QUARTERLY OF APPLIED MATHEMATICS
VOLUME LIX, NUMBER 1
MARCH 2001, PAGES 85–111

STOCHASTIC MODELS FOR GENERIC IMAGES

BY

DAVID MUMFORD AND BASILIS GIDAS

Division of Applied Mathematics, Brown University, Providence, RI

1. Introduction. The idea of using statistical inference for analyzing and understanding images has been used for at least 20 years, going back, for instance, to the work of Grenander [Gr] and Cooper [Co]. To apply these techniques, one needs, of course, a probabilistic model for some class of images or some class of structures present in images. Many models of this type have been introduced. There are stochastic models for image textures [GGGD], [ZMW], for contours in images [Mu], [GCK], for the decomposition of an image into regions [G-G], [M-S], for disparity maps, for grammatical parsing of shapes [Fu], for template matching, and for specific tasks such as face recognition [HGYGM]. The common framework for all these studies is to describe some class of images $I(x, y)$ by means of a set of auxiliary variables $\{x_\alpha\}$ representing the salient structures in the images, e.g., edges, texture statistics, inferred depth values or relations, illumination features, medial axes or shape features, locations of key points such as eyes in a face, labels (as in character recognition), etc. Then i) a *prior* probability model for the "hidden" variables $p(\{x_\alpha\})$ and ii) an imaging model $p(I|\{x_\alpha\})$ for I, given the hidden variables, are defined. Finally, an image is analyzed using Bayes's rule

$$p(\{x_\alpha\}|I) \propto p(I|\{x_\alpha\})p(\{x_\alpha\})$$

which is applied to infer, e.g., the MAP estimate for the hidden variables, given the image. Implicit in this approach is the deduction that there is a well-defined marginal distribution

$$p(I) = \int_{\{x_\alpha\}} p(I, \{x_\alpha\}) \prod dx_\alpha$$

on *all* images that are likely to be seen.

But is there such a thing as a universal stochastic model $p(I)$ for images? Is this a reasonable thing to ask for? What sense would it make—would the model apply equally if we were born in another historical time, if our eyes and bodies were hundreds of times bigger or smaller, if we lived in outer space? Images are so diverse and contain so many distinct types of structure that research has focussed on modeling specific well-defined

Received January 13, 1999.
2000 *Mathematics Subject Classification.* Primary 68T45, 60G60, 60G51, 60G55.
We would like to acknowledge the support of the Army Research Office MURI DAAHO4-96-1-0445; and the National Science Foundation DMS-9615444.

aspects of images rather than looking at the bottom line, $p(I)$, itself. Several discussions and papers have influenced the first author to take seriously the possibility of such a model. Rosenfeld made the remark, about ten years ago, that one seldom encountered white noise, or noise of any standard kind in images: more typically, one encountered what he called "clutter". At that time, the first author was working with a class of models in which the image was assumed to be the sum of Gaussian white noise n and of a cleaned-up piecewise smooth image J (called a "cartoon"): $I = n + J$. But when looking at actual pixel values, one saw instead a random fluctuation caused by small details which one could not resolve. It was the presence of all these small details and small or distant objects rather than the presence of transmission noise or static that made the image pixel values so erratic. More recently, clutter has become an important issue in the design of vision algorithms for object recognition. Here clutter is the mass of irrelevant details in the scene—foliage, houses, roads—in the midst of which the one relevent object, such as a car or a tank, is located. The issue is whether you have to identify and model every one of the mass of objects in the image before finding the car or the tank, or whether there is some statistic that enables you to separate the target from the clutter without explicitly describing the clutter in detail.

A third motivation arose from a joint seminar with S. Shieber where we were comparing stochastic models in vision and language. We studied the beautiful experiments done by Shannon [Sh] using the most naive raw statistical procedures for modeling English language character strings. He counted not merely letter frequencies, but frequencies of letter pairs, letter triples, and letter quadruples; not merely frequencies of words but of word pairs and triples ("bigrams" and "trigrams"). Taking samples from these models, one has the uncanny sense of an almost continuous convergence from models whose samples were random character strings to models whose samples come close to being true English. Can this be done with images? The obvious problem is that to repeat Shannon's experiment, one needs more memory than is even potentially available. For example, if image pixel values are in the range [0,255] and one were to try to compile exhaustively the statistics on image values in 3×3 blocks, one would create a probability table with $256^9 = 2^{72} \approx 5 * 10^{21}$ entries. So some more analysis may be better first!

Shannon's models certainly could not produce fully meaningful English sentences: at best they capture some rudimentary aspects of grammar and reasonable juxtapositions of words with related meanings. What can we expect generic image statistics to capture? We do not want to model any specific class of objects, such as faces, nor specific textures, such as tree bark, nor the physics of the world we live in, such as the effects of specific reflectance functions. The idea behind this paper is that, even when you throw out such specifics, there are commonalities in the statistics of images, striking regularities that can be captured. Our hope is that the models described here are only a start, that much more about the nature of the images that we are used to seeing is contained in very simple low-level statistics. We can formulate this in a conjecture: *there exist simply described stochastic models for images that a) assign high likelihood to any "natural" image of the world we live in and b) whose random samples have the "look and feel" of natural images, i.e., make you look twice to see if you recognize something in them.* For

instance, no Gaussian probability measures on images have anything like the look and feel of the real world—the best one can do is make them look like clouds (see Fig. 1).

The outline of this paper is as follows. In §2, we will introduce the precise mathematical formulation of the problem. In §3, we describe the most striking empirical phenomenon exhibited by the statistics of natural images: their apparent scale-invariance. In §4 we digress to show the problems that scale-invariance creates: there are no scale-invariant probability measures supported on image *functions*. To construct such probability measures, we need their samples to be generalized functions ("distributions" in the sense of Schwartz). In §5, we introduce the basic idea of this paper which is to assume that images can be described by a numerical quantity called clutter and that an image with clutter $c_1 + c_2$ can be constructed by adding independent images of clutter c_1 and c_2. Such a situation is called an infinitely divisible family and we propose that this defines a natural class of image models. Although not exactly satisfied by the "true" probability measure on natural images, the infinite divisibility assumption captures in simple mathematical terms certain essential aspects of this measure. In §6 and §7 we analyze infinitely divisible image models, introducing two further axioms that express a) the idea that objects are local while the image itself is an ergodic field and b) that some parts of scale-space are empty of objects, an assumption we refer to as the "blue-sky" hypothesis. After that, we need to convince ourselves that these axioms can be satisified. It is not at all obvious that there is any probability model satisfying these axioms (which are closely related to what physicists would call a 2D non-Gaussian conformal field theory). We do this by establishing in §8 the convergence of what we call random wavelet expansions. In §9 we review recent experiments with images which support the theory we have described. In §10, however, we describe a basic failure of this class of models: the presence of clouds of tiny objects gives the marginal distribution on filter statistics a smooth density. All experiments, however, have resulted in empirical histograms for such statistics which appear singular at 0.

We would like to thank many people for help with this paper, particularly: Persi Diaconis for introducing the first author to the idea of infinitely divisible distributions; Stuart Geman and Zhiyi Chi for very stimulating conversations on scale-invariance and further ideas on the use of infinitely divisible models; Yves Meyer for help on the convergence of random wavelet expansions; Song-Chun Zhu for many provocative ideas on the modeling of images; and Jinggang Huang for his skill and insight in analyzing the statistics of natural databases.

2. The basic setup. We begin by making precise what we mean by an image. Physically, images arise in a camera or in your eyes. Let (x, y, z) be coordinates in 3 space. Assume the world is viewed from the origin (through a "pin-hole" or lens centered at the origin). Then the 2D manifold of viewed directions is the sphere of rays through the origin, or an open set in this sphere (such as the retina). One can put coordinates (u, v) in this manifold, locally near the ray $x = y = 0, z > 0$ either by spherical coordinates $(x, y, z) = (r\sin(u)\cos(v), r\sin(v), r\cos(u)\cos(v))$ for instance, or projective coordinates $(x, y, z) = (ru, rv, r)$. In either case, a finite set of N sensors is positioned suitably to sample the light energy present around particular rays $(u_\alpha, v_\alpha)_{1 \le \alpha \le N}$. The signal received by

sensor α may be modeled as the convolution $I(\alpha) = \iint_{u,v} K(u-u_\alpha, v-v_\alpha)I(u,v)du\,dv$, where K is the impulse response of the sensor to different directions and I is the energy of the incident light. In this very concrete physical situation, the question addressed in this paper is to construct suitable probability measures on the finite-dimensional vector space $\mathbb{R}^N$ containing the measurements $\{I(\alpha)\}$.

In order to come up with a more mathematically tractable setup, however, we want to simplify the geometry in several ways. First of all, we want to avoid modeling the details of the sensor positioning, modeling the energy I directly. In this case, the simplest mathematical scheme is to consider I as a random *distribution* and specific sensors α as defining *test functions* $K(u-u_\alpha, v-v_\alpha)$, so that $I(\alpha)$ is the inner product of the distribution I with the test sensor α. We therefore seek probability measures $\mu(I)$ on the space $\mathcal{D}'$ of distributions.

Moreover, we want to avoid modeling the details of peripheral vision and the borders of images. The simplest way to do this is to construct probability measures on the space of distributions $I(u,v)$ *defined for all* $(u,v) \in \mathbb{R}^2$. We shall assume the measures we construct are *stationary*, so that their marginals on the distributions in a specific window, i.e., in an open subset $U \subset \mathbb{R}^2$, are independent of translation. The assumption is that physical images $I(u,v)$ for u,v sufficiently small are modeled by the marginals of this probability measure. In this case, it does not matter whether we use spherical or projective coordinates in the manifold of rays, because to first order, in a Taylor expansion:

$$(r\sin(u)\cos(v), r\sin(v), r\cos(u)\cos(v)) \approx (ru, rv, r).$$

To avoid confusion, note that the measures we seek on $I(u,v)$ do *not* model random projective views $I(u,v)$ of the world. This is because projective views distort spheres in (x,y,z)-space near the periphery of sight into elongated ellipses in the (u,v) plane. Nor can they possibly model spherical images $I(u,v)$ of the world because spherical images are only defined for a compact set of values of (u,v). Instead, we are asking for a stationary measure whose windows model actual images of the world locally. The samples from such a stationary measure are more like Chinese landscape scrolls, in which more and more of the world comes into view as the scroll is further unrolled.

