Perceptual Scaling

Wu, Ying N
Guo, Cheng-en
Zhu, Song C.

2004

Perceptual Scaling

Ying Nian Wu, Cheng-En Guo, and Song Chun Zhu
Department of Statistics, UCLA

Introduction

Vision as statistical learning and inference

Vision can be posed as a statistical learning and inference problem. As an over-simplified account, let W be a description of the outside scene in terms of "what is where," let I be the retina image, and let $p(W, I)$ be the joint distribution of W and I.[1] Then visual learning is to learn $p(W, I)$ from training data, and visual perception is to infer W from I based on $p(W|I)$.

There are two major schools on visual learning and perception. One school is operation-oriented and learns the inferential process defined by $p(W|I)$ directly, often in the form of an explicit transformation $W \approx F(I)$. This scheme is mostly used in supervised learning, where W is object category, and is given in training data. The other school is representation-oriented and learns the generative process $p(W)$ and $p(I|W)$ explicitly, then perception is to invert the generative process by maximizing or sampling $p(W|I) \propto p(W)p(I|W)$. In this scheme, $p(W)$ may also be accounted for by a regularization term such as smoothness or sparsity. This scheme is often used in unsupervised learning where W is not available in training data.

In the literature, there are a number of statistical theories proposed for vision. In representation-oriented school, Grenander (1993) and Mumford (1994) proposed pattern theory as a paradigm for vision (see also Geman and Geman, 1984; Amit, Grenander and Piccioni, 1991; Grenander and Miller, 1994; and S. Geman, Potter, and Chi, 2002, for important contributions that

[1] In a philosophically more rigorous formulation, we may assume the existence of an underlying world, which is a functional. When this functional acts on the physical equipments, it gives what we call "W." When this functional acts on the retina cells, it gives what we call "I." A distribution over this "world functional" leads to the joint distribution of W and I. See e.g., Mumford and Gidas (2001).

are related to pattern theory). Olshausen and Field (1996) proposed the sparsity principle as a general strategy employed by primitive visual cortex, and use it to learn linear bases from natural images, and these bases are considered mathematical models for simple visual cells (see also Bell and Sejnowski, 1997, on independent component analysis for learning edge filters from natural images). The sparsity principle was also investigated by Candes and Donoho (1999) in the framework of harmonic analysis on wavelets and curvelets. Zhu, Wu, and Mumford (1997) and Wu, Zhu, and Liu (2000) proposed a class of Markov random field models (Besag, 74; Cressie, 1993) for textures, and studied the minimax entropy principle and the equivalence of ensembles for feature statistics based on linear filters. In operation-oriented school, contributions were made by Amit and D. Geman (1997) and Blanchard and D. Geman (2003), who stressed the importance of computing efficiency in visual perception. Tu and Zhu (2002) proposed data-driven Markov chain Monte Carlo for integrating operation-oriented methods into represented-oriented schemes.

As evidenced by the above theories, to understand visual learning and perceptual inference, it is crucial to identify fundamental visual phenomena and understand the underlying statistical principles. The proposed work is to study a ubiquitous visual phenomenon that we call *perceptual scaling.*

Perceptual scaling

The left column of Figure 1 displays three images of an ivy wall taken at three different distances. For the image at near distance, we perceive individual leafs, including their edges and shapes. For the image at far distance, however, we only perceive a collective foliage impression without discerning individual structures. While the near-distance image looks regular and simple, with sparse structures, the far-distance image appears random and complex, with rich details. Why does the same pattern result in different perceptions at different distances? Can we find a mathematical theory to formally explain this perceptual transition over scale?

This transition from sparse structures to collective textures is ubiquitous in outdoor scenes, and we call such transition perceptual scaling. For instance, the images of branches and twigs in the right column of Figure 1 also exhibit such a scaling effect. More important, perceptual scaling typically presents itself in a single image of a static natural scene, because objects and patterns can appear at a wide variety of distances and depths from the viewer. See Figure 2 for two examples, where the leafs and branches give us different impressions at different scales. While the large and near-distance structures are sparse and perceptible, and provide most crucial information of the scene, the small and far-distance structures are abundant and often not individually perceptible, but they collectively provide us a sense of complexity and richness that is a defining characteristic of realistic natural scenes. Thus, a

