Statistical Edge Detection: Learning and Evaluating Edge Cues

S. Konishi, A.L. Yuille, James M. Coughlan, and Song Chun Zhu *

Smith-Kettlewell Eye Research Institute

2318 Fillmore Street, San Francisco, CA 94115

{konishi,yuille,coughlan}@ski.org

December 26, 2002

* Dept. Computer and Information Sciences, The Ohio State University, Columbus, OH 43210, szhu@cis.ohio-state.edu.

Abstract

We formulate edge detection as statistical inference. This *statistical edge detection* is data driven, unlike standard methods for edge detection which are model based. For any set of edge detection filters (implementing local edge cues) we use pre-segmented images to learn the probability distributions of filter responses conditioned on whether they are evaluated *on* or *off* an edge. Edge detection is formulated as a discrimination task specified by a likelihood ratio test on the filter responses. This approach emphasizes the necessity of modeling the image background (the off-edges). We represent the conditional probability distributions non-parametrically and learn them on two different datasets of 100 (Sowerby) and 50 (South Florida) images. Multiple edges cues, including chrominance and multiple-scale, are combined by using their joint distributions. Hence this cue combination is optimal in the statistical sense. We evaluate the effectiveness of different visual cues using the Chernoff information and Receiver Operator Characteristic (ROC) curves. This shows that our approach gives quantitatively better results than the Canny edge detector when the image background contains

significant clutter. In addition, it enables us to determine the effectiveness of different edge cues and gives quantitative measures for the advantages of multi-level processing, for the use of chrominance, and for the relative effectiveness of different detectors. Furthermore, we show that we can learn these conditional distributions on one dataset and adapt them to the other with only slight degradation of performance *without* knowing the ground truth on the second dataset. This shows that our results are not purely domain specific. We apply the same approach to the spatial grouping of edge cues and obtain analogies to non-maximal suppression and hysteresis.

Index Terms: Edge Detection, Statistical Learning, Performance Analysis, Bayesian Inference.

1 Introduction

Edge detectors, see [9], are intended to detect and localize the boundaries of objects (in this paper we will use "edge" as a shorthand for object boundary or significant albedo change, see figure (1), and later examples in figures (7,8)). In practice, it is clear that edge detection is an ill-posed problem. It is impossible to design an edge detector that will find all the true (i.e. object boundary and significant albedo change) edges in an image and not respond to other image features. Examining real images, it is clear that edge detectors only give ambiguous local information about the presence of object boundaries.

Most conventional edge detectors are designed by assuming models of edges. For example, Canny [9] assumes that edges are step functions corrupted by additive Gaussian noise. But, as has been widely reported [12, 1, 30, 39, 24, 35], natural images have highly structured statistical properties which typically do not agree with the assumptions made by current edge detectors. It makes sense, therefore, to formulate edge detection as statistical inference where the detectability of edges depends both on the statistics of filters *on the edges* but also the statistics of filters *off the edges* (ie. on the background image clutter). These edge and background statistics may be domain specific, and edge detection should take this into account. (An alternative approach would be to learn a classifier [34] without learning probability distributions, but we show there is sufficient data to learn the distributions).

To implement statistical edge detection we make use of ground truth segmentations, see

Figure 1: A typical Sowerby image (top left panel) with its ground truth segmentation (top right panel) and its segmentation using the Canny edge detector (bottom left panel) and by statistical edge detection (bottom centre panel). Statistical edge detection has fewer false negatives in the textured regions and is also better at detecting edges with are partially defined by texture. By contrast, the Canny detector is slightly better at precision of certain edges. The log-likelihood ratios are also shown (bottom right panel).

figures (1,7,8). We first use two pre-segmented datasets, Sowerby and South Florida, in a *learning stage* to determine probability distributions for the response of edge detection filters on and off edges. Edge detection can then be performed using a log-likelihood ratio test, see [11]. (In addition, these log-likelihood ratios, see figure (1) can be use as a local measure of edge strength [14] in formulations such as snakes [18] and region competition [38]). We use standard filters such as the intensity gradient, the Laplacian of a Gaussian, and filterbanks of oriented filter pairs (eg. Gabor filters). To combine different edge cues, we specify the edge filter to be vector-valued, with components corresponding to the different cues (e.g. grey-scale, chrominance, and multi-scale). In other words, we use the joint distributions of the different edge cues (which is the optimal way to combine them).

The probability distributions are represented non-parametrically by multi-dimensional histograms. The bin boundaries are determined adaptively in order to reduce the total number of bins required. This is necessary to ensure that we have sufficient data to learn the probability distributions and to prevent *overlearning* [34]. We use cross-validation [29] to check for overlearning. In addition, we sometimes use decision trees [29] to further reduce the number of bins required.

In our *evaluation stage*, we determine the effectiveness of the edge detection filters by two

criteria: (i) by evaluating the Chernoff information [11] and, (ii) by determining the Receiver Operating Characteristic (ROC) curves [15]. The Chernoff information arises naturally in theoretical studies by Yuille and Coughlan [36] for determining the detectability of roads in aerial images [14]. ROC curves have been used by Bowyer *et al* to empirically evaluate the performance of standard edge detectors on the South Florida dataset [7],[31],[8]. Hence we can use ROC curves to compare the performance of statistical edge detection against more standard edge detectors. In addition, we use the area under the ROC curve, and the Bayes risk.

Most practical edge detectors (eg. Canny [9]) use post-processing techniques such as non-maximal suppression and hysteresis. We therefore extend statistical edge detection to incorporate spatial grouping cues. These grouping cues are also learnt from our image datasets and, not surprisingly, they show analogs of non-maximum suppression and hysteresis.

Our results show that statistical edge detection significantly outperforms the Canny edge detector [9] when evaluated on the Sowerby dataset, see figure (16). On the South Florida dataset, statistical edge detection performs equivalently to the Canny edge detector and the best of the other edge detectors evaluated by Bowyer *et al* [7, 8]. Our results also show that it is significantly harder to detect edges in the Sowerby dataset than in the South Florida dataset. This is because there is far more "clutter" in the Sowerby images which can cause edge detectors to report false positives, see figure (1). We assume that edge detectors should not report edges in cluttered and textured regions. Overall, the Sowerby dataset is more challenging and (arguably) more representative of real world images.

We are also able to adapt our probability distributions between the Sowerby and South Florida datasets with only a small change in performance. In other words, we can perform high quality segmentation on South Florida *without* needing the ground truth (and similarly on Sowerby). Moreover, the success of our adaptation also shows that the image statistics are robust with respect to the ground truth. Inspection of the Sowerby and South Florida datasets shows that the ground truths were determined rather differently, see subsection (3.1). If the statistics were very sensitive to ground truth then it would be impossible to adapt them between the two datasets.

Our approach complements recent work on empirical performance analysis of visual algorithms [6]. Our work was originally inspired by Geman and Jedynak [14], who learnt

statistics responses for filters on and off highways in aerial images. We were also influenced by the work of Balboa and Grzywacz [2], [3],[4], who measured contrast edge statistics on and off occluding boundaries in two image domains which, they argued, corresponded to differences in the receptive field properties of the retinas of animals in the two different environments and propose an alternative adaptation procedure [16]. A recent learning method [27] is rather different from our approach and make use of reinforcement learning with high-level feedback. More recently Sullivan *et al* [33] have learnt statistics for image backgrounds in their work on "Bayesian correlation".

The structure of this paper is as follows. In section (2) we describe the edge filters, the two evaluation criteria, and how we represent and learn the conditional probability distributions. Section (3) gives the results of our edge detection filters on the two datasets using the two evaluation criteria. In Section (4) we describe how we learn spatial grouping as an analogy to non-maximal suppression and hysteresis. Section (5) shows that we can adapt our probability distributions from one dataset to the other illustrating that our results are not purely dataset specific nor overly dependent on the ground truth of the datasets.

2 Representing, Learning, and Evaluating Edge Filters

Statistical edge detection involves learning the conditional probability distributions $P(\phi|\textit{on-edge})$ and $P(\phi|\text{off}-\text{edge})$ for the filter response ϕ conditioned on whether the filter is evaluated *on* or *off* an edge. We can then use the log-likelihood ratio test, $\log \frac{P(\phi(I(x))|\ \textit{on-edge})}{P(\phi(I(x))|\textit{off-edge})} > T$, to determine if a pixel x in image $I(x)$ is an edge, where T is a suitable threshold (visually more pleasing edge maps, however, can be obtained using a further spatial grouping stage, see section (4)). Following the analysis of Geman and Jedynak [14], the log-likelihood ratio can also be used as a measure of edge strength as input to curves detectors such as snakes [18] or region competition [38].

This requires us to specify a set of *edge detection filters* ϕ, see subsection (2.1). We evaluate the effectiveness of different edge filters using *performance criteria*, see subsection (2.2). This requires representing the conditional probability distributions by *adaptive non-parametric representations* (e.g. histograms), see subsection (2.3). The performance criteria are also used to determine the adaptive non-parametric representations by evaluating

the effectiveness of the probability distributions induced by the different possible representations.

Once the non-parametric representations have been chosen then learning the probability distributions reduces to evaluating the filters on the datasets (using the ground truth to determine which pixels are on and off edges) and counting the number of responses in each bin.

2.1 The Two Filter Sets

We consider two sets of edge detection filters. The first set consists of standard edge filters (supplemented by the Nitzberg filter, which turns out to be very effective). The second set consists of oriented filter banks partially inspired by the biology of the human visual system.

2.1.1 The First Filter Set

In this paper, we specify a filter ϕ by a differential (or difference) operator, the scales at which we apply it, and the colour bands we apply it to. The filters in the first set are shown in table (1). The *dimension* of the filter is the product of the dimensions of the operator, the number of scales, and the number of image bands. For example, filter no. 2 in the table is the Laplacian ∇^2 operator at three scales applied to image band Y, and so is a three-dimensional filter.

Filter No.	Operator	Scale	Image Band	Filter No.	Operator	Scale	Image Band
#1	∇^2	$\sigma = 1$	Y	#7	$\lvert\vec{\nabla}\rvert$	$\sigma = 1$	Y, I, Q
#2	∇^2	$\sigma = 1, 2, 4$	Y	#8	$\lvert\vec{\nabla}\rvert$	$\sigma = 1, 2, 4$	I, Q
#3	$\lvert\vec{\nabla}\rvert$	$\sigma = 1$	I, Q	#9	$\lvert\vec{\nabla}\rvert$	$\sigma = 1, 2, 4$	Y
#4	$\lvert\vec{\nabla}\rvert$	$\sigma = 1$	Y	#10	N_1	$\sigma = 1, 2, 4$	Y
#5	N_1	$\sigma = 1$	Y	#11	N_1, N_2	$\sigma = 1, 2, 4$	Y
#6	N_1, N_2	$\sigma = 1$	Y	#12	$\lvert\vec{\nabla}\rvert$	$\sigma = 1, 2$	Y, I, Q

Table 1: There are twelve filters in the first set. For each filter we estimate the joint probability distributions of the differential operators, the scales, and the image bands. See text for definition of $\vec{N}, N_1, N_2$.

For the first filter set, the differential operators are the magnitude of the image gradient

$|\vec{\nabla}|$, the Nitzberg operator $\vec{N}$ [26], and the Laplacian ∇^2 [25]. These are applied at different scales σ by smoothing the image by a Gaussian filter with variance σ^2. There are three colour bands Y, I, Q for Sowerby and one (ie. grey-scale) for South Florida.

More precisely, the modulus of the gradient and the Laplacian operators are specified by the equations $\left|\vec{\nabla}_\sigma I(x)\right| \equiv \left|\vec{\nabla} G(x;\sigma) * I(x)\right|$ and $\nabla^2_\sigma I(x) \equiv \nabla^2 G(x;\sigma) * I(x)$, where $*$ denotes convolution and $G(x;\sigma)$ is a Gaussian at a spatial scale parameterized by the standard deviation σ. The Nitzberg operator involves computing the matrix $\mathbf{N}_\sigma(x) = G(x;\sigma) * \{\vec{\nabla} I(x;\sigma)\}\{\vec{\nabla} I(x;\sigma)\}^T$ where ^{T} denotes transpose. In other words, we take the image gradient at scale σ and then average its outer product by a Gaussian with the same scale (we found it most effective to use the same value of σ for both scales). The output is the two-dimensional vector consisting of both eigenvalues $(N_1(x;\sigma), N_2(x;\sigma))$. This operator is sensitive to image corners (see chps 4,16 by Harris in [5]), which helps it discriminate texture from edges, as we will see in section (3).

Our colour representation is a variant of the NTSC colour space, with $Y = 0.299R + 0.587G + 0.114B$, $I = (0.596R - 0.274G - 0.322B)/Y$, $Q = (0.211R - 0.523G + 0.312B)/Y$. Here Y is interpreted to be the grey-scale image and I, Q are the chrominance vectors. *Unlike NTSC*, we have normalized the chrominance by the greyscale. This normalization enables us to examine the effectiveness of chrominance cues independent of grey-scale. It is important to realize that *the choice of colour space representation is relatively unimportant because we use joint distributions to combine colour cues.* The only reason it matters at all is because we determine the bin boundaries based on the one-dimensional distributions (which do depend on our choice of colour space).

