The g Factor: Relating Distributions on Features to Distributions on Images

James M. Coughlan and A. L. Yuille
Smith-Kettlewell Eye Research Institute,
2318 Fillmore Street,
San Francisco, CA 94115, USA.
Tel. (415) 345-2146/2144. Fax. (415) 345-8455.
Email coughlan@ski.org, yuille@ski.org

April 4, 2001

Abstract

We introduce the g-factor which relates probability distributions on features to distributions on images. It arises when we seek to learn distributions from image data but *it depends only on our choice of features and lattice quantization* and is independent of the training image data. We show that simple, and plausible, approximations of the g-factor can throw light on aspects of Minimax Entropy Learning (MEL) [19], which learns probability distributions on images in terms of Markov Random Fields with clique potentials. Analyzing the g-factor allows us to determine when the clique potentials decouple for different features. Moreover, when the approximations of the g-factor are valid then the clique potentials in MEL can be computed analytically. Finally, we describe ways to extend these approximations by computing approximations to the g-factor offline, thereby enabling rapid methods for computing the clique potentials from new image data. Overall, we seek to give understanding of how MEL relates to alternative methods of learning on images. (In this paper the features we are considering will be extracted from the image by filters – hence we almost always use the terms “features” and “filters” synonymously.)

Draft submitted to SCTV01

1 Introduction

There has recently been a lot of interest in learning probability models for vision. The most common approach is to learn histograms of filter responses or, equivalently, to learn *probability distributions on features.* This has been applied to learning the statistics of textures [15], of images [14], of depth data [12], and foreground and background models for image segmentation [11], [13], [2]. In this paper the features we are considering will be extracted from the image by filters – hence we almost always use the terms “features” and “filters” synonymously.

An alternative approach, however, is to learn probability distributions *on the images themselves.* The Minimax Entropy Learning (MEL) theory [19] is a bold attempt to do this in which the maximum entropy principle is used to learn distributions constrained by the observed histograms of feature responses (with a feature pursuit stage to determine which features should be used to construct the probability distribution). A key aspect of this approach that it learns *clique potentials* on filter outputs to produce a Markov random field [10]. So, for example, when applied to texture it gives a way to unify the filter based approaches (which are often very effective) with the Markov random field approaches (which are theoretically attractive).

As we describe in this paper, distributions on images and on features can be related by a g-factor (such factors arise in statistical physics, see [8]). It can be considered a *phase factor* because it relates different representations of the same physical systems. Understanding the form of the g-factor, and making good approximations to it, enable us to relate distributions on images to distributions on features. This helps determine the tradeoffs between the image and feature based approaches.

In particular, understanding the g-factor helps throw light on MEL and give understanding of some of its more unintuitive aspects. For example, in MEL feature histograms are fed into a stochastic optimization procedure which outputs clique potentials. Can one get understanding of why the clique potentials take the form they do? Moreover, the clique potentials for different filters seem to be decoupled (i.e. the two clique potentials corresponding to two features A and B are identical whether we learn them jointly or independently). When, and why, does this occur? MEL proposes a filter pursuit method to determine which filters are best to use. Can one get some simple intuitive understanding of this?

The g-factor is determined by the form of the features chosen and *the spatial lattice and quantization of the image grey-levels.* It is completely independent of the training image data. Instead we can think of the g-factor as corresponding to learning probability distributions where the training data corresponds to the *uniform distribution* on the set of all images. As we will show, the form of the feature histograms for this uniform image distribution play an important role in determining how easy it is to learn their clique potentials. It should be stressed that the choice of image lattice and grey-level quantization can make a big difference to the g-factor and hence to the probability distributions which are the output of MEL. There are some paradoxical results. For example, our work shows that simple approximations to the g-factor are best when the number of image grey-levels is large (i.e. the quantization is fine) while many practical algorithms for MEL have, for computational reasons, worked with coarse quantization. In other words, the problem gets easier the larger the number of grey-levels we allow.

In this paper we describe approximations to the g-factor and argue for their validity (in particular when the quantization becomes fine). The approximations enable us to obtain analytic expressions for the clique potentials in MEL. We hope this approach will give some insight into MEL and may help guide the construction of effective algorithms. Moreover, our analysis helps show how simpler methods for learning can be obtained as approximations to Minimax Entropy Learning.

Finally, we emphasize the bigger issue here. How do we relate probability distributions on images to probability distributions on features extracted from images? The latter are often far easier to calculate but may *not correspond to consistent distributions* on images.

An early version of this work appeared in NIPS'98, [4], where we introduced the g-factor and investigated ways of approximating it. This paper emphasizes, and extends, the second approach (which was only briefly mentioned in [4]).

In Section (2), we briefly review Minimax Entropy Learning. Section (3) introduces the g-factor and determines conditions for when clique potentials are decoupled. In Section (4) we describe a simple approximation which enables us to learn the clique potentials analytically. Section (5) shows how this approximation can be extended if the size N of the image is sufficiently large.

2 Minimax Entropy Learning

Suppose we have training image data which we assume has been generated by an (unknown) probability distribution $P_{True}(\vec{x})$ where $\vec{x}$ represents an image. The task is to learn a probability distribution that approximates $P_{True}(\vec{x})$.

We attempt to approximate $P_{True}(\vec{x})$ by observing image statistics $\vec{\phi}(\vec{x})$ [7]. Then we apply the maximum entropy principle with the constraint that these statistics have observed (mean) values $\vec{\psi}_{obs}$. This gives:

$$P(\vec{x}|\vec{\lambda}) = \frac{e^{\vec{\lambda}\cdot\vec{\phi}(\vec{x})}}{Z[\vec{\lambda}]}, \tag{1}$$

where $\vec{\lambda}$ is a parameter chosen such that $\sum_{\mathbf{x}} P(\vec{x}|\lambda)\phi(\vec{x}) = \vec{\psi}_{obs}$. Or equivalently, so that $\frac{\partial \log Z[\vec{\lambda}]}{\partial \vec{\lambda}} = \vec{\psi}$.