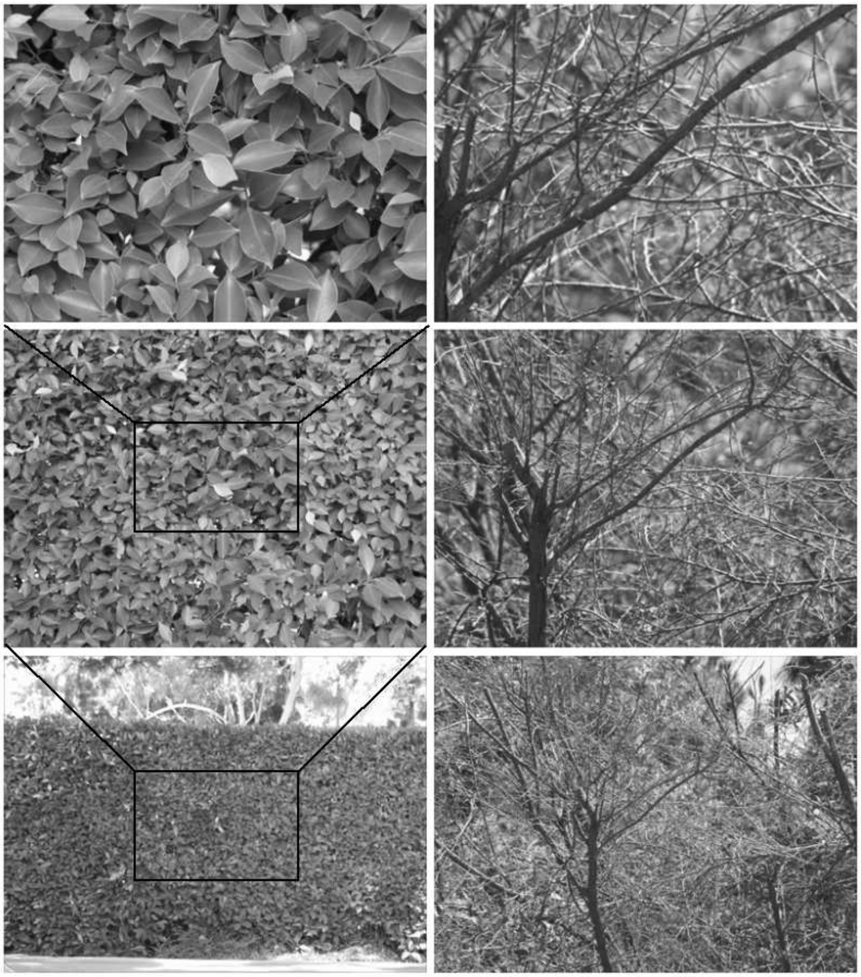


Figure 1: Perceptual scaling: transition from sparse structures to collective textures over distance.

mathematical theory that accounts for this scaling effect is crucial for a visual system to successfully interpret virtually any natural scenes.

As another example of perceptual scaling, in Figure 3, the left image gives us vivid 3D impression of shapes, whereas the right image only gives us an overall impression of roughness.

Perceptual scaling also manifests itself in motion scenes (e.g., Doretto, Chiuso, Wu, and Soatto, 2003). For instance, when we look at sea surface, we perceive the shapes of big waves and we can trace their motions, whereas for

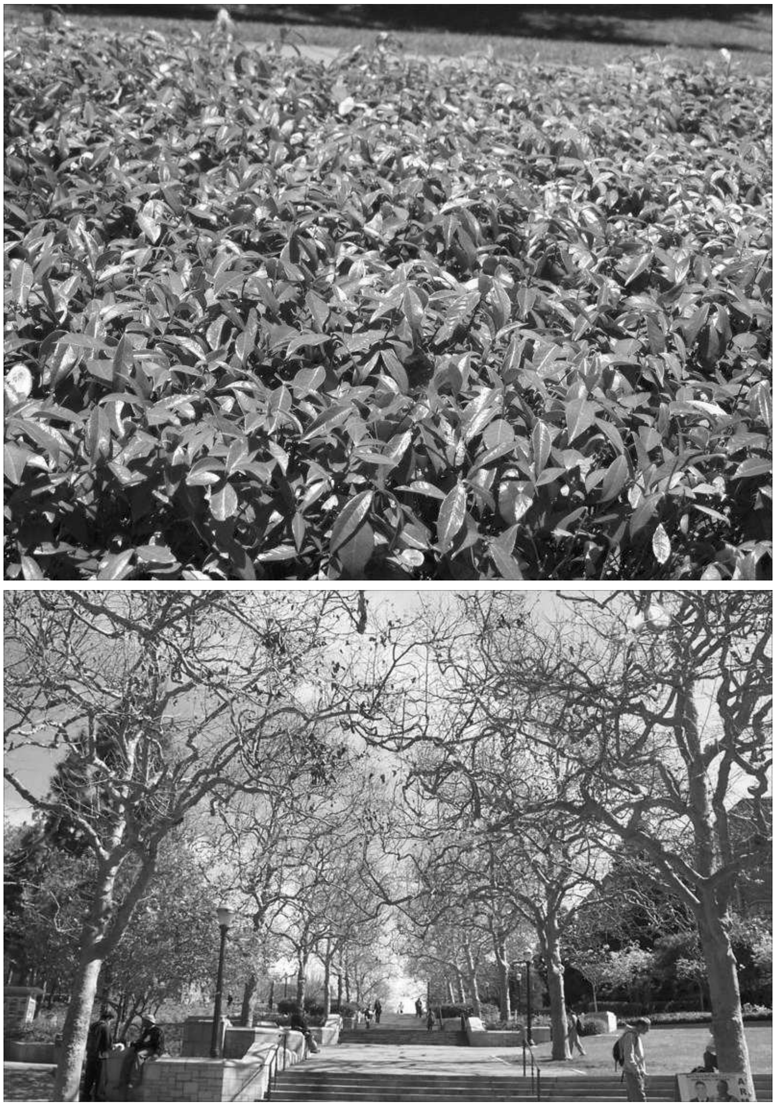


Figure 2: Perceptual scaling: the same patterns can appear at different scales in a single image.

the large number of small ripples, their shapes are not perceptible and their motions are not trackable.