The biology of human vision, combined with more pragmatic motives, strongly suggests that images should be processed at different scales, see [25]. In such "scale-space" approaches it is not always clear how to best combine the information given by the edge detectors at different scales. In statistical edge detection, as described in this paper, *the optimal combination arises naturally* by using the joint distributions of the filters at different scales (subject to the quantization procedure we use).

In the rest of this paper, we represent filters by the operator, the scales it is applied at, and the colour bands it is applied to. For example, $\nabla^2_{\sigma=1,2,4}(Y, I, Q)$ means that the filter is the Laplacian of a Gaussian applied at scales $\sigma = 1, 2, 4$ to the three colour bands

Y, I, Q. This filter is vector-valued with nine dimensions. The effectiveness of these different combinations is shown in section (3.2.1).

2.1.2 The Second Filter Set

The second filter set is a filterbank of orientation-tuned pairs of symmetric (even) and anti-symmetric (odd) filters. It is claimed that the visual cortex uses filterbanks of this type and that edges can be detected by so-called energy filters which sum the squares of even and odd filter pairs. In the computer vision literature, Perona and Malik [28] have advocated filters pairs of this type because of their sensitivity both to step edges (due to the odd filters) and to ridge edges (due to the even filters). See also [17].

In this section, we consider two types of filter pairs. Firstly, we consider even and odd Gabor filter pairs where the even filter is a cosine Gabor (shifted to eliminate the DC term) and the odd filter is a sine Gabor with the same orientation and frequency. We quantize the orientation angles to take 4 values. For each angle the filters are separable with a component in the direction of the angle and in the orthogonal direction. The cross-sections of the Gabor filters in the orthogonal direction is given by the real and imaginary parts of $G(x;\sigma)(e^{2\pi xi/\lambda} - e^{-2(\pi\sigma)^2/\lambda^2})$, where $G(x;\sigma) = \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/(2\sigma^2)}$. Motivated by biological considerations we set $\sigma = \lambda/2$. The Gabors have cross sections of $G(x;\sigma\gamma)$ in the direction of the angle where γ is the aspect ratio. In summary, each Gabor filter is described by an angle θ, a wavelength λ, and the aspect ratio γ.

A well known limitation of Gabor filters is their tendency to "ring" near edges because of their high frequency response. This motivates our second choice, where the filter pairs also occur at a quantized set of angles. The cross sections orthogonal to the angles is the the second derivative of a Gaussian $\frac{d^2}{dx^2}G(x;\sigma)$ and its Hilbert transform defined by $Hf(x) = \frac{-1}{\pi}\int_{-\infty}^{\infty}\frac{f(z)}{z-x}dz$. The cross section in the direction of the angle is also $G(x;\sigma\gamma)$. For comparison to the Gabor filters, we define an effective wavelength $\lambda = \sqrt{2}\pi\sigma$. We refer to them as Hilbert transform filters. (Perona and Malik suggested the use of Hilbert transform pairs [28]). These Hilbert transform pairs are run at 6 orientations (equally spaced).

To represent different ways of combining the filter pairs we use the following notation: S_θ and A_θ denote symmetric and antisymmetric filters at orientation θ respectively, where θ is quantized to take between four and six values (chosen to span the orientation space).

We can represent the filterbank output by a single (high-dimensional) filter $\vec{\phi} = \{S_\theta, A_\theta\}$ (with dimension eight or twelve depending on the number of angles). Alternatively, there are four or six "energy" filters $S_\theta^2 + A_\theta^2$ tuned to the orientations θ. In addition, we test filters which average over angular direction, $S^2 = \sum_\theta S_\theta^2$ and $A^2 = \sum_\theta A_\theta^2$, as well as the two-dimensional filter $\{S^2, A^2\}$. Finally, there is the one-dimensional filter $S^2 + A^2$. Our results, see section (3.2.2), show that a surprising amount of information is given by $S^2 + A^2$.

2.2 Performance Criteria

We use two performance criteria. The first criterion, Chernoff Information [11] is described in section (2.2.1). It is a measure of the ease in determining which of two distributions generates a set of samples (all members of the set must be sampled from the same distribution). It arises in theoretical studies [36] of the difficulty of detecting roads in aerial images [14]. The second criterion, is the Receiver Operating Characteristic (ROC) curve [15] of section (2.2.2).

Two additional measures can be obtained from the ROC curve. The first is the area under the ROC curve, which can be shown to be equal to *one minus the error rate for the 2-alternative forced choice task* [15]. The second measure is the Bayes risk [29] which can also be obtained directly from the ROC curve (with equal prior probability for on-edge and off-edge). Surprisingly, for the edge detectors filters in this paper there is a simple empirical one-to-one mapping between the area under the ROC curve and the Chernoff information, see section (3.3.2). Moreover, the nature of the empirical ROC curves suggests that they can be approximately characterized uniquely by the area under the ROC curves, see section (3.3.2). Hence the ROC curves are also directly related to the Chernoff information.

Both performance criteria are measures of statistical discriminability where the discrimination is done using the log-likelihood ratio test [11]. Therefore both performance measures depend only on the *induced distributions* $\hat{P}(r|\textit{on-edge}), \hat{P}(r|\textit{off-edge})$ on the log-likelihood ratio $r = \log \frac{P(\phi|\textit{on-edge})}{P(\phi|\textit{off-edge})}$. These induced distributions are one-dimensional and empirically are approximately Gaussians with identical variances. This will be important when understanding the empirical relationship between the Chernoff and ROC curves.

Note that both criteria were derived for discrimination formulated as probabilistic inference [11],[15]. It is not straightforward to apply them to edge detectors which are not

formulated in probabilistic terms. For example, the ROC curve assumes that there is a one-dimensional parameter that can be varied. For statistical edge detection, this parameter corresponds to the threshold used for edge detection. But conventional edge detectors can contain several adjustable parameters. For example, the Canny detector [9] contains three adjustable parameters (one scale and two thresholds). Bowyer *et al* [7, 8] obtain ROC curves by choosing the optimal selection of these parameters.

2.2.1 Chernoff Information

Our first performance measure, the Chernoff information [11], is motivated by the following question: suppose we wish to determine whether a *set of samples* is more likely to be on-edge or off-edge. This task is important when determining whether to "group" a set of image pixels to form a continuous edge path. The Chernoff information and the closely related Bhattacharyya coefficient are directly related to the *order parameters* determined by Yuille and Coughlan [36] when analyzing the Geman and Jedynak theory of curve detection [14]. In this theory, the larger the Chernoff information between the probability distributions of filter responses on and off edges, then the larger the order parameter and the easier it becomes to detect the curve.

Let $\vec{y} = \{y(x_1), y(x_2), ..., y(x_N)\}$ be a sequence of independent samples of the responses of the edge detector at positions $x_1, ..., x_N$. Using the Neyman-Pearson lemma [11], the optimal test (e.g., the maximum likelihood test) for determining whether the samples come from $P(.|\textit{on-edge})$ or $P(.|\textit{off-edge})$ depends only on the *log-likelihood ratio*, $r \equiv \log \frac{P(\vec{y}|\textit{on-edge})}{P(\vec{y}|\textit{off-edge})}$. By the assumption of independence, this reduces to $r = \sum_{i=1}^{N} \log\{\frac{P(y(x_i)|\textit{on-edge})}{P(y(x_i)|\textit{off-edge})}\}$.

The larger the log-likelihood ratio, then the more probable that the measurement sample $\vec{y}$ came from the on-edge rather than off-edge distribution (if the log-likelihood ratio is zero then both on-edge and off-edge are equally probable). It can be shown [11] that, for sufficiently large N, the expected error rate of this test decreases exponentially by $e^{-NC(P(.|\textit{on-edge}),P(.|\textit{off-edge}))}$ where $C(p, q)$ is the *Chernoff Information* [11] between two probability distributions p and q, defined by:

$$C(p, q) = -\min_{0 \leq \lambda \leq 1} \log\{\sum_{j=1}^{J} p^{\lambda}(y_j) q^{1-\lambda}(y_j)\}, \tag{1}$$

where $\{y_j : j = 1, ..., J\}$ are the variables that the distributions are defined over (in this paper,

each y_i corresponds to a histogram bin). A closely related quantity is the Bhattacharyya coefficient:

$$B(p,q) = -\log\{\sum_{j=1}^{J} p^{1/2}(y_j)q^{1/2}(y_j)\}. \tag{2}$$

Empirically, however, we find that the Chernoff information for our edge detection filters almost always corresponds to a value of $\lambda \approx 1/2$, see section (3). Therefore the Chernoff information and the Bhattacharyya coefficient give very similar values in our application domain. The only situation where this does not happen is when there is too little data and the model starts to overlearn. In the general case, however, $C(p,q) \geq B(p,q)$ for any p, q (because Chernoff information selects λ to minimize $\log\{\sum_{j=1}^{J} p^{\lambda}(y_j)q^{1-\lambda}(y_j)\}$ with respect to λ while the Bhattacharyya coefficient just sets $\lambda = 1/2$).

To illustrate the Chernoff information, we first calculate it for two univariate Gaussians with variances σ^2 and means μ_1, μ_2. It becomes $(\mu_1 - \mu_2)^2/(8\sigma^2)$ nats (1 nat equals $\log_2 e$ bits), and for the special case when $\mu_2 - \mu_1 = \sigma$, the Chernoff information equals 0.125 nats.

2.2.2 Receiver Operating Characteristic Curves

We also evaluate the edge detection filters using ROC curves [15] for classifying individual pixels.

Pixels are classified as "on-edge *" or "off-edge *" depending on whether the log-likelihood ratio $\log \frac{P(\phi=y|\textit{on-edge})}{P(\phi=y|\textit{off-edge})}$ is above or below a threshold T, respectively. Each threshold T yields a point on the ROC curve corresponding to the proportion of correct responses ($P(on-edge^*|\textit{on-edge})$) and false positives ($P(on-edge^*|\textit{off-edge})$), see figure (5).

We use two additional measures which can be derived from the ROC curve: (i) the area under the ROC curve (which is *one minus the error rate for the 2-alternative forced choice task* (2AFC)), and (ii) the Bayes risk given by $(1/2)\{P(\textit{on-edge}^*|\text{off}-\text{edge})+P(\textit{off-edge}^*|\textit{on-edge})\}$, where pixel x is classified as "on-edge * if $P(\phi(I(x))|\textit{on-edge}) > P(\phi(I(x))|\textit{off-edge})$ and as "off-edge * otherwise.

2.3 Two Non-Parametric Probability Representations

We will consider two non-parametric ways to represent probability distributions. The first uses multi-dimensional histograms with bin boundaries chosen adaptively for each dimension

(one dimension for each visual cue). The number of bins used by this representation increases exponentially with the number of visual cues. Learning such a distribution requires a large amount of training data to avoid overlearning [34], which occurs when we do not have enough data to learn the probability distributions accurately (ie. we can *memorize* the distributions but we cannot *generalize* from them to new data). This motivates our second representation which uses decision trees [29] to select those bin boundary cuts which best help discrimination. This representation enables us to learn distributions for high-dimensional filters.

We use cross-validation [29] to determine if overlearning has occurred. This procedure learns distributions on one part of the dataset and checks for consistency by evaluating them on the rest. For example, suppose we try to learn the distributions for a nine-dimensional filter with six bins for each dimension (ie. 6^9 bins in total). Then cross-validation shows that we cannot accurately learn the distributions, see figure (6). In practice, simple clues are often sufficient to tell us whether overlearning is occurring. Firstly, overlearning only occurs when the number of bins is of the same order of magnitude, or larger, than the number of datapoints. Secondly, the our performance criteria will give suspiciously large values when overlearning is occurring.

The adaptive binning and the decision tree procedure uses performance measures to determine good choices of bin boundaries and decision cuts. These performance measures, Chernoff information and Receiver Operation Characteristic (ROC) curves, were described in the previous subsection (2.2).

2.3.1 Multi-dimensional Histograms with Adaptive Binning

Recall that any edge cue (or combination of cues) is represented by *an operator* $\phi(.)$ which can be a linear, or non-linear, filter with scalar or vector valued output. For example, one possibility is the scalar filter $\left|\vec{\nabla}(.)\right|$, see subsection (2.1) for other filters.

Having chosen an edge operator $\phi(.)$, we have to quantize its response values. This involves selecting a finite set of possible responses $\{y_j : j = 1, ..., J\}$. The effectiveness of the operator will depend on this quantization scheme, so care must be taken to determine that the quantization is robust and close to optimal.

