Order Parameters for Minimax Entropy Distributions: When does high level knowledge help?

A.L. Yuille, and James Coughlan
Smith-Kettlewell Eye Research Institute
2318 Fillmore Street
San Francisco, CA 94115
yuille@ski.org, coughlan@ski.org

Song Chun Zhu and Yingnian Wu [1]
Dept. Computer and Information Science
Ohio State University
2015 Neil Avenue, Columbus, OH 43210
szhu@cis.ohio-state.edu

Abstract

Many problems in vision can be formulated as Bayesian inference. It is important to determine the accuracy of these inferences and how they depend on the problem domain. In recent work, Coughlan and Yuille showed that, for a restricted class of problems, the performance of Bayesian inference could be summarized by an order parameter K which depends on the probability distributions which characterize the problem domain. In this paper, we generalize the theory of order parameters so that it applies to domains for which the probability models can be obtained by Minimax Entropy learning theory. By analyzing order parameters it is possible to determine whether a target can be detected using a general purpose "generic" model or whether a more specific "high-level" model is needed. At critical values of the order parameters the problem becomes unsolvable without the addition of extra prior knowledge.

[1]. **Dept. Statistics. University of California, Los Angeles, CA 90095. ymu@math.ucla.edu**

1. Introduction

Many problems in vision can be formulated as Bayesian inference. Two specific examples we will be concerned with in this paper are texture discrimination and road tracking. We are interested in determining the fundamental limits of vision and, in particular, what are the specific properties of the problem domain which make the visual task easy or hard. For example, consider the task of detecting the three target objects – stop sign, glia monster, and dalmation dog – from the images in figure (1). What is it about the targets and the domains which make these tasks easy or hard?

Coughlan and Yuille [10] introduced the concept of order parameters in an attempt to characterize the difficulty of certain visual tasks, in particular, the Geman and Jedynak model of road detection [5]. It was proved that important properties of this task, such as the expected errors in the MAP (maximum a posteriori) estimates of the road position, depend on an *order parameter K* which is a function of the probability distributions that characterize the problem. For $K < 0$ the task becomes impossible on average. For $K > 0$, quantities such as error rates behave as $\propto \exp^{-NK}$ where N is the size of the problem. Thus K characterizes the difficulty of the problem and can be thought of as analogous to the concept of signal to noise. The convergence rates of A* search algorithms to find the road were also functions of K [3].

Figure 1. Left to right, three detection tasks of increasing degrees of difficulty. The stop sign (left) is easy to find. The glia monster (centre) is harder. The dalmation dog (right) is almost impossible.

The current paper extends the order parameter theory so that it can be applied to problems where the probability distributions of interest can be learnt by Minimax Entropy learning theory [13],[14]. This is a richer and more realistic class of probability models than those previously analyzed. Generalizing order parameters, from those derived in [10], requires a more powerful set of mathematical techniques. The techniques used by Coughlan and Yuille to calculate order parameters were based on Sanov's theorem [4] and required that the probability distributions, which specify the problem, could be factorized. It was then observed by Wu and Zhu, during their work on texture modelling [9], that

more general results could be obtained by using techniques from the large deviation theory literature. In this paper, we build on this observation and determine order parameters for a general class of probability distributions and, in particular, to those which result from Minimax Entropy learning [13],[14]. This enables us, for example, to determine order parameters for more realistic models of curve tracking.

In this paper, we also explore a related problem. How much harder do we make these problems by using a weaker model (i.e. a less accurate probability distribution)? An experienced biologist will doubtless be able to detect the glia monster in figure (1) but a novice may be fooled by the camouflage. Can we quantify how much easier we make the task by using more information? (Which an expert biologist would presumably possess). This is important for three reasons. Firstly, there may not be enough information to obtain accurate probability distributions (or it would cost too much to get this knowledge). Secondly, we may want to search for several different targets and it would seem more economical to use one prior model which would account for all of these targets (at the cost of modelling each of them relatively poorly) rather than having different models for each target. Thirdly, algorithmic considerations may favour using a weaker prior rather than a prior which is more accurate but harder to compute with. Coughlan and Yuille presented results on this problem, which appeared in CVPR'99 [11], but these depended on the factorization assumption used in the Geman and Jedynak model [5].

In this paper, we derive these results for the more general class of probability distributions. Our main result relates the order parameter theory with Minimax Entropy learning [13],[14] by stating that *the information criterion [13] used by the Minimax Entropy theory to measure the improved accuracy of a more complex model p_{i+1} compared with a simpler model p_i, i.e. $D(p_{i+1}||p_i)$, is exactly the increase of the corresponding order parameter due to using p_{i+1} and hence quantifies the improved performance of the better model*, see section (3). This result allows us to determine how accurately (i.e. how many features and statistics) we need to model the problem (and at which level) in order to obtain the desired performance.

