Logarithms of marginal distribution of ∇_x for texture images. Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

still scale well. For this purpose, we collect fourteen texture images from the internet.

We choose $S = \nabla_x$. The result is presented in Figure-1.18, where solid curves represent $\nabla_x I$, dashed curves represent $\nabla_x I^{(2)}$, and dotted curves represent $\nabla_x I^{(4)}$. Figure-1.18 shows that many texture images have good scale invariance property. Moreover, since these texture images do not involve the distance to the lens, this suggests that images without the distance effect can have scale invariance property.

Strips of images

In addition to images without the distance effect, we take horizontal strips out of an image, so that all distances of all pixels in a strip represent objects located at nearly the same distances to the lens.

Among nine pictures in Figure-1.1, we choose picture-2 and take horizontal strips. We take one from row 501 to row 600, and another from row 541 to row 550. Here S is chosen to be $\nabla_x I$, and the result is presented in Figure-1.19. This figure shows that strips of an image have even better scale invariance. The variation of x-distances to the lens is much smaller in strip images than that in the whole image. In strip images, the distance effect discussed in previous section do not provide a good explanation for the scale invariance. Therefore, this suggests that the distance effect may not be the key factor of scale invariance.

Simulated image from LOCO predictor

LOCO (LOw COmplexity LOssless COmpression) is currently the best lossless compression algorithm. In short, it predicts a pixel value according to the past information, and it codes the residue, which is the difference between the prediction and the real value. The predictor is the median of $I(i-1,j)$, $I(i,j-1)$, and $I(i-1,j)+I(i,j-1)-I(i-1,j-1)$, and the residue distribution is two-sided geometric distribution. The density function of TSGD(p) is

$$\frac{p}{2-p}q^{|x|-1} \qquad\qquad x = \ldots,-2,-1,0,1,2,\ldots,$$

where $p = 1-q$. The LOCO algorithm implicitly defines a distribution on images,

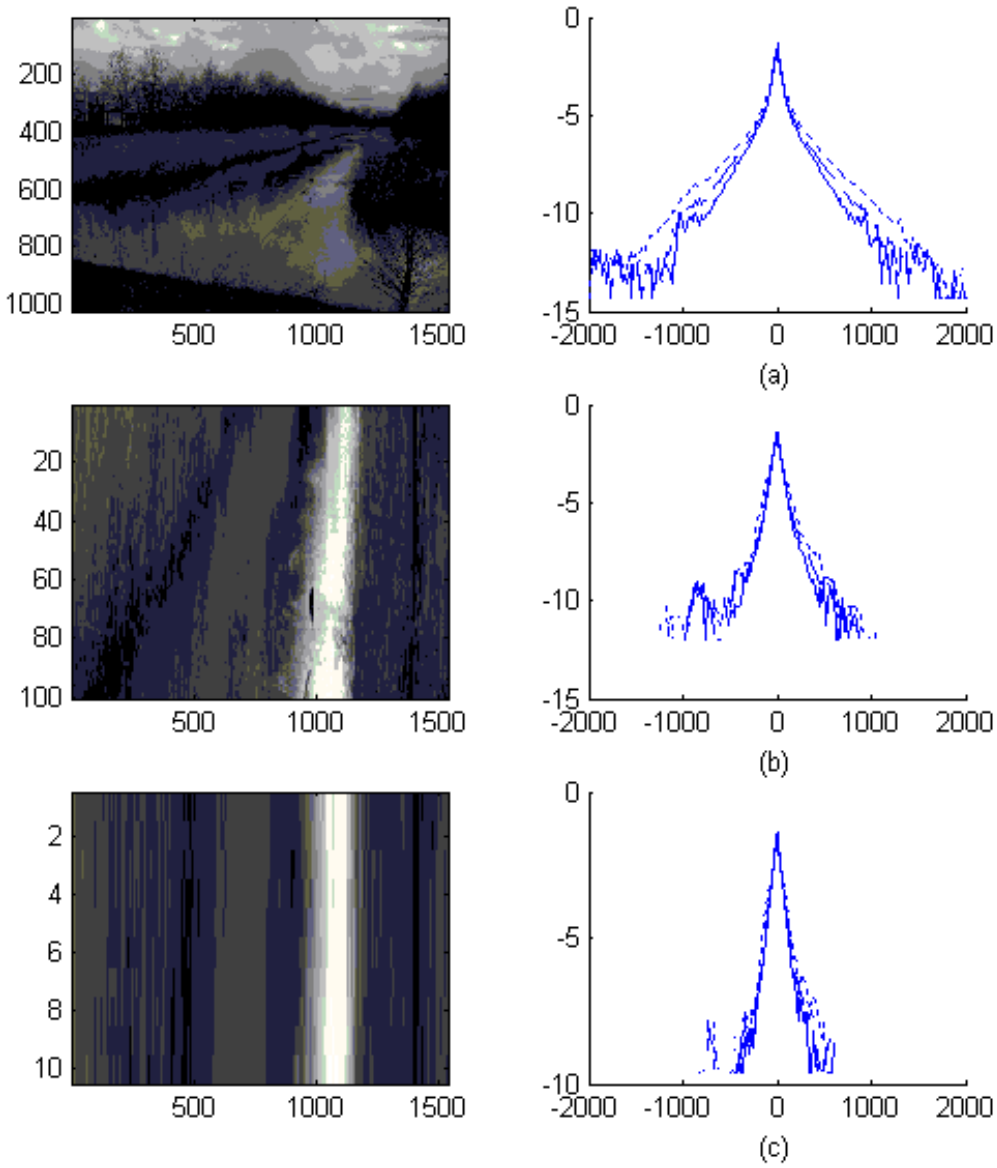

Figure 1.19: Picture-2 from Figure-1.1, and strips from it. Histograms are logarithms of marginal distributions of ∇_x. Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

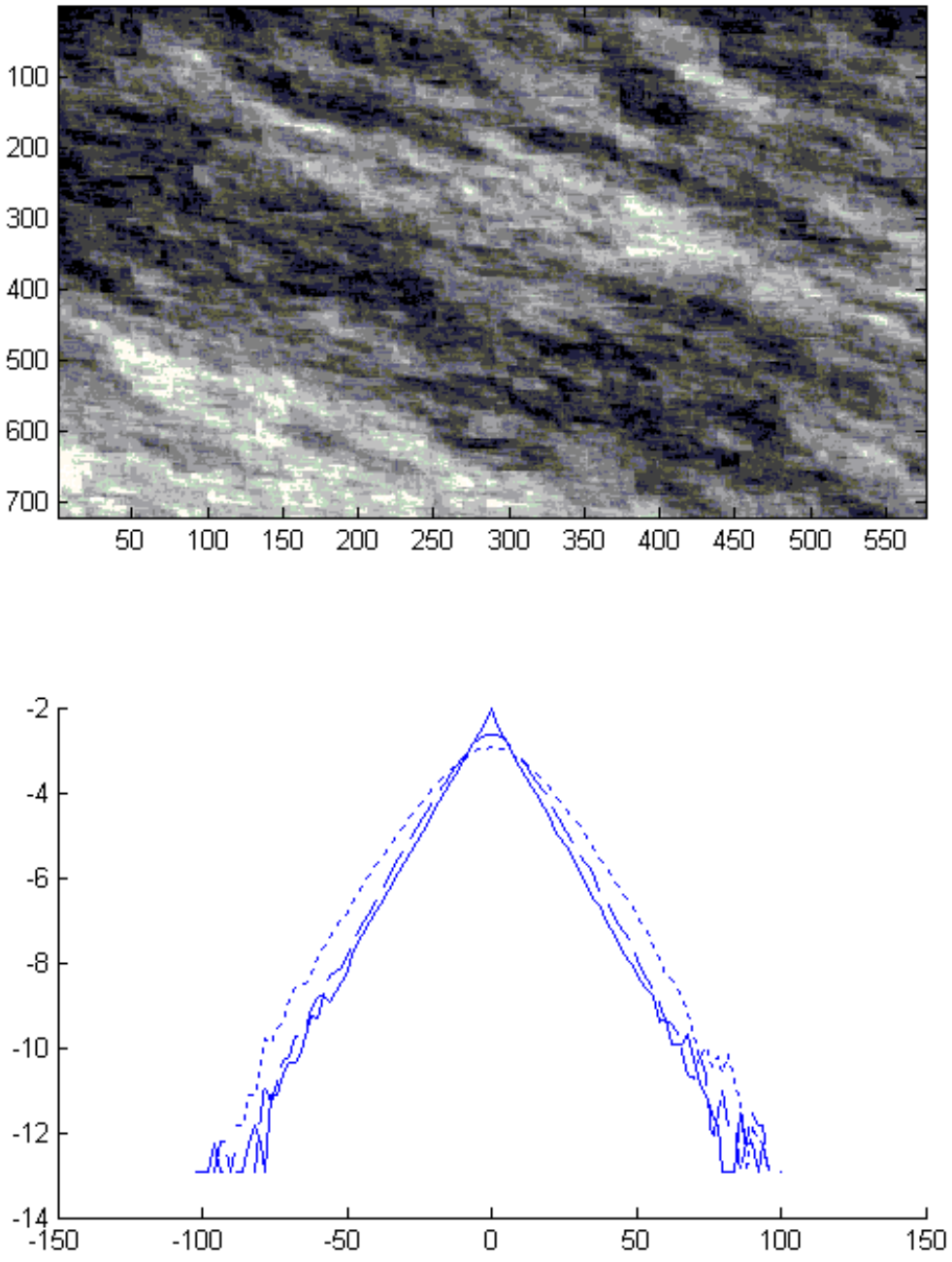

Figure 1.20: Logarithms of marginal distributions of ∇_x for simulated image from LOCO predictor. Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

from which we can simulate. We did this, and then experiment scale invariance in the simulated images.