We illustrate the quantization on the filter $|\vec{\nabla}|_{\sigma=1}(Y)$. For one-dimensional filters there

is always sufficient data to learn histograms with 256 bins for $P(\phi = y|\textit{on-edge})$ and $P(\phi = y|\textit{off-edge})$. Figure (2) shows that the probability distribution for $P(\phi = y|\textit{off-edge})$ is strongly peaked near $y = 0$ (i.e. the image gradient tends to be small away from edges) while the peak of $P(\phi = y|\textit{on-edge})$ occurs at larger values of y (i.e., the image gradient is likely to be non-zero at edges). We compute the Chernoff information between these two distributions to give an upper bound for how well we can discriminate between the distributions. Then we select bin boundaries which maximize the Chernoff information in a greedy manner and compute how the Chernoff information increases towards the upper bound as the number of bins increases. This is plotted in figure (2) and shows that the Chernoff information quickly reaches its asymptotic value with only a small number of bins. It became clear that most of the reliable information can be extracted using only six adaptive bins for each dimension of the filter (this adaptation is performed over the entire dataset and *not* for each individual image).

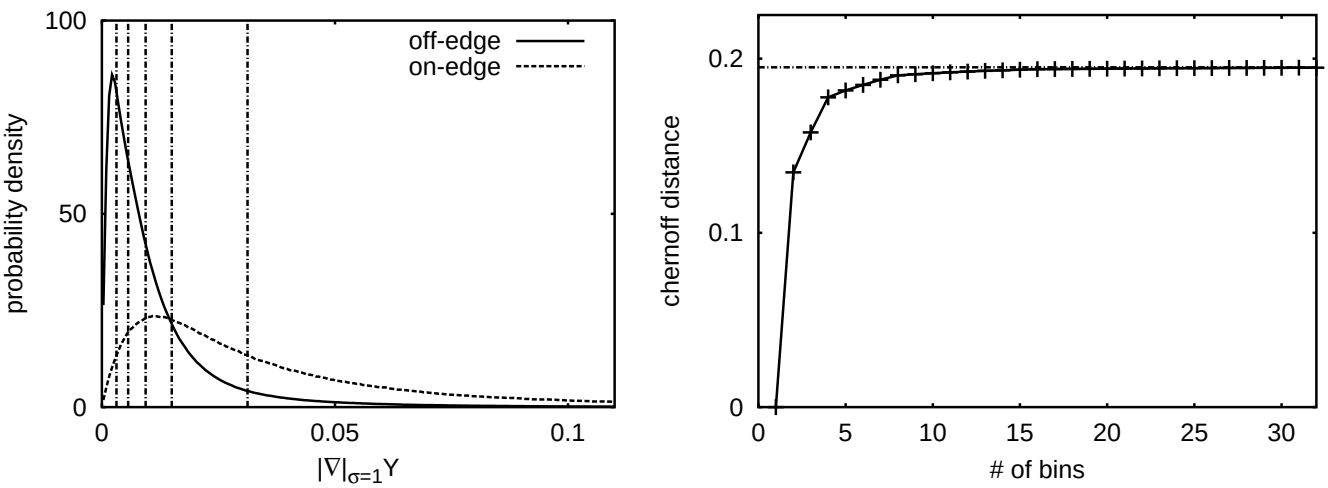

Figure 2: Left Panel: the marginal distributions of the magnitude of the gradient filter on Y at $\sigma = 1$ (evaluated on the Sowerby dataset). The vertical axis labels the probability density and the horizontal axis labels the filter response. The dark line indicates $P(\phi = y|\textit{off-edge})$ and the dotted line shows $P(\phi = y|\textit{on-edge})$. The vertical dashed lines indicate the positions of the bin boundaries which are adaptively chosen. Right Panel: the Chernoff information rapidly reaches an asymptotic value as a function of the number of bins.

For higher-dimensional filters we simply use rectangular bins with the boundaries given by the one-dimensional marginals.

2.3.2 The Decision Tree Representation

The main disadvantage of the multi-dimensional histogram representation is that the number of bins used increases exponentially with the dimensionality of the edge filters, and so the amount of training data required also grows exponentially. This puts limits on the dimensionality of the edge filters that we can use.

The decision tree approach gives a more compact representation. Moreover, it also allows us to learn probabilities in situations where overlearning occurs by adjusting the size of the representation, see subsubsection (2.3.3).

The decision tree procedure consists of adaptively selecting cuts on any of the one-dimensional filter axes so as to maximize the Chernoff information, see figure (3). We use a greedy algorithm to select the best choice of bins. That is, we find the kth cut by adding the bin boundary that maximizes the Chernoff information given the best $k-1$ cuts. More precisely, suppose we have an M-dimensional filter with one-dimensional bin boundaries at $\{y_m^i : \ i = 1, ..., n, m = 1, ..., M\}$ (where n is the number of bins used in the one-dimensional histograms – typically $n = 6$ in this paper). The distributions of the filters are $P(\phi = y|\textit{on-edge})$ and $P(\phi = y|\textit{off-edge})$. With no cuts, the two distributions $P(\phi = y|\textit{on-edge})$ and $P(\phi = y|\textit{off-edge})$ are, of course, indistinguishable. We then find the best cut y_m^i which maximizes the Chernoff information between the two distributions. Then we choose the second best cut (given the first best cut), and so on. This is an alternative way of representing the probability distributions with the number of bins bounded above by 2^k where k is the number of cuts.

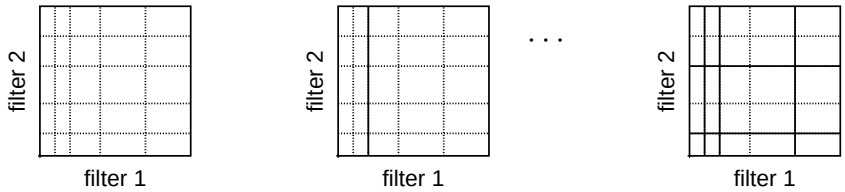

Figure 3: Example of Decision Cuts. No cuts (Left Panel), one cut (Centre Panel), and multiple cuts (Right Panel).

The decision tree procedure, see figure (4), shows that the bulk of the information content can often be obtained using remarkably few decision cuts. For example, with six cuts (ie. $n = 6$) we typically obtain between eighty and ninety percent of the total Chernoff information.

This gives a good approximation to the full histograms using at most $2^6 = 64$ bins instead of $6^9 = 10,077,696$ bins. Indeed, a single cut (i.e. using the marginal distribution of a single filter) typically yields between forty and fifty percent of the total Chernoff information. This shows that there is diminishing returns for adding extra filters of the type we have considered so far and for the binary on-edge versus off-edge decision task.

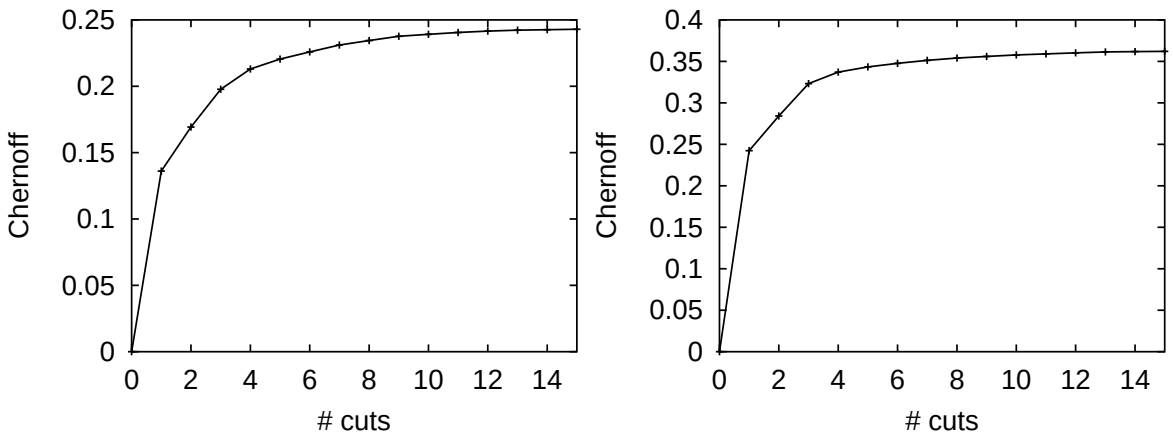

Figure 4: The decision trees for the Sowerby (Left panel) and South Florida (Right panel) datasets. The Chernoff informations approach an asymptote at about 6 cuts and a single cut gives over half the total Chernoff information. The filter is $|\vec{\nabla}|_{\sigma=1,2,4}Y$.

2.3.3 Overlearning, Cross-Validation, and Decision Trees

The decision tree procedure also allows us to learn probability distributions for high-dimensional filters for which overlearning occurs. For each number of decision cuts, we use cross-validation to test whether we are overgeneralizing or not (using either Chernoff or ROC as the performance criterion). This enables us to determine the maximum number of decision cuts we can make while preventing overlearning. The number of on-edge and off-edge pixels are $(2.35 \times 10^6, 34.3 \times 10^6)$ on Sowerby and $(4.31 \times 10^5, 12.1 \times 10^6)$ on South Florida.

To do cross-validation, we randomly divide the dataset (Sowerby or South Florida) into two sets, set0 and set1. We learn the distributions on both datasets as a function of the number of decision cuts. Then we calculate the Chernoff information and/or ROC curves *within* the two datasets (by evaluating set0 on set0 and set1 on set1) and *across* the two datasets by evaluating set0 on set1, and set1 on set0.

For example, we can calculate the ROC curves for the filter $|\vec{\nabla}|_{\sigma=1,2,4}(Y, I, Q)$. The filter

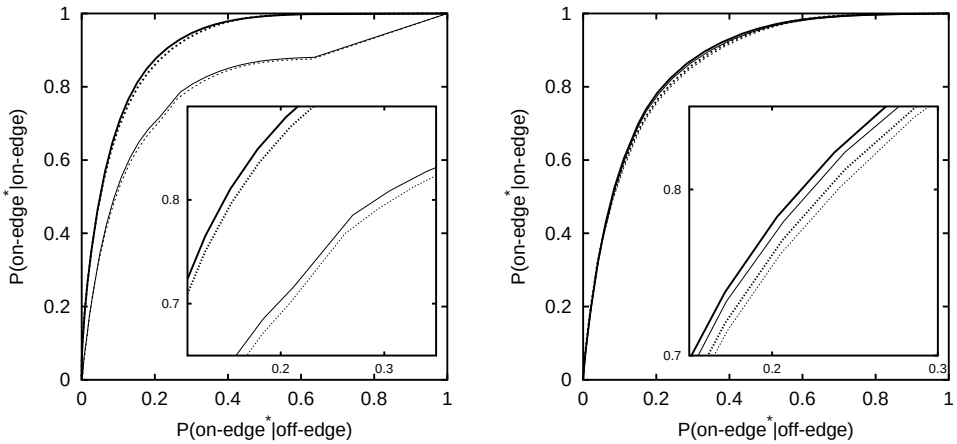

Figure 5: Cross validation for the Sowerby dataset using the filter $|\vec{\nabla}|_{\sigma=1,2,4}(Y, I, Q)$. The inset boxes show blown-up sections of the ROC curves. Left panel shows that the within-set ROC curves (dark bold and dark dashed at top) and across-set ROC curves (light bold and light dashed at bottom) are not consistent (ie. do not overlap) and so overlearning occurs. Right panel, same as above except that we now use decision trees with 20 cuts. The resulting ROC curves are now far more consistent.

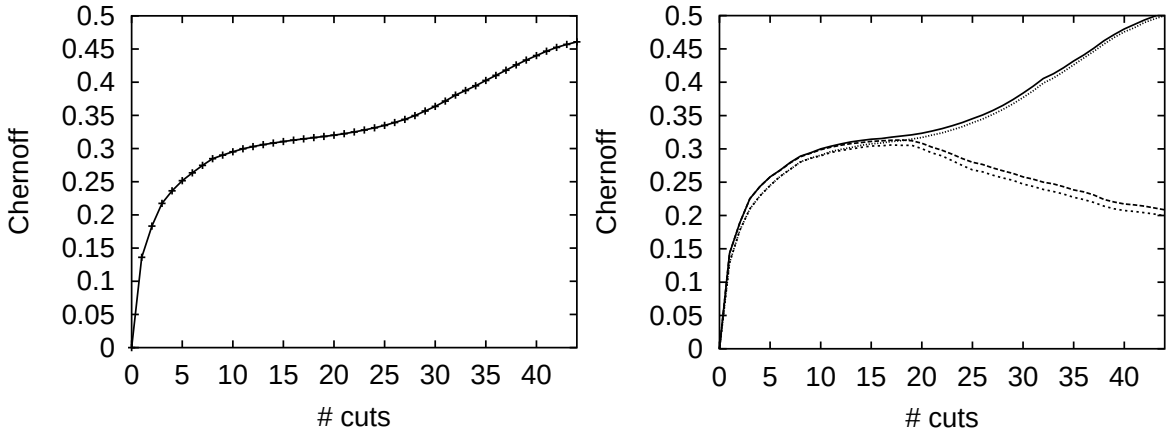

Figure 6: Overlearning for the Sowerby dataset using filter $|\vec{\nabla}|_{\sigma=1,2,4}(Y, I, Q)$. Left Panel: The Chernoff information as a function of the number of decision cuts suggests overlearning. The Chernoff reaches a plateau at 10- 20 cuts but then starts slowly rising again, which is a good (empirical) warning of overlearning. Right Panel: Overlearning is confirmed by Cross-Validation where we plot the two within-set Chernoffs (solid and dotted) and the two between-set Chernoffs (dashed lines). The lack of consistency (overlap) between these curves shows that overlearning occurs if we use more than 20 cuts. The most reliable Chernoff is 0.322.

is nine-dimensional and hence has $6^9 = 10.077696 \times 10^6$ bins which is too large to learn reliably because it is the same order of magnitude as the number of on-edge and off-edge pixels in the Sowerby dataset. If we attempt to learn the distributions using the multi-scale histograms then the within-set ROC curves are not consistent with the between-set ROC's and so we get overlearning, see left panel of figure (5). But if we use a decision tree representation with 20 cuts then all the ROC curves are consistent, see figure (5) (right panel), and there is no overlearning. The decision tree procedure reduces the number of bins to 13.8×10^3 which is far smaller than the amount of on-edge and off-edge Sowerby pixels.