In passing from a model of a bounded part of the world to the idea of images as infinite scrolls, it is natural to assume that distant parts of the image I are more and more independent. In other words, we make it part of our basic assumption that the measure we construct will be *ergodic* in a suitable sense. For some theorems it will be important to formulate this requirement quantitatively, for instance by asking that some covariances decay fast enough. But the independence of 2 random variables is much stronger than having zero covariance and one may also want to assume the decay of various higher-order measures of dependence such as mutual information.

3. Axiom I: Scale-invariance. Vision is quite distinct from hearing and touch in its lack of characteristic scale. In hearing, there are many natural units for measuring time: your heart beat, the frequency of your vocal cords, the length of a day, etc. These units give universal scales in which to measure any time interval; hence units in which

to record the signal received from your ears. Similarly, when you touch an object, its true physical size determines how many tactile sensors in your skin are excited; hence it always evokes a signal in the 2D array of tactile sensors of the same "size". But this is not true of vision: you can see your spouse's face from a 100 foot distance or a 1 inch distance and the resulting signals transmitted by your retina are (approximately) scaled versions of each other. What this means is that any scene that produces an image $I(u, v)$ on your retina or camera focal plane may also be viewed from nearer or farther away, producing (approximately) an image $I(\lambda u, \lambda v)$ *which is a scaled version* of I.

Why have we written "approximately"? The reason is that this ignores perspective effects. In fact, when you get closer to a scene, the nearer objects get larger faster than the farther objects. These effects are not usually very noticeable. Except for unusual views, such as telephoto images down twenty blocks of a straight street or closeup views of a face 1 inch from the nose, the effect is not obvious. The simplification we are proposing to use goes under the name of "weak perspective" in the computer vision literature. The characteristic distance to each part of the viewed scene is fixed at some typical value z_0 and surface points with coordinates (x, y, z) are projected to the image plane via $(u, v) = (x/z_0, y/z_0)$. Then getting closer or moving farther away from the scene simply changes z_0 and precisely rescales the image. Whether this approximation is reasonable depends on the stochastic nature of the world geometry, i.e., what is the natural distribution of objects and their sizes in the world in which we live. We will present below some reasons for believing this weak perspective model is reasonable.

Mathematically, we may express scale-invariance in terms of the basic probability measure μ on $\mathcal{D}'$ as follows. Firstly, the group of diffeomorphisms of $\mathbb{R}^2$ acts on $\mathcal{D}$ and on $\mathcal{D}'$, in two natural ways. Let $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ be a diffeomorphism of $\mathbb{R}^2$. We may define $T_\phi(f)(x) = f(\phi^{-1}(x))$ for $f \in \mathcal{D}$ and define it on $\mathcal{D}'$ by transpose-inverse: $< T_\phi(I), f > = < I, T_{\phi^{-1}}(f) >$, for $I \in \mathcal{D}', f \in \mathcal{D}$. Explicitly, this makes ϕ act on distributions I that are functions by the rule $T_\phi(I)(x) = |D\phi|^{-1} I(\phi^{-1}(x))$. Alternately, we can make it operate this last way on $\mathcal{D}$ and by the first formula on $\mathcal{D}'$. We want to express the invariance of the probability measure μ on $\mathcal{D}'$ by the formula

Axiom I *(scale-invariance):* $\mu(T_\phi(S)) = \mu(S)$

for all measurable subsets $S \subset \mathcal{D}'$ and all ϕ of the form $\phi(\vec{x}) = \lambda\vec{x} + \vec{a}$. But which is right definition of the action?

In terms of measures on $\mathcal{D}'$, the probability of seeing a specific pattern may be described as $\mu(\{I \mid |< I, f_k > - a_k| < \epsilon\})$, where f_k are a set of test functions, e.g., the sensors of a camera, a_k are the expected values for these responses and ϵ allows for noise. The probability of seeing the same pattern at a smaller scale is $\mu(\{I \mid |< I, g_k > - a_k| < \epsilon\})$ where nothing changes but the sensors. We need $g_k(\vec{x}) = \lambda^2 f_k(\lambda\vec{x} + \vec{a})$ where λ is the factor by which the pattern shrinks and $\vec{a}$ is a translation. Note that the factor λ^2 is used so that the sensor has the same sensitivity, i.e., $\iint f_k \, dx \, dy = \iint g_k \, dx \, dy$. These two measurable subsets of $\mathcal{D}'$ differ by the action of the diffeomorphism $\phi(\vec{x}) = \lambda^{-1}(\vec{x} - \vec{a})$, but note that the action is defined *in the second way*, i.e., it acts on test functions with the Jacobian factor and on images by simple substitution.

Unfortunately, as is well known, there are no nontrivial measures on $\mathcal{D}'$ that are invariant under translations and scale changes and that have finite mean and variance. For any such measure μ, the mean $I_0(\vec{x})$ and the covariance $C(\vec{x},\vec{y})$ are distributions defined by

$$\begin{aligned} < I_0, f > &= \text{Exp}(< I, f >) \\ < C, f \otimes g > &= \text{Exp}(< (I - I_0) \otimes (I - I_0), f \otimes g >) \end{aligned}$$

(where "Exp" means expectation). Because of translation invariance, I_0 is a constant and C is a distribution in $\vec{x} - \vec{y}$ only. Because of the scale invariance of μ, C is also invariant under scale changes and hence must be a constant too. Hence the measure μ is supported on the one-dimensional subspace of constant images $\mathbb{R} \cdot 1 \subset \mathcal{D}'$. If these moments do not exist, there are translation and scale-invariant measures. The simplest of these is "Cauchy noise". On a finite grid, this is defined by independent pixels, identically distributed with a Cauchy distribution. The measure $\mu_{\text{cauchynoise}}$ is defined simply by its Fourier transform:

$$\text{Exp}(e^{i<I,f>}) = e^{-\int |f(x)| dx}.$$

This problem stems from "infra-red" blow-up, i.e., scale invariance of the kind we are assuming implies too many extremely large-scale oscillations and these give rise to infinite energy around zero frequency. The solution is to consider images as distributions *modulo constants*. Since the large scale, low frequency contributions to the image are locally nearly constants, they have less and less impact on the image modulo constants. This leads us to look instead for measures on the quotient space

$$\mathcal{D}'_d =_{\text{def}} \mathcal{D}'/\mathbb{R} \cdot 1.$$

For such a measure, the covariance is a distribution $C(\vec{x} - \vec{y})$ that is scale invariant modulo constants. If it is rotationally symmetric too, the only such distributions are $C(\vec{x} - \vec{y}) = c_1 \log(\| \vec{x} - \vec{y} \|) + c_2$. Since the covariance is now only defined on test functions with mean 0, we can ignore c_2.

Measures on $\mathcal{D}'_d$ that are stationary, rotationally symmetric and scale-invariant do exist. The classical one is the Gaussian measure, known as the "two-dimensional free quantum field": the unique Gaussian measure with mean 0 and covariance $-\log(\|\vec{x}-\vec{y}\|)$. It is constructed by approximating $\mathcal{D}'_d$ by finite-dimensional quotients given by finite sets $\{\phi_i\}, 1 \leq i \leq n$ of test functions with $\int \phi_i = 0$. For each such set, we use the Gaussian on $\mathbb{R}^n$ with mean 0 and covariance

$$C_{i,j} = \iint \phi_i(\vec{x})\phi_j(\vec{y}) \log(\| \vec{x} - \vec{y} \|) \, d\vec{x} \, d\vec{y}.$$

By standard arguments (see [G-V], Ch. 4), these define a Gaussian measure on the nuclear space $\mathcal{D}'_d$. In fact, this measure is supported on a suitable Hilbert-Sobolev space defined by a negative degree of differentiability. We may define these spaces by

$$H^{-s}_{\text{loc}} = (I - \triangle)^{s/2} L^2_{\text{loc}}.$$

Then a careful study using the Minlos-Bochner theorem (see [Hi], [Re] as well as [G-V]) shows that it is supported *just outside* the space of locally L^2 functions[1]—in fact in $\bigcap_{\epsilon>0} H^{-\epsilon}_{\text{loc}}$.

Loosely speaking, this measure is given by the probability density function

$$d\mu(I) = e^{-\iint \|\nabla I\|^2 dx\, dy} \prod_{x,y} dI(x,y).$$

This formula is most easily interpreted by rewriting it in terms of the Fourier transform $\hat{I}$ of I. Since

$$\iint \| \nabla I(x,y) \|^2 \, dx\, dy = \iint (\xi^2 + \eta^2) |\hat{I}(\xi,\eta)|^2 d\xi\, d\eta,$$

$d\mu$ can be rewritten, still loosely, in diagonalized form:

$$d\mu(I) = \prod_{\xi,\eta} e^{-(\xi^2+\eta^2)|\hat{I}(\xi,\eta)|^2} d\xi\, d\eta.$$

This shows that samples from this model are simply "colored" noise, white noise with higher frequencies decreased by the factor $\| (\xi,\eta) \|$ and low frequencies amplified by the inverse of this factor. The effect is that, unlike white noise, when it is smoothed, the law of large numbers does not erase all features, but it always retains oscillations of the same contrast. An example of an image sampled from this distribution is shown in Fig. 1.

We will construct non-Gaussian rotation and scale-invariant measures below. At this point note that, even when non-Gaussian, their covariance must be $c_1 \log(\| \vec{x} - \vec{y} \|)$; hence their power spectrum takes the form

$$\text{Exp}(\hat{I}(\vec{\xi}_1)\hat{I}(\vec{\xi}_2)) = \delta_{\vec{\xi}_1,\vec{\xi}_2} \frac{1}{\| \vec{\xi}_1 \|^2}.$$

4. Nonexistence of scale invariant measures on functions. The scale-invariant Gaussian probability measure is well known *not* to be supported on the subspace of measurable functions $L^1_{\text{loc}} \subset \mathcal{D}'$ ([Do], [Re]). This makes such a model seem complicated and inaccessible. Mallat, Meyer, Donoho, Simoncelli, Coifman and others have instead proposed various spaces of true functions as natural models for images (see [Ma]). For example, they propose the space of functions $I(x,y)$ of finite total variation, or the more subtle Besov spaces or specific spaces of wavelet expansions with mother wavelet(s) adapted to image geometry. The point of this section is to prove that there is no scale-invariant probability measure on such spaces (or on them modulo constants). The idea is that scale-invariance automatically implies oscillations everywhere of the same amplitude and measurable functions cannot be so complex. Thus, accepting the stochastic approach to images and their scale invariance forces you to model images by Schwartz distributions

[1] This holds in any dimension. For any d, there is a unique Gaussian probability measure on $\mathcal{D}'_d$ with mean 0 and covariance $C(\vec{x}) = c_1 \log(\| \vec{x} \|) + c_2$. To see where it "lives", we may construct it from "white noise" (see [Hi]) which is given by $C(\vec{x}) = 0$ if $\vec{x} \neq 0$ or equivalently by

$$\text{Exp}_I(e^{i<f,I>}) = e^{-\int |f(\vec{x})|^2 d\vec{x}}.$$

It is well known that white noise is supported in $\bigcap_{\epsilon>0} H^{-d/2-\epsilon}_{\text{loc}}$. The scale-invariant measure is constructed by operating on white noise by convolution with the kernel $\| \vec{x} \|^{-d/2}$ which boosts the differentiability by $d/2$.

