Equivalence of Julesz Ensembles and FRAME Models

Ying Nian Wu, Song Chun Zhu, and Xiuwen Liu

Abstract

In the past thirty years, research on textures has been pursued along two different lines. The first line of research, pioneered by Julesz (1962), seeks essential ingredients in terms of *features and statistics* in human texture perception. This leads us to a mathematical definition of textures in terms of *Julesz ensembles*[26]. A Julesz ensemble is a set of images that share the same value of some basic feature statistics. Images in the Julesz ensemble are defined on a large image lattice (a mathematical idealization being $\mathbf{Z}^2$) so that exact constraint on feature statistics makes sense. The second line of research studies Markov random field (MRF) models that characterize texture patterns on finite (or small) image lattice in a statistical way. This leads us to a general class of MRF models called FRAME (Filter, Random field, And Maximum Entropy)[27]. In this article, we bridge the two lines of research by the fundamental principle of *equivalence of ensembles* in statistical mechanics (Gibbs, 1902). We show that 1). As the size of the image lattice goes to infinity, a FRAME model concentrates its probability mass uniformly on a corresponding Julesz ensemble. Therefore, the Julesz ensemble characterizes the global statistical property of the FRAME model; 2). For a large image randomly sampled from a Julesz ensemble, any local patch of the image given its environment follows the conditional distribution specified by a corresponding FRAME model. Therefore, the FRAME model describes the local statistical property of the Julesz ensemble, and is an inevitable texture model on finite (or small) lattice if texture perception is decided by feature statistics. The key to derive these results is the large deviation estimate of the volume of (or the number of images in) the Julesz ensemble, which we call the entropy function. Studying the equivalence of ensembles provides deep insights into questions such as the origin of MRF models, typical images of statistical models, and error rates in various texture related vision tasks[25]. The second thrust of this paper is to study texture distance based on the texture models of both small and large lattice systems. We attempt to explain the asymmetry phenomenon observed in texture "pop-out" experiments by the asymmetry of Kullback-Leibler divergence. Our results generalize the traditional signal detection theory[8] for distance measures from iid cases to random fields. Our theories are verified by two groups of computer simulation experiments.

Y. N. Wu is with the Department of Statistics, University of California, Los Angeles, CA 90095. Song Chun Zhu and Xiuwen Liu are with the Department of Computer and Information Science, The Ohio State University, Columbus, OH 43210.

1 Motivation and introduction

Recently there is a resurgent interest in texture[1] research inspired by the artful work of (Heeger and Bergen,1995)[10]. A unified texture theory is emerging after nearly four decades of intensive research in computer vision and psychophysics following the philosophical theme erected in (Julesz, 1962)[11]. For the past few years, the authors, together with Mumford, have studied the texture phenomena from a mathematical perspective (Zhu, Wu, and Mumford 1997, Zhu, Liu and Wu, 1999)[27, 26]. Our general goal is to pursue a mathematically sound theory, which can provide self-consistent answers to the following three fundamental questions.

Question I: What is a mathematical definition of texture?

This question has been considered extremely difficult because of the overwhelming diversity of texture patterns and their underlying rendering mechanisms in nature. It turns out that the answer to this question becomes possible once we possess the right perspective and the right tools. Before we approach this question, let's take a look at color theory.

First, optics defines color as an electro-magnetic wave. A visible color is uniquely identified by its wave length $\lambda \in [400, 700]nm$. Second, trichromacy theory states that any visible color is a linear combination of three basic colors: red, green, and blue. One may ask: if we are lucky enough to have a texture theory that is as clean as color theory, then what is the quantity that defines textures uniquely? And what are the basic elements that can generate textures in combination?

Texture is different from color in that it is a spatial phenomenon. A texture definition cannot be based on a single pixel, and one has to deal with spatial statistics averaged over the image. Thus a major theme of texture research is to seek the essential ingredients in terms of *features and statistics*. The objective is to find feature statistics that are the bases for human texture perception. Typical choices of feature statistics include Julesz's 2-gon statistics[11], co-occurrence matrices[6], statistics of texton attributes[24], Fourier transforms[16], rectified functions[17], histograms of Gabor filter responses[10, 27, 4], and correlations of filter responses[21]. To verify the sufficiency of these texture statistics, this research theme also searches for mathematical tools and algorithms that can synthesize texture images that have prescribed statistics. One of our early paper[26] provides a detailed account for the achievements along this research line.

To obtain a quantity that can uniquely identify textures, one needs to define textures on

[1] Throughout the paper, our discussion is focused on homogeneous texture patterns on a 2D plane, and we do not discuss texture deformation on 3D surface.

an infinite lattice $\mathbf{Z}^2$ as a mathematical idealization, where effects of boundary conditions[2] and statistical fluctuations vanish. Therefore the entire image space is partitioned into equivalent classes, within each class all images have identical statistics. We call each equivalent class a *Julesz ensemble*. Like wave length λ for color, a value of feature statistics defines a texture type on $\mathbf{Z}^2$. To study the statistical properties of Julesz ensembles, we attach to each Julesz ensemble a uniform counting measure, or a uniform probability distribution.

When we move from $\mathbf{Z}^2$ to finite lattice, the texture statistics of different Julesz ensembles (or equivalence classes) start to overlap due to statistical fluctuations. As the lattice gets smaller, e.g., in an extreme case the lattice consists of only one pixel, it becomes harder to classify a texture, and boundary condition assumes a more important role. Therefore, on finite lattice, texture is best represented by a *conditional probability distribution* rather than an equivalence class. Very often one calls the conditional probability distribution a texture *model*. Now we naturally come to the second question below.

Question II: What is a legitimate texture model on a finite lattice that is consistent with the texture definition on $\mathbf{Z}^2$*?*

The second major theme of texture research is to pursue statistical models to characterize textures on finite lattice, driven by computer vision tasks such as texture segmentation and classification. Among the studied texture models, Markov random field models, or equivalently the Gibbs distributions, are the most popular and elegant ones. Influential work includes (Besag 1974)[2] and (Cross and Jain 1983)[3]. Recently, a paper by (Zhu, Wu, and Mumford, 1997) has shown that the MRF models can be unified under a minimax entropy learning principle[27]. Given the statistics used in the texture definition, a FRAME (Filter, Random field, And Maximum Entropy) model can be derived as a Gibbs distribution or an exponential family model[27] whose parameters are adjusted in such a way that the expectation of the texture statistics under the Gibbs distribution equal to a prescribed value. The advantage of the FRAME model is that the conditional distribution of any local patch of the image given its environment can be easily specified because of the Markov property. A detailed account for the FRAME model and the minimax entropy principle in selecting statistics is referred to an early paper (Zhu, Wu, and Mumford, 1997)[27].

The theories and methods developed in the two research themes are very different from each other, and thus a crucial question remains unanswered: Are the FRAME models consistent with the texture definition in terms of Julesz ensembles?

[2] We shall discuss phase transition in a later section.

In this article, we unify the two research themes by showing the equivalence between the Julesz ensembles and the FRAME models, using the fundamental principle of equivalence of ensembles in statistical mechanics. The equivalence reveals two interesting facts in texture research.

- For a large image randomly sampled from a Julesz ensemble, any local patch of the image given its environment follows the conditional distribution specified by a corresponding FRAME model. Therefore the FRAME model describes the local statistical property of the Julesz ensemble, and is an inevitable texture model on finite (or small) lattice if texture perception is decided by feature statistics.
- As the image lattice goes to $\mathbf{Z}^2$, a FRAME model concentrates all its probability mass uniformly over a Julesz ensemble (in the absence of phase transition).

The key to the equivalence of ensembles is the large deviation estimate of the volume of the Julesz ensemble (or the number of images in the Julesz ensemble), and we call this estimate the entropy function. The ensemble equivalence provides insights into several basic questions in texture study, such as the origin of the MRF models, typical images of a statistical model, and texture distance measures.

Question III: What is a legitimate texture distance measure?

In computer vision, the distance between two texture images are often defined based on the difference of their feature statistics (see [1, 23] and references therein). These measures are practically very effective and useful, but lose their elegance by ignoring important factors such as the dependence between elements in the feature statistics, boundary conditions, and the effects of lattice sizes.

The second thrust of this paper is to study a legitimate texture distance based on the texture models of both small and large lattice systems. We generalize the traditional distance measures, such as Kullback-Leibler divergence, to random fields. In particular we attempt to explain the asymmetry property observed in texture "pop-out" experiments by the calculation of Kullback-Leibler divergence.

In this paper, our theories on ensemble equivalence and texture distance are verified by two groups of experiments. The first experiment simulates two Monte Carlo Markov chains, one sampling the Julesz ensemble, and the other sampling the corresponding FRAME model. Both chains synthesize typical texture images that have similar visual appearances. The second experiment computes the distance of texture pairs, and demonstrates the asymmetry in texture distance.

The paper is organized as follows. Section (2) explains some background concepts, such as type, ensemble, entropy function, and equivalence between Julesz ensembles and

FRAME models using a simple iid example. Section (3) briefly reviews the Julesz ensembles and FRAME models. Then section (4) proves the equivalence between Julesz ensembles and FRAME models. Some experiments are shown in section (5) to demonstrate the equivalence. Section (6) reviews some important mathematical results and briefly discusses phase transition. Then we study texture distance in section (7) with experiments in section (7.2). Finally we discuss some related issues in section (8).

Throughout this article, we concentrate on understanding basic ideas and important insights while taking a relaxed attitude towards mathematical rigor.

2 Background I: the basic concepts

In this section, we introduce the basic concepts, such as type, ensemble, entropy function, typical images, and equivalence of ensembles, using a simple image model where the pixel intensities are independently and identically distributed (iid).

2.1 Type, ensemble, and entropy function

Let $\mathbf{I}$ be an image defined on a finite lattice $\Lambda \subset \mathbf{Z}^2$, and the intensity at pixel $v \in \Lambda$ is denoted by $\mathbf{I}(v) \in \mathcal{G} = \{1, 2, ..., g\}$. Thus $\Omega_\Lambda = \mathcal{G}^{|\Lambda|}$ is the space of images on Λ, with $|\Lambda|$ being the number of pixels in Λ.

1). The FRAME model for iid images. We consider a simple image model where pixel intensities are independently and identically distributed according to a probability distribution $p = (p_1, ..., p_g), \sum_i p_i = 1$. The distribution of $\mathbf{I}$ can be written as a FRAME model

$$p(\mathbf{I}; \beta) = \prod_{v \in \Lambda} p_{\mathbf{I}(v)} = \prod_{i=1}^{g} p_i^{H_i(\mathbf{I})} = \exp\{< \log p, H(\mathbf{I}) >\} = \exp\{- < \beta, H(\mathbf{I}) >\}, \quad (1)$$

where $H(\mathbf{I}) = (H_1(\mathbf{I}), ..., H_g(\mathbf{I}))$ is the unnormalized intensity histogram of $\mathbf{I}$, i.e., H_i is the number of pixels whose intensities equal to i. $\beta = -(\log p_1, ..., \log p_g)$ is the parameter of $p(\mathbf{I}; \beta)$ – a special case of the FRAME model.

2). Type. Let $h(\mathbf{I}) = (h_1(\mathbf{I}), ..., h_g(\mathbf{I})) = (H_1(\mathbf{I})/|\Lambda|, ..., H_g(\mathbf{I})/|\Lambda|) = H(\mathbf{I})/|\Lambda|$ be the normalized intensity histogram. We call $h(\mathbf{I})$ the *type* of image $\mathbf{I}$, and it is the sufficient statistics for model $p(\mathbf{I}; \beta)$. That is, images of the same type receive the same probabilities from $p(\mathbf{I}; \beta)$.

3). Equivalent class. Let $\Omega_\Lambda(h)$ be the set of images defined on Λ with $h(\mathbf{I}) = h$ [3], i.e., $\Omega_\Lambda(h) = \{\mathbf{I} : h(\mathbf{I}) = h\}$. Then the image space is partitioned into equivalence classes

$$\Omega_\Lambda = \cup_h \Omega_\Lambda(h).$$

As shown in figure 1, each equivalence class $\Omega_\Lambda(h)$ is mapped into one type h on a simplex – a plane defined by $h_1 + \cdots + h_g = 1$ and $h_i \geq 0, \forall i$ in a g-dimensional space.

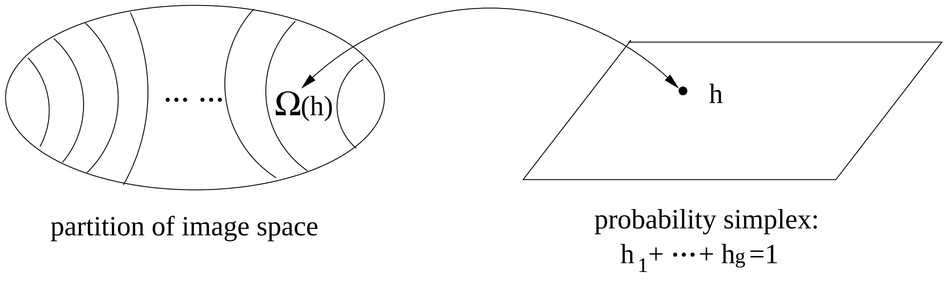


Figure 1: The partition of image space into equivalence classes, where each class corresponds to a type h on the probability simplex.