(This result follows by maximizing the entropy $-\sum_{\vec{x}} P(\vec{x}) \log P(\vec{x})$ subject to the constraints $\sum_{\vec{x}} P(\vec{x})\vec{\phi}(\vec{x}) = \vec{\psi}_{obs}$. Equivalently we can obtain equation (1) by assuming $P(\vec{x}|\vec{\lambda})$ is of exponential form $\sim e^{\vec{\lambda}\cdot\vec{\phi}(\vec{x})}$ where $\vec{\lambda}^* = \arg\max_{\vec{\lambda}} e^{\vec{\lambda}\cdot\vec{\psi}_{obs}}/Z[\vec{\lambda}]$. It is straightforward to show that $\vec{\lambda}^*$ is the Maximum Likelihood estimate of $\vec{\lambda}$.)

The Minimax Entropy Learning (MEL) approach [19] proceeds in two stages. First, it uses the Maximum Entropy principle described above to generate probability distributions, see equation (1), for any choice of sufficient statistics $\vec{\phi}(.)$. Secondly, it uses a Minimum Entropy principle to determine which statistics should be used. Intuitively, statistics which yield Maximum Entropy distributions with *small* entropy are preferred because the smaller the entropy the "sharper" the model. (See Coughlan and Yuille [4] for a discussion of how both aspects of Minimax Entropy have very simple interpretations in terms of Amari's theory of information geometry [1].)

In practice, MEL has usually chosen the sufficient statistics to be the histograms of filter responses. (There is no reason in principle why other statistics should not be considered but, historically, MEL has concentrated on histograms.) The $\vec{\psi}_{obs}$ are therefore the empirical histograms of the filters.

More precisely, if $\vec{x}$ represents an intensity image then if we apply a shift-invariant filter $f(.)$ to the image we obtain a set of responses $\{f_i(\vec{x}) : i = 1, ..., N\}$ where i labels position in the image and N denotes the number of pixels in the image. The empirical histograms $\vec{\psi}$ can be written as $\psi_a = \phi_a(\vec{x}) = \frac{1}{N}\sum_{i=1}^{N} \delta_{a,f_i(\vec{x})}$ where a indicates the (quantized) filter response values. We define Q to be the number of values a can take, so that the components ψ_a of $\vec{psi}$ are indexed in the range $a = 1, ..., Q$. (We also denote ψ_a by $\psi(a)$.) Observe that, by construction, we have $\sum_{a=1}^{Q} \psi(a) = 1$. Moreover, all components of $\vec{\psi}$ are non-negative. Indeed $\vec{\psi}$ can be interpreted as a probability distribution on feature space.

Choosing the statistics to be a filter histogram makes the resulting MEL model into a simple MRF or Gibbs form. To see this, observe that

$$\vec{\lambda} \cdot \vec{\phi}(\vec{x}) = \frac{1}{N} \sum_{a=1}^{Q} \sum_{i=1}^{N} \lambda(a) \delta_{a,f_i(\vec{x})} = \frac{1}{N} \sum_{i=1}^{N} \lambda(f_i(\vec{x})), \tag{2}$$

and so $P(\vec{x}|\vec{\lambda})$ becomes a Gibbs distribution with clique potentials given by $\lambda(f_i(\vec{x}))$. This determines a Markov random field with the clique structure given by the filters $\{f_i\}$.

To make this more concrete, consider a one-dimensional image where the filters are chosen to be difference operators, so that $f_i(\vec{x}) = x_{i+1} - x_i$. The empirical statistics of such filters have been evaluated on many intensity images, and depth images, and typically take the form given in figure (1) (left panel). When the potentials corresponding to these histograms are estimated by MEL they are typically of the form given by figure (1) (right panel). This gives a Markov random field probability distribution of form:

$$P(\vec{x}) = \frac{1}{Z} e^{\sum_{i=1}^{N} \lambda(x_{i+1} - x_i)}, \tag{3}$$

which is similar to models proposed in the early eighties by Blake and Zisserman [3] and Geman and Geman [10]. (These models may appear to be different because they included additional line-process variables but these can be eliminated, see Geiger and Girosi [9], and then the resulting models are very similar to equation (3).)

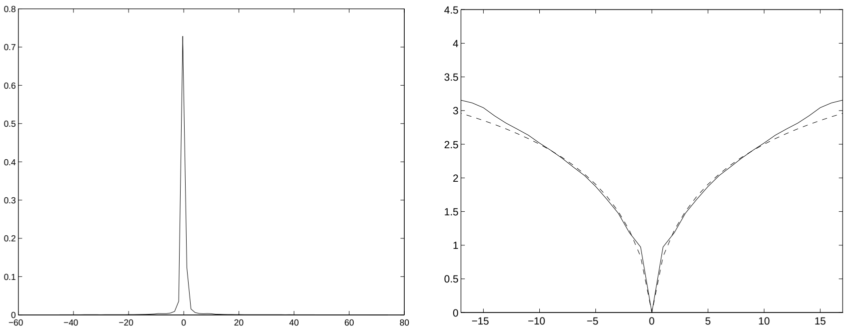

Figure 1: Left panel: The typical histogram $\{\psi_{obs}(a)\}$ of a difference filter when evaluated on image or range data. Right panel: the corresponding clique potentials $\{-\lambda(a)\}$ are similar to those used by Blake and Zisserman or Geman and Geman but sharper at the bottom, leading to less small-scale, fractal, fluctuations.

The Minimum Entropy stage of MEL says that we should evaluate the statistics by computing the entropy $-\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}) \log P(\vec{x}|\vec{\lambda})$ for each choice of statistic (with small entropies being preferred). A *filter pursuit* procedure was described to determine which filters should be considered.

3 The g-Factor

This section defines the g-factor in subsection (3.1) and starts investigating its properties in subsection (3.2). In particular, when, and why, do clique potentials decouple? More precisely, when do the potentials for filters A and B learned simultaneously differ from the potentials for the two filters when they are learnt independently?

3.1 Basic Properties of the g-Factor

We now address these issues by introducing the g-factor $g(\vec{\psi})$. This is defined for any value $\vec{\psi}$ of the statistics $\vec{\phi}(\vec{x})$ by:

$$g(\vec{\psi}) = \sum_{\vec{x}} \delta_{\vec{\phi}(\vec{x}), \vec{\psi}}. \tag{4}$$

It is important to realize that the g-factor is completely *independent* of the observations $\vec{\psi}_{obs}$. It *depends only on the form of the filters* $\{f_i\}$ *used to compute the statistics* $\vec{\phi}$ *and on the choice of lattice and quantization.* It is used to relate probability distributions $\hat{P}(.)$ on feature space to probability distributions $P(\vec{x})$ on image space.