There have been many interesting theories on the issue of scaling in the literature, such as scale space theory (e.g., Lindeberg, 1994), multi-resolution analysis (Mallat, 1989), fractals (Mandelbrot, 1982), spectrum and simple

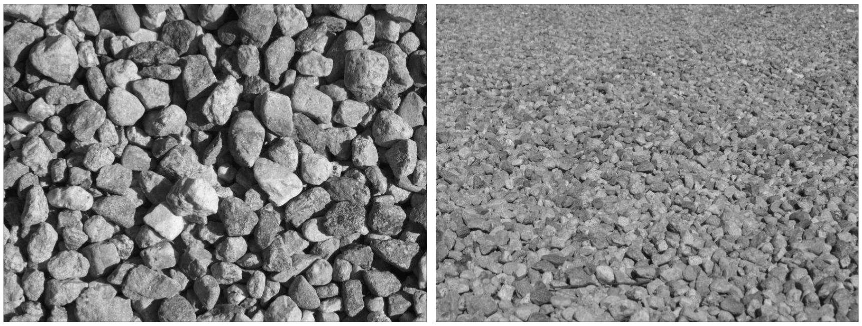


Figure 3 Perceptual scaling: from 3D shapes to texture impression of roughness.

statistics of natural images (Ruderman and Bialek, 1994; Mumford and Gidas, 2001; Chi, 2001; Simoncelli and Olshausen, 2001). However, none of these theories are concerned with the effect of image scaling on our perception of particular patterns such as those in Figure 1.

Given the fact that visual perception is a statistical inference problem, and complexity and randomness must be studied in a statistical framework, we argue that perceptual scaling is a statistical phenomenon.

This paper proves two scaling laws in vision: If we get farther from a visual pattern, then 1) the resulting retina image becomes less sparse, and 2) the underlying pattern becomes less perceptible. The two scaling laws have interesting implications in the possible strategy employed by visual cortex, and reveal the connection between wavelet sparse coding and Markov random fields.

Sparsity and Minimax Entropy

Wavelets and Markov Random Fields

The simple neuron cells in the primitive visual cortex (called V1) are mathematically modeled by a set of localized, oriented, and elongate linear bases/filters, $\{B_{x,y,k}\}$, where (x, y) indexes the location, and k indexes the shape, such as orientation and scale. See Figure 4 for an illustration.

There are two major classes of representations for nature images, both involve the above local bases/filters.

Wavelets and sparse coding: This representation is generative (Lewicki and Olshausen, 1999)

$$c_{x,y,k} \sim p(c), \tag{1}$$

$$I = \sum c_{x,y,k} B_{x,y,k} + \epsilon, \tag{2}$$

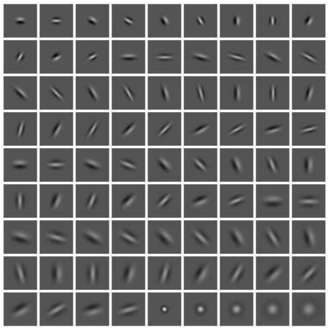


Figure 4 Linear bases/filters as mathematical model of V1 cells.

where $c_{x,y,k}$ are coefficients for representing I in the form of (2), and ϵ is the residual error. The key principle is the sparsity principle (Olshausen and Field, 1996), where $\{B_{x,y,k}\}$ is assumed to be over-complete, i.e., the number of bases exceeds the number of pixels, but for a typical image, only a small number of $c_{x,y,k}$ are significantly different from 0, i.e., the prior distribution $p(c)$ in (1) is a long-tail distribution such as mixture of normals (Olshausen and Millman, 2000; see also George and McCulloch, 1997, for independent but closely related work on Bayesian variable selection in regression). The sparsity assumption can also be expressed in a non-probabilistic form by a regularization or penalty term (Candes and Donoho, 1999). If we treat $\{B_{x,y,k}\}$ as unknown parameters, then we can learn them from natural images (Olshausen and Field, 1996).

Markov random fields (MRFs) and feature statistics: For a homogeneous local image patch, which is still denoted by I for simplicity, we compute filter responses $r_{x,y,k} =< I, B_{x,y,k} >$ for all the filters within this patch (Malik and Perona, 1989), and then for each type of filter k, we compute the histogram $H_k(I)$ by pooling $r_{x,y,k}$ over all (x,y) in this patch. The image patch is then represented by the set of histograms $H_k(I)$ (Heeger and Bergen, 1995; Portilla and Simoncilli, 2000). The basic idea is to consider the ensemble of images (Wu, Zhu, and Liu, 2000):

$$\Omega = \{I : H_k(I) = H_k(I^{\rm obs}), \forall k\}, \quad (3)$$

which collects all the images I that share the same histograms as the observed image $I^{\rm obs}$. This ensemble is called Julesz ensemble by Wu, Zhu, and Liu (2000). One can model I as following the uniform distribution over the Julesz ensemble Ω according to the maximum entropy principle, and this uniform distribution is equivalent to a MRF model or a Gibbs distribution (Wu, Zhu, and Liu, 2000),

$$f(I) = \frac{1}{Z}\exp\{\sum_k < \lambda_k, H_k(I) >\} = \frac{1}{Z}\exp\{\sum_k\sum_{x,y}\lambda_k(< I, B_{x,y,k} >)\}, \quad (4)$$

where λ_k is a vector of the same dimension as $H_k(I)$, so it can also be viewed as a one-dimensional step function over the bins of the histogram $H_k(I)$. Z is the normalizing constant that depends on $\{\lambda_k\}$. This model is called FRAME model (Filter, Random field, And Maximum Entropy) by Zhu, Wu, and Mumford (1997). If $\{H_k(I)\}$ are taken to be other statistics (e.g., moments instead of histograms), then the corresponding $\{\lambda_k()\}$ become other functions (e.g., polynomials instead of step functions). It is just a matter of parametrization.

The set of filters can be learned so that the volume of the Julesz ensemble Ω, i.e., $|\Omega|$, or the entropy of the fitted MRF model $f(I)$ in (4), is minimum. This is the minimum entropy principle. Inferentially, one can estimate $\{B_{x,y,k}\}$ and λ_k in the FRAME model by maximum likelihood. Computationally, this can be accomplished by stochastic gradient algorithm.