Alternatively, we can check for overlearning by using the Chernoff information. In figure (6), left panel, we plot how the Chernoff information increases with the number of cuts. Observe that the Chernoff rapidly increases to a plateau at about 10 cuts but then starts to rise again at 20 cuts. In our experience, this rise from the plateau is always a sign of overlearning. To verify this, observe the results of cross-validation in the right panel of figure (6). This rise from the plateau can be used as a heuristic to check whether overlearning is occurring.

By this technique we can use higher-dimensional filters than is possible with our adaptive histogram approach. This is particularly useful when using the oriented filterbank, see subsubsection (2.1.2). The filterbanks require a lot of data because they involve running filter pairs at 4 or 6 orientations. For example, if we use 4 orientations then the filterbank is eight dimensional and requires 1.679616×10^6 bins which is too large to learn on the South Florida dataset. But the decision tree approach reduces the number of bins to 10^4 and prevents overlearning, see figure (13).

3 Edge Discrimination Results

We now describe our experimental results where the goal is to determine whether a given pixel is *on* or *off* an edge.

We evaluate our approach on both the Sowerby and South Florida datasets. These datasets differ in important respects which we describe in subsection (3.1). Then we evaluate cues using the Chernoff information in subsection (3.2), and ROC curves in subsection (3.3). It is shown in section (3.3.2) that both criteria give similar results.

3.1 The Two Datasets

The Sowerby dataset contains one hundred pre-segmented colour images. The South Florida dataset contains fifty greyscale images. These datasets differ both by the nature of the images and by the methods used to construct the segmentations (the ground truth).

Figure 7: Top row: four typical images from the Sowerby dataset which contains a variety of urban and rural scenes (the original images are in colour). Bottom Row: the ground truth segmentations supplied with the Sowerby image dataset. The ground truth is not perfect; some edges are missing and some are several pixels wide.

The Sowerby images, see figure (7), are outdoor images taken in England which all contain a road or a track. The image background contains a lot of vegetation (eg. grass, brambles, trees) which corresponds to texture in the image. The ground truth include edges which are not very distinct or poorly delineated. They include, for example, the boundary between a footpath and the grass which surround it. Overall, the dataset is a challenge for edge detectors and, in particular, for those which only use greyscale information. By contrast, the South Florida dataset, see figure (8), largely consists of indoor images. There is very little background texture. Moreover, the ground truth edges are often visually salient and spatially localized (eg. only one pixel wide).

We assume that it is far easier to detect edges correctly in the South Florida dataset than in Sowerby. The edges are sharper and the background statistics are less complicated (due to the lack of texture). These assumptions are born out by our experimental results in the rest of this section.

The ground truths in the two datasets were clearly created differently, see figures (7,8).

Figure 8: Left panel: four typical images from the South Florida dataset, which consists primarily of indoor images and man-made objects. Right panel: the ground truth segmentations supplied with the South Florida image dataset.

For example, the South Florida edges are thin and well localized. By contrast, the Sowerby edges are thick (eg. often two pixels wide). Moreover, the South Florida images have a 3-valued ground truth while the Sowerby images have 2-values. For South Florida, the 3-values correspond to three sets: (a) edge, (b) background, and (c) pixels close to edges and some texture regions in the background. By contrast, Sowerby image pixels are labelled either as edge or non-edge. In our experiments we always reclassify South Florida pixels as either edge or non-edge (ie. the non-edge set is the union of sets "b" and "c").

Five images from the Sowerby set (out of a hundred and four) have very poor quality edge maps and so we rejected them. These images are 06-36, 10-19, 13-10, 13-13, 14-22.

It is very useful for us to have two datasets which differ both in their statistics and their criteria for ground truth. Firstly, as we will show in section (5), we are able to learn the statistics on one dataset and then adapt them to the other with only a small loss in performance. This shows that statistical edge detection is robust to errors in the ground truth (because it would be impossible to achieve this level of adaptation if the edge statistics were very sensitive to the rather different ground truth criteria used in the two datasets). Secondly, statistical edge detection gives only slightly better results than standard edge detectors on the (easier) South Florida dataset (as evaluated by the Bayes risk -see section (3.3.2)). But statistical edge detection does better on the (harder) Sowerby dataset. See figures (1, 17)

for visual comparison and then compare the ROC results for Canny detector and statistical edge detection in figure (16).

3.2 Results using Chernoff Information

We show results for the first set of filters in subsection (3.2.1) and for the second set of filters in subsection (3.2.2).

To calibrate the Chernoff information for individual cues, we estimated it to be 0.22 nats for the Geman and Jedynak road tracking application [14]. Recall that it equals 0.125 nats for two univariate Gaussians when the difference between the two means is equal to the variance. These set a baseline and, as we will show, we can obtain Chernoff information significantly higher by combining cues.

To calibrate the Chernoff information for multi-dimensional filters, we need to know how it can change as a function of the dimension. It is guaranteed to never decrease but in principle it could increase by an arbitrarily large amount [11]. For example, consider two distributions $p(i,j) = 1/n^2$ for $i = 1,..,n$ and $j = 1,...,n$, and $q(i,j) = (1/n)\delta_{ij}$. Then the marginal distributions, over i or j, are identical for both distributions, and so the Chernoff information and Bhattacharyya coefficient are zero for the marginals. But the Chernoff information and Bhattacharyya coefficient between p and q are $\log n$ and $(1/2)\log n$, respectively.

If we combine two cues which are independent then the Chernoff information will be less than, or equal to, the sum of the Chernoff informations for each cue. But empirically we always found that the Chernoff information is approximately equal to the Bhattacharyya coefficient (ie. $\lambda \approx 0.5$, see section (2.2.1)). If two independent edge cues are combined, then their Bhattacharyya coefficients will simply add [11]. Hence we expect that the Chernoffs will approximately add if the cues are independent.

In practice, we found that the Chernoff information and Bhattacharyya coefficients of two coupled cues is usually a lot less than the sum for the individual cues, see section (3), so we conclude that cues are rarely independent.

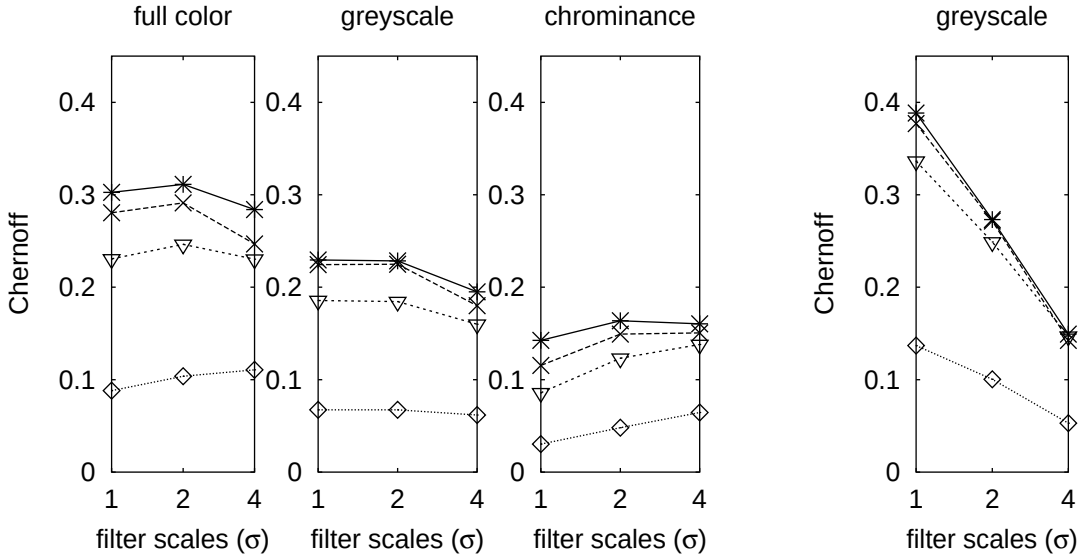

Figure 9: Chernoffs for Sowerby and South Florida. The edge detector operators are labelled by stars for (N_1, N_2), crosses for N_1, triangles for $|\vec{\nabla}|$, and diamonds for ∇^2. The three Leftmost Panels plot the Chernoff Information for Sowerby for full colour, greyscale, and chrominance respectively. The Far Right Panel plots Chernoff for South Florida for greyscale. The horizontal axis shows the filter scale ($\sigma = 1, 2, 4$). Decision trees are not needed.

3.2.1 Results for First Set of Filters

We now show the results on a range of filters, see table (1). Recall from section (2.1) that the basic ingredients are: (I) three differential operators (see below), (II) the three different colours (image bands Y, I, Q), and (III) three scales obtained by convolving the image with a Gaussian at scale $\sigma = 1, 2, 4$ pixels.

Our first result, see figure (9), compares filter performance of $(N_1, N_2), N_1, |\vec{\nabla}|, \nabla^2$ using filters at different scales, different choices of colour bands, and for Sowerby and South Florida. The first two panels illustrate the advantages of colour over greyscale. (The advantage of using colour for edge detection has sometimes been doubted in the computer vision community). It is interesting that the chrominance cues (for which the grey-scale has been factored out) are most effective at large scales, see centre right panel. This corresponds nicely with biological vision (for which the chrominance filters tend to have larger spatial scales than the grey-scale filters). The center left and far right panels show that it is easier to detect edges in South Florida than it is in Sowerby. Moreover, the figure shows that Sowerby edges are easiest to detect at large scales while South Florida edges are easiest at low scales

(ie. South Florida edges are sharply localized).

The Nitzberg filter (N_1, N_2) is good presumably because it can discriminate between edges and textures. Texture is treated as "corners" with two eigenvalues being large. By contrast, at regular edges only one eigenvalue is large. But this means that the Nitzberg filter often treats true edge corners as texture, and so classifies them as off-edge.

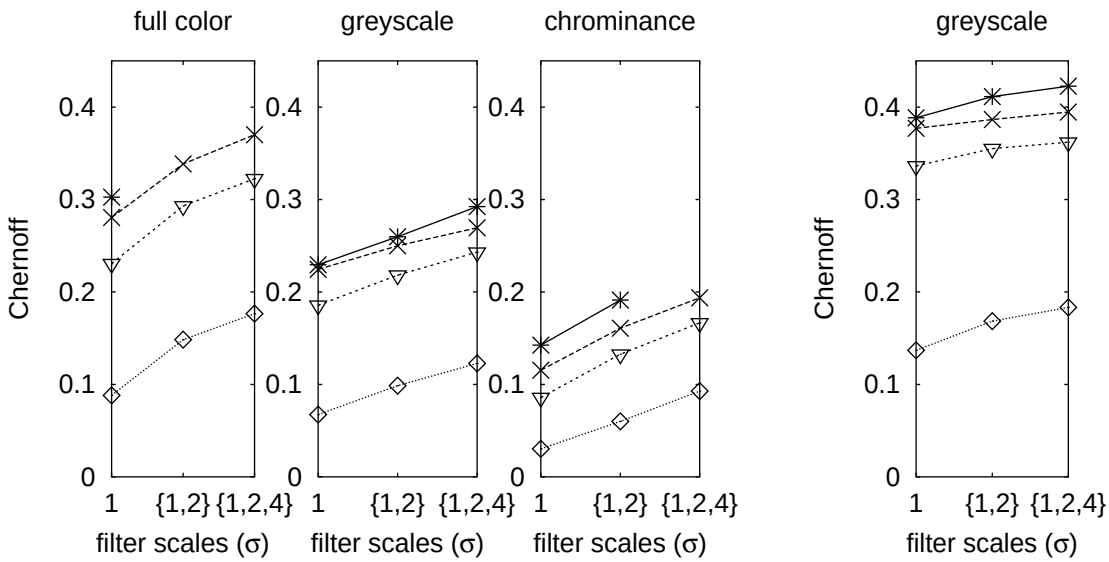

Figure 10: The advantages of using multi-scale filters. The Chernoff information is shown for: 1 the filter at scale $\sigma = 1$, $\{1, 2\}$ the coupled filter for scales $\sigma = \{1, 2\}$, and $\{1, 2, 4\}$ the coupled filter for scales $\sigma = \{1, 2, 4\}$. The Chernoff always increases as we add larger-scale filters. Conventions as in figure (9). Decision trees are required when applying filters $\nabla^2, |\vec{\nabla}|$ to (Y, I, Q) at scales $\sigma = 1, 2, 4$, and when applying (N_1, N_2) to chrominance at scales $\sigma = 1, 2$.