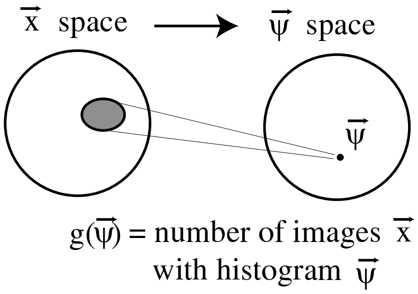
$\vec{x}$ space
$\vec{\psi}$ space
$\vec{\psi}$
$g(\vec{\psi})$ = number of images $\vec{x}$
with histogram $\vec{\psi}$

Figure 2: The g-factor $g(\vec{\psi})$ counts the number of images $\vec{x}$ that have statistics $\vec{\psi}$. Note that the g-factor depends only on the choice of filters and is independent of the training image data.

The g-factor is essentially a combinational factor which counts the number of ways that one can obtain statistics $\vec{\psi}$, see figure (2). One can give it a probabilistic interpretation by dividing it by L^N where L is the number of values that any x_i can take (i.e. L is the number of grey-scale levels, N is the total number of pixels on the lattice, and L^N is the total number of all possible images). Then $\hat{P}_0(\vec{\psi}) = (1/L^N) g(\vec{\psi})$ is the induced distribution on $\vec{\psi}$ where the $\vec{x}$ are assumed to be distributed by *the uniform distribution* $U(\vec{x}) = 1/L^N$ for any image $\vec{x}$. In other words, we can define the g-factor distribution:

$$\hat{P}_0(\vec{\psi}) = \frac{1}{L^N} g(\vec{\psi}), \tag{5}$$

and consider it to be *the default distribution* on $\vec{\psi}$ corresponding to complete lack of structure in the image (i.e. images are generated by the uniform distribution). The top row of figure (3) shows a sample image and histogram from the uniform distribution, along with corresponding samples of a *natural* image for contrast.

More generally, we can use the g-factor to compute the induced distribution $\hat{P}(\vec{\psi}|\vec{\lambda})$ on the statistics determined by MEL, see equation (1), and also the partition function $Z[\vec{\lambda}]$. These are given by:

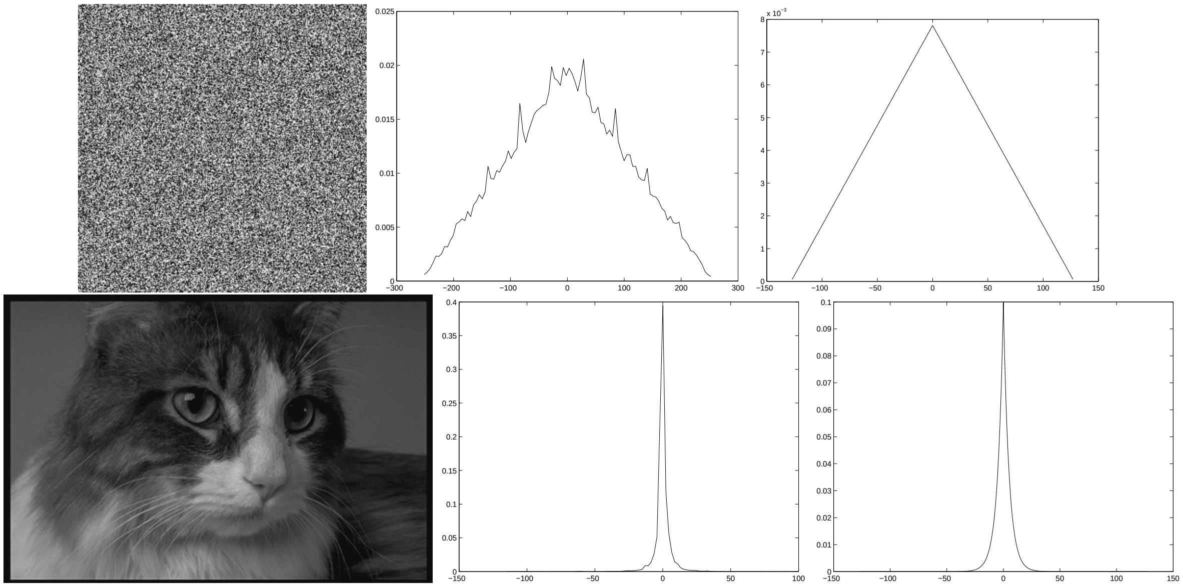

Figure 3: Image samples (left column), the corresponding empirical histograms of $\partial/\partial x$ (middle column), and the mean histograms (right column). The image on the top left was drawn from the uniform distribution $U(\vec{x})$. Its empirical histogram (middle) is close to the mean value $\{\alpha(a)\}$ induced by $U(\vec{x})$, which is calculated exactly [5] and shown on the right. On the bottom left, a natural image with a histogram (middle) that is close to the mean value (right) across a dataset of natural images.

$$\hat{P}(\vec{\psi}|\vec{\lambda}) = \sum_{\vec{x}} \delta_{\vec{\psi},\vec{\phi}(\vec{x})} \frac{e^{\vec{\lambda}\cdot\vec{\phi}(\vec{x})}}{Z[\vec{\lambda}]} = \frac{g(\vec{\psi})e^{\vec{\lambda}\cdot\vec{\psi}}}{Z[\vec{\lambda}]}, \quad Z[\vec{\lambda}] = \sum_{\vec{\psi}} g(\vec{\psi})e^{\vec{\lambda}\cdot\vec{\psi}}. \tag{6}$$

Observe that both $\hat{P}(\vec{\psi}|\vec{\lambda})$ and $\log Z[\vec{\lambda}]$ are sufficient for computing the parameters $\vec{\lambda}$. The $\vec{\lambda}$ can be found by solving either of the following two (equivalent) equations:

$$\sum_{\vec{\psi}} \hat{P}(\vec{\psi}|\vec{\lambda})\vec{\psi} = \vec{\psi}_{obs}, \quad or \quad \frac{\partial \log Z[\vec{\lambda}]}{\partial \vec{\lambda}} = \vec{\psi}_{obs}, \tag{7}$$

which shows that *knowledge of the g-factor and $e^{\vec{\lambda}\cdot\vec{\psi}}$ are all that is required to do MEL.*

Observe from equation (6) that we have $\hat{P}(\vec{\psi}|\vec{\lambda}=0) = P_0(\vec{\psi})$. In other words, setting $\vec{\lambda}=0$ corresponds to a uniform distribution on the images $\vec{x}$.