Although both the sparsity principle and the minimum entropy principle are about representing the image with minimum complexity, the philosophies and the mathematical structures in wavelet model and the FRAME model are very different. Philosophically, the wavelet model is constructive, where I is deterministically constructed by superposition of local bases. The FRAME model is restrictive, where I is defined stochastically by restricting histograms of filter responses. Mathematically, the $\{B_{x,y,k}\}$ in the wavelet model are bases, and the corresponding $c_{x,y,k}$ compete to explain I, so there is lateral inhibition among them, i.e., if one base is active in explaining I, then it will inhibit other overlapping bases. The $\{B_{x,y,k}\}$ in the FRAME model are filters, and there is no lateral inhibition among the filter responses $r_{x,y,k}$.

It is worth of mentioning that, if $\{B_{x,y,k}\}$ is complete, i.e., the number of bases is the same as the number of pixels, then both models reduce to independent component analysis (Bell and Sejnowski, 1997). One may call the latter the "restructive" scheme, because it involves a one to one transformation between I and the coefficients $\{c_{x,y,k}\}$ or the responses $\{r_{x,y,k}\}$. The principle behind independent component analysis is the factorial coding principle, which is closely related to both sparsity principle and the minimum entropy principle.

Complexity regimes

The complexity behavior of the two models are also different.

Proposition 1: *Let $p(I)$ be the true distribution that generates I, let $f(I)$ be the FRAME model (4) where the $\{\lambda_k\}$ are chosen to minimize the Kullback-Leibler divergence $D(p\,||\,f) = \mathrm{E}_p[\log(p(I)/f(I))]$. Then*

$$D(p\,||\,f) = \mathcal{H}(f) - \mathcal{H}(p) \geq 0,$$

where $\mathcal{H}(q(I)) = -\int q(I)\log q(I)dI = -\mathrm{E}_q[\log q(I)]$ is the entropy of a distribution $q(I)$. So it shows that the entropy of the fitted FRAME model f is always no less than the entropy of the true distribution p.

Proof: Setting $\partial D(p||f)/\partial\lambda_k = 0$, we have $\mathrm{E}_f[H_k(I)] = \mathrm{E}_p[H_k(I)]$. Then $\mathrm{E}_p[\log f(I)] = \mathrm{E}_f[\log f(I)]$, and the result follows.

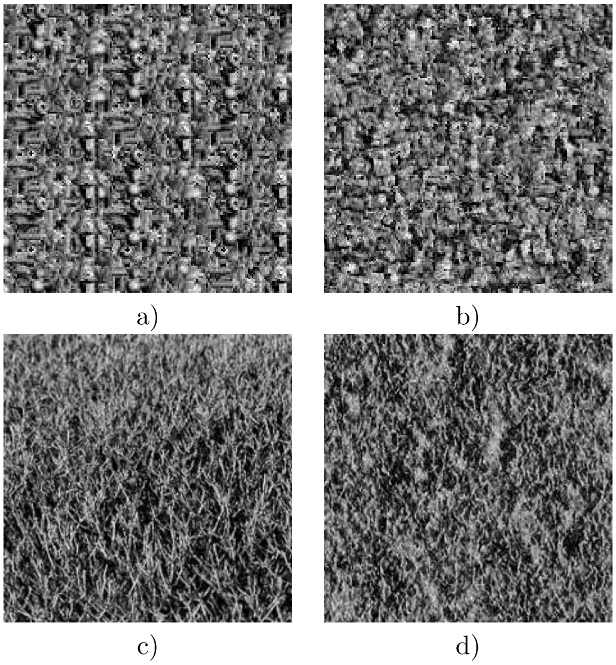


Figure 5: Feature statistics. a) and c) are observed images. b) and d) are "reconstructed" by matching feature statistics.

Figure 5 shows two examples of feature statistics representation. a) and c) are observed images, and b) and d) are respectively the "reconstructed" images. However, the reconstruction is of a statistical nature: b) and d) are sampled from the respective Julesz ensembles Ω (3) by matching feature statistics. We can see that this representation is appropriate for random images such as image a). It captures texture information, but does not do a good job in capturing salient structures.

As to the sparse coding model, we rewrite (2) in a matrix form $J = BW$ and $I = J + \epsilon$. The images I and J are represented by vectors and B is a matrix with each column being a base function $B_{x,y,k}$, and W is the vector that collects the coefficients $\{c_{x,y,k}\}$. Due to sparsity, elements in W are mostly close to zero. Thus $p(W)$ has very low entropy.

Proposition 2: *In sparse coding model with* $J = BW$ *and* $W \sim p(W)$, *then* $\mathcal{H}(p(J)) \leq \mathcal{H}(p(W)) + \frac{1}{2}\log\det(BB')$.

Proof: Let A be a matrix whose rows are orthogonal bases in the null space of the rows of B, and let $K = AW$. Then $\mathcal{H}(p(J,K)) = \mathcal{H}(W) + \frac{1}{2}\log\det(BB')$.

So
$$\mathcal{H}(p(J)) + \mathcal{H}(p(K|J)) = \mathcal{H}(p(W)) + \frac{1}{2}\log\det(BB').$$

Thus the result follows.

That is, the resulting $p(J)$ has low entropy bounded by $\mathcal{H}(p(W))+\log\det(BB')/2$, and it cannot account for the images generated from $p(I)$ whose entropy is larger than this bound. In other words, the sparse coding model puts all the extra complexities into the residue ϵ.