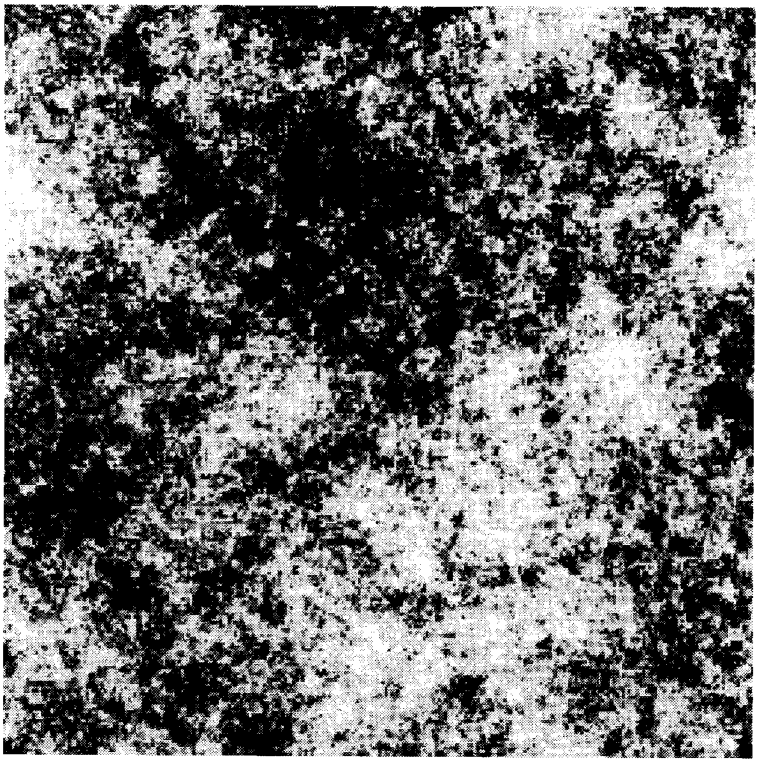

FIG. 1. A computer simulation of a sample of the scale-invariant Gaussian measure on images

that are not measurable functions. It seems that the models of images proposed by the wavelet community are really models of the "cartoon" component J obtained by decomposing an image I into a sum $J + n$, where n is noise or texture or clutter and J are the major salient parts of the image. But, whereas n has been modeled as noise, we propose that it really follows the same statistics as J only scaled down (compare Meyer's discussion [Me]).

THEOREM. Assume μ is a probability measure on $\mathcal{D}'_d$, distributions on $\mathbb{R}^n$ modulo constants. Assume μ is invariant by translations and scale transformations and that μ is not a delta function with support 0. Then μ is not supported on the subspace of locally integrable functions

$$(L^1_{\text{loc}}/\mathbb{R} \cdot 1) \subset \mathcal{D}'_d.$$

Proof. Assume μ is supported on $L^1_{\text{loc}}/\mathbb{R} \cdot 1$. Fix any positive numbers a and r and let $B_r(\vec{x})$ denote the ball of radius r with center $\vec{x}$. Let $|S|$ denote the Lebesgue measure of a subset $S \subset \mathbb{R}^n$. For any $I \in L^1_{\text{loc}}$ and $\vec{x} \in \mathbb{R}^n$, define

$$g_{a,r}(I, \vec{x}) = |\{\vec{y} \in B_r(\vec{x}) \mid |I(\vec{x}) - I(\vec{y})| > a\}|/|B_r(\vec{x})|.$$

This defines a measurable function:

$$g_{a,r} : (L^1_{\text{loc}}/\mathbb{R} \cdot 1) \times \mathbb{R}^n \to [0, 1].$$

In particular, we can integrate $g_{a,r}$ times the measure μ and get $\text{Exp}_I(g_{a,r}) = \overline{g}_{a,r}(\vec{x})$. Because μ is invariant with respect to translations, $\overline{g}_{a,r}$ is constant as a function of $\vec{x}$. Because μ is also scale-invariant, $\overline{g}_{a,r}$ is independent of r too! Thus

$$\text{Exp}_I(g_{a,r}(\cdot, \vec{x})) = p_a$$

for some constant p_a depending only on a.

Next, consider $g_{a,r}(I, \vec{x})$ for fixed I and $r \to 0$. Recall that Lusin's theorem for the function I states that I is "almost everywhere continuous". More precisely, for every $\epsilon > 0$, there is a set $Z_\epsilon \subset \mathbb{R}^n$ with $|Z_\epsilon| \leq \epsilon$ such that $I|_{(\mathbb{R}^n - Z_\epsilon)}$ is continuous. Recall that if $S \subset \mathbb{R}^n$ is measurable, S has density 1 at a point $\vec{x} \in S$ if

$$\lim_{r \to 0} |S \cap B_r(\vec{x})|/|B_r(\vec{x})| = 1$$

and that there is always a set $S_{\text{bad}} \subset S$ with $|S_{\text{bad}}| = 0$ such that $\vec{x} \in S - S_{\text{bad}}$ implies S has density 1 at $\vec{x}$. Combining these two shows that

$$\lim_{r \to 0} g_{a,r}(I, \vec{x}) = 0 \quad \text{if} \quad \vec{x} \in (\mathbb{R}^n - Z_\epsilon) - (\mathbb{R}^n - Z_\epsilon)_{\text{bad}}.$$

But

$$Z_0(I) = \bigcap_\epsilon [Z_\epsilon \cup (\mathbb{R}^n - Z_\epsilon)_{\text{bad}}]$$

has measure 0. Hence, for fixed I and $\vec{x} \notin Z_0(I)$,

$$\lim_{r \to 0} g_{a,r}(I, \vec{x}) = 0.$$

Note that $Z_0 = \bigcup_I Z_0(I)$ is a set of measure zero in $(L^1_{\text{loc}}/\mathbb{R} \cdot 1) \times \mathbb{R}^n$.

Now apply Lebesgue's bounded convergence theorem:

$$\begin{aligned} p_a &= \lim_{r \to 0} \text{Exp}_I(g_{a,r}(I, \vec{x})) \\ &= \text{Exp}_I(\lim_{r \to 0} g_{a,r}(I, \vec{x})) \\ &= 0. \end{aligned}$$

Thus for every a and r, $g_{a,r} = 0$ for almost all I and $\vec{x}$. This can only happen if the set of non-constant I has μ-measure 0, i.e., μ is a delta function supported on 0. This proves the theorem. □

5. Axiom II: Clutter and infinite divisibility. A fundamental fact about the world (or, at least, about the way we think about the world) is that it is not a formless mixture of stuff, but is broken up into discrete *objects*. Individual objects are the things that we name, the things we pick up, and the things that have a specific use. What constitutes an object is never precise: objects typically are made up of parts, which may be thought of as distinct objects, and are part of larger assemblages which can be treated as single objects. The prototypical object is a simple rigid thing made of a single material

with a homogeneous appearance which can be moved independently of the rest of the world, e.g., a knife or a stone. But most objects are more complex and have parts: a body is a single object (e.g., it resists dismemberment), but it is made of parts—limbs, trunk, head, etc.—which move as separate almost rigid objects. Other "objects", referred to in language by so-called mass nouns, break up into tiny parts. Thus sand is made up of a huge number of grains.

Since the 3D world breaks up into objects, the 2D views produced by imaging the world also break up into the viewed surfaces of each object. Visually, simple objects are most readily identified by their motion relative to the rest of the image, e.g., by their simple optic flow fields; but they often appear clearly in static images by virtue of their homogeneous color or texture, separated from the background by sharp intensity or local power spectrum discontinuities. It is natural to break up 2D views of single objects into further parts on the basis of albedo changes as well as its 3D parts. For instance, if the surface of a sweater is variously colored, its pattern breaks its visible surface into distinct 2D surface parts. In other cases there is a mixture of geometric and illumination factors that break a surface into parts. For instance, the visible surface of a lake may break up into vast numbers of ripples. You may also treat shadows and highlights as "parts" of the surface, objects in the 2D world of the image. From the point of view of images, all these effects break up a part U of the image domain into subparts $U_i \subset U$ which we will consider as being the viewed portion of a virtual object, an infinitely flattened object on the surface of another.

Can we express the fact that images depict a world of objects as a mathematical property of the probability measure μ on images? This property is not a simple one to capture, but, as a first approximation, we propose that it means that the measure μ is *infinitely divisible*. Recall that a probability measure μ on $\mathbf{R}$ is infinitely divisible if, for every $n \geq 2$, there is a probability measure $\mu^{(n)}$ such that $\mu = \mu^{(n)} * \mu^{(n)} * \cdots * \mu^{(n)}$ (where $*$ represents convolution). Translating this into the language of random variables, if x is a random variable distributed by μ, then for every n, x can be written as a sum $x = x_1 + x_2 + \cdots + x_n$, where x_i are "iid", independent and identically distributed. It is a theorem that infinitely divisible distributions belong to semi-groups of distributions (see e.g. [Sa]), i.e., for each such μ, there is a family of measures μ_t, defined for all $t \geq 0$, such that $\mu = \mu_1$ and $\mu_s * \mu_t = \mu_{s+t}$. The measures $\mu^{(n)}$ in the definition are just the measures $\mu_{1/n}$ in the semi-group. This gives us the intuitive characterization of infinitely divisible distributions as the marginal distributions on the value X_1 of stationary stochastic processes $\{X_t, t \geq 0, X_0 = 0\}$ with independent increments: i.e., the distribution of $X_{t_1} - X_{t_2}$ depends only on $t_1 - t_2$—it will be $\mu_{t_1 - t_2}$—and it is independent of $X_{s_1} - X_{s_2}$ if the intervals $[t_1, t_2]$ and $[s_1, s_2]$ are disjoint. The same definition works for vector-valued random variables as well as scalar random variables. Thus we define a probability measure μ on a function space $\mathcal{E}$ to be infinitely divisible if for every $n \geq 2$, there is a probability measure $\mu^{(n)}$ such that $\mu = \mu^{(n)} * \mu^{(n)} * \cdots * \mu^{(n)}$. Then there is a semi-group μ_t as before, and a random variable in $\mathcal{E}$ chosen from μ is an iid sum of n random variables in $\mathcal{E}$ chosen from $\mu^{(n)}$.