The structure of this paper is as follows. In section (2) we derive order parameters for Minimax Entropy models. We then describe, in section (3), how the order parameters change if a simplified approximate model is used. Section (4) calculates order parameters for curve detection for Minimax entropy models. Finally, section (5) describes the results of using simplified models for curve tracking.

2 Derivation of Order Parameters

In this section, we briefly summarize the order parameter theory results. The full derivation [12] unifies and extends the work reported in [10],[9].

For concreteness we will assume that there are two probability distributions $P_A(I), P_B(I)$ for texture images I *but the results are general and can be applied directly to other inference problems such as curve detection.*

We require that the probability distributions are of the form resulting from Minimax Entropy learning [13],[14]. These will be a class of Gibbs distributions which are shift-invariant and obey certain scaling results (to be described later). Each distribution is of form:

$$P(I|\vec{\beta}) = \frac{e^{-N\vec{\beta}\cdot\vec{h}(I)}}{Z(\vec{\beta})}, \tag{1}$$

where N is the size of the image I, $\vec{\beta}$ is a parameter (independent of N), $\vec{h}(.)$ are statistics defined on the image, and $Z(\vec{\beta})$ is the partition function (a normalization constant). The statistics can be, for example, (normalized) histograms of filter outputs of an entire image and shift-invariance is assumed in our analysis.

This determines an induced distribution on the *feature space* of all possible values of the statistics:

$$\hat{P}(\vec{h}|\vec{\beta}) = |\Omega_N(\vec{h})|\frac{e^{-N\vec{\beta}\cdot\vec{h}}}{Z(\vec{\beta})}, \tag{2}$$

where $\Omega_N(\vec{h}) = \{I : \vec{h}(I) = \vec{h}\}$ and $|\Omega_N(\vec{h})|$ is the size of this set. Let Q be the number of grayscale levels so that the total number of all possible images is Q^N. Then $|\Omega_N(\vec{h})|/Q^N$ can be considered to be a normalized probability distribution on $\vec{h}$ induced by the uniform distribution on all images (i.e. $\sum_{\vec{h}} |\Omega_N(\vec{h})|/Q^N = 1$).

We want to analyze the chances of misclassification of data generated by models of this form (e.g. texture classification and curve detection). To do this requires determining the probability of rare events such as when a sample of one texture appears to look like a sample of a second (or when random alignments of background clutter appear to look like a contour and hence are confusable with a target contour).

For probability distributions of the form specified by equations (1,2) the analysis becomes simplified as the image, and/or target size, becomes large [6]. Intuitively, this is *because the probability distribution in feature space becomes peaked as the size increases due to ergodicity* (e.g. the law of large numbers). Moreover, the theory gives tight results on how fast the distributions become peaked as N gets large.

This form can be used to put probabilities of the possibility of rare events. For example, H could consist of the set of rare events that would cause misclassification (e.g. by log-likelihood ratio tests) and the theory says that for sufficiently large N *we need only be concerned with the single most likely rare event in H*, see figure (2).

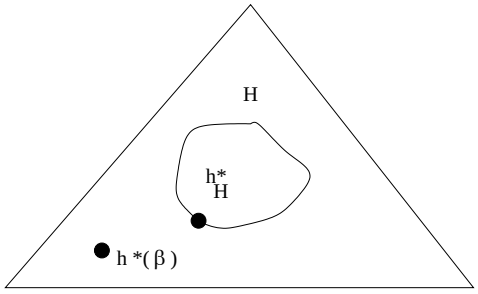

Figure 2. The chance of getting a statistic in H is dominated by the most likely statistic in H (each point in the diagram represents a statistic value $\vec{h}$). More precisely, if the distribution generating the data has mode $\vec{h}^*(\vec{\beta})$ then, for large N, the chances of the data lying in H are dominated by the chances of getting the statistic $\vec{h}^*_H$ in H which is closest to $\vec{h}^*(\vec{\beta})$.

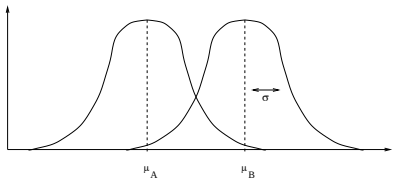

Figure 3. Two Gaussian distributions with the same variance. Kullback-Leibler, Chernoff, and Bhattarcharyya distances are all proportional to the signal to noise ratio $(\mu_A - \mu_B)^2/\sigma^2$.

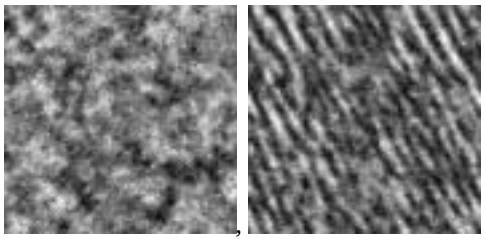

Figure 4. Texture examples, two textures generated by Minimax Entropy learning distributions.