3.2 Decoupling Filters

We now derive an important property of the minimax entropy approach. As mentioned earlier, it often seems that the potentials for filters A and B decouple. In other words, if one applies MEL to two filters A, B simultaneously by letting $\vec{\psi} = (\vec{\psi}^A, \vec{\psi}^B)$, $\vec{\lambda} = (\vec{\lambda}^A, \vec{\lambda}^B)$, and $\vec{\psi}_{obs} = (\vec{\psi}^A_{obs}, \vec{\psi}^B_{obs})$, then the solutions $\vec{\lambda}^A, \vec{\lambda}^B$ to the equations

$$\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}^A, \vec{\lambda}^B)(\vec{\phi}^A(\vec{x}), \vec{\phi}^B(\vec{x})) = (\vec{\psi}^A_{obs}, \vec{\psi}^B_{obs}), \tag{8}$$

are the same (approximately) as the solutions to the equations $\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}^A)\vec{\phi}^A(\vec{x}) = \vec{\psi}^A_{obs}$ and $\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}^B)\vec{\phi}^B(\vec{x}) = \vec{\psi}^B_{obs}$.

We illustrate this decoupling with an example where the features are $\partial/\partial x$ and $\partial/\partial y$. The clique potentials found by MEL are given in figure (4).

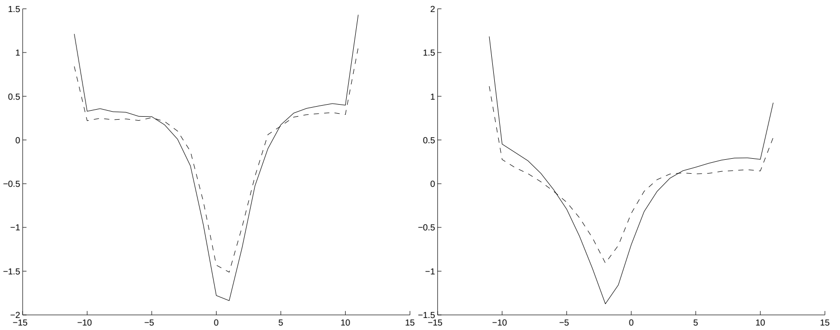

Figure 4: Evidence for decoupling of features. The left and right panels show the clique potentials learnt for the features $\partial/\partial x$ and $\partial/\partial y$ respectively. The solid lines give the potentials when they are learnt individually. The dashed lines show the potentials when they are learnt simultaneously. (The stochastic nature of these computations means that the estimates of the potentials may not have converged to their true values, and so it is possible that the potentials are even more nearly similar.) Figure courtesy of Prof. Xiuwen Liu, USF.

We now show how this decoupling property arises naturally if the g-factor for the two filters factorizes. This factorization, of course, is a property only of the form of the statistics and is *completely independent of whether the statistics of the two filters are dependent for the training data.*

Property I: *Suppose we have two sufficient statistics* $\vec\phi^A(\vec x), \vec\phi^B(\vec x)$ *which are independent on the lattice in the sense that* $g(\vec\psi^A, \vec\psi^B) = g^A(\vec\psi^A) g^B(\vec\psi^B)$, *then we have:*

$$\log Z[\vec\lambda^A, \vec\lambda^B] = \log Z^A[\vec\lambda^A] + \log Z^B[\vec\lambda^B], \quad \hat P(\vec\psi^A, \vec\psi^B) = \hat P^A(\vec\psi^A) \hat P^B(\vec\psi^B), \tag{9}$$

which implies that the parameters $\vec\lambda^A, \vec\lambda^B$ *can be solved from the independent equations*

$$\frac{\partial \log Z^A[\vec\lambda^A]}{\partial \vec\lambda^A} = \vec\psi^A_{obs}, \quad \frac{\partial \log Z^B[\vec\lambda^B]}{\partial \vec\lambda^B} = \vec\psi^B_{obs} \quad \textit{or} \quad \sum_{\vec\psi^A} \hat P^A(\vec\psi^A)\vec\psi^A = \vec\psi^A_{obs}, \ \sum_{\vec\psi^B} \hat P^B(\vec\psi^B)\vec\psi^B = \vec\psi^B_{obs}. \tag{10}$$

Moreover, the resulting distribution $P(\vec x)$ *can be obtained by multiplying the distributions* $(1/Z^A)e^{\vec\lambda^A \cdot \vec\psi^A(\vec x)}$ *and* $(1/Z^B)e^{\vec\lambda^B \cdot \vec\psi^B(\vec x)}$ *together.*

The point here is that the potential terms for the two statistics $\vec\psi^A, \vec\psi^B$ decouple if the phase factor $g(\vec\psi^A, \vec\psi^B)$ can be factorized. *We conjecture that this is effectively the case for many linear filters used in vision processing.* For example, it is plausible that the g-factor for features $\partial/\partial x$ and $\partial/\partial y$ factorizes – and figure (4) shows that their clique potentials do decouple (approximately). Clearly, if factorization between filters occurs then it gives great simplification to the system.

It may, however, be questioned whether this decoupling is desirable. Recall that this "factorization" is purely a property of the filters and the lattice (plus quantization) and is *completely independent* of the training image data. If the g-factor factorizes then MEL (using the feature marginals) will imply that $\hat P(\vec\psi^A, \vec\psi^B) = \hat P^A(\vec\psi^A)\hat P^B(\vec\psi^B)$ and so will *predict* that the joint histograms $\vec\psi^A_{obs}, \vec\psi^B_{obs}$ are statistically independent and uncorrelated. If the observed feature histograms (of the training image data) *are correlated* then MEL is clearly suboptimal (if the marginal histograms are used).

Recall that the g-factor is proportional to the distribution of the features when the input images are uniformly distributed. This enables us to define a diagnostic test which warns us whenever the features are independent for uniformly distributed images but are *dependent* for the training data images. If this warning occurs then we should the joint histograms of the features, or some other statistics, as input into MEL rather than the feature marginals.

4 Approximating the g-factor for a Single Histogram

We now consider the case where the statistic is a single histogram. Our aim is to understand why features whose histograms are of stereotypical shape (see left panel of figure (1)) give rise to potentials of the form given by the right panel of figure (1).

Our results, of course, can be directly extended to multiple histograms if the filters decouple, see subsection (3.2). We first describe the approximation in subsection (4.1) and then explore its relevance for filter pursuit (i.e. the "min" part of Minimax Entropy Learning) in subsection (4.2).