Thus, for images I, we propose:

Axiom II *(infinite divisibility):*

(1) Every image I has associated with it a parameter c, the *clutter* of I,

(2) Images with clutter c are random samples from a probability measure μ_c on $\mathcal{D}'$ and $\mu_c * \mu_d = \mu_{c+d}$.

This is equivalent to saying that an image I with clutter c can be formed as a sum $I = I_1 + I_2 + \cdots + I_n$, where I_k are *independent* images each with clutter c/n. What we have in mind is to create an image with a certain level of clutter as the superposition of images with less clutter. This is clearly a toy version of the way nature makes the real world, starting with bare land, adding rocks, trees, animals, more or less at random. It is not meant to be exactly true of the distribution of generic images, but we propose it as being approximately true, like the axiom of scale invariance.

Let us try to be clearer about "how true" this axiom is for real world images. It seems reasonable to imagine the world as being formed by placing objects in a scene, some simple, some compound, some in large arrays, and by painting their surfaces with patterns and shadows made up of other shapes, simple, compound and textured. This scene is then viewed from a random viewpoint. If we make simpler scenes by leaving out all but a few of its component objects and surface shapes, we can imagine recreating the full scene by adding together these simpler scenes. This will work except for one main phenomenon which will not be captured. This is partial occlusion. Imagine a scene with two objects O_1, O_2 viewed from some point P. If neither occludes the other, the resulting image is the sum of the images of the two separate objects. If O_1 is in front of O_2 and its outline is wholly inside of O_2's, then the resulting image is the sum of O_2 and that of O_1 but painted with the difference of the colors of O_1 and O_2. The hard case is when one object partially occludes the other: this results in "T-junctions" where their contours intersect and the object in front must be painted so as to cancel out the occluded portion of the contour of the object in back. This cannot be done without violating independence of the two simpler images. Thus we propose that T-junctions in images are the simplest structures that violate the infinitely divisible axiom. In addition to T-junctions, partial occlusion produces extended contours which are broken into pieces, which also cannot arise from an infinitely divisible distribution.

6. Axiom III: Locality of objects and ergodicity. The above two axioms are nearly all we want. In fact, a remarkable fact is that infinite divisibility nearly produces objects for us. To see this, we need the famous Levy-Khintchine theorem, which makes very explicit the nature of infinitely divisible distributions. Looking first at the case of scalar random variables x with distributions μ, the Levy-Khintchine theorem, in its usual form, asserts the existence of an auxiliary measure ν, called the Levy measure, on $\mathbb{R} - (0)$ such that $(x^2\nu)([-1,1]) < \infty$ and

$$\int e^{ix\xi}\mu(dx) = e^{ix_0\xi - \sigma^2\xi^2/2 + \hat{\nu}(\xi)},$$

where $\hat{\nu}$ is the Fourier transform of ν, interpreted by defining the principal part of ν as a distribution.

This theorem can be rewritten as an explicit recipe for constructing the random variable x:

$$\begin{aligned} x &= x_0 + \sigma x_1 + \sum_{i \geq 2} x_i, \text{ where} \\ x_0 &= \text{a constant,} \\ x_1 &= \text{a standard normal variable,} \\ x_2, x_3, \cdots &= \text{a Poisson process on } \mathbb{R} - (0) \text{ with density } \nu. \end{aligned}$$

The theorem should be understood to mean that the random variable x has a fixed part, a Gaussian part, and a discrete part that is the sum of a Poisson process, i.e., a countable set of points in $\mathbb{R}-(0)$ distributed randomly with density given by the measure ν. The simple case is where $\nu(\mathbb{R} - (0)) < \infty$, so that the Poisson process consists in a finite set of points and the sum in the discrete part is finite. In this case the Fourier transform $\hat{\nu}$ of the measure ν exists in the usual sense. To include all infinitely divisible distributions, however, ν must be allowed to have infinite measure around 0 so long as $x^2\nu$ assigns finite measure to a neighborhood of 0. In this case, we have to add convergence factors to the series for the discrete part of x (and the series must be summed in the right order). We ignore these technicalities.

The important case for us are those scalar random variables x such that

$$x = x_0 + \sum_i x_i, \qquad \{x_i\} \text{ Poisson for a finite measure } \nu.$$

We may think of these variables as scalar variables resulting from the superposition of a finite number of "objects". It is a standard result that such x's are exactly those with infinitely divisible distributions *and* $\Pr(x = x_0) > 0$ (i.e., their distributions have "atoms").

The Levy-Khintchine theorem generalizes to random variables I with values in Banach spaces $\mathcal{X}$ (see [Li]):

$$\begin{aligned} I &= I_0 + I_1 + \sum_{i \geq 2} I_i, \text{ where} \\ I_0 &= \text{a constant,} \\ I_1 &= \text{a Gaussian random variable,} \\ I_2, I_3, \cdots &= \text{a Poisson process on } \mathcal{X} - (0) \text{ with density } \nu. \end{aligned}$$

As before, the Levy measure ν is a measure on $\mathcal{X} - (0)$ with possible singularities at 0, and the sum has to be interpreted carefully if this singularity is too big.

We want to apply this to the probability measure on images, with $\mathcal{X} \subset \mathcal{D}'_d$ being a Banach subspace of the full space of distributions (modulo constants) which carries the measure μ. The meaning of the samples $I_k, k \geq 2$ from the Levy measure is that these component images are the primitive "objects" out of which images are composed. Thus we add our third locality axiom, which states that these objects are given by functions and are compactly supported:

Axiom III *(locality):* The Levy measure ν for images I is supported on the space L^1_c of measurable functions with compact support.

(Note that L^1_c is a subspace of $\mathcal{D}'_d$ because, among all the functions $f + a, a \in \mathbb{R}$, only one can have compact support.)

What happens when we combine axioms 1, 2 and 3? From infinite divisibility we get the Levy-Khintchine decomposition of I. From scale invariance and the uniqueness of the Levy measure, we conclude that ν must also be translation and scale invariant. In other words, ν is invariant under the three-dimensional group $G = \mathbb{R}^2 \times \mathbb{R}_+$ of maps $(x, y) \to (rx + a, ry + b)$. By axiom 3, ν is supported on L^1_c. Denote by $p : G \times L^1_c \to L^1_c$ the action of the group G on our function space. On the other hand, there is a measurable map $L^1_c - (0) \to \mathbb{R}^2 \times \mathbb{R}_+$ taking every nonzero function f to the center and radius of the unique smallest circle on which it is supported. Let $L^1_u \subset L^1_c$ be the set of functions whose support is contained in the unit circle and in no smaller cirle. Note that acting by $(a, b, r) \in G$ carries the unit circle to the circle with center (a, b) and radius r. So combining these two maps, we get a measurable isomorphism

$$\mathbb{R}^2 \times \mathbb{R}_+ \times (L^1_u - (0)) \cong L^1_c - (0)$$

defined from the LHS to the RHS by the action of the group and from the RHS to the first two factors on the left by the smallest circle construction. The scale invariance of the Levy measure then means that, in this product decomposition, it is a product of Haar measure $dx.dy.dr/r$ on the group[2] and a reduced Levy measure ν_u on L^1_u:

$$\nu = \frac{dx.dy.dr}{r} \times \nu_u.$$

7. Axiom IV: Blue sky. We next want to consider the reduced Levy measure ν_u. A fundamental property of the world and of images of it is that they have blank spaces in them: the blue sky, blank painted walls. This is really a property of scale-space. When things on all scales and at different locations are put together in a scene, there should be parts of scale-space that are not sampled. This effect can arise from the natural fluctuations of the sampling density of the Poisson process, but only if the Levy measure is not too big. In order that our model will produce images with blank regions in them, we assume:

Axiom IV *(blue sky)*
(1) The constant image I_0 is zero,
(2) The Gaussian component I_1 is zero, and
(3) The reduced Levy measure ν_u is finite.

[2]Confusingly, $da.db.dr/r$ is the *right* invariant Haar measure and $da.db.dr/r^3$ is the *left* invariant Haar measure. If the group acts on functions in the usual way, we believe the right-invariant measure on the group makes the product measure invariant.

This axiom implies that the Poisson process I_i sampled from ν can be constructed from a Poisson process (a_i, b_i, λ_i) in the group G with density $cdadbd\lambda/\lambda$ plus independent random choices J_i sampled from ν_u. Thus it gives us the explicit form of the expansion:

$$I(x, y) = \sum_i J_i(\lambda_i x + a_i, \lambda_i y + b_i).$$

We want to call such an expression a *random wavelet expansion.* The terms in this expansion are meant to model the individual "objects" in the scene (where objects is interpreted to include things such as parts of patterns, shadows, textons, etc., as discussed above).

The above axiom implies that for any bounded part K of G there is a nonzero probability that the series for I contains no term with $(a_i, b_i, \lambda_i) \in K$. This means that the resulting image I contains no objects of a certain bounded range of sizes in a certain bounded part of the image plane. Hence images have nearly blank areas, when blurred to eliminate infinitesimal features and considered mod constants to eliminate huge features. It is not clear, however, that any measure of the above type exists. The series clearly converges if we put infra-red and ultra-violet cutoffs, but this is not clear in the full scale-invariant case. The convergence for this case is discussed in the next section.