This result can be re-expressed [12] in terms of the Kullback-Leibler divergence ($D(P_A||P_B) = \sum_I P_A(I) \log P_A(I)/P_B(I)$) between probability distributions. This leads to a direct generalization of Sanov's theorem [4]. Sanov's theorem – which applies only to data which is identically, independently distributed (i.i.d) – and its restriction to log-likelihood tests, was the key tool used by Coughlan and Yuille [10] obtain their previous results on order parameters. The results below enable us to generalize all Coughlan and Yuille's results to this more general class of probability distributions.

For the vision tasks in this paper, the chances of misclassification will behave as e^{-NK} where K is an *order parameter* which summarizes the difficulty of the task. In the tasks we consider, K will be expressed in terms of measures of distance between the distributions P_A, P_B. In particular, for our tasks K will involve the Kullback-Leibler divergence, the Chernoff Information, and the Bhattacharyya bound [4]. *The quantity we obtain will depend on the specific formulation of the task.*

To define Chernoff and Bhattacharyya, we must introduce the *e-geodesic* between $P_A(I)$ and $P_B(I)$. This e-geodesic consists of all distributions of form $P_\lambda(I) = P_A^\lambda(I)P_B^{1-\lambda}(I)/Z[\lambda]$ where $0 \leq \lambda \leq 1$ and $Z[\lambda]$ is a normalization constant. The *Chernoff information* is defined by $C(P_A, P_B) = D(P_{\lambda^*}||P_B)$ where λ^* obeys $D(P_{\lambda^*}||P_A) = D(P_{\lambda^*}||P_B)$. The *Bhattacharyya bound* is defined to be $B(P_A, P_B) = (1/2)(D(P_{1/2}||P_A) + D(P_{1/2}||P_B))$ and results if $\lambda = 1/2$. Our results will be summarized in the next section with detailed proofs given in [12].

One can gain intuition about these quantities by computing them when P_A, P_B are one-dimensional Gaussians with the same variance σ^2 and with means μ_A, μ_B. It can easily be checked that Kullback-Leibler, Chernoff, and Bhattarcharyya are all proportional to the signal to noise ratio $(\mu_A - \mu_B)^2/\sigma^2$.

2.1 Bounds for Log-Likelihood Discrimination Tasks

Suppose we have probability distributions, $P_A(I|\vec{\beta}_A)$ and $P_B(I|\vec{\beta}_B)$, with corresponding potentials $\vec{\beta}_A, \vec{\beta}_B$, see equation (1) (with same function $\vec{h}(.)$). For concreteness, the data I can be thought of as being a texture image *but the results are general*.

We now consider four texture tasks which involve ways of distinguishing between the two textures. Each task will involve the log-likelihood ratio test $R = \log P_A(I)/P_B(I)$. For each task we determine an order parameter.

Theorem 1. *The negative log probability per pixel that a sample from $P_B(I)$ generates a reward R greater than, or equal to, the average reward $< R >_{P_A}$ of a sample from P_A tends to $d(P_A||P_B) = \lim_{N\mapsto\infty}(1/N)D(P_A||P_B)$ as $N \mapsto \infty$. More informally $Pr(R(I) \geq < R >_{P_A} |I\ drawn\ from\ P_B(.)) \sim e^{-Nd(P_A||P_B)}$.*

The second texture task involves determining whether a sample I is generated by P_A or P_B.

Theorem 2. *The negative log probability per pixel that a sample from $P_A(I)$ is misclassified as being from P_B (and vice versa) tends to $c(P_A, P_B) = \lim_{N\mapsto\infty}(1/N)C(P_A, P_B)$ as $N \mapsto \infty$, where $C(P_A, P_B)$ is the Chernoff information. $Pr(R(I) < 0|I\ drawn\ from\ P_A(.)) \sim e^{-Nc(P_A,P_B)}$.*

The third texture task involves two texture samples, one each from P_A and P_B, and requires determining which is which.

Theorem 3. *The negative log probability per pixel that the two samples from $P_A(I)$ and $P_B(I)$ (one form each) are misclassified tends to $b(P_A, P_B) = \lim_{N\mapsto\infty}(1/N)B(P_A, P_B)$ as $N \mapsto \infty$, where $B(P_A, P_B)$ is the Bhattarcharyya information. $Pr(misclassification) \sim e^{-Nb(P_A,P_B)}$.*

The fourth texture task requires determining how easy it is to confuse a sample from P_A with many samples from P_B.

Theorem 4. *Suppose we have e^{NQ} samples of texture from P_B. Then the expected number that have reward higher than $< R >_{P_A}$ is given by $e^{-N\{d(P_A||P_B)-Q\}}$, where $d(P_A||P_B) = \lim_{N\mapsto\infty}(1/N)D(P_A||P_B)$. There is a phase transition and the problem becomes insolvable if $Q > d(P_A||P_B)$.*

Observe that in the first three situations the error rates fall-off as e^{-NK} where K is a non-negative constant and N is the size of the problem. We call K an *order parameter* for the task because it summarizes in a single number the difficulty of the task (in the same way that order parameters, such as magnetization, are used in statistical physics).