4.1 The Multinomial

We assume that the statistic $\vec{\psi}$ is a histogram of form:

$$\psi(a) = \frac{1}{N}\sum_{i=1}^{N}\delta_{f_i,a}, \tag{11}$$

where the $\{f_i : i = 1, ..., N\}$ are filter outputs quantized to take Q discrete values labelled by a. The terms $\psi(a)$ are the components of the vector $\vec{\psi}$.

For statistics of this form, it is convenient to rescale the $\vec{\lambda}$ variables by N so that we have:

$$P(\vec{x}) = \frac{e^{N\vec{\lambda}\cdot\vec{\phi}(\vec{x})}}{Z[\vec{\lambda}]}, \quad \hat{P}(\vec{\psi}) = g(\vec{\psi})\frac{e^{N\vec{\lambda}\cdot\vec{\psi}}}{Z[\vec{\lambda}]}, \tag{12}$$

where we have written $P(\vec{x})$ and $\hat{P}(\vec{\psi})$ as shorthand for $P(\vec{x}|\vec{\lambda})$ and $\hat{P}(\vec{\psi}|\vec{\lambda})$, respectively.

We now consider the approximation that the filter responses $\{f_i\}$ are *independent of each when the images are uniformly distributed.* For this assumption to be valid it does not matter *at all* whether the filter responses are dependent or not for the real data.

One way to verify this assumption is by calculating the Kullback-Leibler divergence between the distribution $\hat{P}_0(f_1, ..., f_N)$ induced by the lattice and the factorized approximation $\prod_{i=1}^{N}\hat{P}_0(f_i)$. We normalize this divergence by the entropy of the distribution $\hat{P}_0(f_1, ..., f_N)$, i.e. yielding the expression $M = D(\hat{P}_0(f_1, ..., f_N)|| \prod_{i=1}^{N}\hat{P}_0(f_i))/H(\hat{P}_0(f_1, ..., f_N))$, and evaluate it for different values of image size (N) and different quantization levels Q. Our computer simulations show that the approximation becomes increasingly better as N and Q becomes large, see figure (5). (We note that computer implementations of MEL have typically required *coarse* quantization of the image lattice to speed up calculations.)

If the filters responses are independent of each when the images are uniformly distributed then we call this the *multinomial approximation*, because it implies that we can express the phase factor as being proportional to a multinomial distribution:

$$g(\vec{\psi}) = L^N \frac{N!}{(N\psi_1)!...(N\psi_Q)!}\alpha_1^{N\psi_1}...\alpha_Q^{N\psi_Q}, \quad \hat{P}_0(\vec{\psi}) = \frac{N!}{(N\psi_1)!...(N\psi_Q)!}\alpha_1^{N\psi_1}...\alpha_Q^{N\psi_Q}, \tag{13}$$

where $\sum_{a=1}^{Q}\psi_a = 1$ (by definition) and the $\{\alpha_a\}$ are the means of the components $\{\psi_a\}$ with respect to the distribution $\hat{P}_0(\vec{\psi})$. As we will describe later, the $\{\alpha_a\}$ will be determined by the filters $\{f_i\}$. See technical report [5] for details of how to compute the $\{\alpha_a\}$.

This approximation enables us to calculate MEL *analytically.*

Theorem *With the multinomial approximation we can compute the log partition function to be:*

$$\log Z[\vec{\lambda}] = N\log L + N\log\{\sum_{a=1}^{Q} e^{\lambda_a + \log\alpha_a}\}, \tag{14}$$

and the "potentials" $\{\lambda_a\}$ can be solved in terms of the observed data $\{\psi_{obs,a}\}$ to be:

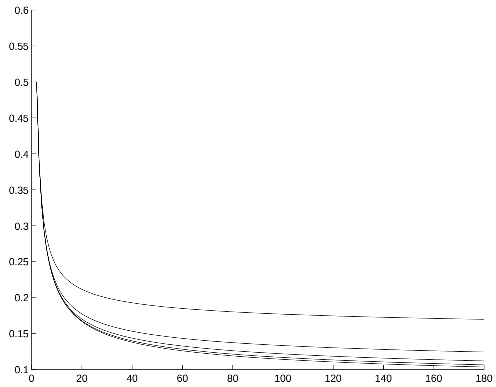

Figure 5: Evidence for the Multinomial approximation. The feature is $\partial/\partial x$ and we plot the Kullback-Leibler divergence normalized by the entropy of the distribution $\hat{P}_0(\{f_i\})$ (i.e. the quanity M, see text for details). This quantity is plotted on the vertical axis as a function of L, the number of grayscale levels (which is related to Q for this particular feature by the equation $Q = 2L - 1$). Five plots are shown, one for each of five image lattice sizes ($N = 10^2, 20^2, 30^2, 40^2, 50^2$ from top curve to bottom curve). The results show that the approximation gets better as the image size N and the quantization levels Q get large. (Note: these calculations are exact and do not use MCMC or other stochastic techniques, see [5].)

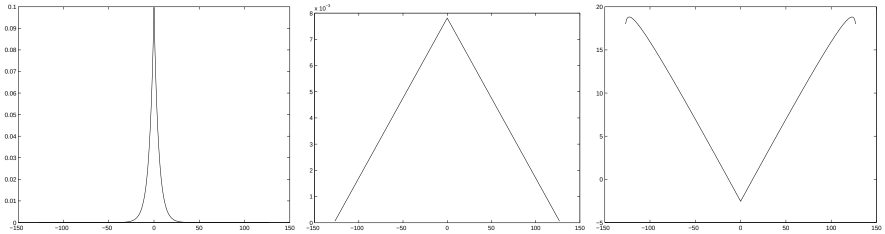

Figure 6: Left to right: $\{\psi_{obs}(a)\}$ measured from a dataset of natural images, $\{\alpha(a)\}$ calculated exactly, and $\{-\lambda(a)\}$ as given by multinomial approximation for $\partial/\partial x$.

$$\lambda_a = \log \frac{\psi_{obs,a}}{\alpha_a}, \quad a = 1, ..., Q. \tag{15}$$

We note that there is an ambiguity $\lambda_a \mapsto \lambda_a + K$ *where* K *is an arbitrary number (recall that* $\sum_{a=1}^{Q} \psi(a) = 1$*). We fix this ambiguity by setting* $\vec{\lambda} = 0$ *if* $\vec{\alpha} = \vec{\psi}_{obs}$ *(in other words,* $\vec{\lambda} = 0$ *if the histogram* $\vec{\psi}_{obs}$ *of the filter on the training data is equal to the histogram* $\vec{\alpha}$ *of the filter for uniformly distributed input data).*