The simulation procedure is as follows. For each pixel (i, j), we make a prediction as median($I(i-1,j)$, $I(i,j-1)$, $I(i-1,j)+I(i,j-1)-I(i-1,j-1)$). Then the pixel value is this prediction plus a noise from TSGD(0.2). We also truncate the pixel value so that it falls on $[0, 255]$. The simulation result is presented in Figure-1.20. It shows that $\nabla_x I$ and $\nabla_x I^{(2)}$ are very close, while $\nabla_x I^{(4)}$ is a little far from them.

This simulated LOCO image has nothing to do with the distance to the lens. However, we still observe good scale invariance. Again, this suggests that images without the distance effect can have scale invariance property .

1.4.3 Proposal

In the following, we will discuss on effects which we deem important in the scale invariance of local statistics.

Noise effect:

For a texture image composed of independent Gaussian noise only, the absolute value of the derivatives will become smaller when we scale down the image. For example, suppose $I(i, j)$ are i.i.d. random variables from $N(\mu, \sigma^2)$, then the distribution of $\nabla_x I$ is $N(0, 2\sigma^2)$. By the definition, $I^{(k)}(i, j)$ are the average of k^2 i.i.d. $N(\mu, \sigma^2)$, which is $N(\mu, \frac{\sigma^2}{k^2})$. Therefore, the distribution of $\nabla_x I^k$ is $N(0, 2\frac{\sigma^2}{k^2})$. Since $2\frac{\sigma^2}{k^2} < 2\sigma^2$, we conclude that ∇_x of the scaled image $I^{(k)}$ is pointier than ∇_x of the original image.

Figure-1.21 presents such an experiment result. Let $I(i, j)$ be i.i.d. $N(128, 900)$. $S = \nabla_x$. We round off $I(i, j)$ to be integers, and truncate them to [0,255]. As expected, Figure-1.21 shows that $\nabla_x I^{(4)}$ is pointier than $\nabla_x I^{(2)}$ and $\nabla_x I$ at 0. This demonstrates that images of independent Gaussian noises do not have scale invariance.

Next, we examine the Cauchy distribution. Cauchy distribution is a special distribution that may present scale invariance. The density function of Cauchy(θ, η)

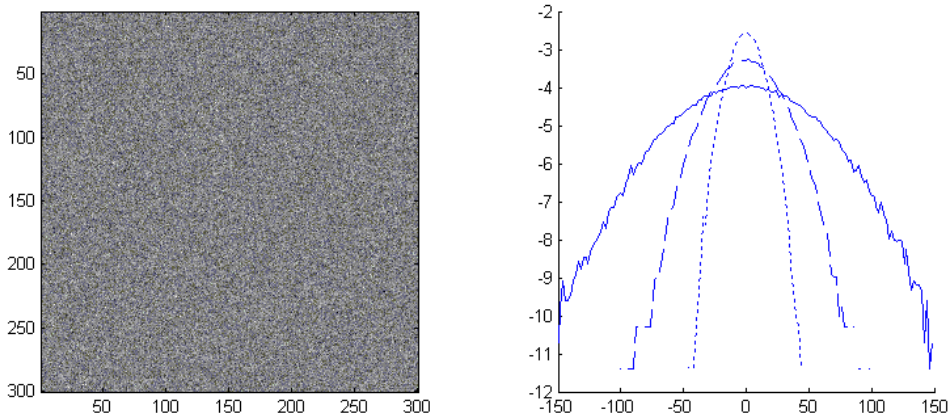

Figure 1.21: Logarithms of marginal distribution of ∇_x for a simulated image from N(128, 900). Solid curve:$\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

is

$$\frac{1}{\theta\pi} \cdot \frac{1}{1 + (\frac{x-\eta}{\theta})^2} \qquad -\infty < x < \infty,$$

and its characteristic function is $\psi(t) = e^{i\eta t - |t|}$. Since $\psi(\frac{t}{n})^n = \psi(t)$, the average of any i.i.d. random variables from Cauchy(θ, η) is still Cauchy(θ, η). Therefore, if $I(i,j)$'s follow i.i.d. Cauchy(θ, η) distribution, then $I^{(k)}(i,j)$'s again follow Cauchy(θ, η) distribution. In other words, an image composed of i.i.d. Cauchy noises should have perfect scale invariance property. However, the key reason for the invariant distribution of the averaged Cauchy lies in its heavy tail property. This heavy tail property does not exist in any image with bounded gray level values.

Figure-1.22 shows the result of our experiments in Cauchy noises. The first experiment uses Cauchy(30,128), and the second one uses Cauchy(3,128). We round off both experiments into integers, and truncate them to [0,255]. Figure-1.22 shows that neither of the Cauchy distributions has scale invariance property.

In fact, for any bounded random variable, the Law of Large Numbers shows that the average of such random variables converges to a constant as the number of random variables tends to infinity. According to this fact, we therefore conclude the following:

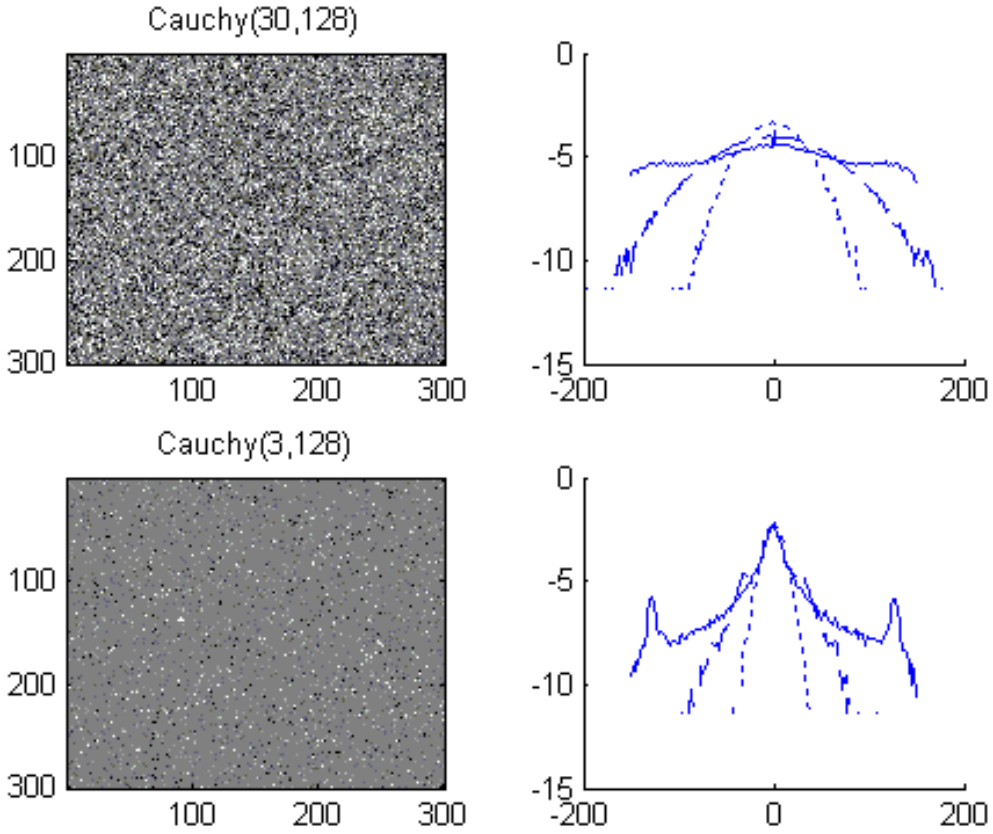

Figure 1.22: Logarithms of marginal distribution of ∇_x for simulated images from Cauchy distributions. Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

under the setup of finite gray levels, images composed of only i.i.d. noises do not have scale invariance property.

Another interesting noise is the two-sided geometric distribution. Figure-1.23 demonstrates an experiment for TSGD(0.1)+128. As expected, it does not present scale invariance property. However, comparing to the result of Gaussian noise in Figure-1.21, Figure-1.23 presents shapes closer to results from real images.

We experiment the noise effect by adding i.i.d. random variables from two-sided geometric distributions to a texture image. The result is presented in Figure-1.24. In

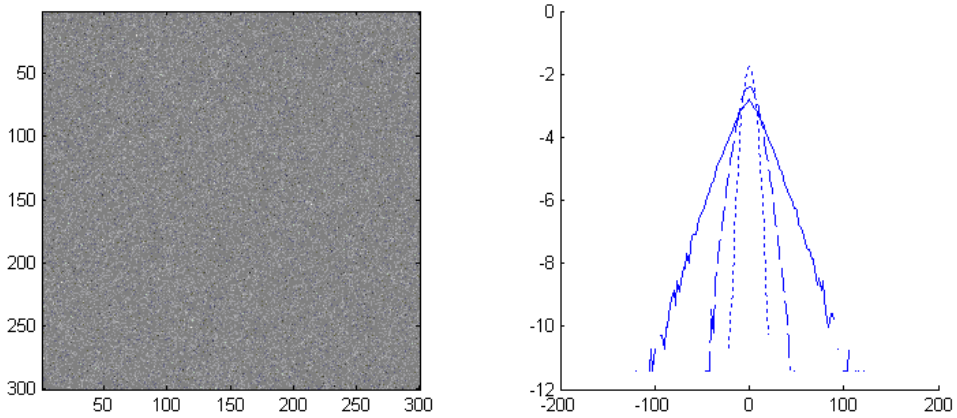

Figure 1.23: Logarithms of marginal distribution of ∇_x for a simulated image from TSGD(0.1). Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

the original image, the absolute values of the derivatives become larger after scaling. After adding noises TSGD(0.1), the image shows good scale invariance property. If adding larger noise TSGD(0.05), it shows that the absolute value of the derivatives are smaller after scaling. Therefore, adding proper noises can help in constructing an image with good scale invariance property.

Ramp effect:

For Ramp effect, we mean that the intensity of a texture image is a linear (with respect to the location) without any noise. When scaling down the image, the absolute value of the derivative becomes larger. For example, if $\nabla_x I = s$, then $\nabla_x I^{(k)} = ks$.

Boundary effect:

In addition to the noise effect and the ramp effect, boundary may also play an important role to scale invariance. For simplicity, we assume that all boundaries are vertical lines and gray level values are constant within each region divided by boundaries. Look at any row of the image[2], a boundary would fall in a pixel or between two pixels [3]. In the following, we use the word "pair" to denote any two

[2]With these assumptions, the whole image is the same as any single row in some sense.