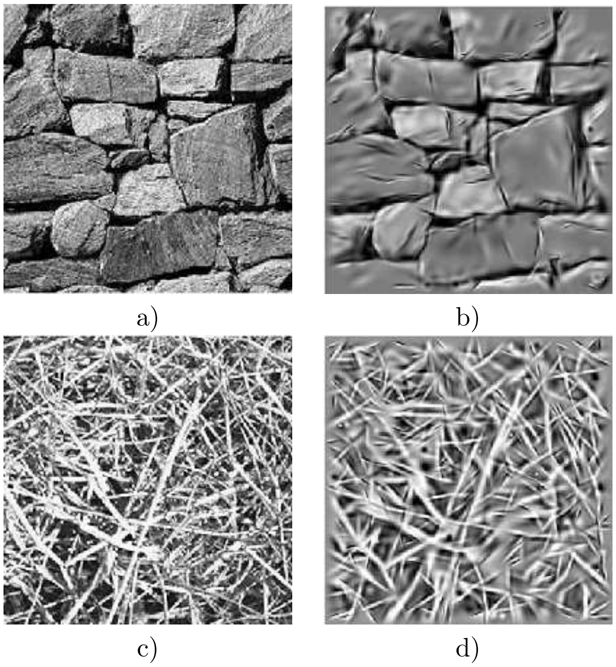


Figure 6: Sparse coding. a) and c) are observed images. b) and d) are respectively the reconstructed images using 300 bases.

Figure 6 shows two examples of sparse coding. a) and c) are observed images, b) and d) are images reconstructed by 300 bases. We used the matching pursuit algorithm of Mallat and Zhang (1993) to select the bases (in a manner very similar to forward stepwise regression). We can see that sparse coding is very effective for images with sparse structures, such as image a). However, the texture information is not well represented.

To summarize, the wavelet sparse coding model is effective in low entropy regime where images have order and structures, such as the shape and geometry. We call this regime as "sketchable." The FRAME model is effective in high entropy regime where images have less structures, such as stochastic texture. We call this regime as "non-sketchable." The competition between these two models in terms of some model selection criterion such as minimum description length (e.g., Hansen and Yu, 2000). This competition may gives us a threshold that tells us when we should stop using sparse coding representation and switch to feature statistics.

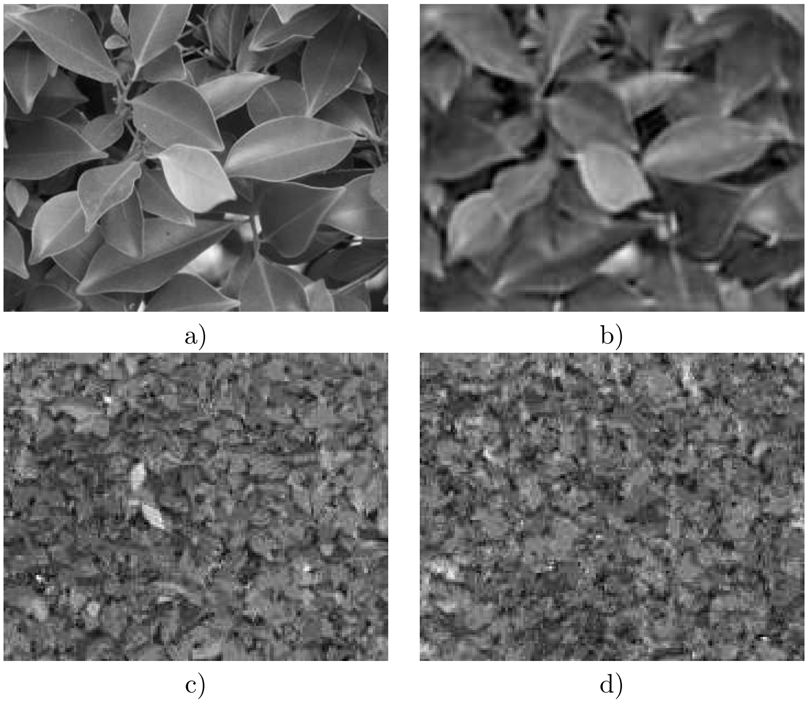


Figure 7: From sparse coding to feature statistics. a) Observed near-distance image. b) Reconstructed by sparse coding with 1,000 bases. c) Observed far-distance image. d) "Reconstructed" by matching feature statistics.

Figure 7 displays results of a pilot study on scale. 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.

The intrinsic connection between the two models are revealed by the following proposition.

Proposition 3: *Consider the FRAME model $f(I)$ in (4) where $\lambda_k()$ are continuous and differentiable, then $f(I)$ is the equilibrium distribution of the following Langevine diffusion*

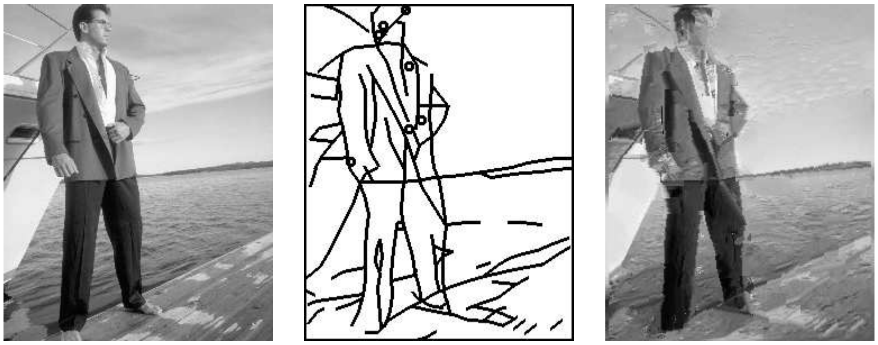
$$dI(t) = \frac{1}{2}\sum_{x,y,k} \lambda_k'(< I(t), B_{x,y,k} >)dt \times B_{x,y,k} + d\epsilon(t),$$

where $\lambda_k'()$ *is the derivative, and* $d\epsilon(t)$ *is Brownian motion.*

In this dynamics, each step is a linear superposition of bases, plus a small Brownian noise $d\epsilon(t)$. This additive form coincides with sparse coding model (2). The difference is that this dynamics is iterative and non-sparse.