What do such random wavelet images look like? We have simulated them for several choices of ν_u and displayed the results in Figs. 2-3. In the first image, ν_u is supported on the characteristic functions of circles and the clutter is low to show the individual terms of the expansion clearly. Each is colored by a Cauchy random variable and the whole image is displayed with a gamma correction, applying a sigmoidal function $1/(1 + \exp(-I/c))$. (This is supposed to mimic real images where the typical ratios of maximum to minimum intensities are 100–1000 and are displayed by film with some gamma correction to compress the dynamic range.) In the next two images, circles are replaced by "ribbons" (also called "worms" or "snakes") obtained by sweeping a circle of varying radius along a curve called its medial axis. In the simulation, the orientation of the medial axis and the log of the radius are given by independent Brownian functions of arc length and the length of the axis is exponentially distributed. In the last image, the support of every function in ν_u is a rectangle but it is not colored with constant intensity but by a sum of a constant and of three random sine-waves.

8. Convergence of random wavelet expansions. In this section, we want to prove that, with mild conditions on the functions in the support of the Levy measure ν, the random wavelet expansions (7.1) converge *almost surely* as distributions. We saw in §4 that they cannot converge almost surely as functions because they would then define a scale-invariant probability measure on functions. However, it turns out that, like samples from the scale-invariant Gaussian model, random wavelet expansions live "just outside" functions.

Let us fix our notation. It is no extra work to consider "images" on $\mathbb{R}^d$ for any dimension d. Define as usual:

$$\begin{aligned} \mathcal{D} &= \mathcal{D}(\mathbb{R}^d) = C^\infty \text{ functions with compact support,} \\ \mathcal{D}' &= \mathcal{D}'(\mathbb{R}^d) = \text{ dual of } \mathcal{D}, \text{ generalized functions,} \end{aligned}$$

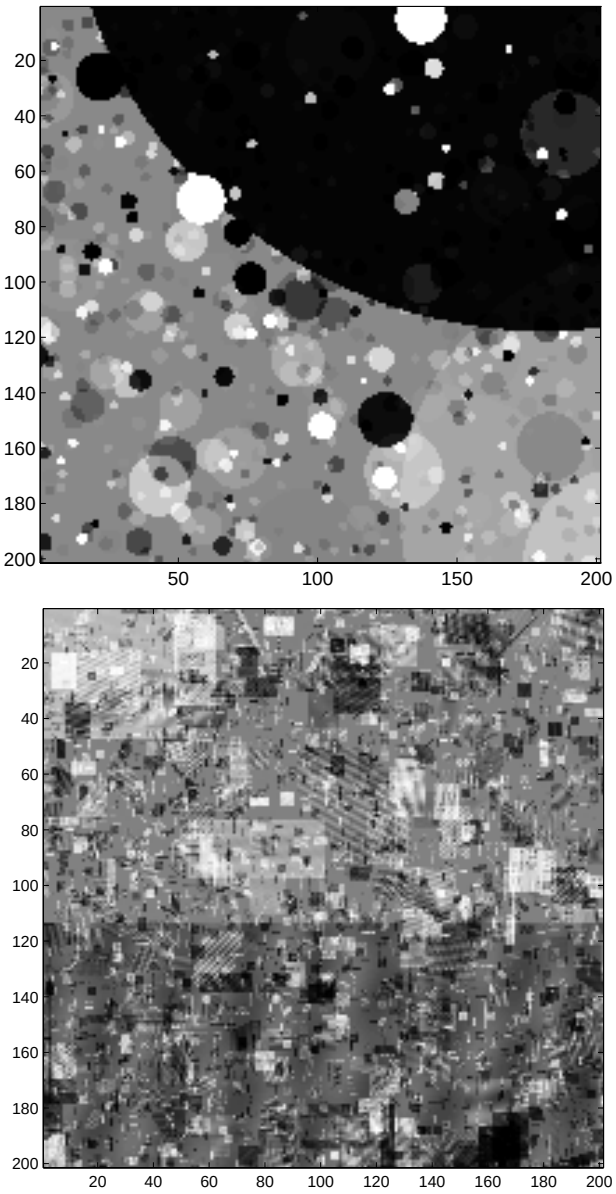

Figure 2: A computer simulated sample of random wavelet images, one with low clutter and disk-like primitives, one with higher clutter and textured rectangular primitives. See text for details.

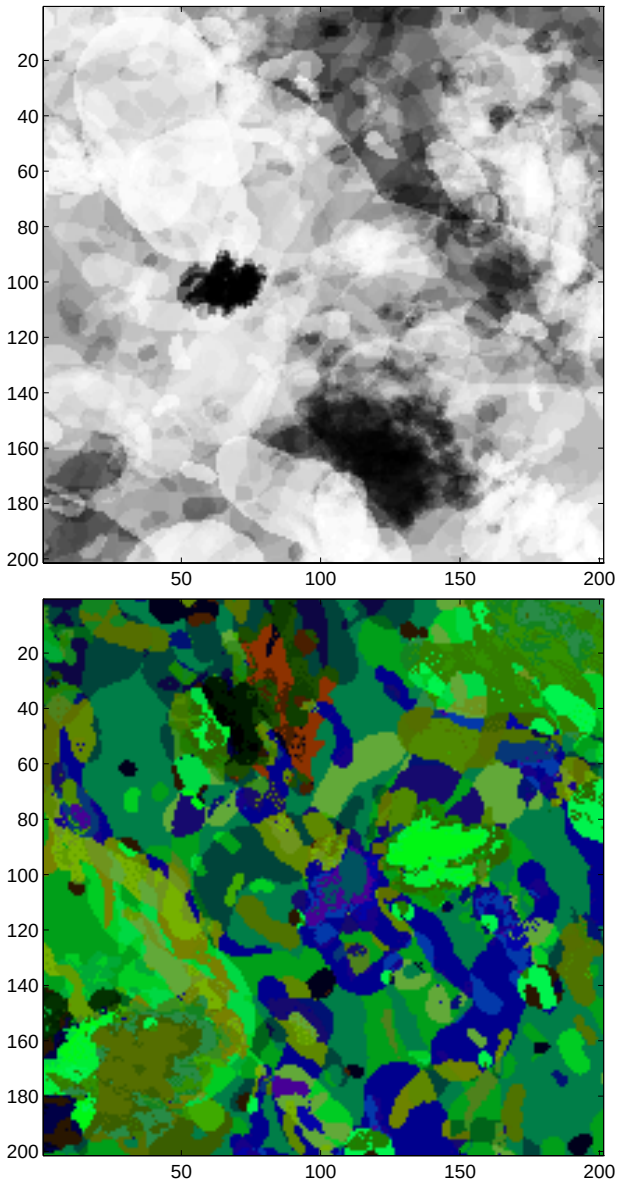

Figure 3: Computer simulated samples of two random wavelet image with ribbon-like primitives. See text for details.

Finally define

$$H^0_{\rm loc} \quad = \text{ functions with } \int_K f^2 < \infty, \text{ all bounded } K.$$

$$H^{-s}_{\rm loc} \quad = (I - \triangle)^{s/2}(H^0_{\rm loc}), s \geq 0.$$

$$
\begin{aligned}
\mathcal{D}_d &= \left\{ f \in \mathcal{D} \,\middle|\, \int_{\mathbb{R}^d} f(x)dx = 0 \right\}, \\
\mathcal{D}'_d &= \text{dual of } \mathcal{D}_d = \mathcal{D}'/\mathbb{R}\cdot 1, \\
H^s &= \text{Hilbert-Sobolev space with norm,} \\
\|f\|_s^2 &= \int (1+\|\xi\|^2)^{s/2} |\hat{f}(\vec{\xi})|^2 d\vec{\xi} \\
&= \int ((1-\Delta)^{s/2} f \cdot f) d\vec{x}, \\
K_s(\vec{x}) &= ((1+\|\vec{\xi}\|^2)^{-s/2})\hat{}\,(\vec{x}), \text{ the semigroup of "Bessel" kernels, } s \geq 0 \\
&= \text{const.} \|\vec{x}\|^{\frac{s-d}{2}} K_{\frac{d}{2}-s}(\|\vec{x}\|), \text{ where } K \text{ is the classical Bessel function.}
\end{aligned}
$$

Thus,

$$K_s * H^t = H^{t+s}, \text{ all } t, s \geq 0.$$

Finally, define

$$
\begin{aligned}
H^0_{\text{loc}} &= \text{functions with } \int_K f^2 < \infty, \text{ all bounded } K, \\
H^{-s}_{\text{loc}} &= (I-\triangle)^{s/2}(H^0_{\text{loc}}), \ s \geq 0.
\end{aligned}
$$

As above, let ν_u be the reduced Levy measure supported on functions whose support is contained in the unit ball (and no smaller ball). We want to assume ν_u is supported in a fractional Sobolev space. The reason this is useful is that natural models for the elementary components of images may include functions that are smooth on a domain K with smooth boundary, but 0 outside K and thus discontinuous on ∂K. Such functions are typically in H^s for all $s < 1/2$. We shall prove:

THEOREM. Assume that for some $\epsilon > 0$

$$\int \|J\|_\epsilon^2 d\nu_u(J) < \infty.$$

Then the random wavelet series:

$$I(\vec{x}) = \sum_i J_i(\lambda_i \vec{x_0} - \vec{x_i}), \ \vec{x} \in \mathbb{R}^d$$

converges almost surely in $H^s_{\text{loc}}/\mathbb{R}\cdot 1$, all $s < 0$.

Proof. We shall show that for all $s > 0$, the series $K_s * I$ converges almost surely in $H^0_{\text{loc}}/\mathbb{R}\cdot 1$. Break up the formal series for I as follows:

$$
\begin{aligned}
I(\vec{x}) &= \sum_{k\in\mathbb{Z}} I_k(\vec{x}), \\
I_k(\vec{x}) &= \sum_{2^k \leq \lambda_i < 2^{k+1}} J_i(\lambda_i \vec{x} - \vec{x_i}).
\end{aligned}
$$

Then I_k is a locally finite sum, i.e., on all bounded K, only a finite number of terms of I_k are nonzero on K. Note that the summands I_k are independent random functions. To show convergence, we use the basic fact (the easy case of Kolmogorov's Two Series theorem):

PROPOSITION. Let $\{f_k\}$ be independent random variables in a Hilbert space H. Then $\text{Exp}(f_k) = 0$ and $\sum \text{Exp}(\|f_k\|^2) < \infty$ implies that $\sum f_k$ converges almost surely in H.