Figure (10) shows that multi-scale processing is very effective. The combination of using operators at scales $\sigma = 1, 2, 4$ always improves the Chernoff significantly. This increase is particularly strong for the Sowerby dataset. Multiscale is better able to discriminate between texture edges (which should be discounted) and the edges which correspond to boundaries. It is also able to detect edges of different widths (which occur in Sowerby but rarely in South Florida).

We analyze the consistency of these results for each image by learning distributions $\{P^i(.|\textit{off-edge})\}$ and $\{P^i(.|\textit{on-edge})\}$ for each image and calculating the Chernoffs. We plot this as a relief map, see figure (11). This shows that although the Chernoff information

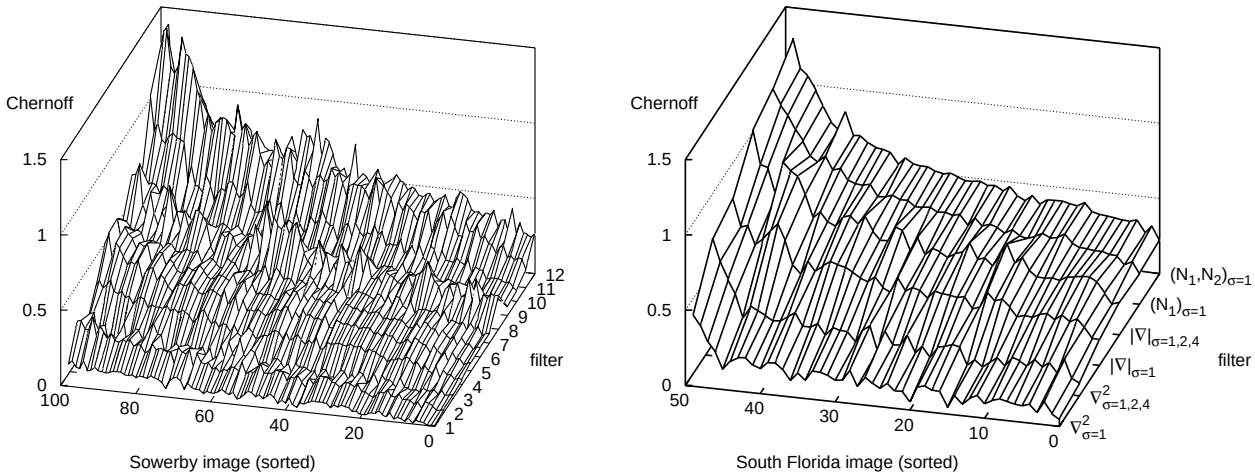

Figure 11: The relative effectiveness of filters is fairly consistent over the entire datasets. We plot the Chernoff information as a function of the filter used and the image number in the dataset (with images sorted by magnitude of Chernoff). For Sowerby (Left Panel) the filters are those from Table 1. For South Florida (Right Panel) the filters are $\nabla^2_{\sigma=1}, \nabla^2_{\sigma=1,2,4}, |\nabla|_{\sigma=1}, |\nabla|_{\sigma=1,2,4}, (N_1)_{\sigma=1}, (N_1, N_2)_{\sigma=1}$.

varies from image to image the *relative effectiveness* of the filters is approximately the same (we order the images so that the Chernoff increases monotonically).

Figure (12) investigates the consistency of the distributions between images. More precisely, we plot the variations of $\{P^i(.|\textit{off-edge})\}$ and $\{P^i(.|\textit{on-edge})\}$ relative to the $P(.|\textit{on-edge})$ and $P(.|\textit{off-edge})$ obtained for the entire dataset. The variations are measured by the Chernoff information. This shows that the $\{P^i(.|\textit{off-edge})\}$ and $\{P^i(.|\textit{on-edge})\}$ separate nicely into two non-overlapping sets. Hence the distributions are fairly consistent between images.

Despite the difference between country road scenes in England (Sowerby dataset) and primarily indoor images in Florida (South Florida dataset), perhaps the most striking observation is that the relative effectiveness of different filters is approximately unchanged, see figure (11).

3.2.2 Oriented Filterbank Results

Overlearning was a significant problem when learning the statistics of the filterbank and so we often used the decision tree representation.

The results we obtained for the filterbanks were slightly surprising, see figure (13). We

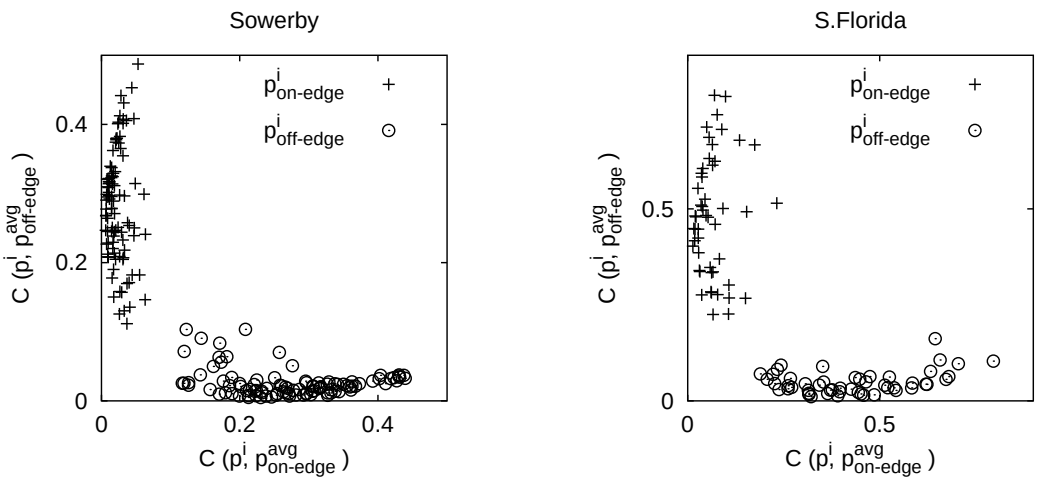

Figure 12: We illustrate that the $P(.|\textit{on-edge})$ and $P(.|\textit{off-edge})$ for all the images cluster nicely into two disjoint sets for Sowerby (Left Panel) and South Florida (Right Panel). The filter is $|\nabla|_{\sigma=1,2,4}Y$. More specifically, we plot $C(P, P(.|\textit{on-edge})), C(P, P(.|\textit{off-edge}))$ for $P = P^i(.|\textit{on-edge})$ (pluses) and $P = P^i(.|\textit{off-edge})$ (circles), where $i = 1, ..., 99$ labels the image.

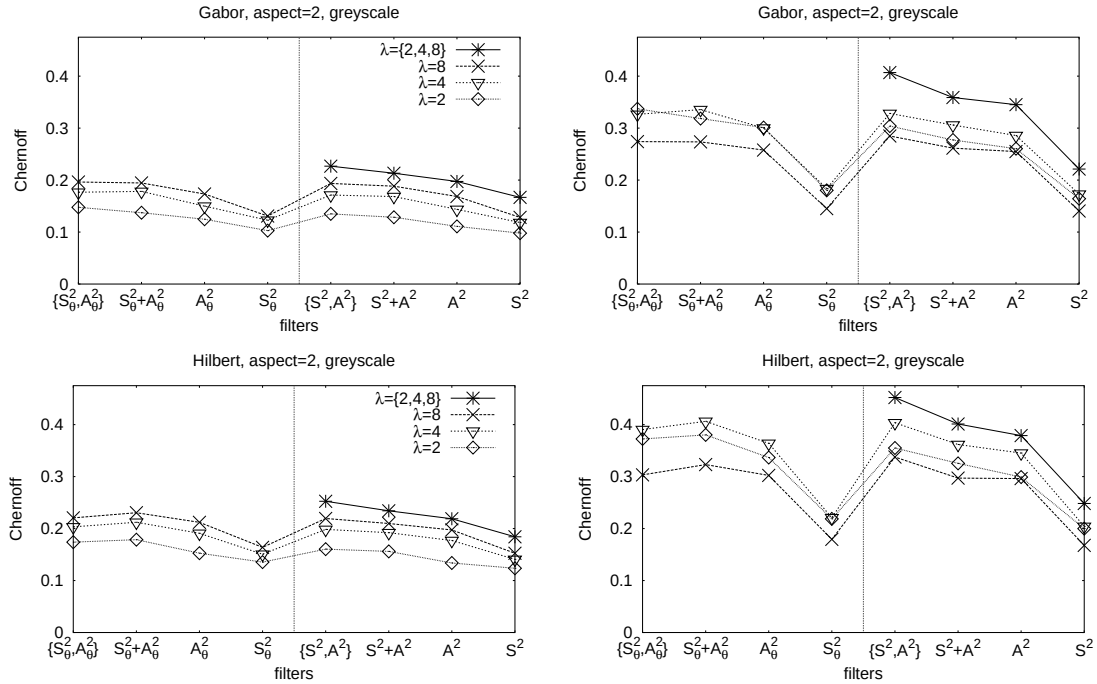

Figure 13: Oriented Filters on Sowerby (left panels) and South Florida (right panels). See section (2.1.2) for the label definitions. Gabor filters (top panels) and Hilbert transform filters (bottom panels). See text for interpretation.

showed that: (I) The energy filters $S^2 + A^2$ were very effective and there was little advantage, as measured by the Chernoff information, in using the joint distributions on all the filters (which is the optimal approach). (II) The Hilbert transform filters yield clearly better performance than Gabor filters, probably due to their lack of "ringing". (III) Summing the energy from all different orientations gave a one-dimensional filter whose performance was close to optimal (a major surprise to some of the authors). (IV) Finally, the Hilbert transform filters including the one dimensional filter (see III) were comparable to the best of the filters previously tested (the Nitzbergs), see greyscale panels in figure (6).

These figures are for aspect ratio $\gamma = 2$ (that is, the filters are twice as long as their envelope in the frequency-tuned direction). For aspect $\gamma = 1$ the Chernoff informations go down by up to ten per cent. Coupling aspects $\gamma = 1$ and $\gamma = 2$ improves performance by about five percent (over $\gamma = 2$).

3.3 ROC results

We can also evaluate the filters using ROC curves, see figure (5). There are two main ROC results. Firstly, see section (3.3.1), there is a simple empirical relationship between the area under the ROC curve and the Chernoff information. Moreover, empirically most of the form of the ROC curve is determined by the area under it. Hence ROC curves and Chernoff information give very similar results. Secondly, see section (3.3.2), we can use ROC curves to compare statistical edge detection to standard edge detectors for South Florida and Sowerby.

3.3.1 Relating Chernoff Information and the ROC areas

In this section we give a formula that, empirically, relates the Chernoff information and the ROC curves for our filters (for both filtersets).

Firstly, when computing the ROC curves for edge discrimination, see right panel of figure (5), we noticed that they looked surprisingly similar to the ROC curves for univariate Gaussian distributions with identical variances. This implies [15] that the form of the ROC curve depends only on the quantity $d' = |\mu_2 - \mu_1|/\sigma$, where μ_1, μ_2 are the means of the Gaussians and σ^2 is their variance. The area under the ROC curve depends only on the same quantity d' and is given by $A(d') = (1/2)\{1 + erf(d'/2)\}$. So knowing the area under

the ROC curve is equivalent to knowing the ROC curve.