4). The Julesz ensemble for iid images. The hard constraint in defining the equivalence class $\Omega_\Lambda(h)$ makes sense only in the limit as $\Lambda \to \mathbf{Z}^2$, where statistical fluctuations vanish. Therefore, we may attempt to define the Julesz ensemble as the limit of $\Omega_\Lambda(h)$ as $\Lambda \to \mathbf{Z}^2$, or even more directly, as the set of images $\mathbf{I}$ defined on $\mathbf{Z}^2$ with $h(\mathbf{I}) = h$.

Unfortunately, the above"definitions" are not mathematically well-defined. Instead, we need to define the Julesz ensemble in a slightly indirect way. First, we associate each equivalence class $\Omega_\Lambda(h)$ a probability distribution $q(\mathbf{I}; h)$, which is uniform over $\Omega_\Lambda(h)$ and vanishes outside. Then, the Julesz ensemble of type h is defined to be the limit of $q(\mathbf{I}; h)$ as $\Lambda \to \mathbf{Z}^2$.

For finite Λ, the equivalence class $\Omega_\Lambda(h)$ may be empty because $|\Lambda|h$ may not be a vector of integers. Thus, to be more rigorous, we should replace h by a small set $\mathcal{H}$ around h, and let $\mathcal{H}$ goes to h as $\Lambda \to \mathbf{Z}^2$. For simplicity, however, we shall neglect this minor complication and simply treat $|\Lambda|h$ as a vector of integers.

The uniform distribution $q(\mathbf{I}; h)$ only serves as a counting measure of the equivalence class $\Omega_\Lambda(h)$, i.e., all the images in $\Omega_\Lambda(h)$ are counted equally. Therefore, any probability statement under the uniform distribution $q(\mathbf{I}; h)$ is equivalent to a statistical or frequency statement of images in $\Omega_\Lambda(h)$. For example, the probability that image $\mathbf{I}$ has a certain

[3] We hope that the notation $h(\mathbf{I}) = h$ will not confuse the reader. The h on the left is a function of $\mathbf{I}$ for extracting statistics, while the h on the right is a specific value of the statistics.

property under $q(\mathbf{I}; h)$ can be interpreted as the frequency or the proportion of images in $\Omega_\Lambda(h)$ that have this property. The limit of $q(\mathbf{I}; h)$ thus essentially defines a counting measure of the set of infinitely large images (defined on $\mathbf{Z}^2$) with histogram h. With a little abuse of language, we sometimes also call the equivalence class $\Omega_\Lambda(h)$ defined on a large lattice Λ a Julesz ensemble, and it is always helpful to imagine a Julesz ensemble as such an equivalence class if the reader finds the limit of probability measures too abstract.

5). Entropy function. We are interested in computing the volume of the Julesz ensemble $\Omega_\Lambda(h)$, i.e., the number of images in $\Omega_\Lambda(h)$. We denote this volume by $|\Omega_\Lambda(h)|$. Clearly

$$|\Omega_\Lambda(h)| = \frac{|\Lambda|!}{\prod_{i=1}^{g}(h_i|\Lambda|)!}.$$

Using the Stirling formula, it can be easily shown that

$$\begin{aligned}\lim_{\Lambda\to\mathbf{Z}^2}\frac{1}{|\Lambda|}\log|\Omega_\Lambda(h)| &= \lim_{\Lambda\to\mathbf{Z}^2}\frac{1}{|\Lambda|}\log\frac{|\Lambda|!}{\prod_{i=1}^{g}(h_i|\Lambda|)!}\\ &= -\sum_{i=1}^{g}h_i\log h_i = \mathrm{entropy}(h).\end{aligned}$$

Thus for large enough lattice, the volume of $\Omega_\Lambda(h)$ is said to be in the order of entropy(h), i.e.,

$$|\Omega_\Lambda(h)| \sim e^{|\Lambda|\mathrm{entropy}(h)}.$$

For notational simplicity, we denote the entropy function by $s(h) = \mathrm{entropy}(h)$.

6). Probability rate function. Now we are ready to compute the total probability mass that $p(\mathbf{I}; \beta)$ assigns to an equivalence class $\Omega_\Lambda(h)$. We denote this probability by $p(\Omega_\Lambda(h); \beta)$. Because images in $\Omega_\Lambda(h)$ all receive equal probabilities, it can be shown that

$$\begin{aligned}\lim_{\Lambda\to\mathbf{Z}^2}\frac{1}{|\Lambda|}\log p(\Omega_\Lambda(h);\beta) &= \lim_{\Lambda\to\mathbf{Z}^2}\frac{1}{|\Lambda|}\log\{|\Omega_\Lambda(h)|\prod_{i=1}^{g}p_i^{|\Lambda|h_i}\}\\ &= -\sum_{i=1}^{g}h_i\log\frac{h_i}{p_i} = -\mathrm{KL}(h||p),\end{aligned}$$

where $\mathrm{KL}(h||p)$ denotes the Kullback-Leibler distance from h to p. $\mathrm{KL}(h||p) \geq 0$ for all h and p, with equality holds when $h = p$.

Thus, on a large enough lattice, the total probability mass of an equivalence class $\Omega_\Lambda(h)$ is said to be in the order of $-\mathrm{KL}(h||p)$, i.e.,

$$p(\Omega_\Lambda(h);\beta) \sim e^{-|\Lambda|\mathrm{KL}(h||p)}. \tag{2}$$

The $-\mathrm{KL}(h||p)$ is the probability rate function of h under model p, and is denoted by $s_\beta(h) = -\mathrm{KL}(h||p)$.

Having introduced the basic concepts, we now explain the basic ideas of ensemble equivalence in the next two subsections by going both directions from one to the other.

2.2 From a FRAME model to a Julesz ensemble on infinite lattice

A simple fact will be repeatedly used in this paper. To see this fact, let's consider the following example. Suppose we have two terms, one is e^{5n}, and the other is e^{3n}. Consider their sum $e^{5n} + e^{3n}$. As $n \to \infty$, the sum $e^{5n} + e^{3n}$ is dominated by e^{5n}, and the order of this sum is still 5, i.e., $\lim_{n\to\infty} \frac{1}{n} \log(e^{5n} + e^{3n}) = 5$. This means that for the sum of many terms, the term with the largest exponential order dominates the sum, and the order of the sum is the largest order among the individual terms.

Now let's study the limit of the FRAME model $p(\mathbf{I}; \beta)$ as $\Lambda \to \mathbf{Z}^2$. According to (2), the probability that $p(\mathbf{I}; \beta)$ assigns to the equivalence class $\Omega_\Lambda(h)$ is of the exponential order $s_\beta(h) = -\mathrm{KL}(h||p)$, which, as a function of type h, achieves the maximum 0 at $h_* = p$. Thus, the equivalence class $\Omega_\Lambda(h_*)$ eventually absorbs all the probability mass of $p(\mathbf{I}; \beta)$ as $\Lambda \to \mathbf{Z}^2$, and for other $h \neq p$, the probability that $\Omega_\Lambda(h)$ receives goes to 0 at an exponential rate $s_\beta(h) = -\mathrm{KL}(h||p) < 0$. Because $p(\mathbf{I}; \beta)$ assigns equal probabilities to images in the same equivalence class, $p(\mathbf{I}; \beta)$ will eventually concentrate its probability mass uniformly on $\Omega_\Lambda(h_*)$, and therefore become a Julesz ensemble of type $h_* = p$.

For statistics $h(\mathbf{I})$, the h_* can be called the *typical value* of $h(\mathbf{I})$ under the model $p(\mathbf{I}; \beta)$ because images with h_* absorb all the probability mass of $p(\mathbf{I}; \beta)$ on large lattice. In other words, if we sample from $p(\mathbf{I}; \beta)$ on large lattice, we will almost always get an image of type h_*. Therefore, as far as statistics $h(\mathbf{I})$ is concerned, images in $\Omega_\Lambda(h_*)$ can be called *typical images*.

It is important to distinguish between typical images and most likely images. To see this point, let's consider the following example. Suppose among $p_1, ..., p_g$, $p_m < 1$ is the largest probability. Consider one extreme type h, with $h_m = 1$, and $h_i = 0, \forall i \neq m$. Then the image in this $\Omega_\Lambda(h)$ is the most likely image under model $p(\mathbf{I}; \beta)$, i.e., it receives the highest probability. However, $\Omega_\Lambda(h)$ has only one constant image, and the probability that $p(\mathbf{I}; \beta)$ assigns to this $\Omega_\Lambda(h)$ is essentially zero for large lattice. In other words, when sampling from the model $p(\mathbf{I}; \beta)$ on large lattice, we will almost never get the most likely images, instead, we will almost always get the typical images (or most common images). Therefore, it is the typical images that a statistical model is intended to characterize.

2.3 From a Julesz ensemble to a FRAME model on finite lattice

In this section, we tight up the notation a little bit. We use $\mathbf{I}_\Lambda$ to denote the image defined on lattice Λ, and we use $\mathbf{I}_{\Lambda_0}$ to denote the image patch defined on $\Lambda_0 \subset \Lambda$. For a fixed type h of feature statistics, consider the uniform distribution $q(\mathbf{I}; h)$ on $\Omega_\Lambda(h)$. Under $q(\mathbf{I}; h)$, the distribution of $\mathbf{I}_{\Lambda_0}$, denoted by $q(\mathbf{I}_{\Lambda_0}; h)$, is well defined.[4] Notice that the rest of the image $\mathbf{I}_{\Lambda/\Lambda_0}$ influences $\mathbf{I}_{\Lambda_0}$ through a global constraint $h(\mathbf{I}_\Lambda) = h$. We shall show that if we fix Λ_0 and let $\Lambda \to \mathbf{Z}^2$, then $q(\mathbf{I}_{\Lambda_0}; h)$ goes to the FRAME model (see equation (1)) with $p = h$.

The number of images in $\Omega_\Lambda(h)$ is

$$|\Omega_\Lambda(h)| = \frac{|\Lambda|!}{\prod_{i=1}^g (h_i|\Lambda|)!}.$$

We fix $\mathbf{I}_{\Lambda_0}$ and calculate the number of images in $\Omega_\Lambda(h)$ whose image value (i.e., intensities) on Λ_0 is $\mathbf{I}_{\Lambda_0}$. Clearly, for every such image, its image value on the rest of the lattice Λ/Λ_0, i.e., $\mathbf{I}_{\Lambda/\Lambda_0}$, must satisfy

$$H(\mathbf{I}_{\Lambda/\Lambda_0}) = h|\Lambda| - H(\mathbf{I}_{\Lambda_0}),$$

where $H(\mathbf{I}_{\Lambda_0}) = |\Lambda_0| h(\mathbf{I}_{\Lambda_0})$ is the unnormalized histogram of $\mathbf{I}_{\Lambda_0}$, and $H(\mathbf{I}_{\Lambda/\Lambda_0})$ is the unnormalized histogram of $\mathbf{I}_{\Lambda/\Lambda_0}$. Therefore

$$\mathbf{I}_{\Lambda/\Lambda_0} \in \Omega_{\Lambda/\Lambda_0}\left(\frac{h|\Lambda| - H(\mathbf{I}_{\Lambda_0})}{|\Lambda/\Lambda_0|}\right).$$

So the number of such images is $|\Omega_{\Lambda/\Lambda_0}((h|\Lambda| - H(\mathbf{I}_{\Lambda_0}))/|\Lambda/\Lambda_0|)|$. Thus,

$$\begin{aligned}
q(\mathbf{I}_{\Lambda_0}; h) &= \frac{|\Omega_{\Lambda/\Lambda_0}(\frac{h|\Lambda| - H(\mathbf{I}_{\Lambda_0})}{|\Lambda/\Lambda_0|})|}{|\Omega_\Lambda(h)|} \\
&= \frac{(|\Lambda| - |\Lambda_0|)!/\prod_{i=1}^g (h_i|\Lambda| - H_i(\mathbf{I}_{\Lambda_0}))!}{|\Lambda|!/\prod_{i=1}^g (h_i|\Lambda|)!} \\
&= \frac{\prod_{i=1}^g (h_i|\Lambda|)(h_i|\Lambda| - 1)...(h_i|\Lambda| - H_i(\mathbf{I}_{\Lambda_0}) + 1)}{|\Lambda|(|\Lambda| - 1)...(|\Lambda| - |\Lambda_0| + 1)} \\
&= \frac{\prod_{i=1}^g h_i(h_i - 1/|\Lambda|)...(h_i - (H_i(\mathbf{I}_{\Lambda_0}) - 1)/|\Lambda|)}{(1 - 1/|\Lambda|)...(1 - (|\Lambda_0| - 1)/|\Lambda|)} \\
&\to \prod_{i=1}^g h_i^{H_i(\mathbf{I}_{\Lambda_0})} \text{ as } |\Lambda| \to \infty.
\end{aligned}$$

Therefore, the distribution of $\mathbf{I}_{\Lambda_0}$ is the FRAME model (see equation (1)) with $p = h$ under the Julesz ensemble of type h.

[4] In the iid case, $q(\mathbf{I}_{\Lambda_0}; h)$ is both the marginal distribution and the conditional distribution of $q(\mathbf{I}; h)$, while in random fields, we only consider the conditional distribution.

