

# Information Scaling Laws in Natural Scenes

Cheng-En Guo<sup>1</sup>, Ying Nian Wu<sup>2</sup>, and Song Chun Zhu<sup>1,2</sup>

Departments of Computer Science<sup>1</sup> and Statistics<sup>2</sup>

University of California, Los Angeles

{cguo, ywu, sczhu}@stat.ucla.edu

## Abstract

In natural scenes, objects and patterns can appear at a wide variety of distances from the viewer. For the same visual pattern viewed at different distances, both the image and our perception of the pattern change over distance. We call the change of the image over distance as *image scaling*, and the change of our perception over distance as *information scaling*. While image scaling can be accounted for by the state space theory, information scaling has not been mathematically studied in computer vision. In this paper, we prove two information scaling laws: 1) the entropy rate of the image changes over distance, and 2) the entropy of the posterior distribution of the pattern also changes over distance. These two information scaling laws have deep implications in computer vision: they call for different models of the same visual pattern at different distances, as well as a model transition mechanism for switching models over different distance/scale regimes.

## 1 Introduction

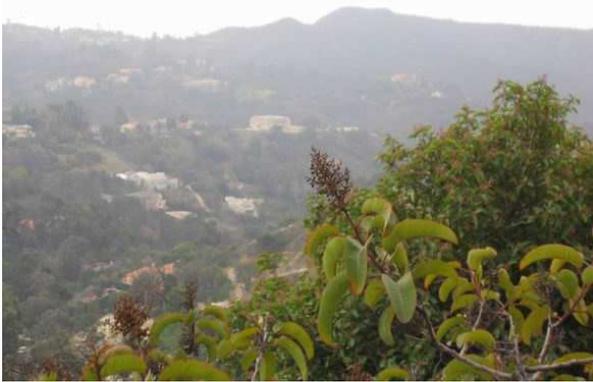


Figure 1: Leaves at different distances lead to different perceptions, from individual structures to foliage texture to flat region.

Visual patterns and objects in natural scenes appear at a wide variety of distances from the viewer, and the same pattern may lead to different perceptions when viewed at

different distances. See Figure 1 for an example, where the leaves nearby can be perceived as individual structures, but the leaves at a distance only give us an impression of foliage texture, and the leaves of the trees on the mountain afar result in a more or less flat region in the image. Similarly, when we look at a person at a far distance, we can only recognize his or her face. When we move closer, we start to see the eyes, nose, and mouth. When we move still closer, we begin to notice facial marks such as wrinkles, etc. Therefore, the same visual pattern or object needs to be described by different models at different distances or scales.

When we study what happens if the viewer changes distance from a visual pattern, it is important to distinguish between the change of the image and the change of our perception. The former can be called *image scaling*, and the latter can be called *information scaling*. From a human vision perspective, image scaling occurs in retina, whereas information scaling occurs in visual cortex. While image scaling can be accounted for by the scale space theory[4][15], to the best of our knowledge, there has not been any mathematical theory for information scaling, even though some theories do take the issue of scaling into account, such as multi-resolution analysis[7], fractals[10], feature statistics of natural images[5][16][11][17].

In this paper, we prove two information scaling laws: 1) the complexity scaling law (Theorems 1&2) – the entropy rate of the image changes over distance, and 2) the perceptibility scaling law (Theorem 4) – the entropy of the posterior distribution of the pattern also changes over distance. These two information scaling laws have deep implications in computer vision: they call for different models of the same visual pattern viewed at different distances, as well as a model transition mechanism for switching models over different distance/scale regimes.

## 2 Complexity Scaling Law

Let  $I$  be the image of a pattern observed at a certain distance, and let's assume that  $I$  follows a probability distribution  $p(I)$ . There are two interpretations of  $p(I)$ .

1.  $p(I)$  is the result of the generative process that produces the images of the visual pattern.

2.  $p(I)$  corresponds to the coding or compression scheme to approximate the Komolgorov algorithmic complexity of an observed image  $I_{\text{obs}}$ .

Let's assume that  $I$  is defined on a lattice  $\Lambda$ . Let  $|\Lambda|$  be the size or the number of pixels of  $\Lambda$ .

**Definition 1: Image complexity**, denoted by  $\mathcal{H}(I)$ , is defined as the entropy of  $p(I)$ , i.e.,  $\mathcal{H}(p(I)) = -\sum_I p(I) \log p(I)$ . The **complexity rate** (or complexity per pixel) of the image is defined as  $\mathcal{H}(I)/|\Lambda|$ .

When we move away from a pattern, the change of image of the pattern involves both local smoothing and down-sampling. Let's first study the effect of down-sampling. To simplify the situation, we assume that we down-sample  $I$  by a factor of 2 along both vertical and horizontal axes. Then there are four down-sampled versions, and let's denote them by  $I_-^{(k)}$ ,  $k = 1, 2, 3, 4$ , each defined on a down-sampled lattice  $\Lambda_-$ , with  $|\Lambda_-| = |\Lambda|/4$ . See Figure 2 for an illustration.