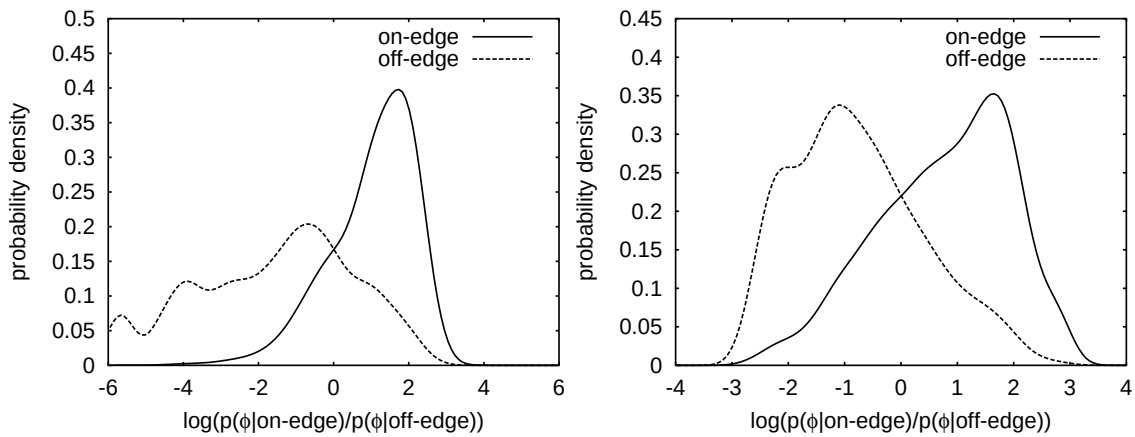

Figure 14: The induced distributions are often approximately Gaussian in the overlap region with identical variances. Probability density as a function of the log-likelihood ratio, for (left panel) $|\vec{\nabla}|_{\sigma=1,2,4}(Y, I, Q)$, (right panel) $|\vec{\nabla}|_{\sigma=1}Y$.

It is paradoxical that the ROC curves look roughly like those of univariate Gaussians with identical variances. The empirical probabilities distributions $P(.|\textit{on-edge})$ and $P(.|\textit{off-edge})$ are not remotely Gaussians. However, the ROC curves depend only on the induced distributions $\hat{P}(r|\textit{on-edge})$ and $\hat{P}(r|\textit{off-edge})$ on the log-likelihood ratio $r = \log\{\frac{P(\phi|\textit{on-edge})}{P(\phi|\textit{off-edge})}\}$ (where $\hat{P}(r|\textit{on-edge}) = \int dy \delta(r - \log\frac{P(\phi=y|\textit{on-edge})}{P(\phi=y|\textit{off-edge})})P(\phi = y|\textit{on-edge})$, $\hat{P}(r|\textit{off-edge}) = \int dy \delta(r - \log\frac{P(\phi=y|\textit{on-edge})}{P(\phi=y|\textit{off-edge})})P(\phi = y|\textit{off-edge})$). Empirically, these induced distributions are often approximately univariate Gaussians with identical variances, at least in the region of overlap of the two distributions, see figure (14). We therefore predict that the area under the ROC curve and the Chernoff information are related as if the edge and non-edge distributions were univariate Gaussians with identical variances. It is straightforward to calculate the Chernoff information to be $C(d') = (1/8)(d')^2$ which, again, only depends on d'.

Figure (15) plots the Chernoff information as a function of the area under the ROC curve. The bold line is the predicted relationship with the assumption of Gaussian distributions with equal variance. The dots correspond to the empirical results of 420 filters on our datasets. All the dots lie very close to the prediction. The right panel of figure (15) plots the ROC curves for the Univariate Gaussians (solid curve), 2-bin symmetric distributions $p = (a, 1-a), q = (1-a, a)$ (dashed line) and 2-bin asymmetrical $p = (1, 0), q = (a, 1-a)$

(line with short dashes and dotted line). The latter has two curves depending on whether we relate the ROC area to the Chernoff information or to the Bhattacharyya coefficient (for the first two distributions these quantities are equal).

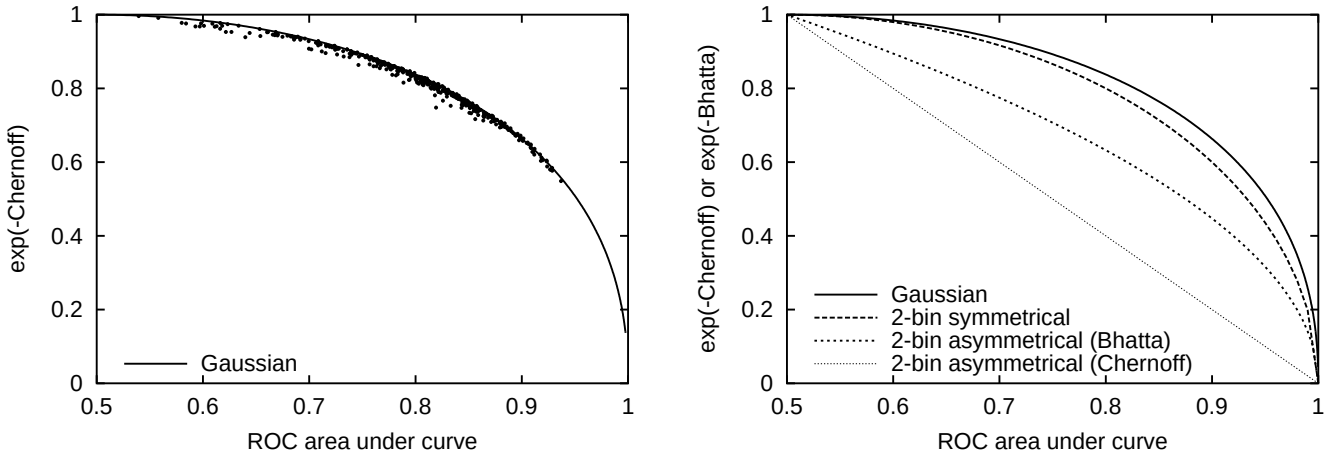

Figure 15: Left Panel: the predicted relationship (solid line) between Chernoff information and the area under the ROC curve fits our experimental data, represented by dots, very well for all of our 420 filters combinations on the Sowerby dataset. Right Panel: the relationship between Chernoff information and the area under ROC curve for three pairs of distributions, see text.

3.3.2 ROC comparison of Statistical and Standard Edge Detectors

We now compare the performance of statistical edge detection with that of the Canny edge detector. In addition, by using the results of Bowyer *et al* [7, 8], we get comparisons of statistical edge detection to other conventional edge detectors on the South Florida dataset.

There are two difficulties in comparing statistical edge detection to conventional edge detectors. Firstly, conventional edge detectors usually have a non-maximal suppression stage (Bowyer *et al* added non-maximal suppression to all of the edge detectors they tested). Secondly, most conventional edge detectors contain several tunable parameters (three for the case of Canny). Both difficulties can cause biases in the ROC curves, see examples in [23], and require non-standard methods for evaluating true positives and false positives of the edge detector responses. We will determine the ROC curves using both the evaluation method proposed by Bowyer *et al* and a new method developed here. It can be argued

that an advantage of statistical edge detection is that it requires a single parameter (the threshold) and is straightforward to evaluate using standard ROC and Chernoff criteria.

Non-maximal suppression causes two types of problem for ROC curves which, unless addressed, can make the curves extremely sensitive to errors in the ground truth. Firstly, non-maximal suppression can create a bias on the true positives by preventing an edge detector from detecting all the ground truth edges. Small errors in ground truth edge location may mean that an edge detector responds correctly at the real position of the edge which *suppresses* its response at the ground truth location. In addition, the ground truth edges may sometimes be two pixels wide and so non-maximal suppression will prevent an edge detector from labelling both pixel points as edges. Secondly, non-maximal suppression can dramatically reduce the number of false positives. This will happen in sections of the ROC curve where the proportion of false positives is high (ie. when many pixels in the image are incorrectly estimated to be edges). This corresponds to very impractical choices of the edge detector parameters and so is *not* representative of the behaviour of the edge detectors with more realistic parameter settings.

On the South Florida dataset, we adjusted our approach so that it can be directly compared with the results of Bowyer *et al*. Firstly, we applied non-maximal suppression to statistical edge detection. Secondly, we used Bowyer *et al*'s evaluation criteria, see next paragraph, to determine the true positive and false positive rates. Thirdly, we compared the edge detectors using the Bayes risk (assuming pixels are equally likely to be on or off edges *a priori*) because the Bayes risk is computed from part of the ROC curve which corresponds to reasonable choices of the edge detector parameter values.

Bowyer *et al*'s criteria for determining true positives and false positives is algorithmic. To evaluate the true positives, a list is constructed of the ground truth pixels. There is a second list consisting of the pixels which the detector labels as edges. The algorithm proceeds by scanning the first list in order. If a pixel in the first list is within three pixels of an element of the second list, then a true positive is counted *and the element in the second list is deleted*. This means that each element in the second list can "validate" at most one element of the first list and hence prevents the algorithm from overcounting the number of true positives. To evaluate the false positives, Bowyer *et al* count the number of pixels that the edge detector labels as edges in region (b) of their three-valued ground truth, see section (3.1). This means

that edge detector responses within a three-pixel distance of a ground truth edge are ignored when counting the false positives (as are edge detector responses in textured regions). These criteria can be criticized, see [23] for an example where they give a misleading measure of the performance of an edge detector, but usually they give intuitively plausible results.

However these criteria only address the first problem of non-maximal suppression (eg. biases on the true positives). There will therefore still be distortions in the ROC curves. Hence we will evaluate the edge detectors by their Bayes risk (with equal prior for pixels being *on* and *off* edge). The Bayes risk can be measured from the ROC curve by finding the point on the curve where the slope is forty five degrees [15] (this is usually close to the point where the number of false negatives equals the number of false positives – and is exactly this point if the distributions are univariate Gaussians with identical variances).

For the edge detectors evaluated by Bowyer *et al.*, we obtain approximate values of the Bayes risks in the range 0.035–0.045 [8]. Our statistical edge detection gives a Bayes risk of 0.0350 using a magnitude of the gradient filter at four scales $\sigma = 0, 1, 2, 4$ (with non-maximal suppression and Bowyer *et al*'s evaluation criteria). Our implementation of the Canny edge detector gave a similar Bayes risk of 0.0352 (which is consistent with Bowyer *et al*'s results and which validates our implementation). Overall, *statistical edge detection performed as well as any edge detector reported in [8] using the identical evaluation criteria.*

We obtained a significant difference between statistical edge detection and the Canny edge detector on the more challenging Sowerby dataset. In this case, we did not apply non-maximal suppression to statistical edge detection but instead used an additional grouping stage, described in the following section. We also modified the evaluation criteria to address both problems of the ROC curve caused by non-maximal suppression. The criteria involved using morphological operators to enlarge the number of pixels labelled as edges by the edge detector being evaluated and to produce a buffer zone around the ground truth edges (Bowyer *et al* used a similar buffer zone). They minimize the bias caused by non-maxmimal suppression while allowing for imprecisions in the ground truth segmentation. More precisely, we defined two binary fields $g(x), g^*(x)$ on the image such that $g(x) = 1$ if pixel x is a ground truth edge, and $g^*(x) = 1$ if an edge detector labels pixel x as an edge ($g(x) = 0$ and $g^*(x) = 0$ otherwise). We defined $\bar{\cdot}$ to be the complement (eg. $\bar{g}(x) = 0$ if $g(x) = 1$). We defined $_n$ to mean a morphological opening on a binary field (eg. $g^*_3(x) = 1$ for any pixel x within

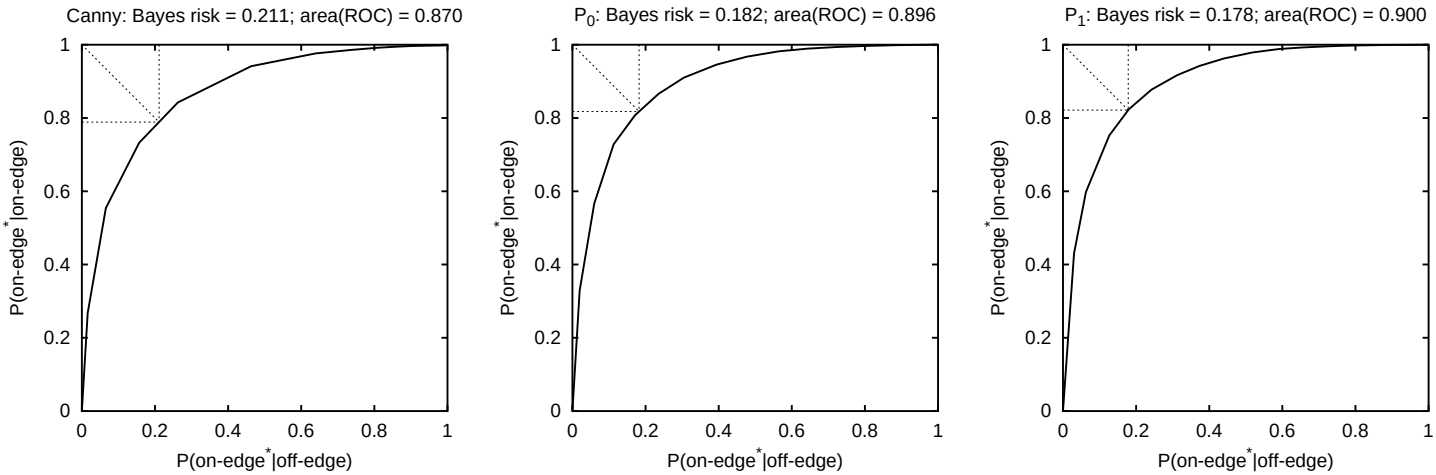

Figure 16: ROC curves for Sowerby show that statistical edge detection outperforms Canny. Left: Canny edge detector with non-maximal suppression and hysteresis. Center: Statistical edge detection without grouping. Right: Statistical edge detection with grouping. (edge-tolerance=3).

a three-pixel distance of a point labelled an edge by our detector). The proportion of true positives is defined to be $\sum_x g(x)g_3^*(x)/\sum_x g(x)$. The proportion of false positives is defined to be $\sum_x \bar{g}_6(x)g_3^*(x)/\sum_x \bar{g}_6(x)$. These criteria also have their limitations, see discussion in [23], but also give plausible results. We tested these criteria by applying them to statistical edge detection and the Canny edge detector on the South Florida dataset and showed, see [23], that they gave similar results to those obtained using Boyer *et al*'s criteria (ie. both edge detectors perform almost identically on the South Florida dataset).