The above calculation can be interpreted in a non-probabilistic way, i.e., $q(\mathbf{I}_{\Lambda_0}; h)$ is the frequency or the proportion of images in $\Omega_\Lambda(h)$ (on large Λ) whose patches on Λ_0 are $\mathbf{I}_{\Lambda_0}$. In other words, if we look at all the images in the Julesz ensemble through Λ_0, then we will find a collection of image patches on Λ_0, and the distribution of this collection is described by the FRAME model. Under the hard constraint on $h(\mathbf{I}_\Lambda)$, $h(\mathbf{I}_{\Lambda_0})$ can still take any possible values.

3 Background II: Julesz ensembles and FRAME models for textures

For this paper to be self-contained, we briefly describe the Julesz ensembles and FRAME models for textures.

3.1 Julesz ensembles – a mathematical definition of textures

To study real world textures, one needs to characterize the dependency between pixels by extracting spatial features and calculating some statistics averaged over the image. One main theme of texture research is to seek the essential ingredients in terms of features and statistics $\mathbf{h}(\mathbf{I})$, which are the bases for human texture perception. From now on, we use the bold font $\mathbf{h}$ to denote statistics of image features. Recently, the search for $\mathbf{h}$ has converged to marginal histograms of Gabor filter responses. We believe that some bins of joint statistics may also be important as long as we can keep the model complexity under check.

Given K Gabor filters $\{F^{(1)}, ..., F^{(K)}\}$ as feature detectors, we convolve the filters with the image $\mathbf{I}$ to obtain the subband filtered images $\{\mathbf{I}^{(1)}, ..., \mathbf{I}^{(K)}\}$, where $\mathbf{I}^{(k)} = F^{(k)} * \mathbf{I}$. Let $h^{(k)}$ be the normalized intensity histogram of $\mathbf{I}^{(k)}$, then the feature statistics $\mathbf{h}$ collects the normalized histograms of these K subband images,

$$\mathbf{h}(\mathbf{I}) = (h^{(1)}(\mathbf{I}), ..., h^{(K)}(\mathbf{I})).$$

We use $\mathbf{H}(\mathbf{I}) = (H^{(1)}(\mathbf{I}), ..., H^{(K)}(\mathbf{I}))$ to denote the unnormalized histograms. We assume that boundary conditions are properly handled (e.g., periodic boundary condition). It should be noted that the conclusions of this paper hold as long as $\mathbf{h}(\mathbf{I})$ can be expressed as spatial averages of local image features. The marginal histograms of Gabor filter responses are only special cases.

Given statistics $\mathbf{h}(\mathbf{I})$, one can partition the image space Ω_Λ into equivalence classes $\Omega_\Lambda(\mathbf{h}) = \{\mathbf{I} : \mathbf{h}(\mathbf{I}) = \mathbf{h}\}$, as we did for the iid case. For finite Λ, the exact constraint

$\mathbf{h}(\mathbf{I}) = \mathbf{h}$ may not be satisfied, so we relax this constraint, and replace $\Omega_\Lambda(\mathbf{h})$ by

$$\Omega_\Lambda(\mathcal{H}) = \{\mathbf{I} : \mathbf{h}(\mathbf{I}) \in \mathcal{H}\}$$

with $\mathcal{H}$ being a small set around $\mathbf{h}$. Then we can define the uniform counting measure or the uniform probability distribution on $\Omega_\Lambda(\mathcal{H})$ as

$$q(\mathbf{I}; \mathcal{H}) = \begin{cases} 1/|\Omega_\Lambda(\mathcal{H})|, & \text{if } \mathbf{I} \in \Omega_\Lambda(\mathcal{H}), \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $|\Omega_\Lambda(\mathcal{H})|$ is the volume of or the number of images in $\Omega_\Lambda(\mathcal{H})$. Now we can define the Julesz ensemble as follows.

Definition *Given a set of feature statistics* $\mathbf{h}(\mathbf{I}) = (h^{(1)}(\mathbf{I}), ..., h^{(K)}(\mathbf{I}))$, *a Julesz ensemble of type* $\mathbf{h}$ *is a limit of* $q(\mathbf{I}; \mathcal{H})$ *as* $\Lambda \to \mathbf{Z}^2$ *and* $\mathcal{H} \to \mathbf{h}$ *with some boundary condition.*[5]

As in the iid example, the Julesz ensemble is defined mathematically as the limit of a uniform counting measure. It is always helpful to imagine the Julesz ensemble of type $\mathbf{h}$ as the image set $\Omega_\Lambda(\mathbf{h})$ on a large Λ. Also, in the later calculation, we shall often ignore the minor complication that the constraint $\mathbf{h}(\mathbf{I}) = \mathbf{h}$ may not be exactly satisfied, and shall simply take $\mathcal{H}$ to be $\mathbf{h}$.

With Julesz ensembles, we are ready to give a mathematical definition for textures.

Definition *A texture pattern is a Julesz ensemble defined by a type* $\mathbf{h}$ *of the feature statistics* $\mathbf{h}(\mathbf{I})$.

Just as the wavelength λ identifies a color, the type $\mathbf{h}$ defines a texture. One of our early paper[26] provides a detailed account for the definition of Julesz ensembles and Markov chain Monte Carlo algorithms for exploring the Julesz ensembles.

3.2 The FRAME models

While a texture is uniquely identifiable by type $\mathbf{h}$ on $\mathbf{Z}^2$, on finite lattice the texture statistics of different Julesz ensembles overlap due to statistical fluctuations, and boundary condition plays an important role. Therefore, on finite lattice, texture is best represented by a conditional probability distribution. Very often one calls the conditional probability distribution a texture model.

Among the studied texture models, Markov random field models, or equivalently the Gibbs distributions, are the most popular and elegant ones. Recently, (Zhu, Wu, and

[5] We assume $\Lambda \to \mathbf{Z}^2$ in the sense of van Hove, i.e., the ratio between the boundary and the size of Λ goes to 0.

Mumford, 1997) [27] proposed a class of MRF models called FRAME (Filter, Random field, And Maximum Entropy). The basic idea is as follows.

Given statistics $\mathbf{h}(\mathbf{I})$ used in the texture definition, we want a model $p(\mathbf{I})$ so that it has the expected statistics $\mathbf{h}$, i.e.,

$$\mathrm{E}_p[\mathbf{h}(\mathbf{I})] = \mathbf{h}.$$

This is a "soft" constraint in comparison with the Julesz ensemble because it only requires that the statistics are matched on average.

Then a maximum entropy distribution, called FRAME, is selected among all distributions that satisfy the constraint. The distribution assumes the following exponential form

$$p(\mathbf{I};\boldsymbol{\beta}) = \frac{1}{Z_\Lambda(\boldsymbol{\beta})} \exp\{- < \boldsymbol{\beta}, \mathbf{H}(\mathbf{I}) >\}, \tag{4}$$

where $\boldsymbol{\beta}$ is the parameter of the model and $Z_\Lambda(\boldsymbol{\beta})$ is the normalizing constant. The parameter $\boldsymbol{\beta}$ is solved from the constraint $\mathrm{E}_{p(\mathbf{I};\beta)}[\mathbf{h}(\mathbf{I})] = \mathbf{h}$. $p(\mathbf{I};\boldsymbol{\beta})$ unifies existing MRF texture models, which are different only in their definitions of feature statistics $\mathbf{h}(\mathbf{I})$. A detailed account of the FRAME models and the minimax entropy principle in selecting statistics $\mathbf{h}(\mathbf{I})$ is referred to an early paper (Zhu, Wu, and Mumford, 1997)[27].

Unlike the Julesz ensembles, the FRAME models assign probabilities to all the images defined on Λ. Although the FRAME models are less straightforward than the Julesz ensembles, they are much more analytically tractable due to the Markov property. That is, for any $\Lambda_0 \subset \Lambda$, the conditional distribution of $\mathbf{I}_{\Lambda_0}$ given the rest of the image $\mathbf{I}_{\Lambda/\Lambda_0}$ only depends on the intensities of the neighboring pixels $\mathbf{I}_{\partial\Lambda_0}$, where $\partial\Lambda_0$ collects all the pixels around Λ_0 that can be covered by the same filters as the pixels in Λ_0. The conditional probability is

$$p(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}) = p(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}) = \frac{1}{Z_{\Lambda_0}(\boldsymbol{\beta})} \exp\{- < \boldsymbol{\beta}, \mathbf{H}(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0}) >\},$$

where $\mathbf{H}(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0})$ collects the unnormalized histograms by filtering inside $\Lambda_0 \cup \partial\Lambda_0$. Note that this conditional distribution is still of the FRAME form with parameter $\boldsymbol{\beta}$, indicating that the FRAME model gives a consistent specification of all the conditional distributions of image patches.

4 Equivalence between Julesz ensembles and FRAME models

In this section, we unify the two research themes by showing the equivalence between the Julesz ensembles and the FRAME models, using the fundamental principle of equivalence of ensembles in statistical mechanics.

4.1 Physics background

In statistical mechanics, there are two major models for physical systems with a large number of degrees of freedom. One is called the "micro-canonical" ensemble, which is the ensemble of all the possible states of a physical system with a fixed energy. The micro-canonical ensemble is used to model a physical system in thermal isolation, i.e., it does not exchange heat with the environment and therefore has a constant energy. When such a system reaches equilibrium, its state is supposed to follow a uniform distribution over the micro-canonical ensemble. The other important model is the Gibbs distribution, or the "canonical ensemble". It is used to model a physical system in thermal equilibrium with an environment of a fixed temperature. As to the equivalence between micro-canonical and canonical ensembles, Gibbs (1902) argued that: 1) If a large physical system is micro-canonically distributed, i.e., following a uniform distribution over the states with a constant energy, then any small part of it follows a Gibbs distribution. 2) A Gibbs distribution for a large physical system is essentially micro-canonically distributed. Gibbs (1902) also proposed other arguments to justify the Gibbs distribution. If we replace the energy of the physical system by the feature statistics of the texture image, then we can identify the micro-canonical ensembles with the Julesz ensembles, and the Gibbs distributions or the canonical ensembles with the FRAME models. So the equivalence between the Julesz ensembles and the FRAME models follow directly from the principle of equivalence of ensembles in statistical mechanics.

Since Gibbs' time, many proofs have been given to the equivalence of ensembles. Recently, Lewis, Pfister, and Sullivan (1995) gave a rigorous proof of the equivalence for lattice systems under very general conditions. However, modern rigorous treatments with large deviation technicalities tend to be too complicated and unapproachable for computer scientists. In this article, therefore, we concentrate on understanding basic ideas and important insights in the context of texture modeling while taking a relaxed attitude towards mathematical rigor. Readers interested in rigorous formalisms are referred to Lewis, Pfister, and Sullivan (1995) and the references therein.

4.2 From the Julesz ensemble to the FRAME model

In this subsection, we derive the local Markov property of the Julesz ensemble, which is globally defined by type $\mathbf{h}$. This derivation is adapted from traditional argument in statistical physics. It is not as rigorous as modern treatments, but is much more revealing.

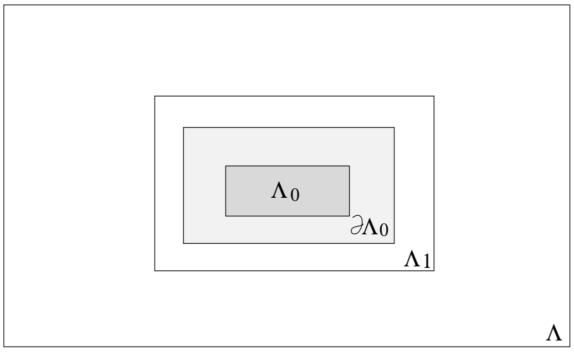


Figure 2: The lattices system: Λ_0 is the local patch, and $\partial\Lambda_0$ is the MRF boundary of Λ_0. Both are inside a fixed lattice Λ_1, and the image lattice Λ goes to $\mathbf{Z}^2$.

Suppose the feature statistics is $\mathbf{h}(\mathbf{I})$ where $\mathbf{I}$ is defined on Λ. For a fixed value of feature statistics $\mathbf{h}$, consider the image set $\Omega_\Lambda(\mathbf{h}) = \{\mathbf{I} : \mathbf{h}(\mathbf{I}) = \mathbf{h}\}$ and the associate uniform distribution $q(\mathbf{I}; \mathbf{h})$. First, we fix $\Lambda_1 \subset \Lambda$, and then fix $\Lambda_0 \subset \Lambda_1$, as shown in figure 2. We are interested in the conditional distribution of the local patch $\mathbf{I}_{\Lambda_0}$ given its local environment $\mathbf{I}_{\Lambda_1/\Lambda_0}$ under the model $q(\mathbf{I}; \mathbf{h})$ as $\Lambda \to \mathbf{Z}^2$. We assume that Λ_0 is sufficiently smaller than Λ_1 so that the neighborhood of Λ_0, $\partial\Lambda_0$, is contained in Λ_1.