For the fourth task, there is more interesting behaviour because the order parameter, $K = d(P_A||P_B) - Q$, can become negative. When this happens there is a phase transition and the task becomes impossible – essentially because there are so many distractors that there is a high probability that one of them looks more like the target than the target itself. Intuitively, the search task then becomes like searching for a needle in a haystack.

3 The wrong reward function

This section analyzes what happens if we have a weaker approximate model of the probability distributions (as will often be the case). How much do we lose by such approximations? Our main result is to show how order parameters change when a weaker model is used and to demonstrate a nice connection to the Minimax Entropy learning model selection criterion.

In many situations, it may be impossible or impractical to use the correct probability models for the problem. We may need to approximate the true prior P_H of the target geometry by a "generic" prior P_G. How much do we lose by using approximations? If P_G is just a minor perturbation of

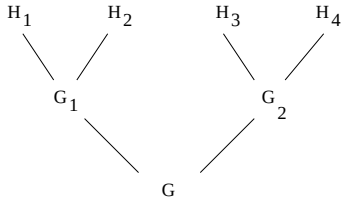

Figure 5. The Hierarchy. Two high-level models H_1, H_2 "project" onto a low-level generic model P_{G_1}. In situations with limited clutter it will be possible to detect either H_1 or H_2 using the single generic model P_{G_1}. This idea can be extended to have hierarchies of projections. This is analogous to the superordinate, basic level, and subordinate levels of classification used in cognitive psychology.

P_H then standard analysis shows that the concavity of the Bayes risk means the system will be stable to such perturbations. A more important case arises when P_G *is a poor approximation to* P_H. In what regimes can we get away with using a poor approximation? We will give results for this case.

A particularly interesting form of "weakness" is when the generic prior P_G is a projection of the high-level prior P_H onto a simpler class of probability distributions. This allows us to formulate the idea of a hierarchy in which the priors for several high-level objects would all project onto the identical low-level prior, see figure (5). For example, we might have a set of priors $\{P_{H_i} : i = 1, ..., M\}$ for different members of the cat family. There might then be a generic prior P_G onto which all the $\{P_{H_i}\}$ project and which is considered the embodiment of "cattiness".

To be more specific, consider the case of discriminating between data from two distributions P_A, P_B. The correct procedure is to use the log-likelihood ratio test $R(I) = \log P_A(I)/P_B(I)$ to perform the classification. *But suppose we only know an approximation $P_{\hat{A}}(I)$ to $P_A(I)$.* How do the order parameters vary if we use the modified reward $\hat{R}(I) = \log P_{\hat{A}}(I)/P_B(I)$ to do classification *but assuming that the samples have been drawn from the correct distribution $P_A(I)$, or from $P_B(I)$?*

We assume that $P_{\hat{A}}(I)$ is related to $P_A(I)$ by what we call an *Amari condition* [1] – i.e. $\sum_I P_A(I) \log P_{\hat{A}}(I) = \sum P_{\hat{A}}(I) \log P_{\hat{A}}(I)$. This condition is motivated by Minimax entropy learning (and the results of Coughlan and Yuille in CVPR)

In particular, Minimax Entropy learning naturally gives rise to a sequence of increasingly accurate Gibbs distributions by pursuing additional features and statistics. The sequence $p_0 = U, p_1, p_2, ..., p_k \rightarrow p_{\rm true}$ (where k is the num-

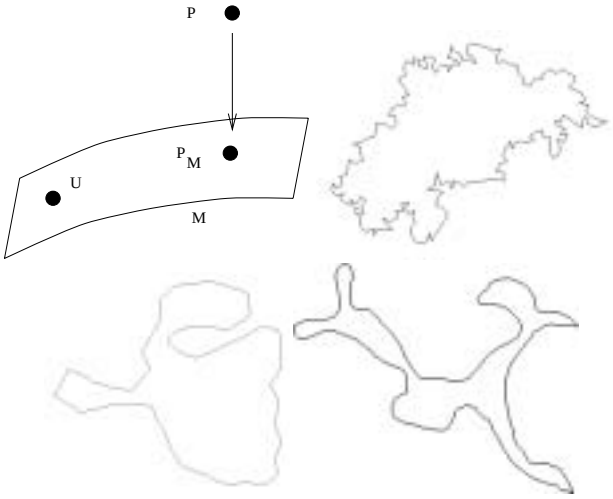

Figure 6. The *Amari projection* and a sequence of prior models for animate object shapes by minimax entropy using an increasing number of feature statistics. See text for interpretation.

ber of features and statistics included in the model p_k) starts with p_0 being a uniform distribution U and approaches the true distribution $p_{\rm true}$ in the limit [13]. The more high-level (i.e. target specific) the model then the more target specific the statistics. Conversely, low-level (i.e. general purpose) models will only use those statistics which are common to many targets. More precisely, each Gibbs distribution p_i is an *Amari projection* [1] of the "true" distribution $p_{\rm true}$ onto the sub-manifold M_i, with p_i being the closest element to $p_{\rm true}$ in M_i, in terms of Kullback-Leibler divergence $D(p_{\rm true}||p_i)$, see figure (6). As shown in figure (6), the first row, from left to right are typical shapes sampled from three minimax entropy models[14]: a uniform model, a model matching contour based statistics, and a model matching both contour and region based statistics.