[3]One may think that a boundary should always fall between two pixels. However, this is not

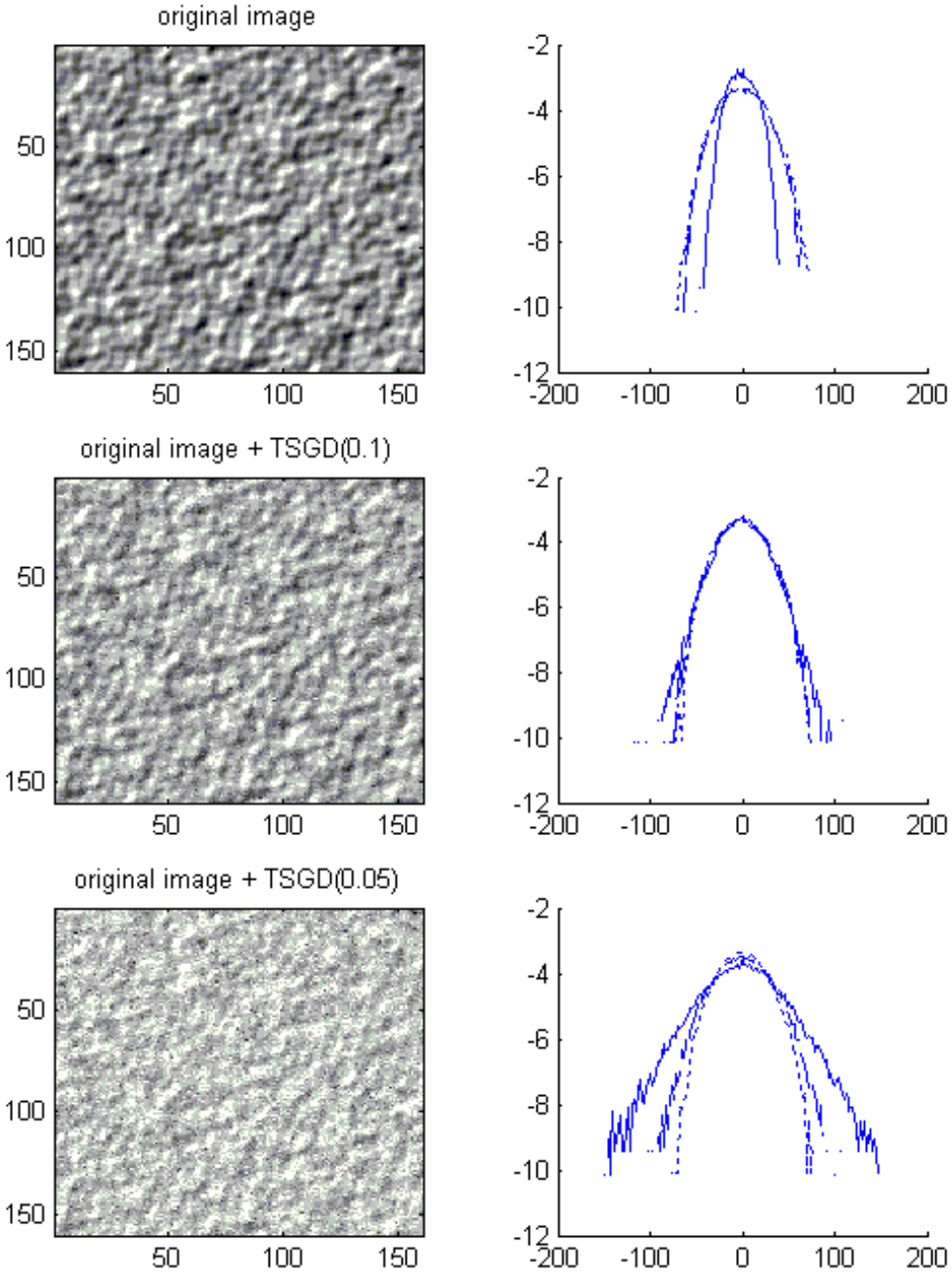

Figure 1.24: Logarithms of marginal distribution of ∇_x for an texture image and that image with TSGD noises. Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

adjacent pixels. Now we only consider pairs that there is a boundary passing through one of the pixels. For each such pair, we calculate ∇_x. Similarly, we calculate ∇_x for the scaled image. Because boundaries appear uniformly within each pair for both the original and the scaled image, we conclude that ∇_x of the original image and ∇_x of the scaled image are the same. Therefore, if we only look at those pixels involving boundaries, we have a perfect scale invariance of local statistics. However, there are more pixels involving boundaries in the scaled image than in the original image. Along with the fact that the absolute values of the derivatives involved boundaries are larger, boundaries increase the absolute values of the derivatives when we scale down an image.

Proposal:

From the discussion above, we summarize our findings in the following. After we scale down an image, The noise effect lower the absolute values of the derivatives, while the ramp effect and the boundary effect increase them.

We claim that a texture image with good scale invariance is the one balancing the noise effect and the ramp effect. In a texture image, where there is no prominent boundary, $\nabla_x I$ can be roughly decomposed into $X + Y$, where X is the ramp effect, and Y is the noise effect. In the scaled image $I^{(k)}$, the ramp effect becomes kX, and the noise effect becomes the average of k^2 i.i.d. Y. For example, if $Y \sim N(0, \sigma^2)$, then $\overline{Y} \sim N(0, \sigma^2/k^2)$. Since $Y/k \sim N(0, \sigma^2/k^2)$, $\overline{Y}$ has the distribution as Y/k. In this case, $\nabla_x I^{(k)} = kX + Y/k$. The image I will have good scale invariance if $X + Y \approx kX + Y/k$.

Assume that the ramp effect X is TSGD(θ_1) and the noise effect Y is TSGD(θ_2). We choose different pairs of (θ_1, θ_2) and compare $X + Y$ and $kX + Y/k$. The result of the experiment with $(\theta_1 = 0.2, \theta_2 = 0.1)$ is presented in Figure-1.25. It shows

---

true. A pixel represents a block, and the gray level value of that pixel is an average of that block. A boundary should fall uniformly in any location of any block. The probability that a boundary fall right between two pixels is zero.

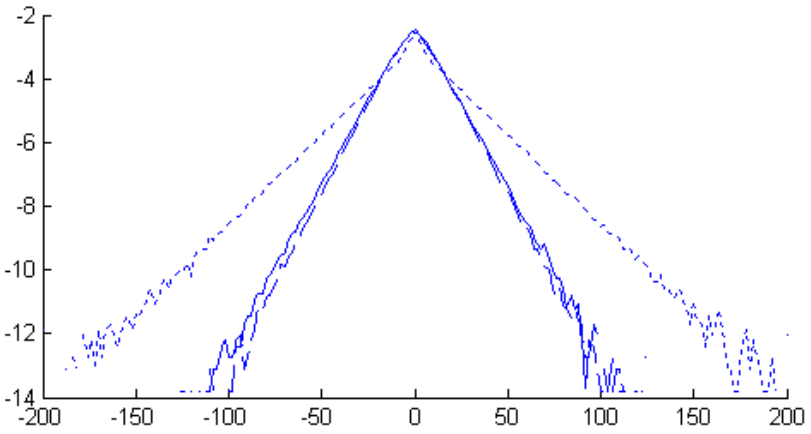

Figure 1.25: Logarithms of marginal distribution of ∇_x for simulations by mixing the ramp effect TSGD(0.2) and the noise effect TSGD(0.1). Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

that with a proper combination of both effects, we get texture images which scale very well. In Figure-1.25, $\nabla_x I$ is very close to $\nabla_x I^{(2)}$, but $\nabla_x I^{(4)}$ is far from them. The reason that the absolute value of $\nabla_x I^{(4)}$ is larger is that the ramp effect of $I^{(4)}$ is overrated. When it is calculated, we assume a local linear plane. Therefore, we have the ramp effect of $I^{(k)}$ is k times of that of I. However, when k becomes larger, the assumption of a linear plane may not be true. The ramp effect of $I^{(k)}$ should be smaller than k times of that of I.

As discussed earlier, Figure-1.24 also shows that, a proper combination of the noise effect and the ramp effect gives us texture images which scale very well .

We claim that an image with good scale invariance is the image which balances the noise, the ramp and the boundary effects. We experiment with adding noises to natural images. Since the noise effect will lower the absolute value of the derivative, we choose one image of which the absolute value of the derivative is larger after scaling. Among nine natural images we used earlier, we chose the one in Figure-1.8 for experiment. With the same reason we mentioned earlier in our experiment regarding to object size, we use the logarithm of the image, instead of the original