In a previous paper (Guo, Zhu, and Wu, 2003), we studied and experimented with a primal sketch model (the name comes from the book by Marr, 1982), where the image I is divided into sketchable part I_{sk} and non-sketchable part I_{nsk}. The model for I is $p(I) = p(I_{\text{sk}})p(I_{\text{nsk}}|I_{\text{sk}})$. I_{sk} is modeled by wavelet sparse coding. $p(I_{\text{nsk}}|I_{\text{sk}})$ is modeled by FRAME model, with I_{sk} being the boundary conditions. Or in other words, I_{nsk} interpolates I_{sk} by matching local feature statistics.

See Figure 8 for an example, where a) is the observed image; b) depicts the sketch version of the image, where each base in representing I_{sk} is replaced by a small line segment (or a circle for center-surround base); c) is the synthesized images, where the structures are reconstructed by sparse coding, and the textures are generated by matching feature statistics. See Figure 9 for two more examples.

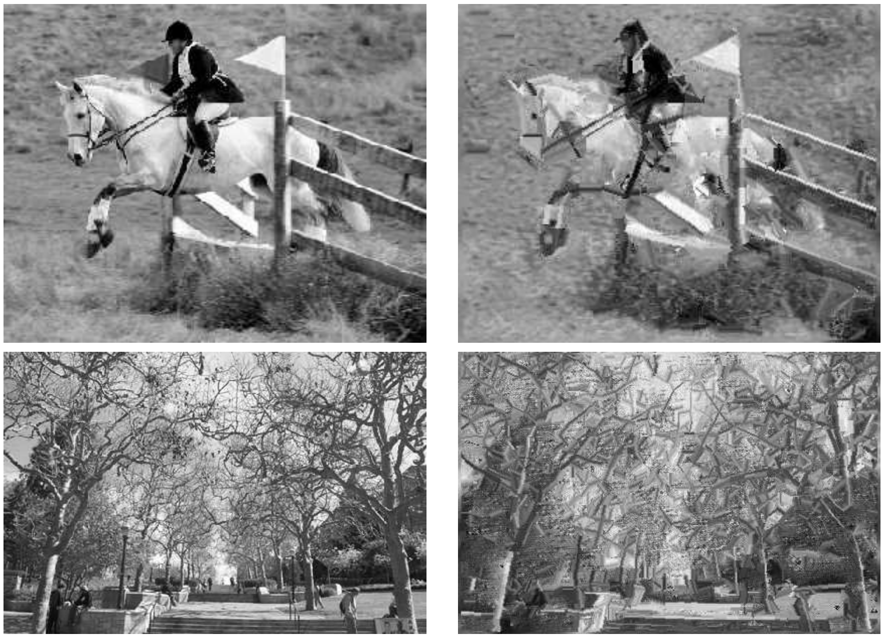


a) Observed image b) Image sketch c) synthesized image

Figure 8: Primal sketch: a) Observed image. b) Image sketch with each base replaced by a line segment (or a circle). c) Synthesized image.

The prior models for the spatial arrangements of local bases is a pair-wise Gibbs point process model (see also Stoyan, Kendall, and Mecke, 1987, Wu, Guo, and Zhu, 2002) that takes care of continuity, joints, and closures of the local bases. We call such model the Gestalt field.

In the next two sections, we will prove two scaling laws that explain the transition from sparse structures to stochastic textures.



a) Observed image

b) synthesized image

Figure 9 Primal sketch: a) Observed image. b) Synthesized image.

Complexity Scaling Law

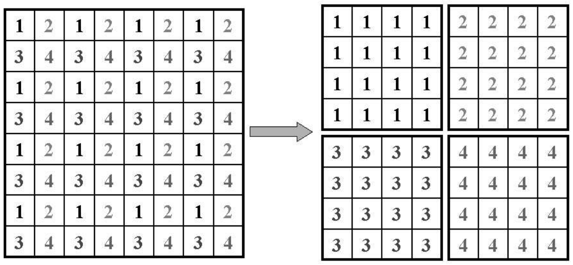
Let I be the image of a pattern observed at a certain distance, and let's assume that I is generated by a physical process that can be summarized by a probability distribution $p(I)$. Let Λ be the lattice on which I is defined.

Definition 1: ***Image complexity,*** *denoted by* $\mathcal{H}(I)$*, is defined as the entropy of* $p(I)$*. The* ***complexity rate*** *is defined as* $\mathcal{H}(I)/|\Lambda|$*.*

When we move away from a scene, the change of image involves both local smoothing and down-sampling. As a first step, we shall only study the effect of down-sampling, while ignoring the effect of local averaging. To simplify the situation even further, let's assume that we down-sample I by a factor of 2 alone 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 Λ_{-}, so that $|\Lambda_{-}| = |\Lambda|/4$. See Figure 10 for an illustration.