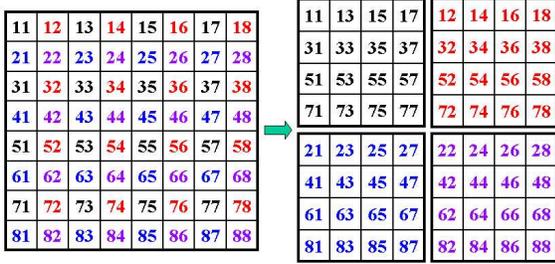


Figure 2: The four down-sampled versions of the original image.

**Theorem 1:** 1) The complexity of  $I_-^{(k)}$  is smaller than the complexity of  $I$ ,  $\mathcal{H}(I_-^{(k)}) \leq \mathcal{H}(I)$ ,  $k = 1, \dots, 4$ , and  $\mathcal{H}(I) - \mathcal{H}(I_-^{(k)}) = \mathcal{H}(I|I_-^{(k)})$ , where  $\mathcal{H}(I|I_-^{(k)})$  is the conditional entropy of  $p(I|I_-^{(k)})$ .

2) The complexity rate of  $I_-^{(k)}$  is larger than the complexity rate of  $I$ ,

$$\frac{1}{4} \frac{1}{|\Lambda_-|} \sum_{k=1}^4 \mathcal{H}(I_-^{(k)}) \geq \frac{1}{|\Lambda|} \mathcal{H}(I), \quad (1)$$

and

$$\frac{1}{4} \frac{1}{|\Lambda_-|} \sum_{k=1}^4 \mathcal{H}(I_-^{(k)}) - \frac{1}{|\Lambda|} \mathcal{H}(I) = \frac{1}{|\Lambda|} \mathcal{M}(I_-^{(k)}), k = 1, \dots, 4,$$

where  $\mathcal{M}()$  denotes mutual information among the four down-sampled versions.

**Proof:** 1)  $p(I|I_-^{(k)}) = p(I)/p(I_-^{(k)})$  since  $I_-^{(k)}$  is fully determined by  $I$ . Thus

$$\mathcal{H}(I) - \mathcal{H}(I_-^{(k)}) = E_I \left[ -\log \frac{p(I)}{p(I_-^{(k)})} \right]$$

$$= E_I [-\log p(I|I_-^{(k)})] = \mathcal{H}(I|I_-^{(k)}) \geq 0.$$

2)

$$\begin{aligned} \sum_{k=1}^4 \mathcal{H}(I_-^{(k)}) - \mathcal{H}(I) &= E \left[ \log \frac{p(I)}{\prod_k p(I_-^{(k)})} \right] \\ &= \mathcal{M}(I_-^{(k)}, k = 1, \dots, 4) \geq 0. \quad \text{QED} \end{aligned}$$

This theorem tells us that if we down-sample an image, the complexity of the image will decrease, but the complexity rate will increase. In other words, inequality (1) tells us that the image looks more random. This can be easily understood from real life experience. For instance, for the leaf pattern in Figure 1, when we move farther away from the leaves, we lose information, so the complexity is decreasing. But we see more leaves within unit area of visual field, so the complexity rate is increasing.

One can also understand this result from the perspective of Komolgorov complexity. The shortest algorithmic coding length of  $I$  must be greater than or equal to the shortest coding length of any of the  $I_-^{(k)}$ , but must be smaller than or equal to the sum of the shortest coding lengths of the four  $I_-^{(k)}$ .

Now let's study the effect of local averaging. Let  $I$  be the original image, and let  $J$  be the image obtained by convolving the image  $I$  with a localized smoothing kernel  $k$ , i.e.,  $J = I * k$ .

**Theorem 2:** As the lattice  $\Lambda \rightarrow Z^2$ ,

$$\frac{1}{|\Lambda|} \mathcal{H}(J) - \frac{1}{|\Lambda|} \mathcal{H}(I) \rightarrow \int \log |\hat{k}(\omega)| d\omega, \quad (2)$$

where the right hand side is smaller than or equal to 0. Here  $\hat{k}$  is the Fourier transform of the kernel  $k$ ,  $\omega \in [-\pi/2, \pi/2] \times [-\pi/2, \pi/2]$  is the spatial frequency.

**Proof:** In the Fourier domain, we have  $\hat{J}(\omega) = \hat{I}(\omega) \hat{k}(\omega)$ , where  $\hat{J}$  and  $\hat{I}$  are Fourier transforms of  $J$  and  $I$  respectively. For finite rectangular lattice  $\Lambda$ , the spatial frequency  $\omega$  takes values in a finite grid. Since the Fourier transform is orthogonal, we have  $\mathcal{H}(I) = \mathcal{H}(\hat{I})$ , and  $\mathcal{H}(J) = \mathcal{H}(\hat{J})$ . Thus

$$\frac{1}{|\Lambda|} \mathcal{H}(J) = \frac{1}{|\Lambda|} \mathcal{H}(I) + \frac{1}{|\Lambda|} \sum_{\omega} \log |\hat{k}(\omega)|. \quad (3)$$

As  $\Lambda \rightarrow Z^2$ , the second term on the right hand side goes to  $\int \log |\hat{k}(\omega)| d\omega$ .