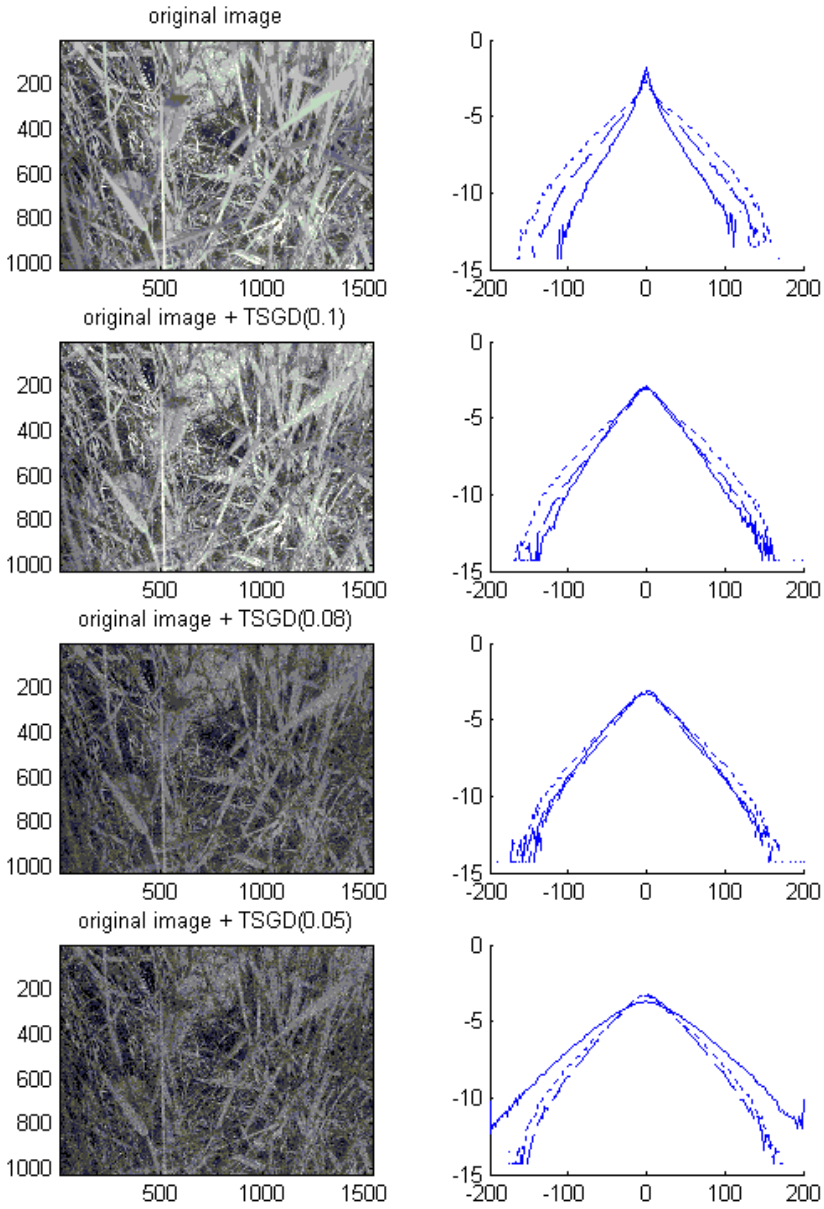

Figure 1.26: An image which does not scale well can have a good scale invariance by adding a proper noise. Solid curve: Logarithms of marginal distribution of $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

image.

The result is presented in Figure-1.26. We take the logarithm of the image, and modify the pixel values to be integers within [0, 255]. Then we add three different noises TSGD(0.1), TSGD(0.8) and TSGD(0.05) into the image, respectively. Figure-1.26 shows that the image mixed with TSGD(0.08) has good scale invariance, suggesting that with a proper combination of the noise, the ramp and the boundary effects, images can have good scale invariance property.

1.4.4 Distance effect

In previous subsections, we gave an explanation of the property of scale invariance based on the distance effect. We also found evidences that the distance effect may not be the key factor in the scale invariance of local statistics. In this subsection, we simulate ∇_x from texture images mixed with the distance effect, in order to investigate how the distance effect affect scale invariance.

Our simulation proceeds as follows. Randomly pick a location from the texture image and a distance r. Imagine that the texture image is placed at distance r, and calculate ∇_x. For convenience, $r \in \{1, 2, 3, 4, 5\}$. Figure-1.27 presents our simulation result. In (b), the density function of the distance is $c \cdot \frac{1}{r}$, where c is a normalization constant. In (c), the density function is from the range data. Figure-1.27 shows that the distance effect helps produce a better scale invariance.

1.4.5 Conclusion

We can explain the property of scale invariance of local statistics using the following two models. The first model is based on the distance distribution derived from the projection effect. In previous sections, we have shown that the property of scale invariance can be explained by this projection effect. However, there is no strong evidence to support that the projection effect is the main cause of the property of

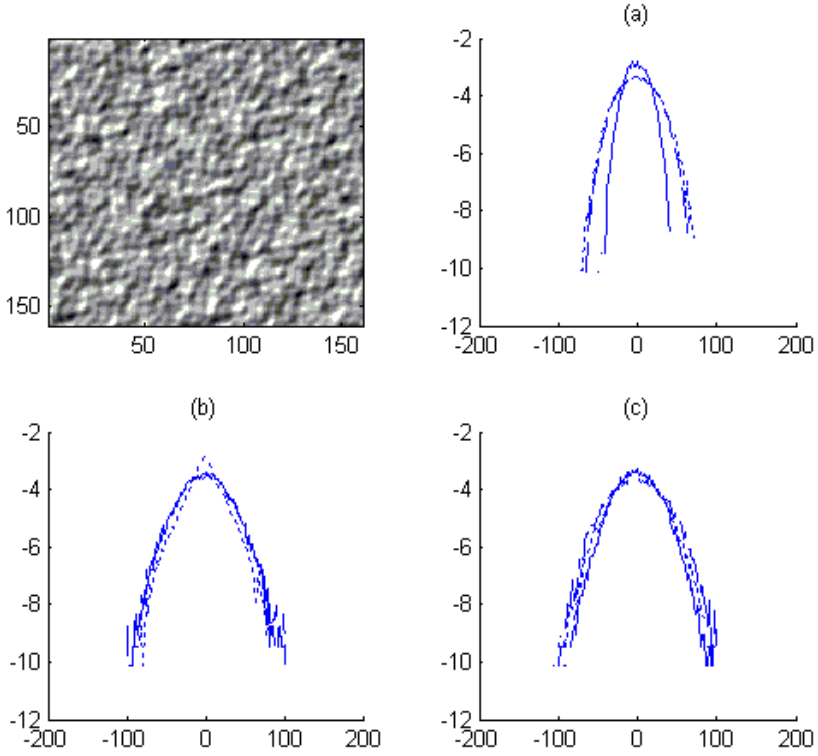

Figure 1.27: (a) ∇_x to the real image. (b) ∇_x to the simulation mixed with distance effect $1/r$. (c) ∇_x to the simulation mixed with the distance effect from range data. Solid curve: $\nabla_x I$, dashed curve: $\nabla_x I^{(2)}$, and dotted curve: $\nabla_x I^{(4)}$.

scale invariance. Moreover, there are images, such as texture images, which do not involve the distance, but still have good scale invariance property.

The second model is based on the noise effect, the ramp effect, and the boundary effect. In previous sections, we have also shown that the property of scale invariance can be explained by a proper combination of these three effects. However, again there is no strong evidence to support that these three effects are the main cause of the property of scale invariance. We can construct an image with good scale invariance by a proper combination of these three effects, but we are not able to claim that

any image with good scale invariance property must have such a proper combination. There is also no explanation why such a proper combination would exist in natural images.

We also explored to understand what types of images are more likely to have better scale invariance property. We observed that images that have more features tend to have better scale invariance. In other words, more textures and boundaries provide images with a better chance to scale well. Assume that all images can be decomposed into different features, and the effect of each feature on scale invariance is roughly additive. Let X be the effect of a feature on scale invariance. Suppose that $X = 0$ represents perfect scale invariance, and the large $|X|$ represents poor scale invariance. Let μ be the mean of X. If $\mu = 0$, the Law of Large Number shows that the average goes to 0 when the number of random variables goes to ∞. While there is no reason that μ should be 0, $|\mu|$ is probably small as many natural images have good scale invariance. The Central Limit Theorem tells that, the larger the sample size is, the smaller variance the average of the random variables has. Therefore, we know that the variance of the effect of images with more features is smaller. Along with the assumption that $|\mu|$ is small, this explains why images having more features tend to scale better.

Our work so far still can not explain why ensemble images have good scale invariance property. However, our work has shown that, either the distance effect or a proper combination of the noise effect, the ramp effect and the boundary effect can improve the scale invariance property. As a result, we suggest that a combination of both models as described above may be the best explanation to the cause of the property of scale invariance.

In the end of this section, we summarize the reasons why natural images have the property of scale invariance.

1. There are three effects in natural images that are important factors to the property of scale invariance. These three effects are the noise effect, the ramp

effect and the boundary effect. Images with proper combinations of these three effects can show good scale invariance property.

2. Images with more features tend to have better scale invariance. This phenomenon can be well explained by the Central Limit Theorem.

3. The distance to the lens is also an important factor to scale invariance. An image that does not have good scale invariance can scale better after the distance effect is involved. Natural images usually have the distance effect, therefore usually show better scale invariance property.

Bibliography

[1] Z. Chi. Probability Models for Complex System: Chapter 7. Scale Invariance of Natural Images. Ph.D. Thesis, Brown University, 1999.

[2] D. L. Ruderman. The statistics of natural images *Computation in Neural Systems* 5(4), November 1994, pp. 517-548.

[3] D. L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 1994, pp. 814-817.

[4] D. L. Ruderman. Origins of Scaling in Natural Images. *Vision Research*, 37(23), 1997, pp. 3385-3398.

[5] D. Mumford and Basilis Gidas. Stochastic Models for Generic Images. *Quarterly Appl. Math*, 59, 2001, pp.85-111.

[6] Rosario M. Balboa, Christopher W. Tyler, Norberto M. Grzywacz. Occlusions contribute to scaling in natural images. *Vision Research* 41, 2001, pp.966-964.

[7] Ann B. Lee, David Mumford and Jinggan Huang. Occlusion Model for Natural Images: A statistical Study of a Scale-Invariant Dead Leaves Model. *International Journal of Computer Vision* 41(1/2), 2001, pp.35-39.

[8] M G A Thomson. Beats, kurtosis and visual coding. *Computation in Neural Systems* 12(3), Aug 2001, pp.271-287

[9] Ulf Grenander and Anuj Srivastava. Probability Models for Clutter in Natural Images. *IEEE Transactions on Pattern Analysis and Machine Intellegence* 23(4), APRIL 2001, pp.424-429

[10] Anuj Srivastava, Xiuwen Liu and Ulf Grenander. Universal Analytical Forms for Modeling Image Probabilities *APPTS Report* #01-4, Novermber 2001

[11] A. Turiel and N. Parga. The multifractal structure of contrast changes in natural images: From sharp edges to textures. *Neural Computation* 12(4), APR 2000, pp.763-793.

[12] A. Turiel, N. Parga, D. L. Ruderman and T. W. Cronin. Multiscaling and information content of natural color images. *Physical Review* E. 62(1), jul 2000, pp.1138-1148.