Theorem 1: ***Complexity Scaling Law.***

$$1) \;\; \mathcal{H}(I_{-}^{(k)}) \leq \mathcal{H}(I), \;\; k = 1, ..., 4.$$



1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4
1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4
1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4
1	2	1	2	1	2	1	2
3	4	3	4	3	4	3	4

1	1	1	1	2	2	2	2
1	1	1	1	2	2	2	2
1	1	1	1	2	2	2	2
1	1	1	1	2	2	2	2
3	3	3	3	4	4	4	4
3	3	3	3	4	4	4	4
3	3	3	3	4	4	4	4
3	3	3	3	4	4	4	4

Figure 10 The four down-sampled versions of the original image.

$$2)\ \frac{1}{|\Lambda_-|}\sum_{k=1}^{4}\mathcal{H}(I_-^{(k)})/4 \geq \frac{1}{|\Lambda|}\mathcal{H}(I).$$

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

$$\mathcal{H}(I) - \mathcal{H}(I_-^{(k)}) = \mathrm{E}_I\left[-\log\frac{p(I)}{p(I_-^{(k)})}\right] = \mathcal{H}(I|I_-^{(k)}) \geq 0.$$

2) Let $\mathcal{M}()$ denote mutual information,

$$\begin{aligned}\sum_{k=1}^{4}\mathcal{H}(I_-^{(k)}) - \mathcal{H}(I) &= \mathrm{E}\left[\log\frac{p(I)}{\prod_k p(I_-^{(k)})}\right]\\ &= \mathcal{M}(I_-^{(k)}, k=1,2,3,4) \geq 0.\end{aligned}$$

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 $I_-^{(k)}$.

In Theorem 1, we only consider the effect of down-sampling, without considering the effect of local averaging. But from information theoretical perspective, the purpose of local averaging is to make the entropy of down-scaled I_- as close to the entropy of I as possible in order to maintain as much information as possible. As a result, the complexity rate of I_- will be even larger if we take into account the local smoothing effect.

This theorem tells us that if we down-sample an image, the image looks more random. This can be easily understood from real life experience. For instance, for the ivy wall pattern in Figure 1, when we move farther away from it, we lose information, so the complexity is decreasing. But we see more leafs within unit area of visual field, so the complexity rate is increasing.

The complexity scaling law we have proved has far reaching implications on sparsity principle (Olahsusen and Field, 1996). At near distance, the complexity rate is very low, so sparsity principle applies. But as the viewer moves farther from the underlying pattern, 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 can only interpret the image by some summaries that cannot determine the image deterministically, and these summaries are feature statistics. This may explain the perceptual transition from sparse coding to feature statistics.

Perceptibility Scaling Law

The purpose of vision is to make inference about the outside world. Now, let's study the issue of perceptual transition in an inferential framework, under the slogan that "vision = inverse graphics."

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 determine I uniquely, 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 leafs 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 W is generated by a physical process that can be summarized by a distribution $p(W)$ (we shall not engage in a philosophical discussion on whether there exists a true $p(W)$). Given $W \sim p(W)$, and $I = g(W)$, we have $p(W|I) = p(W, I)/p(I) = p(W)/p(I)$. $p(W, I) = p(W)$ because I is fully determined by W. This distribution defines an inversion of the graphics equation $I = g(W)$.

Definition 3: ***Scene complexity****, denoted by $\mathcal{H}(W)$, is defined as the entropy of $p(W)$.*

Definition 4: ***Imperceptibility****, denoted by $\mathcal{H}(W|I)$, is defined as the conditional entropy of $p(W|I)$.*

Theorem 2: *Let $W \sim p(W)$, and $I = g(W)$, then $\mathcal{H}(W|I) = \mathcal{H}(W) - \mathcal{H}(I)$. 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.

The imperceptibility $\mathcal{H}(W|I)$ gives a general definition of "ill-posedness" of the inversion problem. Here the concept of imperceptibility only means the possibility of estimating W under a particular physics representation of W.

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 $I_- = R(I)$.

Theorem 3: ***Perceptibility Scaling Law.*** *For $W \sim p(W)$, $I = g(W)$, if $I_- = R(I)$ with $R()$ being any many to one reduction function, then*

$\mathcal{H}(W|I_-) \geq \mathcal{H}(W|I)$. *That is, imperceptibility becomes larger with downscaling.*

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 $\mathcal{H}(W|I_-) - \mathcal{H}(W|I) = \mathcal{H}(I|I_-) \geq 0$.

If $\mathcal{H}(W|I_-)$ is too large, we can only perceive some aspect of W, i.e., $W_- = \rho(W)$, for some many to one reduction $\rho()$, such that $\mathcal{H}(W_-|I_-)$ is small. It is possible to find such a W_-, because of the following theorem.

Theorem 4: *For $W \sim p(W)$, $I = g(W)$, and $I_- = R(I)$, $W_- = \rho(W)$, we have $\mathcal{H}(W_-|I_-) \leq \mathcal{H}(W|I_-)$.*

Here W_- can be a coarser representation of W, where the scale of the elements in W_- may be larger than that of W. It is possible that there still exists a g_-, such that $I_- = g_-(W_-)$, but it is most likely that this is only approximately true. It is also likely that W_- may only correspond to some statistical property of I_-, or in other words, $[I_-|W_-] \sim p(I_-|W_-)$ with a high entropy rate. That is, although W defines I deterministically via $I = g(W)$, W_- may only defines I_- statistically via a probability distribution $p(I_-|W_-)$. While W represents sparse structures, W_- may only represent collective textures.

This perceptibility scaling law provides a possible explanation to the perceptual transition from sparse structures to stochastic textures.