Let $\mathbf{H}_0 = \mathbf{H}(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0})$ be the unnormalized statistics computed for $\mathbf{I}_{\Lambda_0}$ where filtering takes place within $\Lambda_0 \cup \partial\Lambda_0$. Let $\mathbf{H}_{01}$ be the statistics computed by filtering inside the fixed environment Λ_1/Λ_0. Let $\Lambda_{-1} = \Lambda/\Lambda_1$ be the big patch outside of Λ_1. Then the statistics computed for Λ_{-1} is $\mathbf{h}|\Lambda| - \mathbf{H}_0 - \mathbf{H}_{01}$. Let $\mathbf{h}_- = (\mathbf{h}|\Lambda| - \mathbf{H}_{01})/|\Lambda_{-1}|$, then the normalized statistics for Λ_{-1} is $\mathbf{h}_- - \mathbf{H}_0/|\Lambda_{-1}|$.

For a fixed $\mathbf{I}_{\Lambda_0}$, the number of images in $\Omega_\Lambda(\mathbf{h})$ with such a patch $\mathbf{I}_{\Lambda_0}$ and its local environment $\mathbf{I}_{\Lambda_1/\Lambda_0}$ is $|\Omega_{\Lambda_{-1}}(\mathbf{h}_- - \mathbf{H}_0/|\Lambda_{-1}|)|$. Therefore the conditional probability, as a function of $\mathbf{I}_{\Lambda_0}$, is

$$q(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\Lambda_1/\Lambda_0}, \mathbf{h}) \propto |\Omega_{\Lambda_{-1}}(\mathbf{h}_- - \frac{\mathbf{H}_0}{|\Lambda_{-1}|})|.$$

Unlike the iid case, the about volume cannot be computed analytically. However, the volume $|\Omega_\Lambda(\mathbf{h})|$ still shares the same asymptotic behavior as in the iid case, namely,

$$\lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log |\Omega_\Lambda(\mathbf{h})| \to s(\mathbf{h}),$$

where $s(\mathbf{h})$ is a concave entropy function of $\mathbf{h}$.

Like the iid case, in the above derivation, we ignore the minor technical complication that $\Omega_\Lambda(\mathbf{h})$ may be empty because the exact constraint may not be satisfied on finite lattice. A more careful treatment is to replace $\mathbf{h}$ by a small set $\mathcal{H}$ around $\mathbf{h}$, and let $\mathcal{H} \to \mathbf{h}$ as $\Lambda \to \mathbf{Z}^2$. Let $\Omega_\Lambda(\mathcal{H}) = \{\mathbf{I} : \mathbf{h}(\mathbf{I}) \in \mathcal{H}\}$, then we have the following

Proposition 1 *The limit*

$$\lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log \Omega_\Lambda(\mathcal{H}) = s(\mathcal{H})$$

exists. Let $s(\mathbf{h}) = \lim_{\mathcal{H} \to \mathbf{h}} s(\mathcal{H})$*, then* $s(\mathbf{h})$ *is concave, and* $s(\mathcal{H}) = \sup_{\mathbf{h} \in \mathcal{H}} s(\mathbf{h})$.

See Lanford (1973) for a detailed analysis of the above result. The $s(\mathbf{h})$ is a measure of the volume of the Julesz ensemble of type $\mathbf{h}$. It defines the randomness of the texture appearance of type $\mathbf{h}$. The exponential order of $|\Omega_\Lambda(\mathcal{H})|$ is the same as the order of the most random equivalence class. For example, if $\Omega_\Lambda(\mathcal{H}) = \Omega_\Lambda$, then the order is decided by the equivalent class of images whose intensities are uniformly distributed.

With such an estimate, we are ready to compute the conditional probability. Note that the conditional distribution, $q(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\Lambda_1/\Lambda_0}, \mathbf{h})$, as a function of $\mathbf{I}_{\Lambda_0}$, is decided only by $\mathbf{H}_0$, which is the sufficient statistics. Therefore, we only need to trace $\mathbf{H}_0$ while leaving other terms as constants. For large Λ, a Taylor expansion at $\mathbf{h}_-$ gives

$$\begin{aligned}
\log q(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\Lambda_1/\Lambda_0}, \mathbf{h}) &= \text{constant} + \log |\Omega_{\Lambda_{-1}}(\mathbf{h}_- - \frac{\mathbf{H}_0}{|\Lambda_{-1}|})| \\
&= \text{constant} + |\Lambda_{-1}| s(\mathbf{h}_- - \frac{\mathbf{H}_0}{|\Lambda_{-1}|}) \\
&= \text{constant} - < s'(\mathbf{h}_-), \mathbf{H}_0 > + o(\frac{1}{|\Lambda|}).
\end{aligned}$$

Assuming the entropy function $s(\mathbf{h})$ has continuous derivative at $\mathbf{h}$, and let $\boldsymbol{\beta} = s'(\mathbf{h})$, then, as $\Lambda \to \mathbf{Z}^2$, $\mathbf{h}_- \to \mathbf{h}$, and $s'(\mathbf{h}_-) \to \boldsymbol{\beta}$. Therefore,

$$\begin{aligned}
\log q(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\Lambda_1/\Lambda_0}, \mathbf{h}) &\to \text{constant} - < s'(\mathbf{h}), \mathbf{H}_0 > \\
&= \text{constant} - < \boldsymbol{\beta}, \mathbf{H}_0 >,
\end{aligned}$$

so

$$q(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\Lambda_1/\Lambda_0}, \mathbf{h}) \to \frac{1}{Z_{\Lambda_0}(\boldsymbol{\beta})} \exp\{- < \boldsymbol{\beta}, \mathbf{H}(\mathbf{I}_{\Lambda_0} \mid \mathbf{I}_{\partial \Lambda_0}) >\},$$

which is exactly the Markov property specified by the FRAME model. This derivation shows that local computation using the FRAME model is justified under the Julesz ensemble. It also reveals an important relationship, i.e., the parameter $\boldsymbol{\beta}$ can be identified as the derivative of the entropy function $s(\mathbf{h})$, $\boldsymbol{\beta} = s'(\mathbf{h})$.

4.3 From the FRAME model to the Julesz ensemble

In this subsection, we study the statistical properties of the FRAME models as $\Lambda \to \mathbf{Z}^2$.

Consider the FRAME model

$$p(\mathbf{I};\boldsymbol{\beta}) = \frac{1}{Z_\Lambda(\boldsymbol{\beta})} \exp\{-|\Lambda| < \boldsymbol{\beta}, \mathbf{h}(\mathbf{I}) >\},$$

which assigns equal probabilities to images in $\Omega_\Lambda(\mathbf{h})$. The probability that $p(\mathbf{I};\boldsymbol{\beta})$ assigns to $\Omega_\Lambda(\mathbf{h})$ is

$$p(\Omega_\Lambda(\mathbf{h});\boldsymbol{\beta}) = \frac{1}{Z_\Lambda(\boldsymbol{\beta})} \exp\{-|\Lambda| < \boldsymbol{\beta}, \mathbf{h} >\}|\Omega_\Lambda(\mathbf{h})|.$$

The asymptotic behavior of this probability is

$$\begin{aligned} s_\beta(\mathbf{h}) &= \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log p(\Omega_\Lambda(\mathbf{h});\boldsymbol{\beta}) \\ &= -<\boldsymbol{\beta}, \mathbf{h}> + s(\mathbf{h}) - \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log Z_\Lambda(\boldsymbol{\beta}). \end{aligned}$$

For the last term, we have

Proposition 2 *The limit*

$$\rho(\boldsymbol{\beta}) = \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log Z_\Lambda(\boldsymbol{\beta})$$

exists and is independent of the boundary condition. $\rho(\boldsymbol{\beta})$ is convex.

The $\rho(\boldsymbol{\beta})$ is called *pressure* in physics. See Griffiths and Ruelle (1971) for a rigorous analysis of the pressure function.

Therefore, we have

Proposition 3 *For $\Omega_\Lambda(\mathbf{h})$, the probability rate function $s_\beta(\mathbf{h})$ of the FRAME model $p(\mathbf{I};\boldsymbol{\beta})$ is*

$$s_\beta(\mathbf{h}) = \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log p(\Omega_\Lambda(\mathbf{h});\boldsymbol{\beta}) = s(\mathbf{h}) - <\boldsymbol{\beta}, \mathbf{h}> - \rho(\boldsymbol{\beta}).$$

Then the probability mass that $p(\mathbf{I};\boldsymbol{\beta})$ puts on $\Omega_\Lambda(\mathbf{h})$ has an exponential order

$$p(\mathbf{I} \in \Omega_\Lambda(\mathbf{h});\boldsymbol{\beta}) \sim e^{|\Lambda| s_\beta(\mathbf{h})}. \tag{5}$$

$s_\beta(\mathbf{h}) \leq 0$ for any $\mathbf{h}$ and $\boldsymbol{\beta}$, otherwise the probability will go unbounded.

Therefore, $p(\mathbf{I};\boldsymbol{\beta})$ eventually concentrates on $\Omega_\Lambda(\mathbf{h}_*)$ with

$$\mathbf{h}_* = \arg\max_{\mathbf{h}} s_\beta(\mathbf{h}) = \arg\max_{\mathbf{h}} \{s(\mathbf{h}) - <\boldsymbol{\beta}, \mathbf{h}> - \rho(\boldsymbol{\beta})\}.$$

Moreover, the maximum of $s_\beta(\mathbf{h})$, i.e., $s_\beta(\mathbf{h}_*)$, should be 0. Otherwise, if $s_\beta(\mathbf{h}_*) < 0$, then the total probability on Ω_Λ goes to zero, because $\mathbf{h}$ belongs to a compact set. So we have the following

Theorem 1 *If there is a unique* $\mathbf{h}_*$ *where* $s_\beta(\mathbf{h})$ *achieves its maximum 0, then* $p(\mathbf{I};\boldsymbol{\beta})$ *eventually concentrates on* $\mathbf{h}_*$ *as* $\Lambda \to \mathbf{Z}^2$*. Therefore the FRAME model* $p(\mathbf{I};\boldsymbol{\beta})$ *goes to a Julesz ensemble of type* $\mathbf{h}_*$*. Moreover, if* $s(\mathbf{h})$ *is differentiable at* $\mathbf{h}_*$*, then* $s'(\mathbf{h}_*) = \boldsymbol{\beta}$*.*

The uniqueness of $\mathbf{h}_*$ holds under the condition that there is no phase transition at $\boldsymbol{\beta}$.

The above analysis establishes a one to one correspondence between $\boldsymbol{\beta}$ and $\mathbf{h}_*$ on large lattice in the absence of phase transition.

$s_\beta(\mathbf{h})$ can be identified with $-\mathrm{KL}(h||p)$ in the iid case, following the proposition below.

Proposition 4 *Suppose two FRAME models* $p_A = p(\mathbf{I};\boldsymbol{\beta}_A)$ *and* $p_B = p(\mathbf{I};\boldsymbol{\beta}_B)$ *concentrate on* $\mathbf{h}_A$ *and* $\mathbf{h}_B$ *respectively. Then,*

$$kl(p_B||p_A) = \lim_{\Lambda\to\mathbf{Z}^2} \frac{1}{|\Lambda|} KL(p_B||p_A) = -s_{\beta_A}(\mathbf{h}_B),$$

where $kl(p_B||p_A)$ *denotes the Kullback-Leibler divergence rate (per pixel).*

[Proof] By definition, we have

$$\begin{aligned}
\lim_{\Lambda\to\mathbf{Z}^2} \frac{1}{|\Lambda|} \log \mathrm{KL}(p_B||p_A) &= \lim_{\Lambda\to\mathbf{Z}^2} \frac{1}{|\Lambda|} \log \mathrm{E}_{p_B}[\log \frac{p(\mathbf{I};\boldsymbol{\beta}_B)}{p(\mathbf{I};\boldsymbol{\beta}_A)}] \\
&= \lim_{\Lambda\to\mathbf{Z}^2} \frac{1}{|\Lambda|} \log \frac{Z(\boldsymbol{\beta}_A)}{Z(\boldsymbol{\beta}_B)} + <\boldsymbol{\beta}_A, \mathbf{h}_B> - <\boldsymbol{\beta}_B, \mathbf{h}_B> \\
&= <\boldsymbol{\beta}_A, \mathbf{h}_B> - <\boldsymbol{\beta}_B, \mathbf{h}_B> + \rho(\boldsymbol{\beta}_A) - \rho(\boldsymbol{\beta}_B) \\
&= -s(\mathbf{h}_B) + <\boldsymbol{\beta}_A, \mathbf{h}_B> + \rho(\boldsymbol{\beta}_A) = -s_{\beta_A}(\mathbf{h}_B).
\end{aligned}$$

The last step follows from the fact that $s_{\beta_B}(\mathbf{h}_B) = s(\mathbf{h}_B) - <\boldsymbol{\beta}_B, \mathbf{h}_B> - \rho(\beta_B) = 0$.

The above conclusion provides an intuitive explanation for equation (5). The probability mass of $p(\mathbf{I};\boldsymbol{\beta})$ on class $\Omega_\Lambda(\mathbf{h})$ decreases exponentially in an order that is equal to the KL-divergence rate between the two models specified by $\mathbf{h}$ and $\boldsymbol{\beta}$.

4.4 Typical versus non-typical images in a Julesz ensemble

In this section, we discuss typical and non-typical images in a Julesz ensemble.