[13] 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(12), Dec 1987, pp.2379-2394.

[14] D. J. Field. What is the Goal of Sensory Coding? *Neural Computation* 6(4), Jul 1994, pp.559-601.

[15] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Computation in Neural Systems* 7(2), May 1996, p.333-339.

[16] G. J. Burton and Ian R. Moorhead. Color and spatial structure in natural scenes. *Applied Optics* 26(1), Jan 1987, pp.157-170

[17] J. H. van Hateren. A theory of maximizing sensory information. *Biological Cybernetics* 68, 1992, pp.23-29.

[18] 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, 1998, pp. 359-366.

[19] L. Alvarez et J.-M. Morel. The size of objects in natural images. preprint du CMLA 9921, 1999.

[20] Ann B. Lee, Kim S. Pedersen and David Mumford. The Nonlinear Statistics of High-Contrast Patches in Natural Images. *APPTS Report #01-3*, June 2001

Chapter 2

On the Use of of Natural Image Statistics for Compression

2.1 Introduction

The goal of image compression is to create smaller files that use less space to store and less time to send. There are two types of compression: lossless and lossy. In lossless compression, the original image can be decompressed perfectly. In lossy compression, it allows for errors between the decompressed image and the original image. It sacrifices some details in the image to achieve significant gains in the compression ratio. However, in some critical applications (for example, military observation and medical imaging), any loss may not be tolerated. In this paper, we focus on lossless image compression.

To do lossless image compression, we need a good model. Most of the lossless image compression methods code images pixel by pixel in a pre-defined order, which is usually the raster-scan order (from left to right, and top to bottom). At each time $t+1$, past information $x_1^t = x_1 x_2 \cdots x_t$ are observed. Ideally, the best one can do is to code the next pixel x_{t+1} with $-\log \Pr(x_{t+1}|x_1^t)$[1] bits, where the logarithm here and later on is taken to the base 2. But practically, it is hard to model $\Pr(x_{t+1}|x_1^t)$, since the dimension of x_1^t is too large. Instead, many good lossless image compression methods make a prediction $\hat{x}_{t+1}$ based on the past information x_1^t, then code the residual $x_{t+1} - \hat{x}_{t+1}$. Both LOCO (LOw COmplexity LOssless COmpression) [1][2] and CALIC (Context based Adaptive Lossless Image Codec) [3] [4] adopt this approach.

JPEG-LS, based on the LOCO algorithm, is the new lossless/near-lossless compression standard for continuous-tone images, ISO-14495-1/ITU-T.87. The LOCO prediction of a pixel is only based on its three neighboring pixels. In Figure-2.1, x is the pixel to predict, and the prediction is only based on a, b, and c. When c is between a and b, x is predicted by fitting a linear plane on a, b, and c. Therefore, $\hat{x} = a + b - c$. When c is larger than both a and b, LOCO believes that there is an edge and predicts $\hat{x} = \min(a, b)$. Similarly, $\hat{x} = \max(a, b)$ when $c < \min(a, b)$.

[1] We use $\Pr(x_{t+1}|x_1^t)$ to denote $\Pr(X_{t+1} = x_{t+1}|X_1^t = x_1^t)$

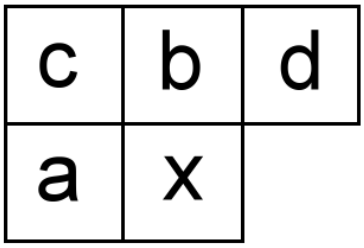

Figure 2.1: LOCO predictor

Equation-(2.1) gives the LOCO predictor:

$$\hat{x}_{LOCO} = \begin{cases} \min(a,b) & \text{if } c \geq \max(a,b) \\ \max(a,b) & \text{if } c \leq \min(a,b). \\ a+b-c & \text{otherwise.} \end{cases} \tag{2.1}$$

This equation can also be rewritten as

$$\hat{x}_{LOCO} = median(a, b, a+b-c).$$

After predicting x as $\hat{x}$, LOCO codes the residual $x - \hat{x}$ by a two-sided geometric distribution (TSGD). It is a widely accepted observation that prediction residuals in continuous-tone images are well modelled by TSGD [5]. LOCO determines the context where x occurs according to its four neighboring pixels, a, b, c and d. They calculate $(d-b, b-c, c-a)$, and classify these triples into 365 different contexts. In each context, a TSGD is estimated to model the residuals.

To summarize, LOCO compression scheme is as follows:

1. predict x_{t+1} according to past information x_1^t.

2. code the residual $x_{t+1} - \hat{x}_{t+1}$ with different distributions depending on the type of past pixels.

In addition, LOCO uses several more steps to improve this scheme. For example, an adaptive correction according to the past prediction errors. In this paper, we will discuss why the scheme which makes predictions then sends residuals works well. Specifically, why is predicting then sending the residual a good approximation to $\Pr(x_{t+1}|x_1^t)$? We will also work on how to construct a good prediction according to the past information.

2.2 Definitions and Assumptions

In this section, we will discuss the general reasoning behind LOCO. We will also introduce some definitions and assumptions which will be used in the following sections.

As mentioned in the previous section, an ideal compression is to code x_{t+1} with $-\log \Pr(x_{t+1}|x_1^t)$ bits. Equivalently, the expected bits per pixel is $H(X_{t+1}|X_1^t)$, where H denotes the entropy function. Let g be any prediction function of X_{t+1} according to X_1^t. Since $g(X_1^t)|X_1^t$ is deterministic, then

$$H(X_{t+1}|X_1^t) = H(X_{t+1} - g(X_1^t)|X_1^t).$$

Therefore, the scheme to make predictions and to code the residuals has the same entropy as to code the actual X_{t+1}. From this observation, this scheme does not have any advantage.

In practice, it is impossible to code X_{t+1} or $X_{t+1} - g(X_1^t)$ according to X_1^t, since we need the distribution on X_{t+1} or $X_{t+1} - g(X_1^t)$ for every X_1^t. For example, to compress gray level images with intensities from 0 to 255, |range of X_1^t| $= 256^t$. Even when $t = 3$, the range is too large. A realistic way is to introduce a function h, and to use $X_{t+1}|h(X_1^t)$ to approximate $X_{t+1}|X_1^t$, where

$$1 < |\text{range of } h(X_1^t)| << |\text{range of } X_1^t|.$$

The function h here is usually referred as "context", as the distribution of $X_{t+1}|X_1^t$ tends to be similar under the same context. For example, LOCO determines h according to $(a-c, c-b, b-d)$. There are 365 contexts in LOCO. After introducing h,

$$H(X_{t+1}|h(X_1^t)) \neq H(X_{t+1} - g(X_1^t)|h(X_1^t)).$$

Therefore, a good prediction function g is to make

$$H(X_{t+1} - g(X_1^t)|h(X_1^t))$$

as small as possible. Note that as h becomes more fine-grained, the efficiency of this scheme approaches optimal for any predictor g, since

$$H(X_{t+1} - g(X_1^t)|h(X_1^t)) \to H(X_{t+1} - g(X_1^t)|X_1^t)$$

as h becomes arbitrarily fine.

More generally, given $X \in R$, $\overrightarrow{Y} \in R^m$, we study how to choose a predictor, $f : R^m \to R$, to minimize

$$H(X - f(\overrightarrow{Y})).$$

Before deriving this "optimal"[2] predictor, we will first give some necessary definitions and assumptions. In the following, we will assume that X and $\overrightarrow{Y}$ have joint density

$$p(X, \overrightarrow{Y})$$

and conditional density

$$p(X|\overrightarrow{Y})$$

for each X and $\overrightarrow{Y}$.

[2]in the sense of minimizing $H(X - f(\overrightarrow{Y}))$

Definition 2.1 *A distribution P(X, $\overrightarrow{Y}$) on R^{m+1} is conditionally symmetric and unimodal (CSUM) if*

$$p(x|\overrightarrow{y})$$

is symmetric and unimodal in x for every $\overrightarrow{y}$.

Definition 2.2 *A distribution P(X, $\overrightarrow{Y}$) on R^{m+1} is shift invariant (SI) if*

$$p(x+s|\overrightarrow{y}+s) = p(x|\overrightarrow{y})$$

for every $(x, \overrightarrow{y}) \in R^{m+1}$, $s \in R$ (where here, and later, a vector plus a scaler is interpreted component by component).

CSUM is often observed in natural images. In fact, it is a widely accepted observation that prediction residuals in continuous-tone images are well modelled by TSGD [5]. The density function of TSGD(p) is

$$\frac{p}{2-p}q^{|x|-1} \qquad\qquad x = \ldots, -2, -1, 0, 1, 2, \ldots,$$

where $p = 1 - q$. It is symmetric and unimodal. Therefore, the distribution of prediction residuals is CSUM.

SI is also widely observed in natural images, especially if work with the log of intensity. For examples, the logarithm of the light intensity is approximately SI.

2.3 Minimizing Residual Entropy

Our goal is to choose $g(X_1^t)$ to minimize

$$H((X_{t+1} - g(X_1^t))|h(X_1^t)).$$

On more general terms, we want to choose $f(\overrightarrow{Y})$ to minimize

$$H(X - f(\overrightarrow{Y}))$$

for two random variables $(x, \overrightarrow{y}) \in R^{m+1}$. Motivated by the statistics and properties of natural images mentioned above, we will prove the following:

Theorem 2.3 *If $P(X, \overrightarrow{Y})$ is CSUM, then*

$$f(\overrightarrow{y}) \equiv median(X|\overrightarrow{Y} = \overrightarrow{y})$$

achieves

$$\min H(X - f(\overrightarrow{Y})).$$

Before we prove this theorem, we make the following observations. $X - f(\overrightarrow{Y})$ is a mixture of $(X - f(\overrightarrow{Y}))|\overrightarrow{Y} = \overrightarrow{y}$. Since $(X, \overrightarrow{Y})$ is CSUM, and since

$$(X - f(\overrightarrow{Y}))|\overrightarrow{y} = X|\overrightarrow{y} - f(\overrightarrow{y}),$$

$(X - f(\overrightarrow{Y}))|\overrightarrow{y}$ is symmetric and unimodal with median at $median(X|\overrightarrow{y} - f(\overrightarrow{y})$. If

$$f(\overrightarrow{y}) = median(X|\overrightarrow{Y} = \overrightarrow{y}),$$

the median of each conditional distribution is 0. Therefore, to prove this theorem, it suffices to show that the entropy of the mixture of a family of symmetric and unimodal distributions is minimized if their peaks are at the same position. We will first prove a discrete and finite version of the above statement in the following lemma, and we will generalize it to the continuous case later.