Texture = Imperceptible Structures

The visual cells in the primitive visual cortex V1 may correspond to various types of local descriptors for local structures appearing at different scales, locations, and orientations. Olshausen and Field (1996) proposed a sparsity principle as a V1 strategy. This principle holds that for a typical image, only a small number of local descriptors need to be selected to interpret the image. We argue that the sparsity principle only accounts for part of V1 representations and activities. This is because the number of local descriptors is much less than the number of all possible image patches. As a result, there are a lot of image patches that cannot be well represented by local descriptors, or there are no sparse representations for such image patches. Such image patches often correspond to patterns viewed at a far distance, so that both the complexity rate and the imperceptibility are high. These image patches cannot be accounted for by the sparsity principle. Then what are the possible representations for them?

One possible choice is to summarize them into feature statistics, i.e., they are interpreted statistically as textures (or more precisely stochastic textures), instead of structures. Then what feature statistics should we use? The next theorem sheds light on this question.

Theorem 6: For $F = F(I)$ be a set of feature statistics, *1) If* $W \sim p(W)$, $I = g(W)$*, then*

$$\begin{aligned} \mathrm{D}(p(W|I)||p(W|F)) &= E_W \left[\log \frac{p(W|I)}{p(W|F)} \right] \\ &= \mathcal{H}(W|F) - \mathcal{H}(W|I) = \mathcal{H}(I|F). \end{aligned}$$

2) If $W \sim p(W)$ *and* $[I|W] \sim p(I|W)$*, then*

$$\begin{aligned} \mathrm{D}(p(W|I)||p(W|F)) &= E_{W,I} \left[\log \frac{p(W|I)}{p(W|F)} \right] \\ &= \mathcal{H}(W|F) - \mathcal{H}(W|I) = \mathcal{M}(W, I|F). \end{aligned}$$

Here $D()$ *denotes Kullback-Leibler divergence, and* $\mathcal{M}()$ *denotes mutual information.*

Result 1) justifies the minimum entropy principle we discussed before. That is, to minimize $\mathcal{H}(W|F)$ over a set of possible $\{F()\}$, we need to minimize $\mathcal{H}(I|F)$. In result 2), $\mathcal{M}(W, I|F)$ measures the sufficiency of F.

This theorem shows that in order to choose good feature statistics, we must have $p(W|F)$ to be close to $p(W|I)$. This makes us believe that F must be derived from some intermediate results in the computation of $p(W|I)$.

We propose the following strategy for primitive visual cortex. For each local patch around pixel (x, y), i.e., $I_{x,y}$, there can be a number of local descriptors to describe it. Let $w_{x,y}$ index the possible local descriptor as well as its parameters. Then by fitting a local model, we compute $p(w_{x,y}|I_{x,y})$. This can be done efficiently in a parallel manner.

For those pixels (x, y) with very low $\mathcal{H}(p(w_{x,y}|I_{x,y}))$, we use sparse coding representation, that is, we select use a small number of local descriptors to represent those pixels, while respecting our prior knowledge for the spatial arrangements of these local descriptors.

For those pixels (x, y) with very high imperceptibility $\mathcal{H}(p(w_{x,y}|I_{x,y}))$, the underlying structures cannot be unambiguously determined. As such, we abort the effort of committing a particular $w_{x,y}$. Instead, we pool the local posterior $p(w_{x,y}|I_{x,y})$ over (x, y) into texture statistics. That is, texture = pooling of imperceptible structures. This should be complimentary to the sparsity principle.

This complementary principle bridges deterministic structures and stochastic textures in a very elegant manner. It also has interesting implications on the two conjectures of Julesz on textures (Julesz, 81), as well as the phenomenon of lateral inhibition in neuroscience.

For the wavelet sparse coding model $I = \sum c_{x,y,k} B_{x,y,k} + \epsilon$, the local model is $I_{x,y} = c_{x,y,k} B_{x,y,k} + \epsilon$. If the bases are not perceptible, we can pool local posterior over (x, y). One can show that the pooled statistics is very close to the histograms of filter responses. If we assume such feature statistics, then we are led to the Markov random field model (4). Thus, we establish an interesting link between wavelet sparse coding theory and Markov random field theory. We shall further investigate this connection, which should be interesting to both wavelet community and spatial statistics community.

Perceptibility and Sparsity

The inferential concept of perceptibility also arises from the coding perspective. 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 $[I|W] \sim f(I|W)$. 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 $-\mathrm{E}_p\left[\mathrm{E}_{f(W|I)}(\log f(W) + \log f(I|W))\right]$.

Theorem 7: *The average coding length is* $\mathrm{E}_p[\mathcal{H}(f(W|I))] + \mathrm{D}(p||f) + \mathcal{H}(p)$. *That is,* ***coding redundancy = imperceptibility + error****. Here* $\mathcal{H}(f(W|I))$ *is the entropy of* $f(W|I)$ *conditional on* I*, and* $D(p||f)$ *is the Kullback-Leibler distance.*

The relationship between perceptibility and sparsity deserves more investigation. To make the idea more concrete, let's consider the sparse coding model $I = \sum c_{x,y,k} B_{x,y,k} + \epsilon$. If the image is very complex, then even the sparsest representation still has a large number of bases, so that sparsity principle is violated. One may ask, what is wrong with a non-sparse representation? This can be answered by perceptibility. That is, if the sparsest representation still has a large number of bases, then there can be a lot of representations that are only slightly less sparse, but can approximate I with equally small error ϵ. Or in other words, there can be a lot of "equivalent" representations, so that there is ambiguity as to which one to use. This ambiguity may be mathematically defined, and clearly it is closely related to imperceptibility. In wavelet sparse coding theory, this issue of ambiguity has not been studied. But it is clearly of fundamental importance to vision applications, because the representation is to be used in later stages of visual processing.