A smoothing kernel  $k$  is a probability distribution function,  $\hat{k}$  is the so-called characteristic function of  $k$ , and  $\hat{k}(\omega) = \sum_x k(x) e^{-i\omega x} = E_k[e^{-i\omega x}]$ , with  $x \sim k(x)$ . So  $|\hat{k}(\omega)|^2 = |E_k[e^{-i\omega x}]|^2 \leq E_k[|e^{-i\omega x}|^2] = 1$ . Thus,  $\int \log |\hat{k}(\omega)| d\omega \leq 0$ . QED

Equation (2) shows that when we smooth an image, the complexity rate will decrease, i.e., we smooth out some randomness. If the image is large, then the decrease in complexity rate approaches a constant.

The *complexity scaling law* is a combination of Theorem 1 and Theorem 2. Let  $J = I * k$ , and let  $J_-$  be the down-sampled version of  $J$  by a factor of 2 on both axes. Then,  $J_-$  can be considered a *down-scaled* version of  $I$ , that is, the image of the underlying visual pattern changes from  $I$  to  $J_-$  if the distance between the viewer and the visual pattern doubles. For large lattice size, we have the *complexity scaling law*:

$$\frac{1}{4} \frac{1}{|\Lambda_-|} \sum_{k=1}^4 \mathcal{H}(J_-^{(k)}) - \frac{1}{|\Lambda|} \mathcal{H}(I) - \left[ \frac{1}{|\Lambda|} \mathcal{M}(J_-^{(k)}, k = 1, 2, 3, 4) + \int \log |\hat{k}(\omega)| d\omega \right] \rightarrow 0.$$

If the mutual information per pixel is greater than  $-\int \log |\hat{k}(\omega)| d\omega$ , we observe an increase in complexity rate. Otherwise, the complexity rate decreases.

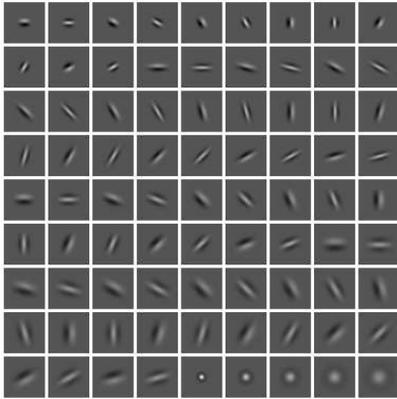


Figure 3: Linear bases/filters of different scales and orientations.

Next, we will demonstrate the complexity scaling law with some experiments. We use three indicators for the entropy of an image:

1) The entropy of the histogram of the intensity gradients (along horizontal direction). This is a very rough indicator, but it appears consistent with the following two indicators.

2) The coding length per pixel of the image using JPEG 2000. This method codes the image using wavelet bases.

3) Code the image (the whitened version where the high frequencies are enhanced, see [12]) by the following model:  $I = \sum_i c_i B_i + \epsilon$ , where  $c_i$  are the coefficients, and  $B_i$  form an overcomplete system of localized, oriented, and elongated bases such as Difference of Offset Gaussian (DOOG) bases [6](or Gabor bases). See Figure 3 for an illustration of such linear bases. We count the number of bases per pixel needed to explain 70% of the intensity variance. We use the matching pursuit algorithm [8] to select bases.

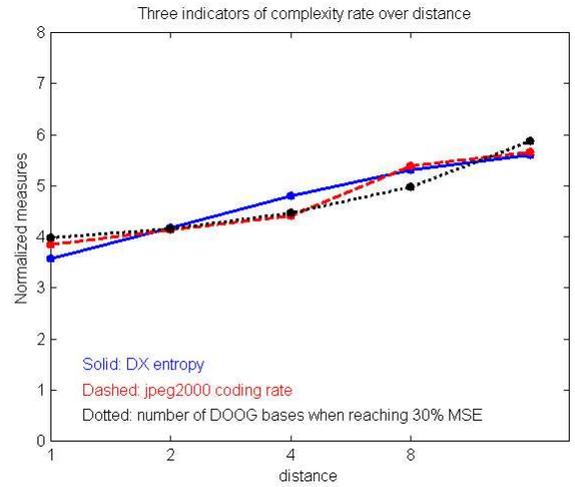


Figure 4: The figure above displays the original image of ivy wall and its down-scaled versions. The plot below shows the change of complexity rate (three indicators) over the cascade order of down-scaling.

We did an experiment on an image of ivy wall in Figure 4. For this image, we keep down-scaling it by a factor of  $2 \times 2$ . Then we plot the entropy rate of the image over the cascade order of down-scaling. Clearly, there is an upward trend in complexity rate within the range of the plot, indicating that the image is getting more random (Here the measures of the three indicators of the image complexity are normalized by linear transforms so that they can be displayed together in one plot). This upward trend continues persistently until the ivy wall is extremely far away from the viewer. Then the local averaging effect will smooth out

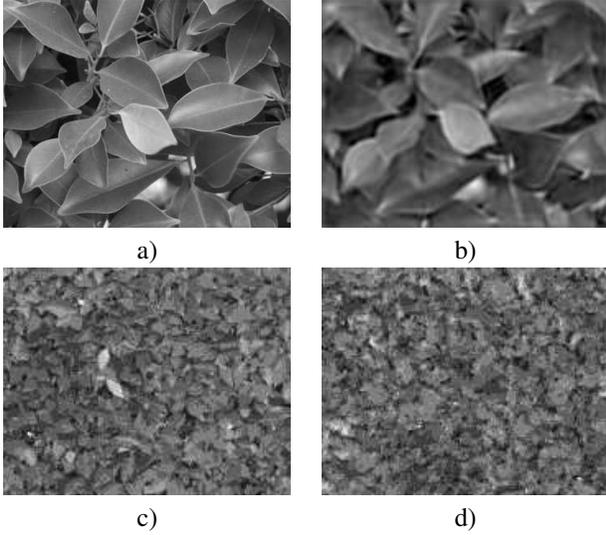


Figure 5: From sparse coding to feature statistics. a) Observed near-distance image. b) Reconstructed by sparse coding with 1,000 bases. c) Observed far-distance image. d) “Reconstructed” by matching feature statistics.