For simplicity, we analyze the first task in section (2). Thus we compute the expected reward $< \hat{R} >_{P_A} = D(P_{\hat{A}}||P_B)$ if the data is generated by $P_A(I)$ and estimate the probability that data generated by P_B will have higher reward. We assume the Amari condition $\sum_I P_A(I) \log P_{\hat{A}}(I) = \sum_I P_{\hat{A}}(I) \log P_{\hat{A}}(I)$ and the additional condition $\sum_I \log P_B(I)\{P_{\hat{A}}(I) - P_A(I)\} = 0$ (for example, this is satisfied if P_B is the uniform distribution). More general conditions are described in [12].

Now we ask, what is the probability that we get a sample I from $P_B(.)$ with reward $\hat{R}(I) > < \hat{R} >_{P_A}$? The problem can be formulated as in Theorem 1 of the previous section. *The only difference is that, because* $< \hat{R} >_{P_A} = D(P_{\hat{A}}||P_B)$*, we can replace* P_A *by* $P_{\hat{A}}$ *everywhere in the calculation.*

We therefore obtain that the probability of error goes like $\sim e^{-D(P_{\hat{A}}||P_B)}$. This means that the order parameter is higher by an amount $D(P_A||P_B) - D(P_{\hat{A}}||P_B)$ when we use the "correct" reward function. This can be expressed as:

$$D(P_A||P_B) - D(P_{\hat{A}}||P_B) = \sum P_A \log \frac{P_A}{P_B} - \sum P_{\hat{A}} \log \frac{P_{\hat{A}}}{P_B},$$
$$= D(P_A||P_{\hat{A}}) + \sum \log P_B \{P_{\hat{A}} - P_A\},$$

where we have used the Amari condition $\sum P_A \log P_{\hat{A}} = \sum P_{\hat{A}} \log P_{\hat{A}}$.

Using the condition $\sum \log P_B \{P_{\hat{A}} - P_A\} = 0$ we see that the order parameter increases by $D(P_A||P_{\hat{A}})$ when we use the correct reward function. *This is precisely the entropy criterion used in Minimax Entropy learning in determining the benefit of using an additional statistic because* $H(P_{\hat{A}}) - H(P_A) = D(P_A||P_{\hat{A}})$*!*. This demonstrates that accurate prior models increase the order parameters.

We can carry through this analysis to the fourth task in section (2) where we have to estimate the chances of confusing a target with one of many distractors. The result is that the phase transition will shift, depending on which reward function is used, by the amount given above.

4 Detecting Curves in Images

We now apply the order parameter theory to the specific problem of detecting curves in images. In previous work, we studied a factorizable model by Geman (D.) and Jedynak, motivated by road tracking from aerial images [5], which assumes a factorizable model. We now consider a generalization of this model to the non-i.i.d. case, allowing spatial coupling, using a Minimax Entropy learning model.

The model is given by Minimax entropy form. To define the likelihood function we first choose three filters:

$$F^1(\vec{x}) = \vec{\nabla} I(\vec{x}) \cdot \vec{\hat{t}}(\vec{x}) \; if \; \vec{x} \epsilon X, \; = \vec{\nabla} I \cdot \vec{i} \; otherwise$$
$$F^2(\vec{x}) = \vec{\nabla} I(\vec{x}) \cdot \vec{\hat{n}}(\vec{x}) \; if \; \vec{x} \epsilon X, \; = \vec{\nabla} I \cdot \vec{j} \; otherwise$$
$$F^3(\vec{x}) = I(\vec{x}) \qquad (3)$$

where $\vec{\hat{t}}(\vec{x}), \vec{\hat{n}}(\vec{x})$ are the tangent and normal to the curve at $\vec{x}$, and $\vec{i}, \vec{j}$ are the horizontal and vertical unit vectors of the image plane. The curve X has M pixels and there are a total of N pixels in the entire image.

We define $\{\vec{h}^{\alpha}_{on}, \vec{h}^{\alpha}_{off} : \alpha = 1, 2, 3\}$ to be the empirical histograms of the $\{F^{\alpha} : \alpha = 1, 2, 3\}$ on and off the road respectively (where α labels the filters $F^1, F^2, ...$). More precisely, $h^{\alpha}_{on,z} = \frac{1}{M} \sum_{\vec{x} \in X} \delta_{z, F^{\alpha}(\vec{x})}$ are the components – indexed by z – of the vector $\vec{h}^{\alpha}_{on}$, and similarly for $\vec{h}^{\alpha}_{off}$, $h^{\alpha}_{off,z} = \frac{1}{N-M} \sum_{\vec{x} \notin X} \delta_{z, F^{\alpha}(\vec{x})}$ are the components – indexed by z – of the vector $\vec{h}^{\alpha}_{on}$. The likelihood function is