Thus we need to show

a) $\text{Exp}(K_s * I_k) =$ constant function a_k, which will be independent of k.

b) Fix a large ball D and let $\overline{f} = \frac{1}{|D|}\int_D f$. Then

$$\sum_k \text{Exp}\left(\int_D (K_s * I_k - \overline{K_s * I_k})^2\right) < \infty.$$

The basic calculation is an application of Campbell's theorem ([Ki], §3.2) to find the mean and variance of $(K_s * I_k)(\vec{x})$. Recall that Campbell's theorem states that if $\{x_i\}$ is a Poisson process with density $\nu(x)$, then

$$\text{Exp}\sum_i f(x_i) = \int f(x)d\nu(x),$$

$$\text{Var}\sum_i f(x_i) = \int f(x)^2 d\nu(x).$$

This gives

$$\begin{aligned}\text{Exp}((K_s * I_k)(\vec{x})) &= \int_{2^k}^{2^{k-1}} \frac{dr}{r}\int d\nu_u(J)\iint K_s(\vec{y})J(r\vec{y}-\vec{x})\,d\vec{x}\,d\vec{y} \\ &= \log(2)\cdot\int K_s(\vec{y})d\vec{y}\cdot\iint J(\vec{x})\,d\vec{x}\,d\nu_u(J) \\ &= \log(2)\cdot\text{Exp}_{\nu_u}\left(\int J\right),\end{aligned}$$

which proves the simple estimate (a). Campbell's theorem also shows

$$\begin{aligned}\text{Var}\left((K_s * I_k)(\vec{x})\right) &= \int\frac{dr}{r}\int d\nu_u(J)\int d\vec{x}\left(\int K_s(\vec{y})J(r\vec{y}-\vec{x})d\vec{y}\right)^2 \\ &= \int\frac{dr}{r}\int d\nu_u(J)\iiint K_s(\vec{y_1})K_s(\vec{y_2})J(r\vec{y_1}-\vec{x})J(r\vec{y_2}-\vec{x})\,d\vec{y_1}\,d\vec{y}\,d\vec{x}.\end{aligned}$$

Let

$$C(\vec{u_1},\vec{u_2}) = \int d\nu_u(J)J(\vec{u_1})J(\vec{u_2}).$$

Substituting $\vec{u_i} = r\vec{y_i} - \vec{x}$ and using the fact that K_s is even, we find

$$\begin{aligned}&\text{Var}\,(K_s * I_k(\vec{x})) \\ &\quad= \int\frac{dr}{r}\iint d\vec{u_1}\,d\vec{u_2}\,C(\vec{u_1},\vec{u_2})r^{-d}\int K_s(\frac{u_1}{r}+\vec{x})K_s(-\frac{u_2}{r}-\vec{x})d\vec{x} \\ &\quad= \int\frac{dr}{r^{d+1}}\iint d\vec{u_1}\,d\vec{u_2}\,C(\vec{u_1},\vec{u_2})K_{2s}(\frac{\vec{u_1}-\vec{u_2}}{r}).\end{aligned}$$

A similar calculation with two points $\vec{x_1},\vec{x_2}$ shows

$$\text{Cov}\,(K_s * I_k(\vec{x_1}), K_s * I_k(\vec{x_2}))$$

$$= \int \frac{dr}{r^{d+1}} \int d\vec{u_1}\, d\vec{u_1}\, C(\vec{u_1}, \vec{u_2}) K_{2s}\left(\frac{\vec{u_1} - \vec{u_2}}{r} + \vec{x_1} - \vec{x_2}\right).$$

The norm we want, however, is an integral:

$$\frac{1}{|D|}\,\mathrm{Exp}\left(\int_D \left(K_s * I_k - \overline{K_s * I_k}\right)^2\right)$$

$$= \frac{1}{2|D|^2}\,\mathrm{Exp}\left(\iint_{D\times D} \left(K_s * I_k(\vec{x_1}) - K_s * I_k(\vec{x_2})\right)^2 d\vec{x_1}\, d\vec{x_2}\right)$$

$$= \mathrm{Var}\left(K_s * I_k(\vec{x})\right) - \int w(\|\vec{z}\|)\,\mathrm{Cov}\left(K_s * I_k(\vec{x}), K_s * I_k(\vec{x} + \vec{z})\right) d\vec{z}$$

where

$$w(\|\vec{z}\|) = \frac{|D \cap (D + \vec{z})|}{|D|^2}.$$

Using our pointwise estimate, we get

$$\frac{1}{|D|}\mathrm{Exp}\left(\int_D \left(K_s * I_k - \overline{K_s * I_k}\right)^2\right) =$$

$$\int_{2^k}^{2^{k+1}} \frac{dr}{r^{d+1}} \iint d\vec{u_1}\, d\vec{u_2}\, C(\vec{u_1}, \vec{u_2}) K^*_{2s}\left(\frac{\vec{u_1} - \vec{u_2}}{r}\right)$$

where

$$K^*_{2s} = K_{2s} - w * K_{2s}.$$

We can now sum this estimate over k. We need to treat $k \to \infty$ and $k \to -\infty$ separately. For k large and positive we use the fact that

$$C(\vec{u_1}, \vec{u_2}) \neq 0 \Rightarrow \|\vec{u_1} - \vec{u_2}\| \leq 2$$

and that

$$K_s(\vec{x}) \approx \text{const.}\|\vec{x}\|^{s-d} \text{ for } \|\vec{x}\| \text{ small}.$$

This gives us

$$\frac{1}{|D|}\mathrm{Exp}\left(\int_D (K_s * I_k - \overline{K_s * I_k})^2\right) \leq \text{ const. } \int_{2^k}^{2^{k+1}} \frac{dr}{r^{1+2s}}$$

whose sum converges as $k \to \infty$. For $|k|$ large, k negative, we pass to the Fourier transform. Let

$$D(\vec{\xi}) = \int d\nu_u(J) |\hat{J}(\vec{\xi})|^2.$$

Then Plancherel's theorem gives us

$$\iint C(\vec{u_1}, \vec{u_2}) K^*_{2s}\left(\frac{\vec{u_1} - \vec{u_2}}{r}\right) d\vec{u_1}\, d\vec{u_2} = \int D(\vec{\xi}/r)\widehat{K^*_{2s}}(\vec{\xi})\, d\vec{\xi}.$$

Now $\widehat{K^*_{2s}}$ is smooth and 0 at 0. Thus for any $\epsilon > 0$,

$$|\widehat{K^*_{2s}}(\vec{\xi})| \leq C_K \|\vec{\xi}\|^{\epsilon}.$$

But

$$\begin{aligned}\int \|\xi\|^{\epsilon} D(\xi) &= \int d\nu_u(J) \int \|\vec{\xi}\|^{\epsilon} |\hat{J}(\vec{\xi})|^2 d\vec{\xi} \\ &< \infty \text{ if } \epsilon \text{ is small enough.}\end{aligned}$$

Combining these, we get

$$\begin{aligned}\frac{1}{|D|}\text{Exp}\left(\int_D (K_s * I_k - \overline{K_s * I_k})^2\right) &= \int_{2^k}^{2^{k+1}} \frac{dr}{r} \int D(\vec{\xi}) C_K r^{\epsilon} \|\vec{\xi}\|^{\epsilon} d\vec{\xi} \\ &\leq \text{const.} \int_{2^k}^{2^{k+1}} \frac{dr}{r^{1-\epsilon}}\end{aligned}$$

whose sum converges as $k \to -\infty$. □

9. Experiments. In this section, we will review the experiments that have been carried on with small and large databases of natural images which address the question of whether the four axioms adopted above are reasonable. The first axiom is that of scale-invariance. For this, there are now quite a substantial number of experimental tests, which, altogether, give quite strong support for the thesis that any reasonably large and representative set of natural images of the world can be viewed as samples from a scale-invariant stochastic model.

9.1. *Scale-invariance.* To test for scale-invariance, one must select specific measurable statistics, which can be estimated from storable databases, and see whether their values are consistent with a scale-invariant model. The statistics that have been examined include:

(1) second-order statistics: the power spectrum and/or auto-correlation;
(2) higher-order statistics on filter responses: moments and histograms;
(3) order statistics of pixel values in small windows;
(4) topological statistics obtained from morphological operations.

The study of the power spectra of natural images originated with television engineers studying band-width issues for the transmission of TV signals in the 1950s ([De],[Kr]). These old results were rediscovered quite recently by Ruderman and Bialek [R-B], who analyzed a small set of images of woods near Princeton finding near scale-invariance. As noted above, the second-order statistics of all scale-invariant models are identical and predict that the power spectrum will fall off like C/ξ^2, where ξ is the spatial frequency. Ruderman and Bialek's results actually gave the best fit as $C/\xi^{1.81}$. This experiment was repeated by many people and it was observed that individual images have a wide range of spectra and that, when fit with power laws, the exponent varied from around 1.5 to around 3. Nonetheless, numbers near 2 seemed to appear whenever the database was large.

An especially careful version of this experiment has been conducted by J. Huang [H-M2] using a database of 214 *calibrated* images collected by British Aerospace. The images are outdoor scenes shot near Bristol England containing 512×768 pixels each and show urban and rural scenes of all kinds. Being calibrated, these images are described by numbers representing energy received by a sensor and have not been subjected to the usual gamma-correction, let alone any free-wheeling histogram manipulation in Adobe

Photoshop. This means that if we take logs of these values and apply any linear filter with mean zero, we get values that are dimension-free and represent objective measurements of light in the world. We have taken this approach. Moreover, British Aerospace has laboriously segmented each of these images into 11 classes of pixels: these include categories such as vegetation, roads, buildings, etc. This database makes possible the examination of second-order statistics for each category separately as well as for the whole ensemble. Since the pixels in each category are not whole images, the approach must be modified to get the exponent. But since all second-order statistics of a stationary process are given by the power spectrum, we must get the same exponent. The method chosen was to look at adjacent pixels in the original image belonging to the same category, adjacent 2×2 blocks of such pixels, adjacent 4×4 blocks and adjacent 8×8 blocks. At each scale, the variance of the difference of the average pixel intensities in the two adjacent blocks was computed. Then the log of these variances was plotted against scale and a linear regression was done.