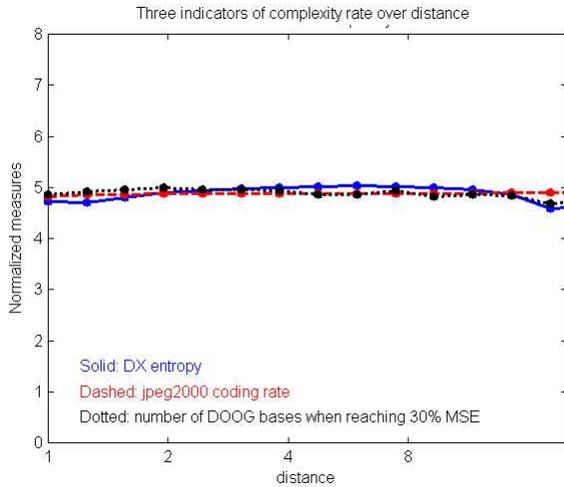


Figure 6: A scale invariant image and the change of complexity rate over cascade order of down-scaling.

the randomness, so that the complexity rate will start to decrease.

This phenomenon has interesting implication in sparsity principle. Sparsity has been a cherished principle in mathematics. It has also been proposed by Olshausen and Field [12] as a principle employed by the primitive visual cortex or V1. They used this principle to learn a set of linear bases that resemble the properties of simple V1 cells. In Figure 4, at near distance, the complexity rate is very low, so sparsity principle applies, and we observe individual structures. But as the viewer moves farther from the underlying pattern, or if we down-scale the image, the complexity rate of the image will increase, so that there may not exist any sparse deterministic representation of the image, and the sparsity principle is violated. As a result, the visual system may only interpret the image by some summaries that cannot determine the image deterministically, and these summaries are feature statistics, and we perceive collective textures instead of individual structures. This may explain the perceptual transition from structures to textures, or from sparse coding to feature statistics. See Guo, Zhu, Wu (2003) [2] for more details, where they call the model regime where sparse coding applies as sketchable regime, and the model regime where we need to use feature statistics as the non-sketchable regime.

Figure 5 displays an example of the transition from sparse coding to feature statistics. a) and c) are images of ivy-wall at near-distance and far-distance respectively. b) is reconstructed near-distance image using sparse coding representation with 1,000 bases selected by the matching pursuit algorithm. d) is statistically reconstructed far-distance image using feature statistics representation by matching histograms of filter responses.

In the second experiment, Figure 6 displays a picture of natural scene, where objects appear at various distances, resulting in a wide variety of scales. If we repeat the first experiment on this picture, we see that the complexity rate does not change much as we down-scale the image. This suggests that the image is scale invariant. See [16][11][17] for more discussions.

In the third experiment, Figure 7 shows a sequence of images of trees taken in Los Angeles area. At the close distance, the image is dominated by one tree in the visual field. At the far distance, we see the forest on the mountain. Figure 8 plots the three complexity indicators of these images ordered by the distance or scale.

When the objects in the visual field are of near distances or large scales, and each object has smooth surface, the mutual information we discussed above can be quite large. As a result, if we down-scale the image, we will observe an increase in complexity rate. When the objects in the visual field are of far distances or small scales, the mutual information can be rather small. If we down-scale the image, we

will observe a decrease in complexity rate.

In the fourth experiment, we study the coding of a sequence of brick wall images (whitened version) using the linear additive model  $I = \sum_i c_i B_i + \epsilon$ , where  $\{B_i\}$  is a set of overcomplete bases. In our experiment, we use two sets of bases. One set consists of Difference of Offset Gaussian (DOOG) bases as we discussed before (see Figure 3). The other set consists of primitives learned from the image. The primitives learned from the brick wall image (left) in Figure 9 are shown in Figure 10. In Figure 9 we plot the mean squared error versus the number of bases for each image under the two vocabularies of bases. For the brick wall image taken at a close distance, the primitives are more efficient. But for the brick wall image taken at a far distance, the DOOG bases win over the primitives. Clearly, there are different distance/complexity regimes where different generative models are needed for the same type of pattern. We shall study the issue of model transition over distance in future work.

### 3 Perceptibility Scaling Law

The purpose of vision is to make inference about the outside world that generates the image. Now, let's study the issue of information scaling in an inferential framework in the context of generative modeling.

Let  $W$  describe the outside world that produces the image  $I$ . Let's assume that both  $W$  and  $I$  are properly discretized, and that  $W$  is detailed enough to produce  $I$ , i.e.,  $I = g(W)$ , where the many to one function  $g()$  can be thought of as a graphics process. For natural patterns such as foliage and grass,  $W$  is typically very complex, including detailed descriptions of all the leaves and strands of grass. Such visual complexity is a defining characteristic of natural scenes and is a key factor for visual realism in graphics and paintings.