Using these criteria, our results show that the statistical edge detector is significantly better than Canny on the Sowerby dataset, see figures (16,17). This applies whether or not we use grouping for statistical edge detection, see section (4). This is not surprising because the Canny detector uses one scale only and statistical edge detection uses many scales which are combined optimally (in the statistical sense). The Sowerby dataset is harder to segment than South Florida because of all the background clutter, and hence multiscale processing gives a big advantage, see figure (10).

For completeness, we also show the log-likelihood ratios, see figure (17), which can be used as measures of edge strength [14].

Figure 17: Top panels shows edges detected using the Canny edge detector. The centre panels shows the output of statistical edge detection on the same images. The bottom panels show the log likelihood ratios which give a measure of edge strength. See figure (7) for the images and the ground truth.

4 Spatial Grouping of Edge Cues

Most standard edge detectors use a form of local spatial grouping. For example, the Canny edge detector [9] uses non-maximal suppression and hysteresis. This grouping exploits prior knowledge of edges in images. Edges are typically spatially contiguous (hysteresis) and one pixel wide (non-maximal suppression). Hysteresis enables low contrast edges to be detected provided they are close to high contrast edges. Alternatively, probabilistic models like Geman and Geman [13] impose prior probabilities so that if there is an edge at one pixel location then this increases the probability of there being edges at neighbouring pixels.

We now apply statistical edge detection to include a form of spatial grouping. Properties similar to hysteresis and non-maximal suppression will arise naturally as part of the learning process. This grouping significantly improves the visual quality of our edge detection results. But paradoxically it only gives a small improvement in our performance criteria.

Our grouping procedure is similar to our method for learning $P(.|\textit{on-edge}), P(.|\textit{off-edge})$. The difference is that we apply a filter bank $\phi_1(.)$ to the posterior distributions $F_0(\vec{x}) =$

$P(edge|\phi_0(Y)|_{\vec{x}})$, where $P(edge|.)$ is the posterior probability that there is an edge at location $\vec{x}$ conditioned on the filter response $\phi_0(Y)$ evaluated at $\vec{x}$. The intuition is that the posterior, like the log-likelihood ratio in figure (17), is a measure of edge strength. (The prior probability for a pixel being an edge is measured as 0.06 from the datasets). Our grouping procedure convolves the old posterior with filterbank and learns a new "posterior" $F_1(\vec{x})$ (using the ground truth) and then repeats the process.

In theory the full procedure is: (i) start with the true posterior $F_0(\vec{x}) = P(edge|\phi_0(Y)|_{\vec{x}})$, (ii) learn $F_1(\vec{x}) = P(edge|\phi_1(F_0)|_{\vec{x}})$, (iii) iterate to learn $F_i(\vec{x}) = P(edge|\phi_1(F_{i-1})|_{\vec{x}})$ for $i = 2, 3, ...$ But in practice, we used a simplified procedure which replaces the third stage by setting $F_i(\vec{x}) = F_1(\phi_1(F_{i-1}(\vec{x})))$ for $i = 2, 3, ...$

In our experiments we used the filters $\phi_0(.) = |\vec{\nabla}|_{\sigma=0,1,2,4,8,16}(.)$ and $\phi_1(.) = (I, |\vec{\nabla}|_{\sigma=2,8}, \nabla^2_{\sigma=0,1,2,4,8})(.)$, where I is the identity filter. The most useful filters for grouping (ie. for ϕ_1) are those that enhance ridges in the posterior (these ridges correspond to edges in the images). These are the Laplacian of a Gaussian, supplemented with gradient filters. The identity filter, of course, is useful (because it gives the posterior).

We give examples of grouping in figure (18). Overall our method is good at hysteresis and enhancing edges between different textures (ie. raising edges above threshold because they lie along ridges and support each other). Edges in texture are suppressed because strong, and weak, edges tend to suppress nearby weak parallel edges. Our method also does well at triple points and corners, where the Canny filter often does poorly. On the other hand, we do not seem to thin edges as well as non-maximal suppression applied to the Canny edge detector. This may be due to the quantization used in our approach which can cause neighbouring pixels to have identical edge strength (non-maximal suppression would not solve this problem).

To quantify the gains by grouping we calculate the Chernoff information. This gives values of 0.263 (without grouping), 0.290 (one level of grouping), 0.282 (two levels of grouping), and 0.274 (three levels of grouping). The improvement with one level of grouping is small (about ten percent), but visually there are definite improvements, see figure (18). The decrease in Chernoff for two and three levels of grouping are presumably caused by our simplified procedure.

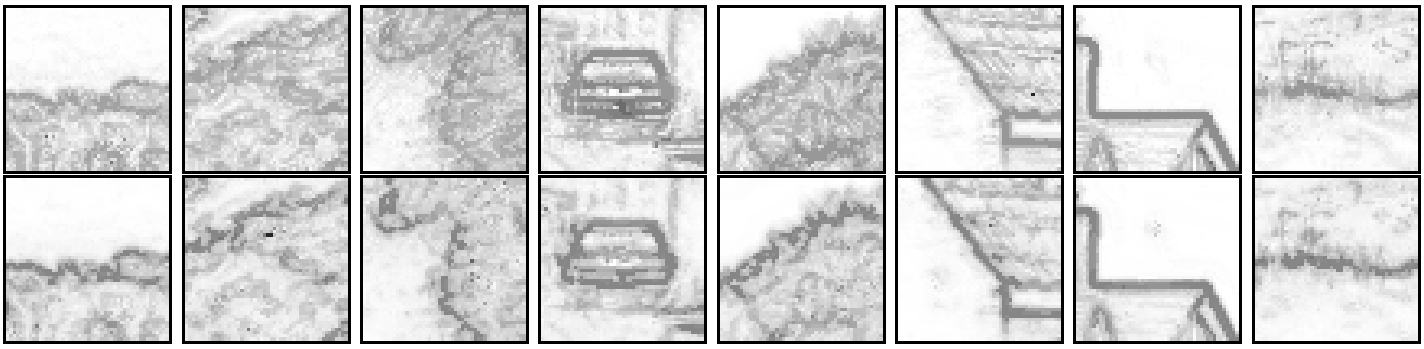

Figure 18: Grouping examples. Top Row: the posterior without grouping: $F_0(x)$. Bottom Row: the posterior after grouping $F_1(x)$. See text.

5 Adaptation Between Datasets

In this section we show that we can learn the conditional distributions on one dataset and adapt them to another with only slight degradation of performance *without* knowing the ground truth on the second. This shows that our results can be adapted from domain to domain. It also illustrates that our results are not overly sensitive to the ground truth, because otherwise such adaptation would cause larger degradation (particularly considering the difference between the ground truths in Sowerby and South Florida).

We note that Canny discusses adaptation [9] and described methods for estimating the amount of noise in images in order to change the parameters of his edge detector dynamically. But this adaptation is not commonly used. More recently, Grzywacz and Balboa [16] have described a method, using Bayesian probability theory, for how biological vision systems may adapt their receptive fields from domain to domain based on edge statistics.

Formally, we define rules to estimate distributions $P^{S|F}(\phi = y|\textit{on-edge}), P^{S|F}(\phi = y|\textit{off-edge})$ for the Sowerby dataset using only knowledge of the edge statistics in the South Florida dataset. Similarly, we use these rules to estimate distributions $P^{F|S}(\phi = y|\textit{on-edge}), P^{F|S}(\phi = y|\textit{off-edge})$ for Florida using edge statistics from Sowerby. (We use the superscripts $^{S|F}$ to indicate the distributions estimated on the Sowerby dataset using the segmentations from South Florida – and vice versa for $^{F|S}$.)

Our adaptation approach is based on using different strategies for estimating the off statistics $P^{S|F}(\phi = y|\textit{off-edge}), P^{F|S}(\phi = y|\textit{off-edge})$ and the on edge statistics $P^{S|F}(\phi = y|\textit{on-edge})$,$P^{F|S}(\phi = y|\textit{on-edge})$.

The strategy for the off statistics is to exploit the fact that most pixels in an image are not edges. Thus, for each domain, we calculate the probability distributions $P(\phi = y|all)$ of the filter responses for all the pixels (which doesn't require us to know the segmentation) to yield our estimate of $P(\phi = y|\textit{off-edge})$. (More formally, we can express $P(\phi = y|all) = (1-\epsilon)P(\phi = y|\textit{off-edge}) + \epsilon P(\phi = y|\textit{on-edge})$ where $\epsilon \approx 0.06$ is the proportion of edges in the image. Our strategy sets $\epsilon = 0.0$ and, by calculating the Chernoff information we verify that little information is lost.)

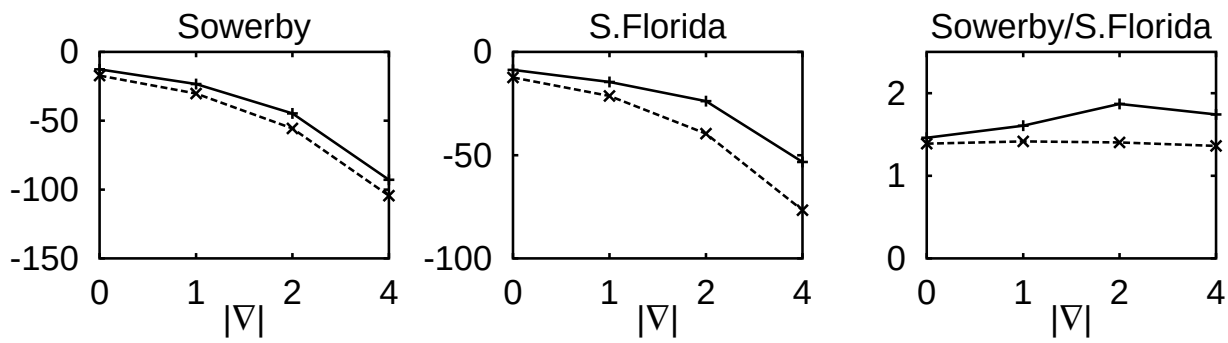

Figure 19: These figures show that for both Sowerby (left panel) and South Florida (centre panel) the asymptotic slope of $\log P(\phi|\textit{on-edge})$ (solid line) and $\log P(\phi|all)$ (dotted line) are practically identical independent of scale. The horizontal axis labels the scale of the filters and the vertical axis is the asymptotic slope of the log probability. The right panel shows that the ratios of the asymptotic slopes of $\log P(\phi|\textit{on-edge})$ for Sowerby divided by South Florida (solid line) and the ratios of $\log P(\phi|all)$ (dotted line) all have (approximately) the same value $k = 1.5$.

To adapt for $P(\phi(\vec{x})|\textit{on-edge})$ between datasets, we note that for most of our marginal filters $\phi(\vec{x})$, the distribution $P(\phi(\vec{x})|all)$ approximates the *on-edge* distribution $P(\phi(\vec{x})|\textit{on-edge})$ at large $\phi(\vec{x})$, see the left and centre panels of figure (19). We therefore have access to $P(\phi(\vec{x})|\textit{on-edge})$ (up to a scaling factor) for large $\phi(\vec{x})$, *without knowledge of the ground truth*. Empirically, we find that, for large $\phi(\vec{x})$, $P(\phi(\vec{x})|all)$ drops approximately exponentially, so if we take $\log P(\phi(\vec{x})|all)$, and calculate its asymptotic slope for large $\phi(\vec{x})$, it approximates the asymptotic slope of $\log P(\phi(\vec{x})|\textit{on-edge})$. Furthermore, if the statistics of both datasets drops exponentially, the ratio of the asymptotic slopes of $\log P(\phi(\vec{x})|all)$ yields a constant scaling factor k which relates the $\phi(\vec{x})$ of the two datasets. For adapting from

South Florida to Sowerby, we measure $k = 1.5$ for the magnitude of the gradient filter, see right panel of figure (19). We therefore take the distributions $P^S(\phi = y|\textit{on-edge})$ measured on the Sowerby dataset and adapt them by a linear scaling $y \mapsto ky$ (where k is the scaling factor) so that the fall-off rate for large y is similar to that of $P^F(\phi = y|\textit{all})$ in the South Florida dataset. This yields an estimate $P^{F|S}(\phi = y|\textit{on-edge})$ of the on edge statistics in South Florida, see figure (20). Similarly, we can estimate the edge distributions in Sowerby from those measured in South Florida. It can be shown [23] that similar results hold for other filters and, moreover, the performance is fairly insensitive to the value of k.