Consider a Julesz ensemble $\Omega_\Lambda(\mathbf{h})$ of type $\mathbf{h}$. Images in $\Omega_\Lambda(\mathbf{h})$ all share the same statistics of type $\mathbf{h}$, however, they may differ in terms of other statistical properties. Suppose we introduce an arbitrary new statistics $h^{(0)}(\mathbf{I})$ which measures additional image features (e.g., marginal histogram of a new Gabor filter). Then, images in $\Omega_\Lambda(\mathbf{h})$ may differ in their $h^{(0)}(\mathbf{I})$. This suggests that we can partition $\Omega_\Lambda(\mathbf{h})$ into finer equivalence classes (or sub-classes) according to $h^{(0)}(\mathbf{I})$, i.e.,

$$\Omega_\Lambda(\mathbf{h}) = \cup_{h^{(0)}} \Omega_\Lambda(\mathbf{h}, h^{(0)}),$$

where

$$\Omega_\Lambda(\mathbf{h}, h^{(0)}) = \{\mathbf{I} : \mathbf{h}(\mathbf{I}) = \mathbf{h},\ h^{(0)}(\mathbf{I}) = h^{(0)}\}.$$

Now let's study the volumes of these finer Julesz ensembles. Let

$$s(\mathbf{h}, h^{(0)}) = \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log |\Omega_\Lambda(\mathbf{h}, h^{(0)})|$$

be the entropy function of the subclass $\Omega_\Lambda(\mathbf{h}, h^{(0)})$. For a fixed $\mathbf{h}$, if there is a unique $h_*^{(0)}$ such that $s(\mathbf{h}, h^{(0)})$ achieves its maximum as a function of $h^{(0)}$, then the volume $|\Omega_\Lambda(\mathbf{h})|$ is dominated by the volume $|\Omega_\Lambda(\mathbf{h}, h_*^{(0)})|$, and

$$s(\mathbf{h}) = s(\mathbf{h}, h_*^{(0)}),$$

because the order of the sum equals to the largest order among individual terms.

Proposition 5 *For a fixed* $\mathbf{h}$*, if there is a unique* $h_*^{(0)}$ *that maximizes* $s(\mathbf{h}, h^{(0)})$ *as a function of* $h^{(0)}$*, then the Julesz ensemble of type* $\mathbf{h}$ *concentrates on* $h_*^{(0)}$*, i.e., almost all the images in the Julesz ensemble of type* $\mathbf{h}$ *have statistics* $h^{(0)}(\mathbf{I}) = h_*^{(0)}$*. We call* $h_*^{(0)}$ *typical value of* $h^{(0)}(\mathbf{I})$ *for the Julesz ensemble of type* $\mathbf{h}$*.*

Therefore, the Julesz ensemble of type $\mathbf{h}$ is essentially the Julesz ensemble of type $(\mathbf{h}, h_*^{(0)})$. All images in the other sub-classes are non-typical and have zero probability mass as $\Lambda \to \mathbf{Z}^2$. The uniqueness of $h_*^{(0)}$ holds in the absence of phase transition.

Because $h^{(0)}(\mathbf{I})$ is arbitrary, we can let it collect as many statistical properties as possible. The above proposition then tells us that almost all the images in the same Julesz ensemble share the same typical statistical properties $h_*^{(0)}$ and therefore the same typical visual appearance. As a result, if we can sample just *one* typical image from $\Omega_\Lambda(\mathbf{h})$ on large lattice, then we should be able to tell the visual appearances of almost all the images in $\Omega_\Lambda(\mathbf{h})$. Obtaining a typical image can be accomplished by sampling from $q(\mathbf{I}; \mathbf{h})$, i.e., the uniform distribution over $\Omega_\Lambda(\mathbf{h})$, or sampling from the corresponding FRAME model $p(\mathbf{I}; \boldsymbol{\beta})$. See the next section for some experiments. The non-typical subclasses include images such as human faces and office scenes, which may not be considered as texture in perception.

5 Equivalence of ensembles: experiments and its significance

In this section, we demonstrate some experimental results on sampling the Julesz ensembles and their corresponding FRAME models, and discuss practical implications of ensemble equivalence in modeling visual patterns beyond textures.

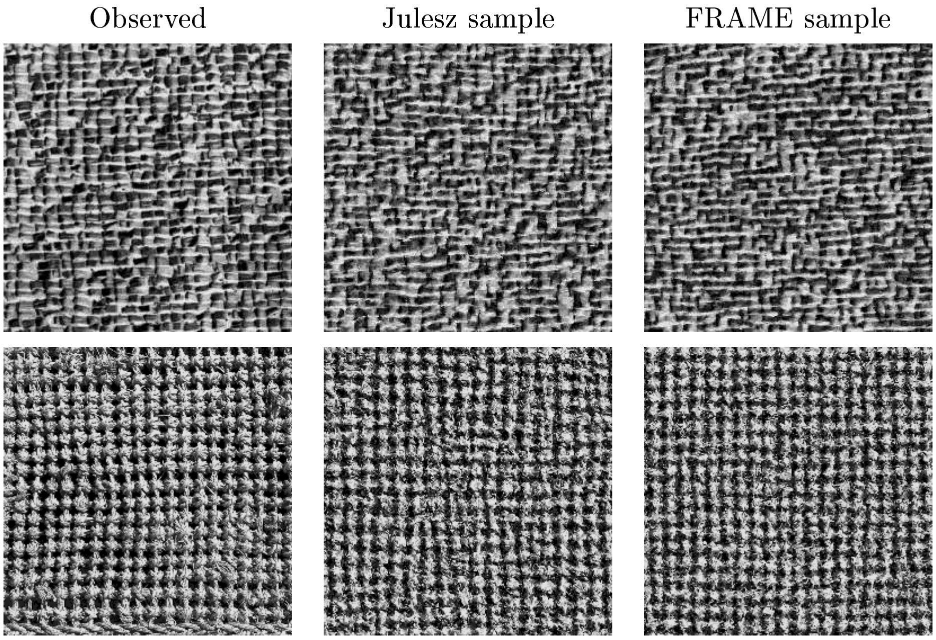


Figure 3: For each row, the left image is observed as the training image, the middle image is a typical sample from the Julesz ensemble, and the right image is a typical sample from the corresponding FRAME model.

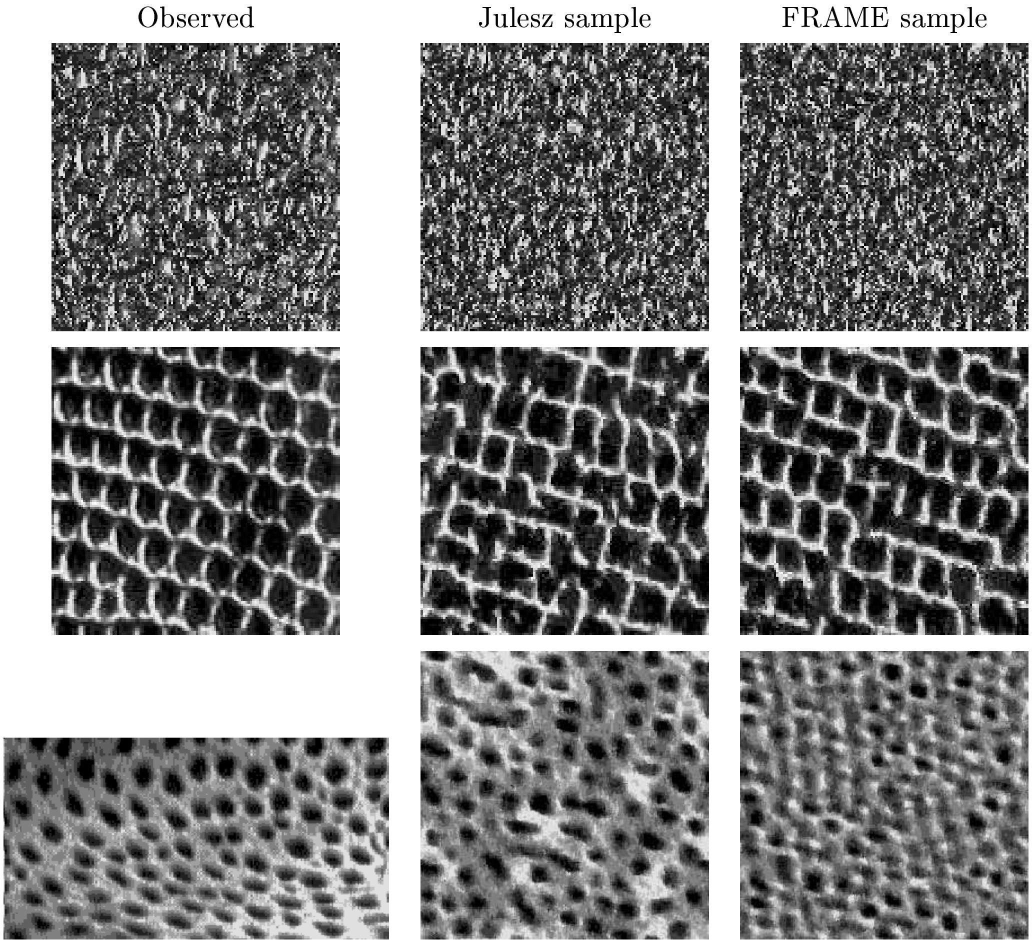


Figure 4: For each row, the left image is observed as the training image, the middle image is a typical sample from the Julesz ensemble, and the right image is a typical sample from the corresponding FRAME model.

We conduct our experiments on a set of 20 texture images, five of which are shown in figures 3 and 4. For ease of computation, these images are quantized into 8 gray levels only. We do not implement the filter pursuit process used in our early work[27]. Instead we fix a set of 34 filters for all 20 images: one for intensity, four gradient filters for the horizontal and vertical directions, five Laplacian of Gaussian filters at various scales, and 24 Gabor filters at 4 scale and six different orientations. The statistics $\mathbf{h}(\mathbf{I})$ collects the histograms of the 34 filters.

For each of the 20 images, we simulate three Monte Carlo Markov Chains (MCMC) for stochastic sampling.

MCMC I: it starts from a white noise image, and samples from the uniform distribution $q(\mathbf{I}; \mathbf{h}_{\text{obs}})$ using a simple annealing precess, where $\mathbf{h}_{\text{obs}} = \mathbf{h}(\mathbf{I}_{\text{obs}})$ is the type of the observed image. This process simulates typical images from the Julesz ensemble of type $\mathbf{h}_{\text{obs}}$. A detailed account is given in [26].

MCMC II: it simulates an inhomogeneous Markov chain to learn the parameters $\boldsymbol{\beta}$ in the FRAME model $p(\mathbf{I}; \boldsymbol{\beta})$ from the observed statistics $\mathbf{h}_{\text{obs}}$, as is discussed in our early work[27].

MCMC III: it starts from a white noise image, and simulates a homogeneous Markov chain sampling from the model $p(\mathbf{I}; \boldsymbol{\beta})$ learned using MCMC II. This process synthesizes typical images from the FRAME model, which, as we have shown, is equivalent to the Julesz ensemble on large image lattice.

MCMC I and MCMC III provide two different ways to explore the typical images of the Julesz ensemble of type $\mathbf{h}_{\text{obs}}$. It is worth mentioning that the convergence of MCMC III is practically much slower and harder than that of MCMC II.

The results of MCMC I and III are shown in the middle and right columns of figures 3 and 4 respectively. The visual similarity of their appearances demonstrates that both the Julesz ensemble and the FRAME model focus on the same set of typical images that share identical statistical properties subject to minor statistical fluctuations on finite lattice.

The ensemble equivalence has a broad implication for modeling general visual patterns beyond textures, for example, shapes, flow patterns, speech signals and natural languages.

Figure 5 summarizes a unified paradigm for modeling general visual patterns using feature statistics from a dictionary shown on the right side. Given a natural pattern generated by some unknown stochastic process, we have as observation a set of samples, such as a set of images. The natural process is shown by the dotted lines. The goal is to characterize these samples in computer applications. There are two methodologies as shown by the two paths in figure 5. The solid line (path 1) represents the research theme that pursues a Gibbs model based on a minimax entropy learning scheme[27]. The dashed

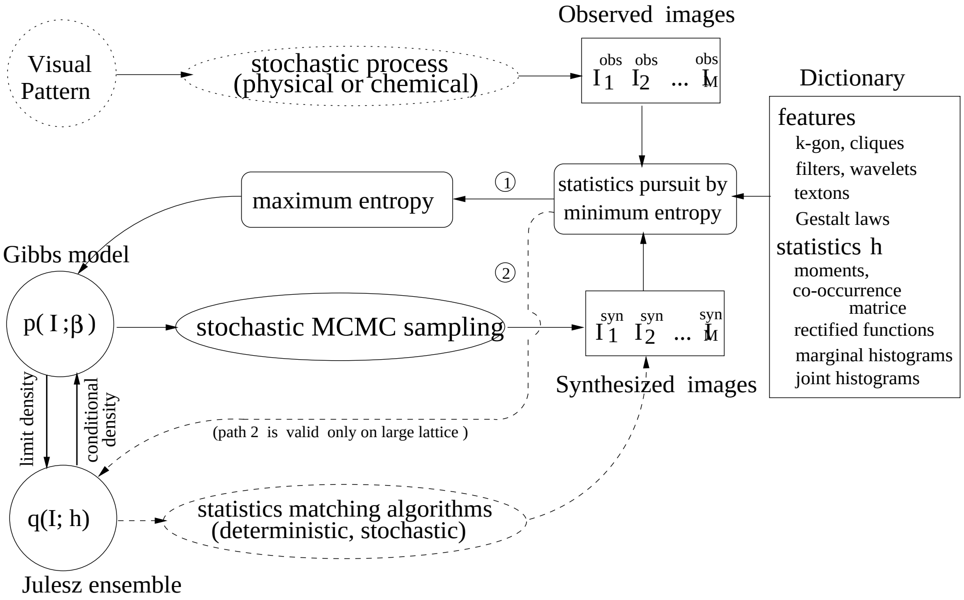


Figure 5: A global picture for theories of stochastic modeling.

line (path 2) represents the research theme that seeks the definition of the pattern on large lattice systems, i.e. the Julesz ensemble. Both the Gibbs model and the Julesz ensemble are verified through stochastic sampling using Markov chain Monte Carlo as a general engine. The two lines are connected by the equivalence of ensembles.