Suppose our prior knowledge about  $W$  can be represented by a prior distribution  $p(W)$ . Again, we can either interpret  $p(W)$  as result of the stochastic process that gives rise to  $W$ , or frequencies of an ensemble of scenes, or a scheme of coding  $W$  to approximate Komolgorov complexity. Given  $W \sim p(W)$ , and  $I = g(W)$ , we have the marginal distribution of  $I$ :  $p(I) = \sum_{W:g(W)=I} p(W)$ . The posterior distribution of  $W$  given  $I$  is  $p(W|I) = p(W, I)/p(I) = p(W)/p(I)$  when  $I = g(W)$ .  $p(W, I) = p(W)$  because  $I$  is fully determined by  $W$ . This posterior distribution defines our perception of  $W$  based on  $I$ , and is an inversion of the graphics equation  $I = g(W)$ .

**Definition 2: Scene complexity**, denoted by  $\mathcal{H}(W)$ , is defined as  $\mathcal{H}(p(W)) = -\sum_W p(W) \log p(W)$ .

**Definition 3: Imperceptibility**, denoted by  $\mathcal{H}(W|I)$ , is defined as  $\mathcal{H}(p(W|I)) = -\sum_W p(W) \log p(W|I)$ .

**Theorem 3:** Let  $W \sim p(W)$ , and  $I = g(W)$ , then

$$\mathcal{H}(W|I) = \mathcal{H}(W) - \mathcal{H}(I). \quad (4)$$

That is, **imperceptibility = scene complexity - image complexity**.

**Proof:**  $p(W|I) = p(W)/p(I)$ , by taking log on both sides, and then taking expectation, Theorem 2 follows. QED

One may interpret equation (4) from the perspective of inverting the graphics equation  $I = g(W)$ . The scene complexity  $\mathcal{H}(W)$  counts the number of the unknowns (in terms of bits), and the image complexity  $\mathcal{H}(I)$  counts the number of equations. The imperceptibility  $\mathcal{H}(W|I)$  then tells us the degrees of freedom that are left undetermined from this equation. Thus the imperceptibility  $\mathcal{H}(W|I)$  gives a general definition of "ill-posedness" of the inversion problem.

For an image  $I$ , its down-scaled version  $I_-$  can be obtained by local smoothing and down-sampling, and the process can be represented by a many to one reduction function  $R()$ , such that  $I_- = R(I)$ . Then we have the following *perceptibility scaling law*.

**Theorem 4:** For  $W \sim p(W)$ ,  $I = g(W)$ , if  $I_- = R(I)$  with  $R()$  being any many to one reduction function, then

$$\mathcal{H}(W|I_-) \geq \mathcal{H}(W|I). \quad (5)$$

That is, *imperceptibility becomes larger with down-scaling*.

**Proof:**  $\mathcal{H}(W|I_-) = \mathcal{H}(W) - \mathcal{H}(I_-)$ ,  $\mathcal{H}(W|I) = \mathcal{H}(W) - \mathcal{H}(I)$ , and  $\mathcal{H}(I) - \mathcal{H}(I_-) = \mathcal{H}(I|I_-)$ . So  $\mathcal{H}(W|I_-) - \mathcal{H}(W|I) = \mathcal{H}(I|I_-) \geq 0$ . QED

Inequality (5) tells us that if we get farther from the scene  $W$ , it will become less perceptible, and if  $\mathcal{H}(W|I_-)$  is above a threshold, then our perception may experience a transition. We can only perceive some aspect of  $W$ , i.e.,  $W_- = \rho(W)$ , for some many to one reduction  $\rho()$ , such that  $\mathcal{H}(W_-|I_-)$  remains small. It is possible to find such a  $W_-$ , because of the following theorem.

**Theorem 5:** For  $W \sim p(W)$ ,  $I = g(W)$ , and  $I_- = R(I)$ ,  $W_- = \rho(W)$ , we have

$$\begin{aligned} 1) \quad & \mathcal{H}(W_-|I_-) \leq \mathcal{H}(W|I_-). \\ 2) \quad & p(I_-|W_-) = \frac{\sum_{W:\rho(W)=W_-;R(g(W))=I_-} p(W)}{\sum_{W:\rho(W)=W_-} p(W)}. \end{aligned}$$

**Proof:** 1)  $p(W|I_-)/p(W_-|I_-) = p(W|W_-, I_-)$ , thus  $\mathcal{H}(W|I_-) - \mathcal{H}(W_-|I_-) = \mathcal{H}(W|W_-, I_-) \geq 0$ . 2)  $p(I_-|W_-) = p(I_-, W_-)/p(W_-)$ . QED

Result 2) tells us that although  $W$  defines  $I$  deterministically via  $I = g(W)$ ,  $W_-$  may only define  $I_-$  statistically via a probability distribution  $p(I_-|W_-)$ . While  $W$