Lemma 2.4 *Suppose $\{p_{i,j}\}$ is a positive measure ($p_{i,j} \geq 0$) defined on $[-N, N] \times$*

$[-n, n]$. *And* $p_{\cdot,j}$ *is unimodal and symmetric with median at 0. Let*

$$\begin{aligned} q_i &= \sum_j p_{i,j} \\ r_i &= \sum_j p_{i+g(j),j} \end{aligned}$$

where g *is any function mapping from* Z *onto* Z*. Then*

$$H(\{q_i\}) \leq H(\{r_i\}),$$

where

$$H(\{q_i\}) \equiv \sum_i (\log \frac{1}{q_i}) \cdot q_i.$$

Note that $\{q_i\}$ and $\{r_i\}$ may not be probability measures. Therefore, we extend the domain of entropy function H by giving the same definition on positive measures.

For convenience, we will prove the case that $\{p_{i,j}\}$ is a probability measure. The proof for the case that $\{p_{i,j}\}$ is a positive measure is exactly the same. To prove Lemma 2.4, we will begin with another two lemmas. The first one in the following is very intuitive: If we move some probability from a state to another state with higher probability, we lower the entropy.

Lemma 2.5 *Let* $\{p_i\}$ *be a probability distribution. Define another probability distribution* $\{q_i\}$ *by* $q_1 = p_1 + s$, $q_2 = p_2 - s$, $q_i = p_i$ $\forall i \neq 1, 2$. *If* $p_1 \geq p_2$ *and* $0 < s < p_2$, *then*

$$H(\{q_i\}) < H(\{p_i\}).$$

Proof:

Since $q_i = p_i$ $\forall i \neq 1, 2$, it suffices to prove

$$q_1 \log \frac{1}{q_1} + q_2 \log \frac{1}{q_2} < p_1 \log \frac{1}{p_1} + p_2 \log \frac{1}{p_2}.$$

Let $F(s) = (p_1 + s)\log\frac{1}{p_1+s} + (p_2 - s)\log\frac{1}{p_2-s}$. Then

$$\begin{aligned}\frac{\partial F(s)}{\partial s} &= \log\frac{1}{p_1+s} + \log e - \log\frac{1}{p_2-s} - \log e \\ &= \log\frac{p_2-s}{p_1+s} \\ &< 0 \qquad \forall s \in (0, p_2).\end{aligned}$$

Hence, $F(s)$ is decreasing for $s > 0$. That is

$$q_1\log\frac{1}{q_1} + q_2\log\frac{1}{q_2} < p_1\log\frac{1}{p_1} + p_2\log\frac{1}{p_2}.$$

□

In Lemma 2.4, we claim that the entropy of $q_i = \sum_j p_{i,j}$ is the smallest among

$$\Omega_0 = \{r = (r_1, r_2, \cdots) | r_i = \sum_j p_{i+g(j),j} \text{ for some function } g\}.$$

For any probability distribution $r \in \Omega_0$ which is not the same as q defined in Lemma 2.4, we attempt to apply Lemma 2.5 to create a new probability distribution with a lower entropy. However, this new probability distribution may not belong to Ω_0. Therefore, we define a larger family of probability distributions $\Omega_1 \supset \Omega_0$ to make sure that the new probability distribution we create is always in Ω_1. If $|\Omega_1| < \infty$, and if we can always find a new probability distribution with a lower entropy, the entropy of q is the smallest among Ω_1. We present the above argument in the following lemma.

Lemma 2.6 *Suppose $\{p_{i,j}\}$, $1 \le i \le N$, $1 \le j \le n$, is a probability distribution with $p_{i_1,j} \ge p_{i_2,j}$ whenever $i_1 < i_2$. Define Ω_1 as a set of probability distributions by*

$$\begin{aligned}\Omega_1 &= \{s = (s_1, \cdots, s_m) | s_1 \ge s_2 \ge \cdots \ge s_m; \sum s_k = 1; \\ &\qquad s_k = \sum_j p_{\tau_{k,j},j}; \tau_{k_1,j} \ne \tau_{k_2,j} \text{ whenever } k_1 \ne k_2\},\end{aligned}$$

where $\tau_{k,j} \in \{1, \cdots, N, \phi\}$, $\tau_{k,j}$ *can be* ϕ *for more than one* k, *and* $p_{\phi,j} \equiv 0$. *In this setup, for any* $s \in \Omega_1$, *each* s_k *is the summation of at most one element from each* $p_{\cdot,j}$, *and each* $p_{i,j}$ *appears in exactly one* s_k. *Let*

$$q_i = \sum_j p_{i,j},$$

then

$$H(q) = \min_{s \in \Omega_1} H(s).$$

Proof:

Since $|\Omega_1| < \infty$, it suffices to prove that $\forall s \neq q$, there exists r such that $H(r) < H(s)$. Suppose that $k_1 < k_2$, $p_{\tau_{k_1,j},j} = p_{\tau_{k_2,j},j}$ and $\tau_{k_1,j} > \tau_{k_2,j}$. then we will exchange the values of $\tau_{k_1,j}$ and $\tau_{k_2,j}$. Therefore, we always have $\tau_{k_1,j} < \tau_{k_2,j}$ whenever $k_1 < k_2$ and $p_{\tau_{k_1,j},j} = p_{\tau_{k_2,j},j}$, without changing the value of each s_k.

If $\tau_{1,j} \neq 1$ for some j, then $p_{\tau_{1,j},j} < p_{1,j}$, and then $s_1 < q_1$. Hence, if $q_1 = s_1$, we will have $\tau_{1,j} = 1$. Now define $\alpha = \min\{k | q_k \neq s_k\}$, then we have $\tau_{k,j} = k \quad \forall k < \alpha$ and $q_\alpha > s_\alpha$. Since $q_\alpha > s_\alpha$, there exists j_1 such that $\tau_{\alpha,j_1} \neq \alpha$ and $p_{\tau_{\alpha,j_1},j_1} < p_{\alpha,j_1}$. As p_{α,j_1} is not in any of $s_k \quad \forall k \leq \alpha$, it is in some s_δ with $\delta > \alpha$. Now, we define σ as

$$\begin{aligned} \sigma_{\alpha,j_1} &= \alpha \\ \sigma_{\delta,j_1} &= \tau_{\alpha,j_1} \\ \sigma_{i,j} &= \tau_{i,j} \qquad \text{otherwise,} \end{aligned}$$

we also define another probability distribution $r = (r_1, \cdots, r_m)$ as

$$r_k = \sum_j p_{\sigma_{k,j},j}.$$

Then we have

$$\begin{aligned} r_\alpha &= s_\alpha + p_{\alpha,j_1} - p_{\tau_{\alpha,j_1},j_1} \\ r_\delta &= s_\delta - p_{\alpha,j_1} + p_{\tau_{\alpha,j_1},j_1} \\ r_j &= s_j \qquad\qquad\qquad \text{otherwise.} \end{aligned}$$

Since $p_{\tau_{\alpha,j_1},j_1} - p_{\alpha,j_1} < 0$, by Lemma 2.5, we have $H(r) < H(s)$. Note that $r(r_1, \cdots, r_m)$ may not satisfy

$$r_1 \geq r_2 \geq \cdots \geq r_m.$$

However, we can always change the order of r_i so that the condition above holds and the entropy stays unchanged. As a result, we have a new probability distribution in Ω_1 with a lower entropy.

□

From the setup of Ω_1, for any $s \in \Omega_1$, s_k is the summation of at most one element from each $p_{\cdot,j}$. This lemma tells that, by grouping the largest $p_{i,j}$ in each $p_{\cdot,j}$ together and grouping the second largest ones together, and so on, the entropy of the distribution will be the smallest among all in Ω_1.

Note that $p_{\cdot,j}$ in Lemma 2.6 is assumed to be decreasing, instead of unimodal and symmetric as in Lemma 2.4. It is easier to prove Lemma 2.6 under this assumption of decreasing $p_{\cdot,j}$.

Proof of Lemma 2.4:

To prove Lemma 2.4 using Lemma 2.6, we define a new distribution p^* on $[1, 2N + 1] \times [-n, -n]$ by

$$p^*_{i,j} = \begin{cases} p_{i/2,j} & \text{if } i \text{ is even} \\ p_{-(i-1)/2,j} & \text{if } i \text{ is odd.} \end{cases}$$

That is, we rearrange $p_{\cdot,j}$ in the order of 0, 1, -1, 2, -2, $\cdots$, then $p^*_{i_1,j} \geq p^*_{i_2,j}$ whenever $i_1 < i_2$. Similarly, define q^* as

$$q^*_i = \{ \begin{array}{ll} q_{i/2} & \text{if } i \text{ is even} \\ q_{-(i-1)/2} & \text{if } i \text{ is odd.} \end{array}$$

We have

$$q^*_i = \sum_j p^*_{i,j}.$$

From lemma 2.6, we know that the entropy of q^* is the smallest among all probability distributions in Ω_1. For any r in Lemma 2.4, rearrange r to r^* so that

$$r_1 \geq r_2 \geq \cdots \geq r_N.$$

Then r^* is also in Ω_1. By Lemma 2.6, we have

$$H(q^*) \leq H(r^*).$$

Then

$$H(q) \leq H(r).$$

□

In next lemma, we generalize the result of Lemma 2.4 from the finite case to the countable case.