then given by:

$$P(I|X) = \frac{1}{Z} e^{\sum_{\alpha=1}^{3} \{M\vec{\beta}^{\alpha}_{on} \cdot \vec{h}^{\alpha}_{on} + (N-M)\vec{\beta}^{\alpha}_{off} \cdot \vec{h}^{\alpha}_{off}\}}$$
$$\propto e^{\sum_{\alpha} \sum_{\vec{x} \epsilon X} \{\beta^{\alpha}_{on, F^{\alpha}(\vec{x})} - \beta^{\alpha}_{off, F^{\alpha}(\vec{x})}\}}, \quad (4)$$

where $\beta^{\alpha}_{on,z}, \beta^{\alpha}_{off,z}$ are the components of $\vec{\beta}^{\alpha}_{on}, \vec{\beta}^{\alpha}_{off}$. In our simulations we typically allow F_3 to have eight components (i.e. the images have eight grey-level values) and F_1, F_2 to have six components.

Similarly, we define the prior model for the road by $P(X) = p(\vec{x}_1) \prod_{i=2}^{N} p(\vec{x}_i|\vec{x}_{i-1})$ (the prior is chosen to prevent the curve from ever intersecting itself). In some cases we extend this to a second order Markov chain prior determined by distributions such as $p(\vec{x}_i|\vec{x}_{i-1}, \vec{x}_{i-2})$.

This gives an overall reward function:

$$R(X|I) = \sum_{i} \log p(\vec{x}_i|\vec{x}_{i-1})$$
$$+ \sum_{\alpha} \sum_{\vec{x} \epsilon X} \{\beta^{\alpha}_{on, F^{\alpha}(\vec{x})} - \beta^{\alpha}_{off, F^{\alpha}(\vec{x})}\}. \quad (5)$$

To learn this model from observed data $\vec{d}^{\alpha}_{on} =< \vec{h}^{\alpha}_{on} >_{observer}$ and $\vec{d}^{\alpha}_{off} =< \vec{h}^{\alpha}_{off} >_{observer}$ Minimax Entropy learning [13],[14] requires us to estimate the potentials $\vec{\beta}^{\alpha}_{on}, \vec{\beta}^{\alpha}_{off}$. In [12], we describe a recursive algorithm for learning the model from real data.

We obtain order parameters for these models by applying Theorem 4. These order parameters will have contributions both from the geometry and the pixel intensity information, see [12] for details. Figure (7) shows the results of simulating from the curve model for different distributions.

5 High-Low information

In this section we consider the effects of using the wrong prior. More specifically, we will consider two possible geometry priors P_H and P_G related by an Amari projection, $\sum_X P_H(X) \log P_G(X) = \sum_X P_G(X) \log P_G(X)$. We call $P_H(X)$ the *high-level* model and it is used to generate the data (i.e. it is the "true prior"). By contrast, $P_G(X)$ is called the *generic prior* (i.e. it is the "wrong prior").

We will perform inference on the data in two ways. Firstly, we use the high-level prior in the reward function (i.e. standard Bayesian inference). Secondly, we will use the generic prior in the reward function. The theory predicts there will be three regimes, *ultra*, *challenging*, and *easy*, see caption of figure (8).

In figure (9), we consider two high-level models, second order Markov chains, which we call roman road and english road. They are both approximated by the same generic, first order Markov, road model. We illustrate the three different regimes.