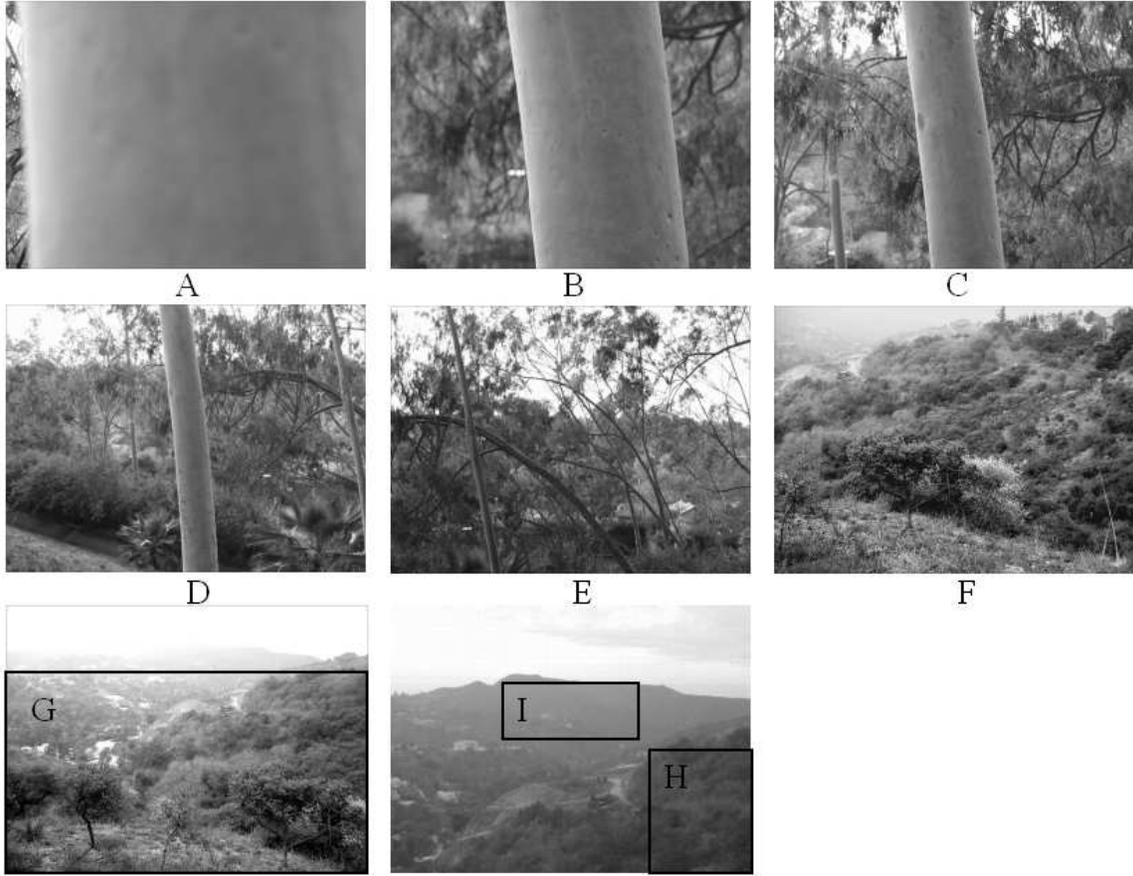


Figure 7: Natural images taken at different distances from the trees.

represents sparse structures,  $W_-$  may only represent collective textures. Figure 11 illustrates the image and the world which are described at two scales.  $W$  is not perceptible from  $I_-$ . While  $I$  is a deterministic function of  $W$ ,  $W_-$  may define  $I_-$  statistically.

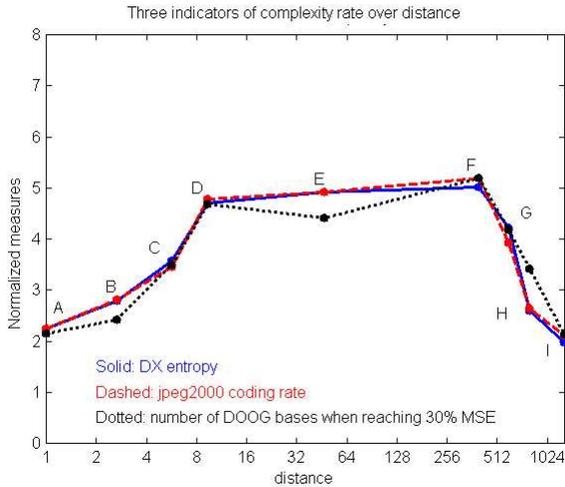


Figure 8: The change of the complexity rate of the images over distance in Figure 7.

For patterns like hair, fur, sands and soil, the physical  $W$ , which includes all individual strands of hairs or individual grains of sands, is perhaps never fully perceptible. This is also true with water, fire, smoke, and clothes, for which the fluid dynamics or mechanics models of  $W$  consists of a large number of particles. In that case, we are always in the  $(W_-, I_-)$  regime, and the physics model is not relevant to vision.

What is the implication of the perceptibility scaling law in generative modeling? Let  $I_d$  be the image of a visual pattern at distance  $d$ . If we model  $I_d$  by a single generative model  $W \sim p(W)$ , and  $I_d | W \sim p_d(I_d | W)$ , then the imperceptibility  $\mathcal{H}(W|I_d)$  is increasing as  $d$  increases, and if it is above a certain threshold, the imperceptibility will be too large for the generative model to be useful. In that case, we need to use a set of generative models  $W_d \sim p(W_d)$ , and  $I_d | W_d \sim p_d(I_d | W_d)$ , and for different distance regime, we may need different  $W_d$ .