Lemma 2.7 *Suppose $\{p_{i,j}\}$ is a positive measure on Z^2. And $p_{\cdot,j}$ is unimodal and symmetric with 0. Let*

$$\begin{aligned} q_i &= \sum_j p_{i,j} \\ r_i &= \sum_j p_{i+g(j),j}, \end{aligned}$$

where g is any function mapping from Z onto Z. Then

$$H(\{q_i\}) \leq H(\{r_i\}).$$

Proof:

Suppose there exists a r such that

$$H(r) < H(q) < \infty.$$

Let $\epsilon = H(q) - H(r)$. Define $p^{(n)}$ as

$$p_{i,j}^{(n)} = \begin{cases} p_{i,j} & \text{if } -n \leq i,j \leq n \\ 0 & \text{otherwise.} \end{cases}$$

Define $q^{(n)}$ and $r^{(n)}$ as

$$\begin{aligned} q_i^{(n)} &= \sum_j p_{i,j}^{(n)} \\ r_i^{(n)} &= \sum_j p_{i+g(j),j}^{(n)}. \end{aligned}$$

Then,

$$\begin{aligned} \lim_{n\to\infty} H(q^{(n)}) &= H(q) \\ \lim_{n\to\infty} H(r^{(n)}) &= H(r). \end{aligned}$$

From the continuity of entropy function H, there exists an $N = N(\epsilon)$ such that

$$q^{(n)} \neq r^{(n)},$$

and

$$\begin{aligned} |H(q^{(N)}) - H(q)| &< \frac{\epsilon}{2} \\ |H(r^{(N)}) - H(r)| &< \frac{\epsilon}{2}. \end{aligned}$$

Then

$$\begin{aligned} H(q^{(N)}) - H(r^{(N)}) &= (H(q^{(N)}) - H(q)) + (H(q) - H(r)) + (H(r) - H(r^{(N)})) \\ &> -\frac{\epsilon}{2} + \epsilon - \frac{\epsilon}{2} \\ &= 0. \end{aligned}$$

On the other hand, apply Lemma 2.4 to $p^{(N)}$, $q^{(N)}$ and $r^{(N)}$, we have

$$H(q^{(N)}) \leq H(r^{(N)}),$$

which leads to a contradiction. Therefore, there does not exist r such that

$$H(r) < H(q) < \infty.$$

□

Proof of Theorem 2.3:

As discussed earlier, it suffices to prove a continuous version of Lemma 2.4 or Lemma 2.7. Without loss of generality, we assume that the dimension of $\overrightarrow{y}$ is 1. The proof for a higher dimension is almost the same. Suppose $p(x, y)$ is symmetric, unimodal with median at 0 in x for each y. Define $\{q(x, f)\}$ as

$$q(x, f) \equiv \int_y p(x - f(y), y)dy$$

where $f : R^m \to R$. We need to show that

$$H(\{q(x,0)\}) \leq H(\{q(x,f)\})$$

for any function f.

Suppose the statement above is not true, then there exists a function g such that

$$H(\{q(x,g)\}) < H(\{q(x,0)\}) < \infty.$$

Let $\epsilon = H(\{q(x,0)\}) - H(\{q(x,f)\})$. Define $p^{(n)}(x)$ and $q^{(n)}(x,g)$ as

$$\begin{aligned} p^{(n)}(x) &\equiv p(\frac{[x \cdot 2^n]}{2^n}, \frac{[y \cdot 2^n]}{2^n}) \\ q^{(n)}(x,f) &\equiv \int_y p^{(n)}(x - f(y), y)dy, \end{aligned}$$

where here and later, $[x]$ denotes the nearest integer to x. Since

$$\lim_{n\to\infty} p^{(n)}(x) = p(x),$$

then

$$\lim_{n\to\infty} q^{(n)}(x,f) = q(x,f).$$

Along with the continuity of entropy function, we have

$$\lim_{n\to\infty} H(\{q^{(n)}(x,f)\}) = H(\{q(x,f)\}).$$

Similarly, define $f^{(n)}$ as

$$f^{(n)} \equiv \frac{[f \cdot 2^n]}{2^n}.$$

Then

$$\lim_{n\to\infty} H(\{q(x,f^{(n)})\}) = H(\{q(x,f)\}).$$

Therefore, there exist $N = N(g)$ such that

$$\begin{aligned} |H(\{q^{(N)}(x,0)\}) - H(\{q(x,0)\})| &< \frac{\epsilon}{3} \\ |H(\{q^{(N)}(x,g)\}) - H(\{q(x,g)\})| &< \frac{\epsilon}{3} \\ |H(\{q^{(N)}(x,g^{(N)})\}) - H(\{q^{(N)}(x,g)\})| &< \frac{\epsilon}{3}. \end{aligned}$$

Moreover,

$$\begin{aligned} q^{(N)}(x,f^{(N)}) &= \int_y p^{(N)}(x - f^{(N)}(y), y)dy \\ &= \int_y p(\frac{[(x - f^{(N)}) \cdot 2^N]}{2^N}, \frac{[y \cdot 2^N]}{2^N})dy \\ &= \int_y p(\frac{[x \cdot 2^N]}{2^N} - f^{(N)}, \frac{[y \cdot 2^N]}{2^N})dy \\ &= \sum_j p(\frac{[x \cdot 2^N]}{2^N} - f^{(N)}, \frac{j}{2^N}) \cdot \frac{1}{2^N} \end{aligned}$$

$$\begin{aligned} H(\{q^{(N)}(x,f^{(N)})\}) &= \int_x (\log \frac{1}{q^{(N)}(x,f^{(N)})}) \cdot q^{(N)}(x,f^{(N)})dx \\ &= \sum_i (\log \frac{1}{q^{(N)}(\frac{i}{2^N},f^{(N)})}) \cdot q^{(N)}(\frac{i}{2^N},f^{(N)}) \cdot \frac{1}{2^N} \end{aligned}$$

Define $p_{i,j}$ and $q_i(f^{(N)})$ as

$$\begin{aligned} p_{i,j} &\equiv p(\frac{i}{2^N}, \frac{j}{2^N}) \cdot \frac{1}{2^{2N}} \\ q_i(f^{(N)}) &\equiv q^{(N)}(\frac{i}{2^N}, f^{(N)}). \end{aligned}$$

Then

$$\begin{aligned} q_i(f^{(N)}) &= q^{(N)}(\frac{i}{2^N}, f^{(N)}) \cdot \frac{1}{2^N} \\ &= \sum_j p(\frac{i}{2^N} - f^{(N)}, \frac{j}{2^N}) \cdot \frac{1}{2^N} \cdot \frac{1}{2^N} \\ &= \sum_j p_{i-2^N f^{(N)},j} \end{aligned}$$

$$\begin{aligned} H(\{q^{(N)}(x, f^{(N)})\}) &= \sum_i (\log \frac{1}{q^{(N)}(\frac{i}{2^N}, f^{(N)})}) \cdot q^{(N)}(\frac{i}{2^N}, f^{(N)}) \cdot \frac{1}{2^N} \\ &= \sum_i (\log \frac{2^N}{q_i(f^{(N)})}) \cdot q_i(f^{(N)}) \\ &= N + \sum_i (\log \frac{1}{q_i(f^{(N)})}) \cdot q_i(f^{(N)}) \\ &= N + H(\{q_i(f^{(N)})\}). \end{aligned}$$

Then

$$\begin{aligned} H(\{q_i(0)\}) - H(\{q_i(g^{(n)})\}) &= H(\{q^{(N)}(x,0)\}) - H(\{q^{(N)}(x, g^{(N)})\}) \\ &= (H(\{q^{(N)}(x,0)\}) - H(\{q(x,0)\})) \\ &\quad +(H(\{q(x,0)\}) - H(\{q(x,g)\})) \\ &\quad +(H(\{q(x,g)\}) - H(\{q^{(N)}(x,g)\})) \\ &\quad +(H(\{q^{(N)}(x,g)\}) - H(\{q^{(N)}(x,g^{(N)})\})) \\ &> -\frac{\epsilon}{3} + \epsilon - \frac{\epsilon}{3} - \frac{\epsilon}{3} \\ &= 0. \end{aligned}$$

On the other hand, Lemma 2.7 shows that

$$H(\{q_i(0)\}) \leq H(\{q_i(g^{(n)})\}),$$

which leads to a contradiction. Therefore, there does not exist g such that

$$H(\{q(x,g)\}) < H(\{q(x,0)\}) < \infty.$$

□

The next theorem shows how shift invariance can reduce the complexity of computing the conditional median.

Theorem 2.8 *If $P(X, \overrightarrow{Y})$ is CSUM and SI, then for every $\overrightarrow{y} \in R_m$,*

$$median(X|\overrightarrow{Y} = \overrightarrow{y}) = \arg\max_x p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x),$$

where $p_{\overrightarrow{Y}-X}$ is the density of the distribution on $\overrightarrow{Y} - X$.

Proof:

$$\begin{aligned}
& p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x) \\
= & \sum_t \Pr(X = x + t, \overrightarrow{Y} = \overrightarrow{y} + t) \\
= & \sum_t \Pr(X = x + t | \overrightarrow{Y} = \overrightarrow{y} + t) \cdot \Pr(\overrightarrow{Y} = \overrightarrow{y} + t) \\
= & \sum_t \Pr(X = x | \overrightarrow{Y} = \overrightarrow{y}) \cdot \Pr(\overrightarrow{Y} = \overrightarrow{y} + t) \qquad (SI) \\
= & \Pr(X = x | \overrightarrow{Y} = \overrightarrow{y}) \cdot \sum_t \Pr(\overrightarrow{Y} = \overrightarrow{y} + t) \\
= & \Pr(X = x | \overrightarrow{Y} = \overrightarrow{y}).
\end{aligned}$$

Since $X|\overrightarrow{Y} = \overrightarrow{y}$ is symmetric and unimodal,

$$\begin{aligned}
median(X|\overrightarrow{Y} = \overrightarrow{y}) &= \arg\max_x \Pr(X|\overrightarrow{Y} = \overrightarrow{y}) \\
&= \arg\max_x p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x)
\end{aligned}$$

□

Remarks:

1. In applications, construction of the conditional median requires first estimating an $(m+1)$-dimensional density $p(x, \overrightarrow{y})$, whereas $p_{\overrightarrow{Y}-X}$ has only m dimensions. This difference will be important in our experiments (see §4).