Proof. *We have* $Z[\vec{\lambda}] = \sum_{\vec{\psi}} e^{N\vec{\lambda}\cdot\vec{\psi}} g(\vec{\psi})$*. We use the multinomial approximation for* $g(\vec{\psi})$ *and the fact that* $e^{N\vec{\lambda}\cdot\vec{\psi}} \times \prod_{a=1}^{Q} \alpha_a^{N\psi_a} = \prod_{a=1}^{Q} e^{N\psi_a\{\lambda_a + \log\alpha_a\}}$*. We now sum over the* $\{\psi_a\}$ *using properties of multinomial distribution to get* $Z[\vec{\lambda}] = \{\sum_{a=1}^{Q} e^{\lambda_a + \log\alpha_a}\}^N$*. We can then solve the equations* $(\partial \log Z[\vec{\lambda}])/(\partial\vec{\lambda}) = N\vec{\psi}_{obs}$ *(recall we rescaled* $\vec{\psi}$ *above) to determine the* $\{\lambda_a\}$ *analytically.*

We see at once that this simple approximation gives the typical potential forms generated by Markov Chain Monte Carlo (MCMC) algorithms for Minimax Entropy Learning. Compare the multinomial approximation results of figure (6) to those of figure (1).

4.2 Filter Pursuit for the Multinomial Approximation

Filter pursuit is required to determine which filters carry most information. MEL [19] prefers filters (statistics) which give rise to low entropy distributions (this is the "Min" part of Minimax).

It is straightforward to show that the entropy for distributions generated by MEL are of form:

$$H(P) = -\sum_{\vec{x}} P(\vec{x}|\vec{\lambda}) \log P(\vec{x}|\vec{\lambda}) = \log Z[\vec{\lambda}] - \sum_{a=1}^{Q} \lambda_a \psi_a. \tag{16}$$

For the multinomial approximation it is straightforward to show that the entropy is:

$$N \log L - N \sum_{a=1}^{Q} \psi_a \log \frac{\psi_a}{\alpha_a}. \tag{17}$$

This gives a very simple interpretation of the MEL feature pursuit procedure. It says, very intuitively, that we should prefer to pick filters whose statistical response to the image training data is *as large as possible* from their responses to uniformly distributed images. This is measured by the Kullback-Leibler divergence $\sum_{a=1}^{Q} \psi_a \log \frac{\psi_a}{\alpha_a}$.

Recall that if the multinomial approximation is used for multiple filters then we should simply add together the entropies of different filters.

5 Beyond Multinomial: Large N Approximations

The multinomial is a nice approximation because it gives very simple results (and is plausible in some cases). We now discuss how to go beyond it.

5.1 Large N Behaviour

First, we observe that $\hat{P}(\vec{\psi}) = g(\vec{\psi}) e^{N\vec{\lambda}\cdot\vec{\psi}}$. We can write $g(\vec{\psi}) = e^{N\rho(\vec{\psi})}$ where the scaling argument comes from Wu *et al* [17]. Then we have:

$$\hat{P}(\vec{\psi}) = e^{N\{\rho(\vec{\psi}) + \vec{\lambda}\cdot\vec{\psi}\}}. \tag{18}$$

For large N, the mean $\sum_{\vec{\psi}} \vec{\psi} \hat{P}(\vec{\psi})$ will be dominated by $\vec{\psi}^* = \arg\max_{\vec{\psi}} \{\rho(\vec{\psi}) + \vec{\lambda}\cdot\vec{\psi}\}$. In other words, $\vec{\psi}^*$ satisfies:

$$\frac{\partial \rho(\vec{\psi})}{\partial \vec{\psi}}(\vec{\psi}^*) + \vec{\lambda} = 0. \tag{19}$$

Now to relate this to real data $\vec{\psi}_{obs}$ we need to find the parameter $\vec{\lambda}$ which satisfies equation (19) with $\vec{\psi}^* = \vec{\psi}_{obs}$. This gives us a simple equation for $\vec{\lambda}$:

$$\vec{\lambda} = -\frac{\partial \rho(\vec{\psi})}{\partial \vec{\psi}}(\vec{\psi}_{obs}). \tag{20}$$

The difficulty is in estimating $\rho(\vec{\psi})$. If we can do this analytically then MEL would simply reduce to evaluating the derivative of $\rho(\vec{\psi})$ (at least in the large N limit).

The question is how to approximate $\rho(\vec{\psi})$. The factorization approach gives one possibility which we will explore in the next subsection, see subsection (5.2), where it can be obtained as the limit of the multinomial case as $N \mapsto \infty$. An alternative quadratic approximation was presented by the authors in [4]. We showed that this approximation gave reasonable values for the potentials for some image statistics. As an approximation, however, it was limited by having no clear intuition behind it (unlike the multinomial approximation).

The large N analysis also clears up an apparent paradox about the distribution $P(\vec{x}|\vec{\lambda}) = \frac{e^{\vec{\lambda}\cdot\vec{\phi}(\vec{x})}}{Z[\vec{\lambda}]}$. The paradox is that the *most probable state $\vec{x}^*$ is not necessarily one that obeys the condition that $\vec{\phi}(\vec{x}^*) = \vec{\psi}_{obs}$.*

To check this observe that the conditions say that the *mean* value of $\vec{\psi}$ is equal to $\vec{\psi}_{obs}$ but the mode of $\vec{\psi}$ might be quite different. (In fact they will be the same when the multinomial approximation is made, see subsection (5.2).) But for $N \mapsto \infty$ the mean and the mode become identical. In this limit MEL becomes equivalent to the Julesz ensemble formulated in [17], in which all images $\vec{x}^*$ such that $\vec{\phi}(\vec{x}^*)$ is close to $\vec{\psi}_{obs}$ become equally probable and all other images have zero probability.

5.2 The Large N Limit for Multinomials

We first investigate the large N behaviour of the multinomial approximation (or factorization assumption). This is helpful for developing more advanced approximations (even though we can solve the multinomial case exactly).