Interestingly, the inferential concept of perceptibility also arises from the representation and coding perspective, even if we do not assume an objective  $W$ . That is, we only assume  $I \sim p(I)$ , and  $W$  is an augmented variable purely for the purpose of coding  $I$ , via a model  $W \sim f(W)$  and

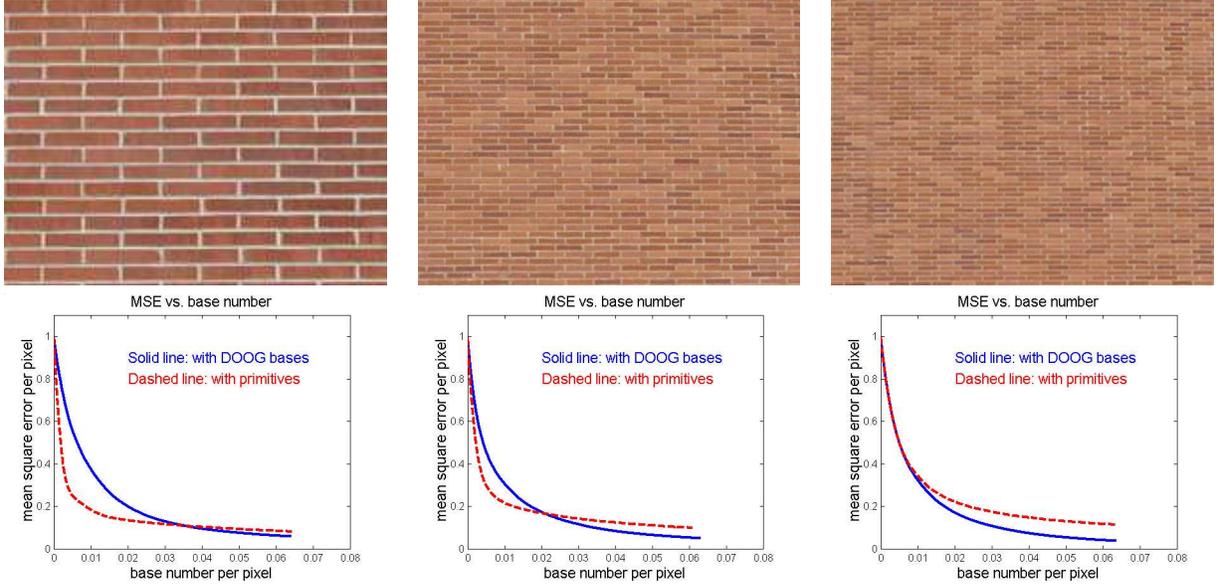


Figure 9: The figures above display brick wall images taken at different distances. The plots below show the change of mean square error over the number of bases under two different vocabularies: one consists of DOOG bases, and the other consists of primitives learned from images.

$[I|W] \sim f(I|W)$ . In this model, the marginal distribution  $I$  is  $f(I) = \sum_W f(W)f(I|W)$ , and the posterior distribution  $f(W|I) = f(W)f(I|W)/f(I)$ . In this coding scheme, for an image  $I$ , we first estimate  $W$  by a sample from the posterior distribution  $f(W|I)$ , then we code  $W$  by  $f(W)$  with coding length  $-\log f(W)$ . After that, we code  $I$  by  $f(I|W)$  with coding length  $-\log f(I|W)$ . So the average coding length is

$$-\mathbb{E}_p [\mathbb{E}_{f(W|I)}(\log f(W) + \log f(I|W))].$$

**Theorem 6:** *The average coding length is*

$$\mathbb{E}_p[\mathcal{H}(f(W|I))] + D(p||f) + \mathcal{H}(p). \quad (6)$$

That is, **coding redundancy = imperceptibility + error**. Here  $\mathcal{H}(f(W|I)) = -\sum_W f(W|I) \log f(W|I)$ , and  $D(p||f)$  is the Kullback-Leibler distance.

**Proof:** The theorem follows from

$$\begin{aligned} & \mathbb{E}_p \{ \mathbb{E}_{f(W|I)}[\log f(W) + \log f(I|W)] - \log p(I) \} \\ &= \mathbb{E}_p \{ \mathbb{E}_{f(W|I)}[\log f(W) + \log f(I|W) - \log f(I)] \} \\ &- \mathbb{E}_p[\log p(I) - \log f(I)]. \quad \text{QED} \end{aligned}$$

Equation (6) tells us that we not only want the error to be small, we also want  $W$  to be perceptible, in the sense that  $\mathbb{E}_p[\mathcal{H}(f(W|I))]$  is small.

As another example of imperceptibility, let's go back to the sparse coding model using linear bases, with  $I = \sum_i c_i B_i + N(0, \sigma^2)$ , and  $c = \{c_i\} \sim p(c)$  (see, e.g., Pece [13]). We suspect that if  $\mathcal{H}(c)$  is large, then the imperceptibility  $\mathcal{H}(c|I)$  should also be large. However, we have not

established any general result on this issue. If the result is true, then it has interesting implication. That is, if the image has a high complexity rate, then we may not want to pursue a sparse coding model of the above form, because even the sparsest representation may need a lot of bases, i.e.,  $\mathcal{H}(c)$  is large, then  $\mathcal{H}(c|I)$  can also be large, so there is a lot of ambiguities as to how to code  $I$  using the linear bases. If this is the case, we may switch to the model based on feature statistics.