Practically, the ensemble equivalence enables us to utilize the advantages of both methodologies. Path 2 is more effective for model verification and model selection, since it does not have to learn the expensive Gibbs model explicitly. Path 1 is useful for local computation in vision tasks, such as image segmentation and discrimination.

Conceptually, the unification helps us link mathematical concepts such as *probability models* and *entropy* in finite lattice systems to intuitive concepts such as *ensembles* and *volumes* on large lattice systems.

In a broader sense, figure 5 represents a self-consistent paradigm based on the philosophy dated back to (Julesz, 1962)[11]: perception is a process that computes essential features and statistics. In recent papers, this paradigm has been applied to modeling other visual patterns, such as 2D object shapes[29] and generic images and clutter[28].

6 Geometric interpretation and phase transition

In this section, we review the geometric interpretation of the relationship between $s(\mathbf{h})$ and $\rho(\boldsymbol{\beta})$, and discuss phase transition briefly.

So far, we have introduced three important concepts in the limit $\Lambda \to \mathbf{Z}^2$.

1. Given statistics $\mathbf{h}$ and its Julesz ensemble $\Omega_\Lambda(\mathbf{h})$, we have the entropy function $s(\mathbf{h})$ that is the exponential order of the volume of $\Omega_\Lambda(\mathbf{h})$,

$$s(\mathbf{h}) = \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log |\Omega_\Lambda(\mathbf{h})|.$$

2. Given parameters $\boldsymbol{\beta}$ and its FRAME model $p(\mathbf{I}; \boldsymbol{\beta})$, we have the pressure function $\rho(\boldsymbol{\beta})$ that is the exponential order of the partition function $Z_\Lambda(\boldsymbol{\beta})$,

$$\rho(\boldsymbol{\beta}) = \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log Z_\Lambda(\boldsymbol{\beta}).$$

3. The probability rate function $s_\beta(\mathbf{h})$ links $\boldsymbol{\beta}$ and $\mathbf{h}$. $s_\beta(\mathbf{h})$ is the exponential order of the probability mass that $p(\mathbf{I}; \boldsymbol{\beta})$ assigns to $\Omega_\Lambda(\mathbf{h})$,

$$\begin{aligned} s_\beta(\mathbf{h}) &= \lim_{\Lambda \to \mathbf{Z}^2} \frac{1}{|\Lambda|} \log p(\mathbf{I} \in \Omega_\Lambda(\mathbf{h}); \boldsymbol{\beta}) \\ &= s(\mathbf{h}) - < \boldsymbol{\beta}, \mathbf{h} > - \rho(\boldsymbol{\beta}). \end{aligned}$$

When $s_\beta(\mathbf{h})$ achieves its maximum zero, we have the relationship between $\mathbf{h}$ and $\boldsymbol{\beta}$.
Definition *If $s_\beta(\mathbf{h}) = 0$, i.e., $s(\mathbf{h}) - < \boldsymbol{\beta}, \mathbf{h} > - \rho(\boldsymbol{\beta}) = 0$, then $\boldsymbol{\beta}$ and $\mathbf{h}$ are said to correspond to each other.*

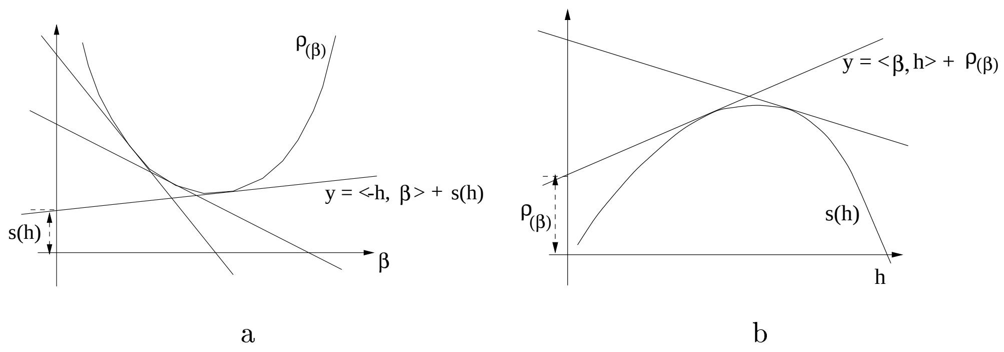


Figure 6: Convex conjugate between $\rho(\boldsymbol{\beta})$ and $s(\mathbf{h})$. a) The tangent $\rho'(\boldsymbol{\beta}_0) = \mathbf{h}_0$ and all planes are below $\rho(\boldsymbol{\beta}_0)$. b) The tangent $s'(\mathbf{h}_0) = \boldsymbol{\beta}_0$ and all planes are above $s(\mathbf{h}_0)$.

From the definition, one can derive the interesting geometric relationship between $\boldsymbol{\beta}$ and $\mathbf{h}$ as displayed in figure 6.

Proposition 6 *If $\boldsymbol{\beta}_0$ and $\mathbf{h}_0$ correspond to each other, and if $\rho(\boldsymbol{\beta})$ is differentiable at $\boldsymbol{\beta}_0$, then*

$$s'(\mathbf{h}_0) = \boldsymbol{\beta}_0, \quad \text{and} \quad \rho'(\boldsymbol{\beta}_0) = \mathbf{h}_0.$$

That is, $\boldsymbol{\beta}_0$ is the tangent of $s(\mathbf{h})$ at $\mathbf{h} = \mathbf{h}_0$ and $\mathbf{h}_0$ is the tangent of $\rho(\boldsymbol{\beta})$ at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. Furthermore, because $\rho(\boldsymbol{\beta})$ is convex, all the planes $\rho = s(\mathbf{h}) - < \boldsymbol{\beta}, \mathbf{h} >$ are below the point $(\boldsymbol{\beta}_0, \rho(\boldsymbol{\beta}_0))$, i.e.,

$$\rho(\boldsymbol{\beta}_0) \geq s(\mathbf{h}) - < \boldsymbol{\beta}_0, \mathbf{h} > \quad \forall \mathbf{h}, \forall \boldsymbol{\beta}_0.$$

In a similar way, because $s(\mathbf{h})$ is concave, all the planes $s = \rho(\boldsymbol{\beta}) + < \boldsymbol{\beta}, \mathbf{h}_0 >$ are above $(\mathbf{h}_0, s(\mathbf{h}_0))$

$$s(\mathbf{h}_0) \leq \rho(\boldsymbol{\beta}) + < \mathbf{h}_0, \boldsymbol{\beta} > \quad \forall \boldsymbol{\beta}, \forall \mathbf{h}_0,$$

This is formally expressed by the following proposition, illustrated in Figure 6. It holds even when $\rho(\boldsymbol{\beta}_0)$ is not differentiable.

Proposition 7 *$s(\mathbf{h})$ and $\rho(\boldsymbol{\beta})$ are convex conjugates, i.e.,*

$$\rho(\boldsymbol{\beta}) = \max_{\mathbf{h}} \{ s(\mathbf{h}) - < \boldsymbol{\beta}, \mathbf{h} > \}, \tag{6}$$

$$s(\mathbf{h}) = \min_{\beta} \{ p(\boldsymbol{\beta}) + < \boldsymbol{\beta}, \mathbf{h} > \}. \tag{7}$$

If one of (6) and (7) is true, then the other must be true.

The equalities in (6) and (7) holds when $\boldsymbol{\beta}$ and $\mathbf{h}$ correspond to each other. See Lanford (1973) for a detailed analysis.

The differentiability of $\rho(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_0$ determines whether there is a phase transition at $\boldsymbol{\beta}_0$. Recall that

$$\frac{\partial}{\partial \boldsymbol{\beta}} \frac{1}{|\Lambda|} \log Z_\Lambda(\boldsymbol{\beta}) = -\mathrm{E}_\beta[\mathbf{h}(\mathbf{I})].$$

Although $\rho(\boldsymbol{\beta})$ as the limit of $\log Z_\Lambda(\boldsymbol{\beta})/|\Lambda|$ always exists, it may not be differentiable at $\boldsymbol{\beta}_0$, indicating that a phase transition occurs at $\boldsymbol{\beta}_0$. So $\mathrm{E}_{\beta_0}[\mathbf{h}(\mathbf{I})]$ may go to multiple limits under different boundary conditions. Meanwhile, the probability rate function $s_{\beta_0}(\mathbf{h})$ or $s(\mathbf{h}) - < \boldsymbol{\beta}_0, \mathbf{h} >$ may achieve its maximum at multiple $\mathbf{h}$. Because $s(\mathbf{h})$ is a concave function, this can happen only when $s(\mathbf{h})$ is not strictly concave, i.e., $s(\mathbf{h})$ has a linear piece. Figure 7 illustrates the concept. In a), a cusp appears at point $\boldsymbol{\beta}_0$, so the convex function $\rho(\boldsymbol{\beta})$ can be above multiple planes at $\boldsymbol{\beta}_0$. In b), there is a flat linear piece in $s(\mathbf{h})$ so that many $\mathbf{h}$ share the same tangent $\boldsymbol{\beta}_0$.

If there is a phase transition at $\boldsymbol{\beta}_0$, then when we sample from the FRAME model $p(\mathbf{I}; \boldsymbol{\beta}_0)$ on a large lattice, we may get images of different statistical properties $\mathbf{h}(\mathbf{I})$ and

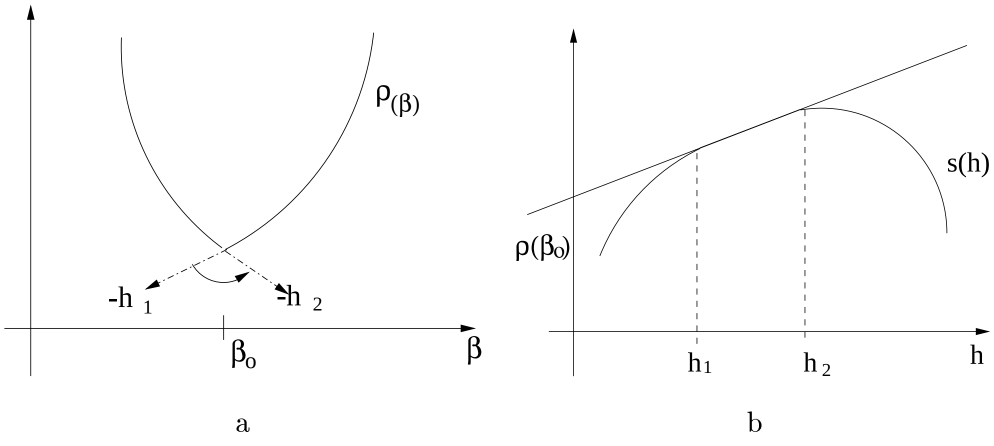


Figure 7: a). If $\rho(\boldsymbol{\beta})$ is not differentiable at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, a phase transition occurs, and there exists a convex set of expected statistics $\mathbf{h}$, as shown by the interval $[\mathbf{h}_1, \mathbf{h}_2]$. b). The entropy function $s(\mathbf{h})$ has a constant tangent $\boldsymbol{\beta}_0$ over a set of $\mathbf{h}$.

therefore different visual appearances if we use different boundary conditions. This indicates that the effect of boundary conditions does not vanish on large lattice.

If for an $\mathbf{h}_0$, the corresponding $\boldsymbol{\beta}_0$ leads to a phase transition, then when we sample from $q(\mathbf{I}; \mathbf{h})$, i.e., the uniform distribution over $\Omega_\Lambda(\mathbf{h})$ on a large lattice, we may get images consisting of several large pieces of different statistical properties (and visual appearances), and each piece can arise from the FRAME model $p(\mathbf{I}; \boldsymbol{\beta}_0)$ under suitable boundary conditions. See Martin-Lof (1979) for a more discussion.

In our experiments, we have not captured a definitive phase transition phenomenon described above. We will leave this issue for future investigation.

7 Measure of texture distance and asymmetry

In this section, we study model-based texture distance that extends the traditional signal detection theory[8] from iid signals to random fields.