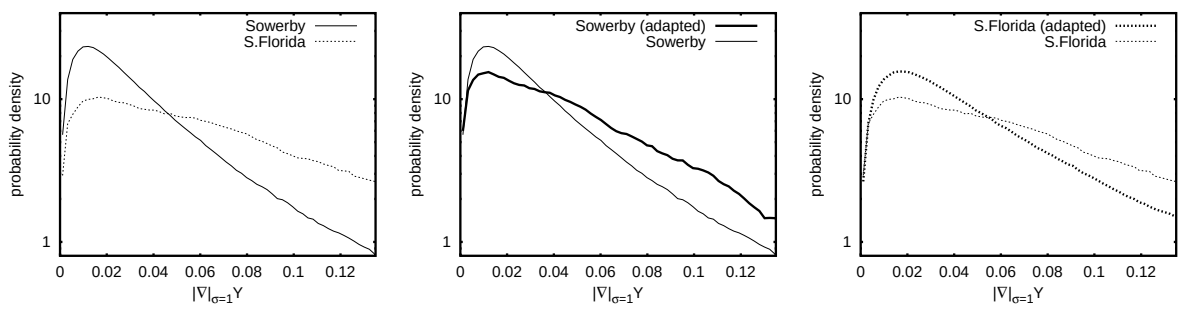

Figure 20: Adaption of $P(.|\textit{on-edge})$ from South Florida to Sowerby for $|\nabla|_{\sigma=1}Y$. The left panel shows (un-adapted) $P(.|\textit{on-edge})$ on Sowerby (dotted line) and South Florida (thin line). The centre panel shows $P(.|\textit{on-edge})$ for Sowerby (thin line) and the estimate of $P(.|\textit{on-edge})$ for Sowerby (bold line) by adapting from South Florida. The right panel shows $P(.|\textit{on-edge})$ for Soth Florida (thin dashed line) and the estimate of $P(.|\textit{on-edge})$ for South Florida (bold dashed line) by adapting from Sowerby. The adaptation is done by scaling the filter responses $y \mapsto ky$, using the method described in the previous figure.

We have tested this process by adapting the multiscale filter $|\vec{\nabla}|_{\sigma=1,2,4}(Y)$ from Sowerby to South Florida and vice versa. The figures show that the adaptation is very close despite the very different nature of datasets (and the different ground truths). On the Sowerby dataset, we get ROC area and Chernoff information of $(0.827, 0.223)$ for the true distributions (i.e. using distributions $P^S(\phi|\textit{on-edge}), P^S(\phi|\textit{off-edge})$) and $(0.825, 0.219)$ for the adapted distributions (i.e. using $P^{S|F}(\phi|\textit{on-edge}), P^{S|F}(\phi|\textit{off-edge})$). Similarly, we get ROC area and Chernoff information of $(0.877, 0.336)$ for the true South Florida distributions $(P^F(\phi|\textit{on-edge}), P^F(\phi|\textit{off-edge}))$ and $(0.867, 0.322)$ for the adapted distributions

$P^{F|S}(\phi|\textit{on-edge}), P^{F|S}(\phi|\textit{off-edge})$.

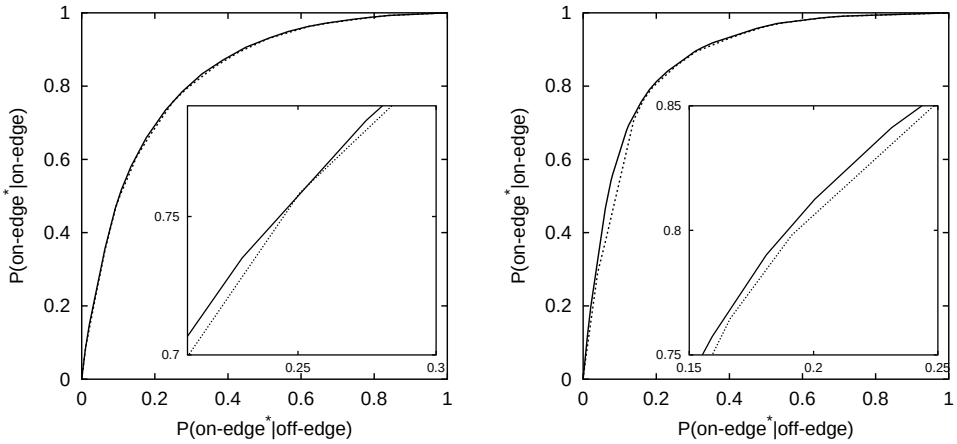

Figure 21: The effectiveness of adaptation shown by ROC curves on Sowerby (Left Panel) and South Florida (Right Panel). The bold and dashed lines show the ROC curves trained on the appropriate dataset and adapted (respectively). The similarity between the bold and dashed curves shows the success of the adaptation. The filter is $|\vec{\nabla}|_{\sigma=1,2,4}(Y)$.

6 Discussion and Conclusion

It has recently been argued [19], that perception should be formulated as Bayesian inference. This paper has taken this argument literally and applied it to the most basic vision task of edge detection. We learn the probability distributions of edge filter responses *on* and *off* edges from pre-segmented datasets, detect edges using the log-likelihood ratio test, and evaluate different edge cues using statistical measures (Chernoff information and ROC curves).

This approach enables us to study the effectiveness of different edge cues and how to combine cues optimally (from a statistical viewpoint). This allows us to quantify the advantages of multi-scale processing, and the use of chrominance information. We use two very different datasets, Sowerby and South Florida, and demonstrate a way to adapt the edge statistics from one dataset to the other.

We compare the results of statistical edge detection to those of standard edge detectors. On the South Florida dataset our results are comparable to those reported by Bowyer *et al* [7],[31],[8] for standard edge detectors. On the Sowerby dataset statistical edge detection

outperforms the Canny edge detector [9] significantly. We note that the Sowerby dataset is significantly harder to segment than the South Florida dataset (we assume that edge detectors should not respond to texture edges).

Our work was first published as a conference paper [20]. Subsequent work by Sidenblath applied this approach to motion tracking [32]. We have extended our studies of statistical cues for regional segmentation [21]. In addition, we have applied the approach to the task of edge localization and to quantify the amount of information lost when the image is decimated [22]

Acknowledgements

We want to acknowledge funding from NSF with award number IRI-9700446, from the Center for Imaging Sciences funded by ARO DAAH049510494, from the National Institute of Health (NEI) with grant number RO1-EY 12691-01, from the Smith-Kettlewell core grant, and the AFOSR grant F49620-98-1-0197 to ALY. We gratefully acknowledge the use of the Sowerby image dataset from Sowerby Research Centre, British Aerospace. We thank Andy Wright for bringing it to our attention. We also thank Prof. K. Bowyer for allowing us to use the South Florida dataset.

References

[1] J.J. Atick and A.N. Redlich, "What does the retina know about natural scenes?", *Neural Computation*, 4:196-210, 1992.

[2] R. Balboa. PhD Thesis. Department of Computer Science. University of Alicante. Spain. 1997.

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

[4] R. Balboa and N.M. Grzywacz. "The Distribution of Contrasts and its Relationship with Occlusions in Natural Images". *Vision Research.* In press. 2000.

[5] A. Blake and A.L. Yuille (Eds). **Active Vision**. MIT Press, Boston. 1992.

[6] K. W. Bowyer and J. Phillips, (editors), *Empirical evaluation techniques in computer vision*, IEEE Computer Society Press, 1998.

[7] K. Bowyer, C. Kranenburg, and S. Dougherty. "Edge Detector Evaluation Using Empirical ROC Curves". In *Proc. Computer Vision and Pattern Recognition*. CVPR'99, Fort Collins, Colorado. pp 354-359. 1999.

[8] K. W. Bowyer, C. Kranenburg, and S. Dougherty. "Edge detector evaluation using empirical ROC curves." *Computer Vision and Image Understanding*. **84**, 10, pp 77-103. 2001.

[9] J.F. Canny. "A Computational Approach to Edge Detection". *IEEE Transactions of Pattern Analysis and Machine Intelligence*. 8(6), pp 34-43. 1986.

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

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

[12] D.J. Field, "Relations between the statistics and natural images and the responses properties of cortical cells", *J. Optical Society of America*, A 42379-2394, 1987.

[13] S. Geman and D. Geman. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images". *IEE Trans. PAMI* **6**. pp 721-741. 1984.

[14] 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.

[15] D. M. Green and J. A. Swets. *Signal Detection Theory and Psychophysics*. 2nd Edition. Peninsula Publishing P.O. Box 867. Los Altos, California 94023. 1988.

[16] N.M. Grzywacz and R.M. Balboa. "A Bayesian Theory of Adaptation and its Application to the Retina". Submitted for publication. 2000.

[17] F. Heitger, L. Rosenthaler, R. von der Heydt, E. Peterhans and D. Kubler. "Simulation of neural contour mechanisms: From simple to end-stopped cells". *Vision Research*, vol 32, pages 963-981, 1992.

[18] M. Kass, A. Witkin, and D. Terzopoulos. "Snakes: Active Contour models". In *Proc. 1st Int. Conf. on Computer Vision*. 259-268. 1987.

[19] D.C. Knill and W. Richards. (Eds). **Perception as Bayesian Inference**. Cambridge University Press. 1996.

[20] S. M. Konishi, A.L. Yuille, J.M. Coughlan and S. C. 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.

[21] S. Konishi and A.L. Yuille. "Bayesian Segmentation of Scenes using Domain Specific Knowledge". In Proceedings *Computer Vision and Pattern Recognition*. CVPR'00. 2000.

[22] S. Konishi, A.L. Yuille, and J.M. Coughlan. "A Statistical Approach to Multi-Scale Edge Detection". In *Workshop on Generative-Model-Based Vision: GMBV 2002*. Organizer: A. Pece. 2002.

[23] S. Konishi. **PhD Thesis.** Department of Biophysics. University of California at Berkeley. 2002.

[24] A.B. Lee, J.G.Huang, and D.B. Mumford, "Random collage model for natural images", *Int'l J. of Computer Vision*, oct. 2000.

[25] D. Marr. **Vision**. W.H. Freeman and Company. San Francisco. 1982.

[26] M. Nitzberg, D. Mumford, and T. Shiota, **Filtering, Segmentation and Depth,** Springer-Verlag, 1993.

[27] J. Peng and B. Bhanu. "Closed-Loop Object Recognition Using Reinforcement Learning". *Pattern Analysis and Machine Intelligence*. Vol. 20, No. 2. pp 139-154. February. 1998.

[28] P. Perona and J. Malik. "Detecting and Localizing Edges Composed of Steps, Peaks, and Roofs". In *Proc. 3rd Int. Conf. on Computer Vision*, pp 52-57. 1990.

[29] B.D. Ripley. **Pattern Recognition and Neural Networks.** Cambridge University Press. 1996.

[30] D.L. Ruderman and W. Bialek, "Statistics of natural images: scaling in the woods", *Phy. Rev. Letter*, 73:814-817, 1994.

[31] M.C. Shin, D. Goldof, and K.W. Bowyer. "Comparison of Edge Detectors Using an Object Recognition Task". In *Proc. Computer Vision and Pattern Recognition*. CVPR'99, Fort Collins, Colorado. pp 360-365. 1999.

[32] H. Sidenblath. **Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences**. PhD Thesis. Royal Institute of Technology. Stockholm. 2001.

[33] J. Sullivan, A. Blake, M. Isard and J. MacCormick. "Object Localization by Bayesian Correlation". In *Proc. Int. Conf. Comp. Vision*. Corfu, Greece. pp 1068-1075. 1999.

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

[35] M.J. Wainwright and E.P. Simoncelli, "Scale mixtures of Gaussian and the statistics of natural images", *NIPS*, 855-861, 2000.

[36] A. L. Yuille and J. 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.

[37] A.L. Yuille, James M. Coughlan, and S.C. Zhu. "A Unified Framework for Performance Analysis of Bayesian Inference.". In *Proceedings SPIE*. Orlando, Florida. April. 2000.

[38] S. C. Zhu, T.S. Lee and A.L. Yuille. "Region Competition: Unifying Snakes, Region Growing, and Bayes/MDL for Multi-band Image Segmentation.". In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 9, Sept. 1996.

[39] S.C. Zhu and D.B. Mumford, "Prior learning and Gibbs reaction-diffusion", *IEEE Trans. PAMI*, vol.19, no.11, Nov. 1997.

[40] 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.