The results show that

(1) for the vegetation category, the power spectrum scales like $C/\xi^{1.8}$ similar to Bialek and Ruderman's results,
(2) for manmade categories, the power spectrum scales like $C/\xi^{2.3}$,
(3) for road surfaces, the power spectrum scales like $C/\xi^{1.4}$,
(4) for the sky (including clouds), the power spectrum scales like C/ξ,
(5) for all categories together, the power spectrum scales like C/ξ^2.

These results confirm that there is great local variability in the second-order statistics, with blank and/or white-noise-like regions in some parts of some images at some scales, shifting power to higher frequencies; and with large objects and less clutter, e.g., in manmade settings, shifting the power to lower frequencies. Taking the whole database, there is very good fit with scale-invariance.

Looking beyond second-order statistics, Mallat, Meyer and others have proposed that Besov spaces are natural spaces for images, Besov norms being computable from suitably scaled p-norms on wavelet coefficients. This leads you to the statistics given by higher moments and by quantile measures of filter responses. To get to the heart of all potential statistics derived from single filter responses, it seems best to consider the whole histogram of filter responses and ask whether this entire histogram remains the same when the filter is scaled. If this histogram is scale-invariant, then so are all expected filter moments, quantiles, etc. The basic test, then, is to select a wavelet expansion of the image and measure the histogram of wavelet coefficients for a fixed wavelet at different scales. Huang and one of the authors have carried this out for a) the simplest possible Haar-type filter, namely the difference of adjacent pixels and b) a sophisticated wavelet filter from Simoncelli and Freeman's steerable pyramid [S-F]. To do this, we used an even larger, though unsegmented database, of 4000 1024×1536 calibrated images of Holland, assembled by J. H. van Hateren [vH]. In both cases, the histograms show an amazingly precise scale-invariance, out 8 or more standard deviations. Moreover, the histograms are consistent between the Dutch and the British databases. In Fig. 4, we show the logs of the histograms for the steerable pyramid wavelet at scales 1,2,4 and 8 (to "see" what is happening in the tails, it is useless to plot the frequencies themselves: one must plot

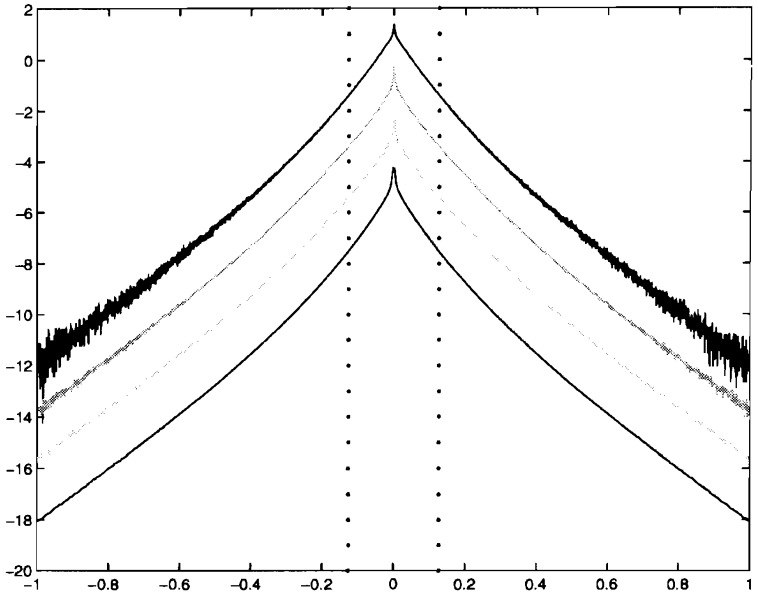

FIG. 4. Histograms of wavelet coefficients on four different scales. The vertical axis is log probability and the four curves have been shifted vertically to separate them. The wavelet scheme used is Simoncelli and Freeman's steerable pyramid [S-F] and the images are from the database of van Hateren [vH].

the logs of the frequencies). These curves are completely on top of each other; so to see this clearly, we have shifted them vertically.

Going beyond linear filters altogether, D. Geman and A. Koloydenko [G-K], have proposed analyzing 3×3 blocks in images by a modified order statistic. They first order the 9 pixel values $a_1 < a_2 < \cdots a_9$ (assumed to be in [0,255]) and then map them to small numbers by mapping a_1 to 0 and a_k either to the same or one more than the image of a_{k-1} depending on whether $a_k - a_{k-1} > 16$ or not. The result is a simplified 3×3 block of small numbers, which most often is either all 0's (background blocks with intensity variation less than 16: about 65%) or all 0's and 1's (blocks showing edges or corners with roughly two grey levels present: about 20%). They look at the following two statistics: a) z defined by the range $[0, z]$ of the simplified block and b) conditional on $z = 1$ and the block being divided into a connected set of 0's and a connected set of 1's, the number y of pixels in the component not containing the center pixel. They calculate the distributions of z and y for their database of 80 images and for downscaled (2×2 block averaged) images. The two histograms appear identical to within experimental fluctuations.

9.2. *Infinite divisibility.* We have performed some experiments to see whether the infinite divisibility axiom holds approximately for real data. In one experiment, 18 scenes were acquired around a house, garden and the nearby streets using an Apple Quick-Take camera. The camera's response was calibrated using an optical gray card. These images were first tested for scale-invariance. The full images were 480 by 640 and seem to be

smoothed by the hardware set-up; hence all measurements were done on block averaged 2×2, 4×4 and 8×8 blow-downs. If the images had scale-invariant statistics, the gradients of these 3 images would have identical histograms. To measure the departure from scale-invariance, we fit the log of the variances of the gradients as above. Expressed in terms of power spectrum fall-off, we found scaling exponents C/ξ^α with α's in the range $[1.94, 2.12]$ for the 14 out of the 18 images showing vegetation and in the range $[2.18, 2.3]$ for 4 images of interior scenes without complex textured objects.

According to the infinite divisibility axiom, we should interpret this as meaning that the 4 interior scenes are samples from the prior with less clutter, while the other 14 are samples from more cluttered priors in the same infinitely divisible family. The four interior scenes can be identified by 3 properties: the variances of the image gradient were smallest; the histograms of the image gradient were most sharply peaked; and they all represented clean clutter-free interior scenes. A second subset of 4 images was chosen from the remainder by the opposite properties: the variances of the image gradients were the largest; their histograms were broadest; and they all represented cluttered garden scenes.

We then formed the composite histogram of nearest neighbor pixel differences for each set. If we have sampled two points in a semi-group of infinitely divisible distributions, we should be able to reconstruct approximately one histogram from the other by the following procedure. Taking one histogram h_1, form its Fourier transform, raise it to a suitable positive real power and take the inverse Fourier transform. If the power is greater than one, this operation smooths the histogram and hence is stable; but when it is less than one, it is unstable. So for powers less than one, we introduce a high frequency cutoff, by multiplying the Fourier transform by a Gaussian (or, equivalently, convolving the original histogram with a Gaussian). The results are shown in Fig. 5. The best fitting powers turned out to be 3.8 and 1/3.8, i.e., the garden scenes were 3.8 times as cluttered as the interior scenes. Although this is a rather weak test for infinite divisibility, it does lend some credence to our Axiom II.

9.3. *Blue sky.* We have no experimental tests for the locality axiom! It is hard to imagine how you could have a sensible model of the real world with infinite divisibility and without locality. The samples from the Levy measure are meant to represent elementary objects or parts of objects and these should be local.

However, the blue sky axiom has one very strong piece of evidence supporting it: this is the presence of sharp peaks in the probability distribution of filter responses at 0. In every case we have examined, for every database and every filter with mean 0, this peak seems to be present. In the cases where the clutter is less and the filter is matched to typical image features (like edges), the peak is much more pronounced. If the clutter is greater or the filter has no geometric significance (e.g., a random set of +1's and -1's of equal number), the peak is less pronounced.

This has a clear interpretation for infinitely divisible distributions. Note that if $\nu = \nu' + \nu''$, then the corresponding distributions satisfy $p = p' * p''$. The basic idea is that the bigger the Levy measure, the smoother the distribution. Thus i) p is C^∞ whenever the Levy measure has a Gaussian component, and on the other extreme ii) p has a delta function component if and only if the Levy measure is finite. A simple

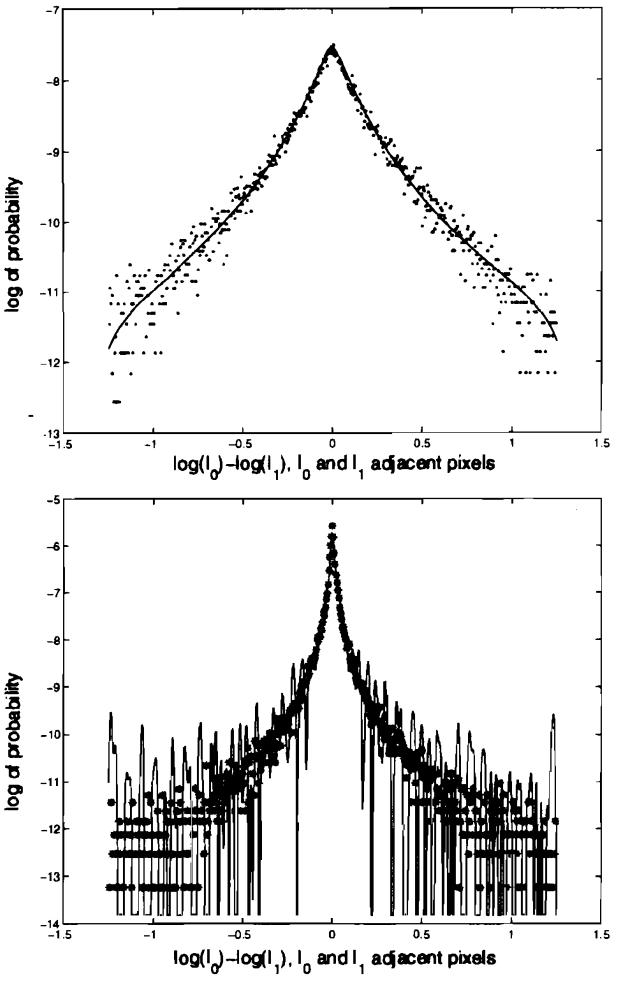

FIG. 5. Top: Adjacent pixel statistics for the 4 garden scenes (represented by dots) versus the 3.8-th convolutional power of the adjacent pixel statistics for the 4 interior scenes (represented by the solid line). Below: Adjacent pixel statistics for the 4 interior scenes (represented by stars) versus the 3.8-th convolutional root of the adjacent pixel statistics for the 4 garden scenes (represented by the solid line).

example to keep in mind is the infinitely divisible gamma family of distributions. Here $\nu = \frac{1}{x}e^{-x}dx, x \geq 0$ and $p_t = c_t x^{t-1}e^{-x}$ for suitable constants c_t. Note that p_t is infinite at 0 if $t < 1$ and gets more and more differentiable at 0 as $t \to \infty$, but never becomes C^∞. A symmetric version of this is given by the family with even Levy measure $\nu = \frac{1}{|x|}e^{-|x|}dx$ and $p_t = c_t|x|^{t-0.5}K_{t-0.5}(|x|)$, where K_t are the modified Bessel functions. Then $p_1 = 0.5 \cdot e^{-|x|}$ and $p_t(0) = \infty$ if $t \leq 0.5$. These two examples are included in the general theory of self-decomposable distributions. These can be defined by requiring $\nu = |x|f(x)dx$ where f, restricted to the positive axis, is decreasing, and restricted to

the negative axis, is increasing. In this case, the p_t are all unimodal with maxima at 0 and $p_t(0) = \infty$ if and only if $f(0) < t$.