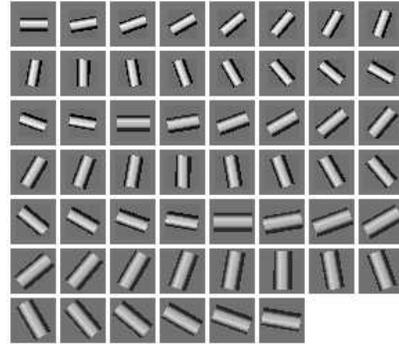


Figure 10: The primitives learned from the brick wall image (left) in Figure 9.

## 4 Future Work

In this paper, we established the complexity scaling law and the perceptibility scale law. The two information scaling laws imply that we need to use different models for the same visual pattern viewed at different distance/scale regimes.

In the future work, we shall study concrete models for visual patterns at different distances, and study the forms of

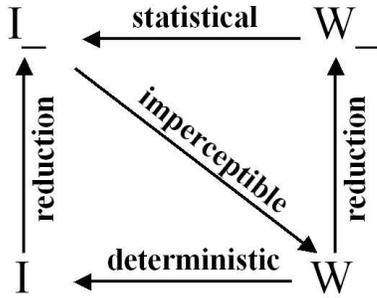


Figure 11: Illustration of down-scaling and imperceptibility.

the models at different distance/scale regimes, the underlying vocabularies of different models, and the criterion for model switching over distance. In the context of concrete models and patterns, we shall be able to compute complexity rate and imperceptibility for different models, and the comparison between different models will lead to a mechanism for model switching.

## Acknowledgements

This work is supported by NSF IIS-0244763, NSF-0240148, a sloan fellowship, and ONR N00014-02-1-0952. We thank Arthur Pece for his advice on presentation, and the four reviewers for their helpful comments.

## References

- [1] E. J. Candes and D. L. Donoho, "Ridgelets: a Key to Higher-dimensional Intermittency?" *Phil. Trans. R. Soc. Lond. A.*, Vol. 357, pp. 2495-509, 1999.
- [2] C. E. Guo, S. C. Zhu, and Y. N. Wu, "Towards a Mathematical Theory of Primal Sketch and Sketchability," *Proc. of Int'l Conf. on Computer Vision*, Nice France, 2003.
- [3] D. J. Heeger and J. R. Bergen, "Pyramid Based Texture Analysis/Synthesis," *Computer Graphics Proc.*, 1995.
- [4] T. Lindeberg, "Scale-space Theory: a Basic Tool for Analyzing Structures at Different Scales," *J. of Applied Stat.*, Vol. 21(2), pp. 225-270, 1994.
- [5] A. Lee, D. Mumford, and J. Huang, "Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model," *Int'l Journal of Computer Vision*, Vol. 41(1/2), pp. 35-59, 2001.
- [6] J. Malik and R. Perona, "Preattentive Texture Discrimination with Early Vision Mechanisms," *J. Opt. Soc. AM*, Vol. 7(5), pp. 923-932, 1990.
- [7] S. Mallat, "A Theory of Multiresolution Signal Decomposition: the Wavelet Representation," *IEEE Trans. PAMI*, Vol. 11(7), pp. 674-693, 1989.
- [8] S. Mallat and Z. Zhang, "Matching Pursuit in a Time-Frequency Dictionary," *IEEE Sig. Proc.*, Vol. 41, pp. 3397-415, 1993.
- [9] D. Marr, *Vision*, W. H. Freeman and Company, 1982.
- [10] B.B. Mandelbrot, *The fractal Geometry of Nature*, S.F. CA. Freeman, 1982
- [11] D. Mumford and B. Gidas, "Stochastic models for generic images," *Quarterly of Applied Math.*, Vol. 59(1), pp. 85-111, 2001.
- [12] B. A. Olshausen and D. J. Field, "Emergence of Simple-cell Receptive Field Properties by Learning a Sparse Code for Natural Images," *Nature*, Vol. 381, pp. 607-609, 1996.
- [13] A. Pece, "The Problem of Sparse Image Coding," *Journal of Mathematical Imaging and Vision*, vol. 17(2), pp. 89-108, 2002.
- [14] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int'l Journal of Computer Vision*, Vol. 40(1), pp. 49-71, October, 2000.
- [15] B. M. ter Haar Romeny, *Front-End Vision and Multiscale Image Analysis: Introduction to Scale-Space Theory*, Dordrecht, the Netherlands: Kluwer Academic Publishers, 1997.
- [16] D. L. Ruderman and W. Bialek, "Statistics of Natural Images: Scaling in the Woods," *Phy. Rev. Lett.*, Vol. 73, 1994.
- [17] E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," *Annual Review of Neuroscience*, Vol. 24, pp. 1193-1216, May 2001.
- [18] Y. N. Wu, S. C. Zhu, and X. Liu, "Equivalence of Julesz and Gibbs Ensembles," *Proc. of ICCV*, Corfu, Greece, 1999.
- [19] S. C. Zhu, Y. N. Wu, and D. Mumford, "Minimax Entropy Principle and Its Applications in Texture Modeling," *Neural Computation*, Vol. 9(8), pp. 1627-1660, 1997.