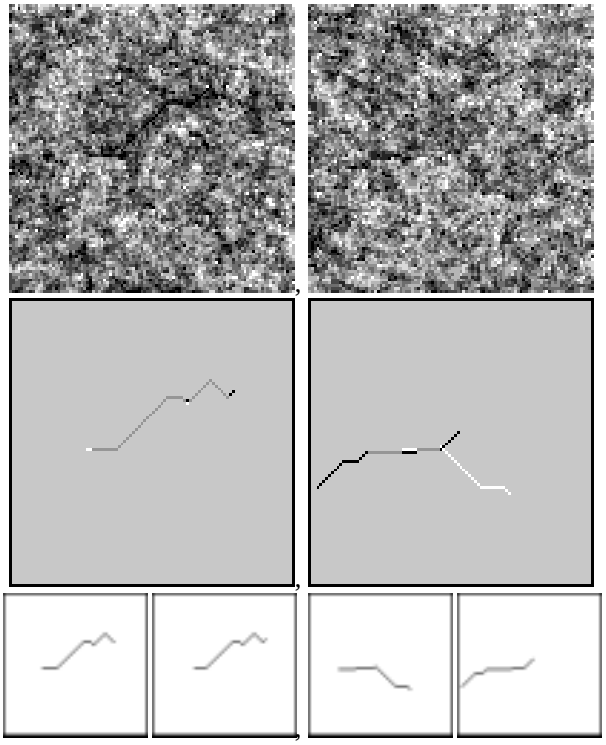

Figure 7. (Top) Samples from the Minimax Entropy curve model, $K = 1.00$ on left and $K = -0.43$ on right. (Middle) The true curve positions for the corresponding samples are shown in white. The solution path, found by dynamic programming, is in black. Places where the solution overlaps with the true path are shown in grey. (Bottom) The true path and the solution for $K = 1.0$ (far left, and left). The true path and the solution for $K = -0.43$ (right, and far right). Observe that for positive K, on the left, the solution is very close to the true path. But if K is negative, on the right, then the solution is very different from the true path – i.e. the task becomes impossible. The order parameters calculated for the models are consistent with the results. The best paths are determined by optimizing the reward functions using a dynamic programming algorithm that does not require known starting point [2].

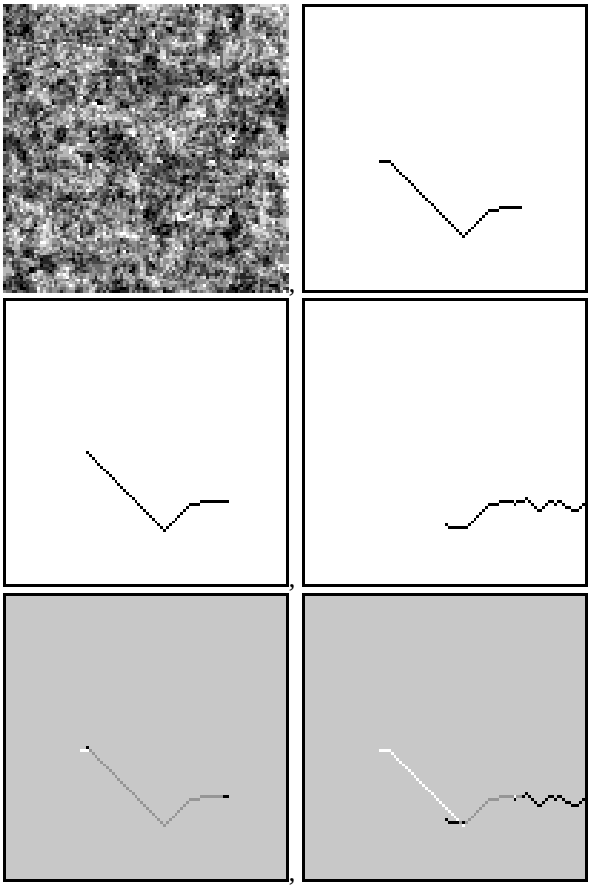

Figure 8. The Challenging regime figure. In the *ultra regime*, detection of the curve is impossible even if the high-level model is used. In the *challenging regime* we will be able to detect the curve if we use the high-level model but *not* if we use the generic model. In the *easy regime*, both models are adequate to detect the curve. The data is shown in the top left square and the true path is shown in the top right square. The results of estimation using the high-level and generic models are shown in the left and right middle squares respectively. Their overlaps with the true path are shown in the bottom two squares (similar conventions to the previous figures). Observe that the high-level model correctly finds the true path (with a few pixels of error) but the generic model fails (apart from finding one small subsection).