7.1 Distance measure on random fields

In search of texture statistics $\mathbf{h}(\mathbf{I})$ to which pre-attentive vision is sensitive, psychophysicists use texture discrimination experiments to see how effortlessly a foreground texture patch B can "pop out" from a background texture A and vice versa.

One widely observed phenomenon in the pop-out experiments is asymmetry. For example, it is easier for a moving dot to pop out from a background of static dots than for a static dot to pop out from a background of moving dots. A curve pops out easily from a background of straight lines, whereas it is harder to detect a straight line from a background of curves. The perceptual distances between two texture images are also found to be asymmetrical (see [20] and references therein).

This asymmetry can be explained by the asymmetry of the Kullback-Leibler distance between the statistical models of the two signals. In the case where elements in each signal, such as moving dots, are independently Gaussian distributed, this KL-divergence is reduced to the Mahalanobis distance[22] or signal to noise ratio (SNR) in traditional signal detection theory[8]. To our knowledge, there has been no rigorous work for computing distance for signals that are not independently distributed, such as textures on random fields.

The basic scenario is as follows. A background of texture A defined on a large lattice Λ is generated from a FRAME model $p(\mathbf{I};\boldsymbol{\beta}_A)$. Within the background, a small patch $\mathbf{I}_{\Lambda_0}$ of texture B with $\Lambda_0 \subset \Lambda$ is generated from a model $p(\mathbf{I};\boldsymbol{\beta}_B)$. There are two ways to generate B in A. One is to generate the foreground patch $\mathbf{I}_{\Lambda_0}$ from the conditional distribution $p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_B)$ with $\mathbf{I}_{\Lambda/\Lambda_0} \sim p(\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_A)$ being the boundary condition. The other method crops a patch $\mathbf{I}_{\Lambda_0}$ from $\mathbf{I}_\Lambda \sim p(\mathbf{I};\boldsymbol{\beta}_B)$, and pastes it to the background of texture A by occlusion, so $\mathbf{I}_{\Lambda_0}$ is generated from the marginal distribution of $p(\mathbf{I};\boldsymbol{\beta}_B)$ with the boundary condition integrated out according to $p(\mathbf{I};\boldsymbol{\beta}_B)$. The second case often generates sharp edges, which constitute a strong artificial cue for discrimination, thus we only discuss the first case where the background is used as the boundary condition.

We formulate the problem in a Bayesian inference framework. The easiness of pop-out is measured by the ratio of the posterior probabilities of pop-out versus no pop-out.

$$
\begin{aligned}
r(\mathbf{I}_\Lambda) &= \frac{\Pr(\text{pop-out}|\mathbf{I}_\Lambda)}{\Pr(\text{no pop-out}|\mathbf{I}_\Lambda)} \\
&= \frac{\Pr(\text{pop-out})}{\Pr(\text{no pop-out})}\frac{\Pr(\mathbf{I}_\Lambda|\text{pop-out})}{\Pr(\mathbf{I}_\Lambda|\text{no pop-out})} \\
&= \frac{\Pr(\text{pop-out})}{\Pr(\text{no pop-out})}\frac{p(\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_A)p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_B)}{p(\mathbf{I}_\Lambda;\boldsymbol{\beta}_A)}, \\
&= \frac{\Pr(\text{pop-out})}{\Pr(\text{no pop-out})}\frac{p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_B)}{p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_A)}, \quad \mathbf{I}_{\Lambda/\Lambda_0} \sim p(\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_A),
\end{aligned}
$$

where Pr(pop-out) and Pr(no pop-out) are prior probabilities of pop-out and no pop-out respectively. Therefore, the log of posterior ratio

$$\log r(\mathbf{I}_\Lambda) = \log \frac{p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_B)}{p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_A)} + \log \frac{\Pr(\text{pop-out})}{\Pr(\text{no pop-out})}$$

is decided by the first term with fixed prior probabilities. In the following, we assume $\Pr(\text{pop-out}) = \Pr(\text{no pop-out})$, so the second term in the above vanishes. Averaging over $\mathbf{I}_{\Lambda_0} \sim p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_B)$ and $\mathbf{I}_{\Lambda/\Lambda_0} \sim p(\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_A)$, the easiness of pop-out is

$$\begin{aligned} M_{AB} &= \mathrm{E}_{p(\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_A)}\{\mathrm{E}_{p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_B)}[\log r(\mathbf{I}_\Lambda)]\} & (8) \\ &= \mathrm{E}_{p(\mathbf{I}_{\Lambda/\Lambda_0};\boldsymbol{\beta}_A)}\{\mathrm{KL}(p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_B)||p(\mathbf{I}_{\Lambda_0}|\mathbf{I}_{\partial\Lambda_0};\boldsymbol{\beta}_A))\}. & (9) \end{aligned}$$

The larger M_{AB} is, the easier for patch B to pop out from background A.

Given $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$, M_{AB} only depends on the shape of the foreground patch Λ_0. We now briefly study the behavior of M_{AB} when Λ_0 is sufficiently large such that the effect of the boundary condition diminishes. Let $\mathbf{h}_A$ and $\mathbf{h}_B$ be the statistics corresponding to $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$ respectively. We have

$$\begin{aligned} M_{AB} &\approx \mathrm{E}_{p(\mathbf{I}_{\Lambda_0};\boldsymbol{\beta}_B)}[\log \frac{p(\mathbf{I}_{\Lambda_0};\boldsymbol{\beta}_B)}{p(\mathbf{I}_{\Lambda_0};\boldsymbol{\beta}_A)}] \\ &\approx \log Z_{\Lambda_0}(\boldsymbol{\beta}_A) + |\Lambda_0| < \boldsymbol{\beta}_A, \mathbf{h}_B > - \log Z_{\Lambda_0}(\boldsymbol{\beta}_B) - |\Lambda_0| < \boldsymbol{\beta}_B, \mathbf{h}_B > \\ &\sim |\Lambda_0|\{\rho(\boldsymbol{\beta}_A) + < \boldsymbol{\beta}_A, \mathbf{h}_B > - \rho(\boldsymbol{\beta}_B) + < \boldsymbol{\beta}_B, \mathbf{h}_B >\} \\ &= -|\Lambda_0| s_{\beta_A}(\mathbf{h}_B) = |\Lambda_0| \mathrm{kl}(p_B||p_A) = -\log p(\mathbf{I} \in \Omega_{\Lambda_0}(\mathbf{h}_B);\boldsymbol{\beta}_A). \end{aligned}$$

$p(\mathbf{I} \in \Omega_{\Lambda_0}(\mathbf{h}_B);\boldsymbol{\beta}_A)$ measures how likely a texture patch of type A has typical statistics $\mathbf{h}_B$ of texture B. If this probability is large, then the background A is very distracting, and it is hard for B to pop out.

From the above derivation, for large patch Λ_0, M_{AB} increases in proportion to $|\Lambda_0|$ with a rate $-s_{\beta_A}(\mathbf{h}_B) \geq 0$, which is the Kullback-Leibler divergence rate. Because $\mathrm{kl}(p_B||p_A) \neq \mathrm{kl}(p_A||p_B)$ in general, $M_{AB} \neq M_{BA}$, which leads to the asymmetry in pop-out easiness. Also, for large lattice, the task of texture discrimination becomes trivial, that is, *the foreground texture must pop out effortlessly unless* $\mathrm{kl}(p_B||p_A) = 0$. This is why psychologists can use pop-out experiments to test what kind of $\mathbf{h}(\mathbf{I})$ are essential in the pre-attentive visual processing stage.

7.2 Experiments on texture distance

The KL-divergence rate between two FRAME models are not analytically computable, so we seek numerical approximation.

First, we synthesize a large image $\mathbf{I}_A \sim p(\mathbf{I};\boldsymbol{\beta}_A)$. Then we dig a number of N holes in $\mathbf{I}_A$, each hole has $m \times m$ pixels. We denote by Λ_m the lattice for $m \times m$ pixels, and we label the boundary images for each hole as $\mathbf{I}_A^{(i)}, i = 1, ..., N$. They are typical samples from $p(\mathbf{I}_{\Lambda/\Lambda_m};\boldsymbol{\beta}_A)$. Then within each hole, we sample L patches from $p(\mathbf{I}_{\Lambda_m}|\mathbf{I}_A^{(i)};\boldsymbol{\beta}_B)$, and we denote these L patches by $\mathbf{I}_B^{(i,j)}$ for $j = 1, ..., L$. Then we can approximate the M_{AB} for the foreground of shape Λ_m by Monte Carlo integration,

$$\begin{aligned}
M_{AB}^{(m)} &= \mathrm{E}_{p(\mathbf{I}_{\Lambda/\Lambda_m};\boldsymbol{\beta}_A)}\{\mathrm{E}_{p(\mathbf{I}_{\Lambda_m};I_{\partial\Lambda_m},\boldsymbol{\beta}_B)}[\log r(\mathbf{I}_\Lambda)]\} \\
&\approx \frac{1}{N}\sum_{i=1}^{N}\frac{1}{L}\sum_{j=1}^{L}\log\frac{p(\mathbf{I}_B^{(i,j)}|\mathbf{I}_A^{(i)};\boldsymbol{\beta}_B)}{p(\mathbf{I}_B^{(i,j)}|\mathbf{I}_A^{(i)};\boldsymbol{\beta}_A)}, \\
&= \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L}[\log\frac{Z(\mathbf{I}_A^{(i)},\boldsymbol{\beta}_A)}{Z(\mathbf{I}_A^{(i)},\boldsymbol{\beta}_B)}] < \boldsymbol{\beta}_A - \boldsymbol{\beta}_B, \mathbf{H}(\mathbf{I}_B^{(i,j)}|\mathbf{I}_A^{(i)}) > .
\end{aligned}$$

In practice, we set $L = 100$ and $N = 200$. The key difficulty is to compute the ratio $Z(\mathbf{I}_A^{(i)},\boldsymbol{\beta}_A)/Z(\mathbf{I}_A^{(i)},\boldsymbol{\beta}_B)$.

We estimate the ratio by importance sampling[19]. We choose an intermediate model $\boldsymbol{\beta}_0$ between $\boldsymbol{\beta}_A$ and $\boldsymbol{\beta}_B$, for example, $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_A + \boldsymbol{\beta}_B)/2$, and generate $\mathbf{I}_{\Lambda_m}^{(1)}, ..., \mathbf{I}_{\Lambda_m}^{(n)}$ from $p(\mathbf{I}_{\Lambda_m}|\mathbf{I}_A^{(i)};\boldsymbol{\beta}_0)$, and then compute the ratio as

$$\begin{aligned}
\frac{Z(\boldsymbol{\beta}_A)}{Z(\boldsymbol{\beta}_B)} &= \frac{\sum_{\mathbf{I}_{\Lambda_m}}\exp\{- < \boldsymbol{\beta}_A, H(\mathbf{I}_{\Lambda_m}) >\}}{\sum_{\mathbf{I}_{\Lambda_m}}\exp\{- < \boldsymbol{\beta}_B, H(\mathbf{I}_{\Lambda_m}) >\}} \\
&= \frac{\sum_{\mathbf{I}_{\Lambda_m}}\exp\{- < \boldsymbol{\beta}_A - \boldsymbol{\beta}_0, H(\mathbf{I}_{\Lambda_m}) >\}p(\mathbf{I}_{\Lambda_m}|\mathbf{I}_{\partial\Lambda_m};\boldsymbol{\beta}_0)}{\sum_{\mathbf{I}_{\Lambda_m}}\exp\{- < \boldsymbol{\beta}_B - \boldsymbol{\beta}_0, H(\mathbf{I}_{\Lambda_m}) >\}p(\mathbf{I}_{\Lambda_m}|\mathbf{I}_{\partial\Lambda_m};\boldsymbol{\beta}_0)} \\
&\approx \frac{\sum_{i=1}^{n}\exp\{- < \boldsymbol{\beta}_A - \boldsymbol{\beta}_0, H(\mathbf{I}_{\Lambda_m}^{(i)}) >\}}{\sum_{i=1}^{n}\exp\{- < \boldsymbol{\beta}_B - \boldsymbol{\beta}_0, H(\mathbf{I}_{\Lambda_m}^{(i)}) >\}}.
\end{aligned}$$

For small hole size m, e.g., $m < 40$, the model $p(\mathbf{I}_{\Lambda_m};\boldsymbol{\beta}_0)$ has enough overlap with $p(\mathbf{I}_{\Lambda_m};\boldsymbol{\beta}_A)$ and $p(\mathbf{I}_{\Lambda_m};\boldsymbol{\beta}_B)$. Thus we can obtain reasonable approximations.

Given the distance computed for small lattices of $m \times m$ pixels, we compute M_{AB} as

$$M_{AB} \approx \frac{|\Lambda_0|}{m^2}M_{AB}^{(m)}.$$

Figure 8 shows a pair of images $\mathbf{I}_A$ and $\mathbf{I}_B$. We synthesize three images for A inside B with $\Lambda_0 = 32 \times 32$, 64×64, and 96×96 pixels respectively. In comparison we also synthesize three images for texture B in A. The estimated divergence $M_{AB}^{(m)}$ and $M_{BA}^{(m)}$ are plotted in figure 9. The KL-divergence rates per pixel are also plotted in the same figure.