10. A problem: Small objects and the smoothness of filter marginals. The main result of this section is that, when images are formed by a scale-invariant process, there will be clouds of tinier and tinier objects everywhere and a kind of central limit theorem will take over. The effect turns out to be that images will be the sum of a Cauchy-like component and a second component independent of this; and the Cauchy-like piece will have a smooth (C^∞) distribution, hence so will the sum. Here is how we make this precise: first assume that the reduced Levy measure ν_u is *not* supported entirely on functions with mean 0. We will return later to remove this restrictive hypothesis. We introduce the following notation: for any test function $f \in \mathcal{D}_d$, the filter response $I(f)$ is also infinitely divisible and its Levy measure is the image of that of the Levy measure $\nu(J)$ of I under the map $J \mapsto J(f)$ (excluding any atom at 0). We call this Levy measure ν_f. Then we have the theorem:

THEOREM. If the reduced Levy measure is not supported in $L^1_d = \{f \in L^1 | \int f = 0\}$, and if f is any test function that is constant in some small open set U, then the Levy measure satisfies:

$$\nu_f \geq \frac{C_1}{|x|^2} dx\Big|_{[0,a]},$$

for some positive constant C_1 and nonzero a.

Recall from §6 that $\nu = \frac{dx\,dy\,dr}{r} \times \nu_u$. Hence ν_f is the image of the measure $\frac{dx\,dy\,dr}{r} \times \nu_u(J)$ under the map:

$$(x, y, r, J) \mapsto \iint J(ru - x, rv - y) f(u, v)\, du\, dv \in \mathbb{R}.$$

Now choose $(u_0, v_0) \in U$ and assume that U contains a disk around (u_0, v_0) of radius r_0. Then we are interested in the translated and scaled versions of J whose support lies entirely in this small disk: this holds if $r > 2/r_0$ and $\| (\frac{x}{r}, \frac{y}{r}) + (u_0, v_0) \| < r_0/2$. Let V denote this set of triples (x, y, r). When this holds, f is a constant on the support of the translate of J and we get

$$\begin{aligned} \iint J(ru - x, rv - y) f(u, v) du\, dv &= r^{-2} \iint J(u', v') f(\frac{u' + x}{r}, \frac{v' + y}{r}) du' dv' \\ &= r^{-d} f(u_0, v_0) \iint J(u', v') du' dv'. \end{aligned}$$

Thus

$$\nu_f \geq \phi_* \left(\frac{dx\, dy\, dr}{r} \times \nu_u(J)\Big|_V \right),$$

where $\phi(x, y, r, J) = r^{-2} f(u_0, v_0) \iint J$. Now the area of the allowed circle in the (x, y) plane is $\pi(r r_0/2)^2$; so we have

$$\nu_f \geq \phi'_* \left(\pi(r_0/2)^2 r\, dr \times \nu_u(J)\Big|_{r \geq 2/r_0} \right),$$

where $\phi'(r, J) = r^{-2} f(u_0, v_0) \iint J$. But the image of the measure $r\,dr|_{r \geq 2/r0}$ under the map $x = r^{-2}$ is just $\frac{2dx}{x^2}\big|_{0<x<(r_0/2)^2}$. So finally ν_f is at least as big as a sum of scaled versions of the measure dx/x^2 on intervals $[0, a]$, and this proves the theorem. □

Recall that the Cauchy distribution is infinitely divisible with Levy measure $dx/|x|^2$. This is why we call the random variables defined by $\frac{C_1}{|x|^2}dx\big|_{[0,a]}$ "Cauchy-like". In fact, it is easy to see that these have C^∞ distribution functions. This follows because they are self-decomposable and $|x|d\nu/dx$ goes to ∞ as $x \to 0$ ([Sa]). Thus we have:

COROLLARY. If the reduced Levy measure is not supported in $L^1_d = \{f \in L^1 | \int f = 0\}$, and if f is any test function that is constant in some small open set U, then the distribution function of $I(f)$ is C^∞.

It is easy to remove the hypothesis of ν_u if we change the selection of f appropriately. In fact, choose the smallest m such that functions J with nonzero m^{th} moments have positive measure. Thus there is a set of J's of ν_u-positive measure for which $\iint x^i y^j J(x, y)\,dx\,dy \neq 0$. We choose the test function f to be equal to $x^i y^j$ on some open set. Then a similar argument goes through and proves that

$$\nu_f \geq \frac{C_1}{|x|^{(4+m)/(2+m)}}dx\Big|_{[0.a]}.$$

This still imples that $I(f)$ has a C^∞ distribution function.

The problem is that, although not contradicted by finite data, this result clearly shows that the models satisfying our axioms do not correspond well to the data, for which these distribution functions look very non-differentiable at 0. We believe this is an important clue about what the true stochastic model for generic images must look like. We believe that the axioms introduced in this paper are a natural model for images, one of which is closer to the truth than Gaussian models but is still short of capturing all the basic qualitative properties.

REFERENCES

[AGM] L. Alvarez, Y. Gousseau, and J. M. Morel, *The size of objects in natural images*, CMLA preprint, Ecole Normale Sup., Cachan, 1999

[Co] D. Cooper, *Maximum likelihood estimation of Markov process blob boundaries in noisy images*, IEEE Trans PAMI **1**, 372-384 (1979)

[De] N. G. Deriugin, *The power spectrum and the correlation function of the television signal*, Telecommunications **1**, 1-12 (1956)

[Do] R. L. Dobrushin, *Generalized random fields and their renormalization group*, in Multicomponent Random Systems, Dekker, NY, pp. 153-198, 1978

[Fu] K. S. Fu, *Syntactic Pattern Recognition*, Prentice-Hall, 1982

[G-V] I. Gelfand and N. Vilenkin, *Generalized Functions IV*, Academic Press, 1964

[G-G] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and Bayesian restoration of images*, IEEE Trans. PAMI **6**, 721-741 (1984)

[GGGD] D. Geman, S. Geman, C. Graffigne, and P. Dong, *Boundary detection by constrained optimization*, IEEE Trans. PAMI **12** (1990)

[G-K] D. Geman and A. Koloydenko, *Invariant Statistics and Coding of Natural Microimages*, CVPR 99, SCTV workshop

[Gr] U. Grenander, *Lectures in Pattern Theory I, II and III*, Springer, 1976-1981

[GCK] U. Grenander, Y. Chow, and D. Keenan, *HANDS, A Pattern-Theoretic Study*, Springer, 1991

[HGYGM] P. Hallinan, G. Gordon, A. Yuille, P. Giblin, and D. Mumford, *Two- and Three-Dimensional Patterns of the Face*, A K Peters, Ltd., Natick, MA, 1999

[H-M1] J. Huang and D. Mumford, *Statistics of Natural Images and Models*, CVPR 99

[H-M2] J. Huang and D. Mumford, *Image Statistics for the British Aerospace Segmented Database*, MPTC preprint, http://www.dam.brown.edu/mptc

[Hi] T. Hida, *Stationary Stochastic Processes*, Princeton Univ. Press, 1970

[Ki] J. Kingman, *Poisson Processes*, Oxford Univ. Press, 1993

[Kr] E. R. Kretzmer, *Statistics of television signals*, Bell System Tech. J. **31**, 751-763 (1952)

[Li] W. Linde, *Infinitely divisible and stable measures on Banach spaces*, Wiley, 1986

[Ma] S. Mallat, *Applied Mathematics meets Signal Processing*, in Proc. Internat. Congress Math., Vol. 1, 1998, Documenta Mathematica

[Me] Y. Meyer, *Wavelets and functions with bounded variations from image processing to pure mathematics*, preprint, Ecole Normale-Cachan, 1999

[M-S] D. Mumford and J. Shah, *Optimal approximation by piecewise smooth functions*, Comm. Pure and Appl. Math. **42**, 577-685 (1989)

[Mu] D. Mumford, *Elastica and Computer Vision*, in Algebraic Geometry and its Applications, ed. C. Bajaj, Springer-Verlag, 1993, pp. 507-518

[Re] M. Reed, *Functional Analysis and Probability*, in Constructive Quantum Field Theory, ed. G. Velo and A. Wightman, Springer Lecture Notes in Physics **25**, 1973

[R-B] D. Ruderman and W. Bialek, *Statistics of Natural Images*, Phys. Rev. Letters **73**, 814-817 (1994)

[Sa] Kenichi Sato, *Levy Processes and Infinitely Divisible Distributions*, Cambridge University Press, 1999

[Sh] C. Shannon, *Prediction and Entropy of Printed English*, Bell System Technical Journal, 50-64, 1951

[S-F] E. Simoncelli and W. Freeman, *The Steerable Pyramid*, IEEE Internat. Conference on Image Proc., 1995

[vH] J. H. van Hateren and A. van der Shaaf, *Independent Component Filters of Natural Images*, Proc. Royal Soc. London B **265**, 359-366 (1998)

[ZWM] S. C. Zhu, Y. Wu, and D. Mumford, *Minimax Entropy Principle and its Application to Texture Modeling*, Neural Computation **9**, 1627-1660 (1997)