2. If we think of each component of $\overrightarrow{Y} = (Y_1, \cdots, Y_m)$ as a predictor of X, then Theorem 2 can be interpreted as a recipe for combining m predictors into 1 predictor, based on a likelihood principle.

2.4 Application

In this section, we apply Theorem 2 to construct a good predictor.

Recall that

$$\hat{x}_{LOCO} = median(a, b, a + b - c)$$

depends only on three pixels. In this case, the dimension of $\overrightarrow{Y}$ is 3. By Theorem 2, the best predictor g based on $\overrightarrow{Y} = (a, b, c)$ is

$$g(a, b, c) = \arg\max_x p_{\overrightarrow{Y}-X}(a - x, b - x, c - x).$$

In practice, it costs too much to send the entire empirical joint residual distribution. Therefore, we estimate $\arg\max_x p_{\overrightarrow{Y}-X}(a - x, b - x, c - x)$ directly. We looked at the empirical joint residual distributions of several different images, and tried to figure out where the peak of $p_{\overrightarrow{Y}-X}(a-x, b-x, c-x)$ is for every (a, b, c). We found that the predictor of LOCO is usually near the peak. This explains why LOCO has a great success. However, we also found that $\hat{x}_{LOCO}$ has a better prediction when $|a - b|$ is large than when it is small. To make the prediction closer to the peak, we modified the predictor a little bit when $|a - b|$ is small. This should lower the entropy of prediction residuals.

Figure 2.2: test images

$\hat{x}_{LOCO}$ can be written as

$$\hat{x}_{LOCO} = \begin{cases} \min(a,b) & \text{if } c \geq \max(a,b) \\ \max(a,b) & \text{if } c \leq \min(a,b) \\ a+b-c & \text{otherwise} \end{cases}$$

We constructed a new predictor $\hat{x}$, such that $\hat{x} = \hat{x}_{LOCO}$ when $|a-b| > 15$. When $|a-b| \leq 15$,

$$\hat{x} = \begin{cases} [\frac{a+b+\min(a,b)}{3}] & \text{if } c \geq \max(a,b) \\ [\frac{a+b+\max(a,b)}{3}] & \text{if } c \leq \min(a,b) \\ [0.6*a+0.6*b-0.2*c] & \text{otherwise,} \end{cases}$$

where $[t]$ denotes the nearest integer to t. We compared the modified predictor with LOCO predictor by calculating the entropy of the empirical residual distribution on six of ISO/JPEG test images. Figure-2.2 show the six images we used, and the result is presented in Table-2.1.

From Table-2.1, we can see that the modified predictor does better than the orig-

Image	LOCO	Modified	Optimum
Gold	4.46	4.42	4.36
Hotel	4.40	4.34	4.19
Water	3.59	3.49	3.43
Woman	4.45	4.41	4.36
Cmpnd1	1.13	1.13	1.03
Tools	5.25	5.23	5.07

Table 2.1: Entropies of Empirical Residual Distributions

inal predictor of LOCO. The column "Optimum" is the result of using the best predictor we constructed in the previous section. However, this predictor is based on the empirical joint residual distribution. In practice, we do not have the empirical joint residual distribution for free. Therefore, this is a bound which can never be achieved. Table-2.1 shows that the result from LOCO predictor is close to the bound. By creating a new prediction based on (a, b, c), one will not make a significant improvement in lowering the entropy.

2.5 Extension

In §3, we prove that

$$f(\overrightarrow{y}) = \arg\max_x p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x)$$

achieves

$$H(X - f(\overrightarrow{Y})),$$

if $(X, \overrightarrow{Y})$ is CSUM and SI. In this section, we will show that weaker assumptions can achieve similar results. We begin this section with more definitions. For any $\overrightarrow{y_0}$, define

$$D(\overrightarrow{y_0}) \equiv \{\overrightarrow{y_0} + s; s \in R\}.$$

For any $\overrightarrow{y} \in D(\overrightarrow{y_0})$, if $\overrightarrow{y'} = \overrightarrow{y} + s$, define

$$d_{\overrightarrow{y_0}}(\overrightarrow{y}) \equiv s.$$

Definition 2.9 *A distribution P(X, $\overrightarrow{Y}$) on R^{m+1} is diagonally symmetric and unimodal (DSUM) if*

$$p(x - d_{\overrightarrow{y_0}}(\overrightarrow{y}) | \overrightarrow{y} \in \overrightarrow{y_0})$$

is symmetric and unimodal in x for every $\overrightarrow{y_0}$.

Define

$$\mathcal{F} \equiv \{f : f(\overrightarrow{y} + s) = f(\overrightarrow{y}) + s \quad \forall \overrightarrow{y}, s\}.$$

We have a theorem similar to Theorem 2.3:

Theorem 2.10 *If $P(X, \overrightarrow{Y})$ is DSUM, then*

$$f(\overrightarrow{y}) \equiv median((X - d_{\overrightarrow{y}}(\overrightarrow{Y}))|\overrightarrow{Y} \in D(\overrightarrow{y}))$$

achieves

$$\min_{f \in \mathcal{F}} H(X - f(\overrightarrow{Y})).$$

Proof:

First, we check that $f \in \mathcal{F}$:

$$\begin{aligned} f(\overrightarrow{y} + s) &= median((X - d_{\overrightarrow{y}+s}(\overrightarrow{Y}))|\overrightarrow{Y} \in D(\overrightarrow{y} + s)) \\ &= median((X - d_{\overrightarrow{y}}(\overrightarrow{Y}) + s)|\overrightarrow{Y} \in D(\overrightarrow{y})) \\ &= median((X - d_{\overrightarrow{y}}(\overrightarrow{Y}))|\overrightarrow{Y} \in D(\overrightarrow{y})) + s \\ &= f(\overrightarrow{y}) + s \end{aligned}$$

Then the proof is almost the same as that of Theorem 1. $X - f(\overrightarrow{Y})$ is a mixture of $(X - f(\overrightarrow{Y}))|\overrightarrow{Y} \in D(\overrightarrow{y})$. It suffice to check that $meadian((X - f(\overrightarrow{Y}))|\overrightarrow{Y} \in D(\overrightarrow{y})) =$

0.

$$\begin{aligned} & X - f(\overrightarrow{Y}) \\ = \; & X - (f(\overrightarrow{y}) + d_{\overrightarrow{y}}(\overrightarrow{Y})) \\ = \; & X - d_{\overrightarrow{y}}(\overrightarrow{Y}) - f(\overrightarrow{y}) \end{aligned}$$

Since

$$f(\overrightarrow{y}) = median(X - d_{\overrightarrow{y}}(\overrightarrow{Y}) | \overrightarrow{Y} \in D(\overrightarrow{y})),$$

the median of $X - f(\overrightarrow{Y})$ is 0. □

Next, present a result similar to Theorem 2.8, but with assumption DSUM, instead of CSUM + SI.

Theorem 2.11 *If $(X, \overrightarrow{Y})$ is DSUM, then for every $\overrightarrow{y} \in R_m$,*

$$median((X - d_{\overrightarrow{y}}(\overrightarrow{Y})) | \overrightarrow{Y} \in D(\overrightarrow{y})) = \arg\max_x p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x),$$

where $p_{\overrightarrow{Y}-X}$ is the density of the distribution on $\overrightarrow{Y} - X$.

Proof:

$$\begin{aligned} p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x) &= \Pr(\overrightarrow{Y} - X = \overrightarrow{y} - x) \\ &= \Pr(X - \overrightarrow{Y} + \overrightarrow{y} = x) \\ &= \Pr(X - d_{\overrightarrow{y}}(\overrightarrow{Y}) = x). \end{aligned}$$

Since $(X - d_{\overrightarrow{y}}(\overrightarrow{Y}))|D(\overrightarrow{y})$ is symmetric and unimodal, we have

$$\begin{aligned} \arg\max_x p_{\overrightarrow{Y}-X}(\overrightarrow{y} - x) &= \arg\max_x \Pr(X - d_{\overrightarrow{y}}(\overrightarrow{Y}) = x) \\ &= \arg\max_x \Pr(X - d_{\overrightarrow{y}}(\overrightarrow{Y}) = x | \overrightarrow{Y} \in D(\overrightarrow{y})) \\ &= median(X - d_{\overrightarrow{y}}(\overrightarrow{Y}) = x | \overrightarrow{Y} \in D(\overrightarrow{y})) \end{aligned}$$

□

Remark: CSUM + SI → DSUM, but the converse is not necessarily true.

Bibliography

[1] M. Weinberger, G. Seroussi, G. Sapiro. LOCO-I: A Low Complexity, Context-Based, Lossless Image Compression Algorithm. Proc. IEEE Data Compression Conference, Snowbird, Utah, March-April 1996.

[2] M. Weinberger, G. Seroussi, G. Sapiro. The LOCO-I Lossless Image Compression Algorithm: Principles and Standardization into JPEG-LS. Hewlett-Packard Laboratories Technical Report No. HPL-98-193R1, November 1998, revised October 1999. *IEEE Trans. Image Processing*, Vol. 9, August 2000, pp.1309-1324.

[3] X. Wu and N. D. Memon. Context-based, adaptive, lossless image coding. *IEEE Trans. Commun.* Vol. 45 (4), pp. 437-444, Apr. 1997.

[4] X. Wu. Efficient Lossless Compression of Continuous-tone Images via Context Selection, Quantization, and Modeling. *IEEE Trans. Image Processing*, Vol. IP-6, pp. 656-664, May 1997.

[5] A. Netravali and J. O. Limb. Picture coding: A review. *Proc. IEEE*, Vol. 68, pp. 366-406, 1980

[6] P. G. Howard and J. S. Vitter. Fast and efficient lossless image compression. *Proc. 1993 Data Compression Conference*, pp.351-360, Mar. 1993.