We observed that the KL-divergence rate per pixel become almost a constant as the patch size increases. This indicates that the computed distance is a valid estimation. $M_{BA}^{(m)} > M_{AB}^{(m)}$, indicating that A in B should be easier to discriminate than B in A.

It would be interesting to compare the numbers with human perception.

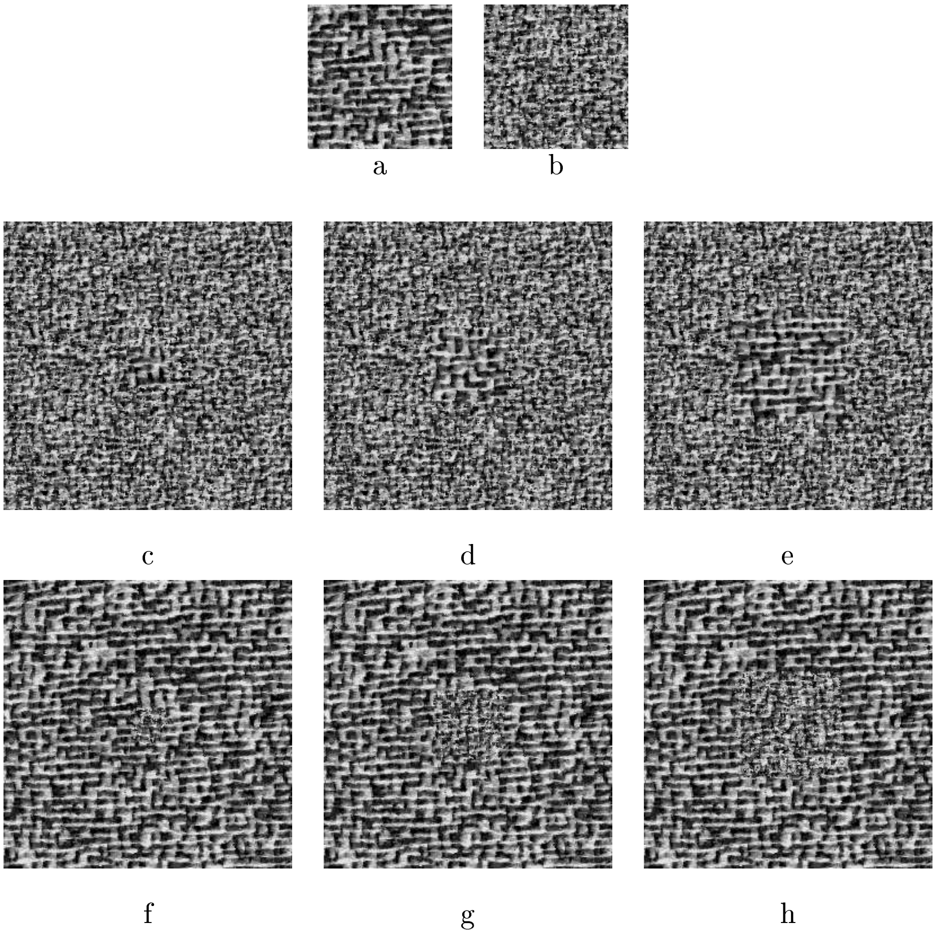
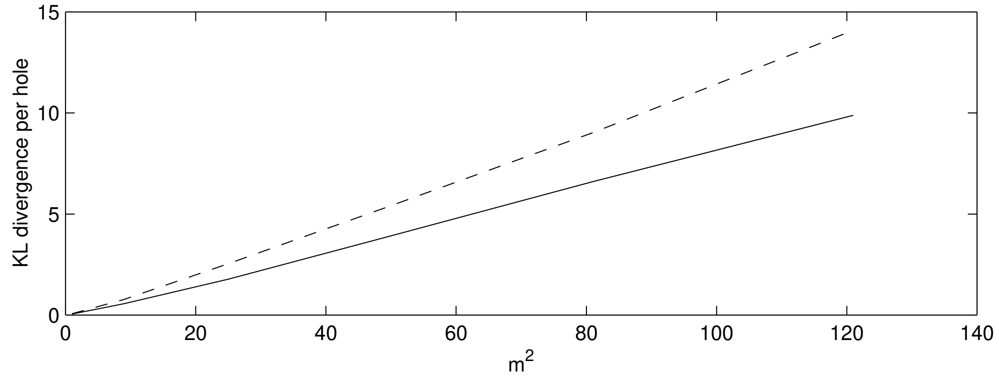
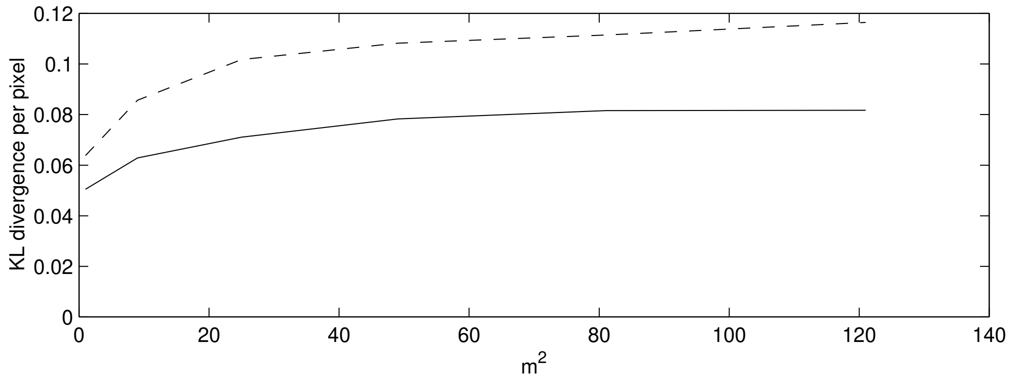


Figure 8: Pop-out experiments with different foreground patch sizes. The size of the whole images is 256×256. a). Texture A. b). Texture B. c)-e). A in B with $|\Lambda_0| = 32 \times 32, 64 \times 64,$ and 96×96 pixels respectively. f)-g). B in A with $|\Lambda_0| = 32 \times 32, 64 \times 64,$ and 96×96 pixels respectively.



a



b

Figure 9: Estimated KL-divergence $M_{AB}^{(m)}$ for the image pair shown in figure 8 plotted against the hole size m^2. a). The dashed curves is $M_{BA}^{(m)}$: texture A in texture B, and the solid curves is $M_{AB}^{(m)}$: B in A. b). The average KL-divergence per pixel $\frac{1}{m^2}M_{BA}^{(m)}$ (dashed) and $\frac{1}{m^2}M_{AB}^{(m)}$ (solid).

8 Discussion – remaining issues

This paper and two of our previous papers[27, 26] study texture phenomena from a mathematical perspective. The proposed paradigm (see figure 5) is quite powerful judged from the recent successes of texture synthesis experiments by ourselves[27, 26] and others[10, 4, 21]. In the following, we pose two major questions and challenges which may lead to further development of texture research.

Question 1. What if texture perception is not bottom-up computation?

The texture theory (both definition and model) is self-consistent and mathematically sound. This theory is built on the philosophy expressed implicitly in the fundamental question asked by Julesz[12]:

> what features and statistics are characteristic of a texture pattern, so that texture pairs that share the same features and statistics cannot be told apart by pre-attentive human visual perception?

Two important assumptions are implied in Julesz's question. One is that textures are "subjective" notion defined by a particular visual system, such as pre-attentive vision. The other assumption is that this notion is determined by *computing a set of feature statistics*. Thus by definition, statistics are extracted deterministically in a bottom-up process. In other words, these statistics are considered as attributes of the observed texture images. This notion is also adopted in recent work on extracting textons[14]. However, if these assumptions are not exactly right, then we may have to investigate texture models of other forms.

Question 2. What are the other factors in texture perception?

Textures should also be studied for attentive vision in a broad context of visual perception, and many other factors may influence our perception of textures. As in color perception, one needs to study texture categorization and mental dimensions in human texture perception. We notice that some interesting non-metric scaling techniques such as multi-dimensional scaling[23] and trajectory mapping[20] have been used in some exploratory studies.

We leave these questions for future investigation.

Acknowledgments

We thank Prof. C.-E. Pfister and Dr. S. Golowich for insightful advice on statistical mechanics, Prof. P. Diaconis, Prof. W.H. Wong, and Dr. A. Yuille for helpful comments, and anonymous reviewers for useful suggestions that greatly improve the presentation of the paper. This work was supported by a grant DAAD19-99-1-0012 from the Army Research Office and a grant NSF-9877127 from National Science Foundation.

References

[1] Azencott, R. Wang, J. P. and Younes, L. 1997. "Texture classification using windowed Fourier filters", *IEEE Trans on PAMI*, Vol. 19, No.2.

[2] Besag, J. 1974. "Spatial interaction and the statistical analysis of lattice systems (with discussion)". *J. Royal Stat. Soc., B*, **36**, 192-236.

[3] Cross, G. R. and Jain, A. K. 1983. "Markov random field texture models". *IEEE Trans. PAMI,* **5**, 25-39.

[4] De Bonet, J. S. and Viola, P. 1997. "A non-parametric multi-scale statistical model for natural images", *Advances in Neural Information Processing*, **10**.

[5] Geman, S. and Geman, D. 1984. "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images", *IEEE Trans. PAMI*, **6**, 721-741.

[6] Gagalowicz, A and Ma, S. D. 1986. "Model driven synthesis of natural textures for 3D scenes". *Computers and Graphics*, **10**, 161-170.

[7] Gibbs, J. W. 1902. *Elementary Principles of Statistical Mechanics.* Yale University Press.

[8] Green, D. M. and Swets, J. A. 1988. *Signal Detection Theory and Psychophysics.* 2nd Edition. Peninsula Publishing. Los Altos, CA.

[9] Griffiths, R. and Ruelle, D. 1971. "Strict convexity ("continuity") of the pressure in lattice system", *Comm. Math. Phys.*, **23**, 169.

[10] Heeger, D. J. and Bergen, J. R. 1995. "Pyramid-based texture analysis/synthesis". *Proceedings of ACM SIGGRAPHS.*

[11] Julesz, B. 1962. "Visual pattern discrimination." *IRE Transactions of Information Theory* IT-8, pp84-92.

[12] Julesz, B. 1995. *Dialogues on Perception.* Bradford Books.

[13] Lanford, O. E. 1973. "Entropy and equilibrium states in classical mechanics". In *Statistical Mechanics and Mathematical Problems*, A. Lenard ed. Springer, 1-113.

[14] Leung, T. and Malik, J. 1999. "Recognizing surfaces using three dimensional textons", *Proc. 7th Int'l Conf. on Computer Vision*, Corfu, Greece.

[15] Lewis, J. T. Pfister, C. E. and Sullivan, W. G. 1995. "Entropy, concentration of probability and conditional limit theorems". *Markov Processes Relat. Fields*, **1**, 319-396.

[16] Liu, F. and Picard, R. W. 1996. "Periodicity, directionality, and randomness: Wold features for image modeling and retrieval", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 7.

[17] Malik, J. and Perona, P. 1990. "Preattentive texture discrimination with early vision mechanisms", *Journal of the Optical Society of America A*, Vol. 7, No. 5.

[18] Martin-Lof, A. 1979. "The equivalence of ensembles and Gibbs' phase rule for classical lattice systems". *J. Stat. Phys.*, **20**, 557-569.

[19] Meng, X. L. and Wong, W. H. 1996. "Simulating ratios of normalizing constants via a simple identity: a theoretical exploration", *Statistica Sinica*, **6**, 831-860.

[20] Richard, W. and Koenderink, J. J. 1995. "Trajectory mapping: a new non-metric scaling techniques", *Perception*, vol. 24, pp1315-1331.

[21] Portilla, J. and Simoncelli, E. P. 1999. "Texture representation and synthesis using correlation of complex wavelet coefficient magnitudes", *Prof. IEEE workshop on Statistical and Computational Theories of Vision*, Fort Collins, CO.

[22] Rosenholtz, R. 1999. "A simple saliency model predicts a number of motion popout phenomena", *Vision Research*, 39, pp3157-3163.

[23] Rubner, Y. Tomasi, C and Guibas, L. 1998. "The Earth Mover's Distance as a metric for image retrieval". *Technical Report STAN-CS-TN-98-86*, Computer Science Department, Stanford University.

[24] Vistnes, R. 1989. "Texture models and image measures for texture discrimination", *Int'l Journal of Computer Vision*, 3, 313-336.

[25] Yuille, A. L. and Coughlan, J. "The limit of Bayes inference: fundamental bounds, order parameters, phase transition", *IEEE Trans. on PAMI*, (To appear in 2000).

[26] Zhu, S. C. Liu, X. W. and Wu, Y. N. "Exploring Texture Ensembles by Efficient Markov Chain Monte Carlo", *IEEE Trans. on PAMI*, (To appear in 2000).

[27] Zhu, S. C. Wu, Y. N. and Mumford, D. B. 1997. "Minimax entropy principle and its application to texture modeling", *Neural Computation*, **9**, 1627-1660.

[28] Zhu, S. C. and Mumford, D. B. 1997. "Prior learning and Gibbs reaction-diffusion." *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol.19, no.11, Nov.

[29] Zhu, S. C. 1999. "Embedding Gestalt laws in Markov random fields", *IEEE Trans. on PAMI*, Vol. 21, No.11.