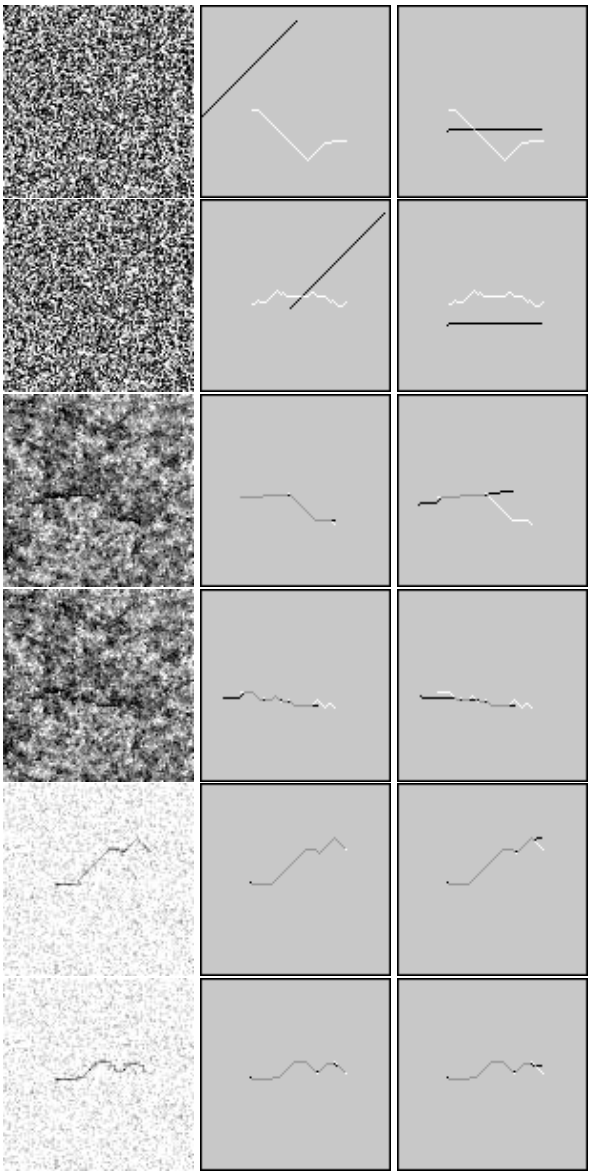

Figure 9. Three panels, of two rows each, top to bottom giving examples of ultra, challenging, and easy regimes. For each panel, the top row gives a sample generated by an *roman road model* (left), the best path found using the *roman road model* (center), and the best path found using the *generic road model* (right). Similarly, for each panel, the bottom row gives a sample generated by an *english road model* (left), the best path found using the *english road model* (center), and the best path found using the *generic road model* (right). In the ultra regime, top panel, no method works. In the challenging regime (centre panel), the high-level models (roman and english) find their targets but the generic models make errors. In the easy regime, everything works.

6. Conclusion

This paper demonstrated how to determine order parameters for visual problems for which the probability models can be learnt by Minimax Entropy learning theory [13],[14]. It builds on the work of Coughlan and Yuille [10] on factorizable distributions and the work on texture by Wu and Zhu [9].

This paper also investigated how much prior knowledge is needed to detect a target road or curve in the presence of clutter. We used order parameters to determine whether a target could be detected using a general purpose "generic" model or whether a more specific high level model was needed. At critical values of the order parameters the problem becomes unsolvable without the addition of extra prior knowledge. Results of this type were presented in CVPR'99 [11] for the restricted class of factorized probability distributions.

We observe that our results are in a similar spirit to the theoretical analysis by Tsotsos on the complexity of vision [8]. The techniques used, however, are quite different and the relationship between these two approaches is a topic for further study. In addition, our results on phase transitions are reminiscent of those obtained by Selman and Kirkpatrick [7] but who also use completely different techniques for analysis.

Hopefully, analysis of the type performed in this paper can help quantify when high-level knowledge is needed for visual tasks. This may throw light into the development of efficient algorithms for segmentation and recognition.

Acknowledgments

We want to acknowledge funding from NSF with award number IRI-9700446, from the Center for Imaging Sciences funded by ARO DAAH049510494, and from the Smith-Kettlewell core grant.

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] James M. Coughlan, A.L. Yuille, D. Snow and C. English. "Efficient Optimization of a Deformable Template Using Dynamic Programming". In *Proceedings Computer Vision and Pattern Recognition. CVPR'98*. Santa Barbara. California. 1998.

[3] James M. Coughlan and A.L. Yuille. "Bayesian A* tree search with expected O(N) convergence rates for road tracking". In *Proceedings EMMCVPR'99*. Springer-Verlag Lecture Notes in Computer Science 1654. 1999.

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

[5] 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.

[6] J.T. Lewis, C.E. Pfister, and W.G. Sullivan. "Entropy, Concentration of Probability, and Conditional Limit Theorems". *Markov Processes Relat. Fields*, **1**, pp 319-396. 1995.

[7] B. Selman and S. Kirkpatrick. "Critical Behaviour in the Computational Cost of Satisfiability Testing". Artificial Intelligence. 81(1-2); 273-295. 1996.

[8] J.K. Tsotsos. "Analyzing Vision at the Complexity Level". *Behavioural and Brain Sciences*. Vol. 13, No. 3. September. 1990.

[9] Y. Wu and S.C. Zhu. "Equivalence of Image Ensembles and Fundamental Bounds". International Journal of Computer Vision (Marr Prize special issue). To appear. 1999.

[10] A. L. Yuille and James M. Coughlan. "Fundamental Limits of Bayesian Inference: Order Parameters and Phase Transitions for Road Tracking" . *Pattern Analysis and Machine Intelligence* PAMI. Vol. 22. No. 2. February. 2000.

[11] A.L. Yuille and James M. Coughlan. "High-Level and Generic Models for Visual Search: When does high level knowledge help?" In *Proceedings Computer Vision and Pattern Recognition CVPR'99*. Fort Collins, Colorado. 1999.

[12] A.L. Yuille, James M. Coughlan, Y.N. Wu, and S.C. Zhu. "Order Parameters for Minimax Entropy Distributions: When does high level knowledge help?". Submitted to *International Journal of Computer Vision*. 1999.

[13] S-C Zhu, Y-N Wu and D. Mumford. FRAME: Filters, Random field And Maximum Entropy. *Int'l Journal of Computer Vision* 27(2) 1-20, March/April. 1998.

[14] S.C. Zhu. "Embedding Gestalt Laws in Markov Random Fields". To appear in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI). 1999.

[15] S.C. Zhu, X.W. Liu and Y. Wu. "Exploring Texture Ensembles by Efficient Markov Chain Monte Carlo". To appear in *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI). 1999.