Consider the large N limit of $\hat{P}(\vec{\psi})$ assuming the multinomial approximation. By applying Stirling's approximation $\log N! \approx N \log N - N$ to the expression for the phase factor $g(\vec{\psi})$ we obtain $\log g(\vec{\psi}) \approx -N \sum_{a=1}^{Q} \psi_a \log \psi_a + N \sum_{a=1}^{Q} \psi_a \log \alpha_a$. This gives us the expression:

$$\hat{P}(\vec{\psi}) = \frac{e^{-N \sum_{a=1}^{Q} \psi_a \log \frac{\psi_a}{\alpha_a e^{\lambda_a}}}}{Z[\vec{\lambda}]}. \tag{21}$$

Observe that the exponent of equation (21) is of Kullback-Leibler divergence between the $\{\psi_a\}$ and the $\{\alpha_a e^{\lambda_a}\}$.

As $N \mapsto \infty$, the distribution of equation (21) will become sharply peaked at $\vec{\psi}^*$ given by $\psi_a^* = k\alpha_a e^{\lambda_a}, \quad a = 1, .., Q$ where k is a normalization constant (to ensure that $\sum_{a=1}^{Q} \psi_a = 1$). We can now solve for the potentials $\{\lambda_a\}$ by requiring that $\vec{\psi}^* = \vec{\psi}_{obs}$ (the observed data). This gives the same result as before. This can be considered a saddle point or Laplace approximation which is valid as $N \mapsto \infty$.

Sanov's theorem [6] can be used to put bounds on the probabilities of the errors which are caused by this large N approximation. It can be used to determine how much data is required in order to obtain accurate estimates of the potentials $\{\lambda_a\}$. It will also tell us how fast the asymptotic results kick in.

5.3 Beyond Multinomial

We now want to go beyond the multinomial approximation for $g(\vec{\psi})$.

To do this, we reformulate the problem in terms of the $\{f_i\}$ filter responses which are used to construct the histogram. This is exactly analogous to the derivation which we used for the g-factor. See figure (7) for a schematic showing the relationships among the g-factor and two other related combinatorial factors we introduce in this section.

We define an h-factor on the $\{f_i\}$ by counting the number of images consistent with specified filter responses:

$$h(\{f_i\}) = \sum_{\vec{x}} \prod_{i=1}^{N} \delta_{f_i(\vec{x}), f_i}. \tag{22}$$

We can then get an induced distribution on the $\{f_i\}$ to be:

$$\hat{P}(\{f_i\}) = h(\{f_i\}) \frac{e^{N\vec{\lambda} \cdot \vec{\psi}(\{f_i\})}}{Z[\vec{\lambda}]}. \tag{23}$$

From this we can obtain a distribution on $\vec{\psi}$ by:

$$\hat{P}(\vec{\psi}) = \sum_{\{f_i\}} \delta_{\vec{\psi}, \vec{\phi}(\{f_i\})} \hat{P}(\{f_i\}). \tag{24}$$

In this picture, the multinomial approximation for $g(\vec{\psi})$ is equivalent to a *factorizable* assumption for $h(\{f_i\})$. In other words, if we assume that $h(\{f_i\}) = \prod_{i=1}^{N} p(f_i)$ (for some choice of $p(.)$) then $g(\vec{\psi})$ is a multinomial. (We obtain $g(\vec{\psi})$ from $h(\{f_i\})$ by using equation (24) in the special case where $\vec{\lambda} = 0$.)

We also observe that the factorization assumption on the $h(\{f_i\})$ is equivalent, after normalization, to applying a maximum entropy principle to estimate $h(\{f_i\})$ using sufficient statistics $\phi_a(\{f_i\}) = (1/N)\sum_i \delta_{a,f_i}$. In other words, this uses the same sufficient statistics on the filter responses to estimate $h(\{f_i\})$ *before we have any data* as we do *after we get the data.*

To go further, we estimate the $h(\{f_i\})$ by maximum entropy using additional statistics. The mathematics greatly simplifies if we use statistics of form $\vec{F}(\vec{\phi}(\{f_i\}))$ and apply maximum entropy to this. This will give:

$$h(\{f_i\}) = \frac{e^{N\vec{\mu}\cdot\vec{F}(\vec{\phi}(\{f_i\}))}}{\hat{Z}[\vec{\mu}]}, \tag{25}$$

where $\vec{\mu}$ is determined by MEL where the input are the filter responses $\{f_i\}$ with image data generated by the uniform distribution $U(\vec{x})$. Determining $\vec{\mu}$ can be done off-line (i.e. it is independent of the training image data).

We can now calculate the distribution on $\{f_i\}$ caused by MEL. It is of form:

$$\hat{P}(\{f_i\}) = \frac{e^{N\vec{\lambda}\cdot\vec{\phi}(\{f_i\})}e^{N\vec{\mu}\cdot\vec{F}(\vec{\phi}(\{f_i\}))}}{Z[\vec{\lambda}]\hat{Z}[\vec{\mu}]}. \tag{26}$$

We now induce a distribution on $\vec{\psi}$ by using the relationship $\psi_a = \frac{1}{N}\sum_i \delta_{a,f_i}$. We compute:

$$\hat{P}(\vec{\psi}) = \sum_{\{f_i\}:\vec{\phi}(\{f_i\})=\vec{\psi}} \hat{P}(\{f_i\}) = \frac{e^{N\vec{\lambda}\cdot\vec{\psi}}e^{N\vec{\mu}\cdot\vec{F}(\vec{\phi})}}{Z[\vec{\lambda}]\hat{Z}[\vec{\mu}]} m(\vec{\psi}) \tag{27}$$

where

$$m(\vec{\psi}) = \sum_{\{f_i\}:\vec{\phi}(\{f_i\})=\vec{\psi}} 1 \tag{28}$$

can be thought of as another g-factor that counts the number of combinations of filter responses $\{f_i\}$ (across the entire lattice) having histogram $\vec{\psi}$. After normalization, this term becomes the distribution on the $\vec{\psi}$ induced by assuming that the $\{f_i\}$ are generated by a *uniform distribution on the* $\{f_i\}$. In this case, the distribution is simply the multinomial distribution with mean value equal to $1/Q$.

This gives

$$\hat{P}_0(\vec{\psi}) = \frac{1}{\hat{Z}[\vec{\mu}]} e^{\vec{\mu}\cdot\vec{F}(\vec{\psi})} e^{-\sum_{a=1}^{Q}\psi_a \log\psi_a}. \tag{29}$$

Hence we can make the approximation:

$$\rho(\vec{\psi}) = -\sum_{a=1}^{Q} \psi_a \log\psi_a + \vec{\mu}\cdot\vec{F}(\vec{\psi}). \tag{30}$$

It should be appreciated that $\vec{\mu}$ are completely independent of the image and hence can be estimated off-line.

Using this form of $\rho(.)$ we can use equation (20) to solve for the clique potentials $\vec{\lambda}$ in closed form.

It may, however, be questioned whether we should be using different statistics to estimate $h(\{f_i\})$ from the uniform data than we use to estimate the distribution from the true data. In other words, *if the histogram statistic $\vec{\phi}(\vec{x})$ is not sufficient to estimate the distribution from the uniform data, then why should it be adequate to estimate the statistic from the real data?* Surely the additional dependencies which occur in the real data should require even more statistics? In short, we probably should use the same statistics to learn $h(\{f_i\})$ as we do to learn the true distribution.

$$\begin{tikzcd}
\vec{x} \arrow[r] \arrow[rr, dashed, no head, bend left, "g(\vec{\psi})"] \arrow[r, dashed, no head, bend right, "h(\{f_i\})"'] & \{f_i\} \arrow[r] \arrow[r, dashed, no head, bend right, "m(\vec{\psi})"'] & \vec{\psi}
\end{tikzcd}$$

Figure 7: Schematic summarizing the combinatorial factors in MEL that relate a probability distribution in one space (at the left end of each dotted line) to the *induced* distribution in another space (at the other end of each dotted line). $g(\vec{\psi})$ counts the number of images having a specified histogram, $h(\{f_i\})$ counts the number of images having specified filter responses across the entire lattice, and $m(\vec{\psi})$ counts the number of combinations of filter responses across the entire lattice consistent with a specified histogram (and hence is proportional to a multinomial distribution).

6 Discussion

This paper introduced the g-factor which depends on the lattice and quantization and is independent of the training image data. Alternatively it can be thought of as being proportional to the feature responses when the input images are uniformly distributed.

We showed that the g-factor can be used to relate probability distributions on features to distributions on images. In particular, we described approximations which, when valid, enable MEL to be computed analytically. In addition, we can determine when the clique potentials for features decouple.

These approximations throw light on MEL and help relate it to alternative ways of learning image statistics. Moreover, they also give guidelines, or diagnostic tests, to determine whether marginal histograms should be used as input to MEL (or whether more complicated statistics such as joint distributions are needed).

Our approach also emphasizes the importance of understanding the feature properties *independent of the dataset* and, in particular, to determine what the feature histograms are when the input images are uniformly distributed. This depends strongly on the quantization procedure used to describe the images. We also point out that the problem of estimating clique potentials may get simpler for fine quantization (because the approximations become more accurate) although empirical tests of MEL have usually been done using coarse quantization, see [19], for computational reasons.

Acknowledgements

We thank Anand Rangarajan, Xiuwen Liu, and Song Chun Zhu for helpful conversations. Sabino Ferreira gave useful feedback on the manuscript. This work was supported by the National Institute of Health (NEI) with grant number RO1-EY 12691-01.

References

[1] S. Amari. "Differential Geometry of curved exponential families – Curvature and information loss. Annals of Statistics, vol. 10, no. 2, pp 357-385. 1982.

[2] R. Balboa and N.M. Grzywacz. "The Minimal Local-Asperity Hypothesis of Early Retinal Lateral Inhibition". *Neural Computation.* bf 12, pp 1485-1517. 2000.

[3] A. Blake and A. Zisserman. Visual Reconstruction. MIT Press. 1987.

[4] J.M. Coughlan and A.L. Yuille. "A Phase Space Approach to Minimax Entropy Learning; The Minutemax approximation". In *Proceedings NIPS'98.* 1998.

[5] J.M. Coughlan and A.L. Yuille. “The g Factor: Relating Distributions on Features to Distributions on Images”. Technical Report. Smith-Kettlewell Eye Research Institute. San Francisco, CA 94115. 2001.

[6] T.M. Cover and J.A. Thomas. Elements of Information Theory. Wiley Interscience Press. New York. 1991.

[7] M.H. DeGroot. **Optimal Statistical Decisions**. McGraw-Hill. 1970.

[8] C. Domb and M.S. Green (Eds). **Phase Transitions and Critical Phenomena**. Vol. 2. Academic Press. London. 1972.

[9] D. Geiger and F. Girosi. “Parallel and Deterministic Algorithms from Mrfs: Surface Reconstruction”. *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 13. 1991.

[10] S. Geman and D. Geman. “Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images”. IEEE Trans. on PAMI 9(7), pp 721-741. 1984.

[11] D. Geman. and B. Jedynak. “An active testing model for tracking roads in satellite images”. *IEEE Trans. Patt. Anal. and Machine Intel.* Vol. 18. No. 1, pp 1-14. January. 1996.

[12] J. Huang, A.B. Lee, and D.B. Mumford. “Statistics of Range Images”. In *Proceedings Computer Vision and Pattern Recognition CVPR'2000.* pp 324-329. Hilton Head Island. North Carolina. 2000.

[13] S. M. Konishi, A.L. Yuille, J.M. Coughlan and Song Chun Zhu. “Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues.” In *Proceedings Computer Vision and Pattern Recognition CVPR'99.* Fort Collins, Colorado. 1999.

[14] A.B. Lee, D.B. Mumford, and J. Huang. “Occlusion Models of Natural Images: A Statistical Study of a Scale-Invariant Dead Leaf Model”. *International Jouranl of Computer Vision.* Vol. 41, No.s 1/2. January/February. 2001.

[15] J. Portilla and E. P. Simoncelli. “Parametric Texture Model based on Joint Statistics of Complex Wavelet Coefficients”. *International Journal of Computer Vision.* October, 2000.

[16] B. Ripley. “Pattern Recognition and Neural Networks”. Cambridge University Press. 1996.

[17] Y. Wu, S.C. Zhu, and X. Liu. ”Equivalence of Julesz texture ensembles and FRAME models”, *International Journal of Computer Vision*, 38(3), 247-265. 2000.

[18] V.N. Vapnik. **Statistical Learning Theory**. John Wiley and Sons, Inc. New York. 1998.

[19] S.C. Zhu, Y. Wu, and D. Mumford. “Minimax Entropy Principle and Its Application to Texture Modeling”. Neural Computation. Vol. 9. no. 8. Nov. 1997.