Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues.

S. M. Konishi
James M. Coughlan
Smith-Kettlewell Eye Research Institute
San Francisco, CA 94115

A.L. Yuille
Song Chun Zhu**
** Computer Science and Info. Dept.
Ohio State University. Ohio.

Abstract

We treat the problem of edge detection as one of statistical inference. Local edge cues, implemented by filters, provide information about the likely positions of edges which can be used as input to higher-level models. Different edge cues can be evaluated by the statistical effectiveness of their corresponding filters evaluated on a dataset of 100 pre-segmented images. We use information theoretic measures to determine the effectiveness of a variety of different edge detectors working at multiple scales on black and white and colour images. Our results give quantative measures for the advantages of multi-level processing, for the use of chromaticity in addition to greyscale, and for the relative effectiveness of different detectors.

Submitted to CVPR'99

1 Introduction

Edge detectors are intended to detect and localize the boundaries of objects. 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) 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.

A standard approach in computer vision runs as follows: combine local edge cues by generic (i.e. general purpose) grouping laws based on the assumption that edges of objects tend to be spatially smooth locally. Such grouping rules allow the local ambiguous edge cues to be disambiguated (to some extent). In some cases, however, grouping techniques may not be enough and object specific knowledge may be required to detect the object boundary.

This approach is plausible but, so far, there has not been a rigourous theory for it. Such a theory would, for example, attempt to answer: (i) precisely how much information is available in the local edge cues?, (ii) how independent is the information between different cues (e.g. intensity and colour)?, (iii) how do the grouping rules interact with the edge cues?, (iv) what types of grouping rules are most effective?, and (v) in what situations do these grouping rules need to be supplemented by object specific knowledge?

Recent work by Yuille and Coughlan [9] gives a theoretical basis for addressing some of these questions. In particular, they studied the problem of road tracking [5] and determined *order parameters* which characterized the difficulty of the problem. These order parameters combine measures of the effectiveness of: (i) local cues (implemented by edge detector operators), and (ii) geometrical grouping rules (e.g. the assumption that the road is spatially smooth). The order parameters have two purposes: (i) they determine *fundamental limits* (i.e. algorithm independent) on whether the road can be detected at all , (ii) they determine the expected convergence rates, and expected errors, of A* algorithms designed to detect the road.

In this paper, we build on the work of Yuille and Coughlan [9] to determine good cues for edge detection. In this approach, the effectiveness of edge cues is determined by how much information they provide. This can only be evaluated by the statistics of their effectiveness evaluated over a calibrated dataset. We therefore have: (i) a learning phase where we learn the statistics of edge detector responses both on and off edges, and (ii) an evaluation phase where we evaluate the effectiveness of different edge detectors (i.e. edge cues).

Both aspects of our work have interesting similarities and difference to the Minimax Entropy learning theory recently developed by Zhu, Wu, and Mumford [10] When this theory is applied to learning probability distributions on images [11] it measures the statistics of filter responses over an ensemble of images and uses the maximum entropy principle to determine probability distributions on the underlying images. The choice of filters to use is determined by a *minimum entropy* criteria which selects those which make the resulting *maximum entropy* distributions as close to the true distribution as possible. By contrast, in this paper we aim at *discrimination* rather than *representation* which means that the minimum entropy selection procedure is not appropriate. Instead we will use measures of discriminability such as the Chernoff and Bhattarcharrya bounds [4],[8].

Figure 1: Four typical images from the Sowerby dataset. These images contain a variety of urban and rural scenes.

Our approach presupposes that the statistics of edge cues are relatively constant within certain image domains. Of course, some individual images may be completely unrepresentative (the statistics of an outdoor scene with the camera pointed vertically at the sky will be clearly different from one where the camera points horizontally). On balance, however, we might expect that the effectiveness of edges cues varies comparatively little from image to image. As well as the results demonstrated in this paper, we should emphasize the pioneering work of Balboa and Grzywacz [1] who studied the on-edge and off-edge statistics of natural outdoor scenes and underwater images, demonstrated qualitative similarity between the statistics in these two different domains, and related them to the receptive field properties of the retinas of animals.

The work described in this paper, therefore, attempts a statistical study of edge detection cues which helps evaluate the information provided by the cues. To perform this analysis, we make use of the Sowerby image database which contains several hundred interactively segmented natural images. As we will show, we obtain robust quantative measures of the effectiveness of different edge detection cues including colour and multi-scale approaches.

In section (2) we introduce our statistical evaluation criteria. Section (3) describes, and justifies, the edge cue filters we use. In section (4) we discuss the stability of the statistics and adaptive techniques we employed. Section (5) gives the results on the dataset.

2 Statistical Basics

This section provides the statistical basis of our approach. It describes how, for any edge detector operator, we can obtain empirical probability distributions for its response to edges and to non-edges. From these *learned* distributions we determine whether a *set of samples* is more likely to be on-edge or off-edge (i.e. we assume that the samples in the set are either all on an edge or all off an edge). This can be evaluated using the Chernoff and Bhattacharyya bounds, which determine asymptotic error rates, see [4]. The order parameters derived in [9] were sums of local edge information with global geometric information. The local edge information was given by the Chernoff and Bhattacharyya bounds, depending on the precise formulation of the problem. We can therefore consider the Chernoff and Bhattacharyya bounds as *order parameters* for the edge cues.

We have also considered the related task of determining whether an *individual sample* is more likely to be on or off an edge. This can be evaluated by the Bayes risk (a weighted average of the expected number of false positives and false negatives). The Bayes risk is closely linked to the Chernoff bound and it can be shown, see [8], that the Bayes risk can be given an upper and lower bound in terms of functions of the Chernoff and Bhattacharyya bounds. In this paper, however, we concentrate on the first task (for reasons of space).

2.1 Determining empirical probability distributions.

Any edge cue (or combination of cues) is represented by *an operator* $\phi(.)$ which can be evaluated at each position in the image. The operator $\phi(.)$ will typically be a linear, or non-linear filter, and can have a scalar or vector valued output. For example, one possibility is the scalar valued filter $\left|\vec{\nabla}(.)\right|$ for which $\phi(I(x)) = \left|\vec{\nabla}I(x)\right|$. Another possibility is to combine edge filters at different spatial scales to give a vector valued output $\phi(I(x)) = (\left|\vec{\nabla}G(x;\sigma_1) * I(x)\right|, \left|\vec{\nabla}G(x;\sigma_2) * I(x)\right|)$, where $G(x;\sigma)$ is a Gaussian with standard deviation σ and $*$ denotes convolution. Yet another possibility is to apply filters to the different colour bands of the image. We will develop the basic theory at an abstract level so that it can apply directly to all these cases.

Having chosen an edge operator $\phi(.)$ we have to quantize its response values. This involves selecting a finite set of possible responses $\{y_i : i = 1, ..., J\}$. 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, see section (4). The operator is run over the image and its empirical statistics (histograms) are evaluated for the operator's responses on and off edge. These histograms are then normalized to give two *marginal distributions* $P_{on}(y), P_{off}(y)$ specified by:

$$P(y_x) = P_{on}(y_x), \;\; if \text{ “}x\text{''} \; lies \; on \; an \; edge$$
$$P(y_x) = P_{off}(y_x), \;\; if \text{ “}x\text{''} \; lies \; off \; an \; edge \; path. \quad (1)$$

For example, for the filter $\phi(.) = \left|\vec{\nabla}\right|$ we would anticipate that the probability distribution for P_{off} is strongly peaked at $y = 0$ (i.e. the image gradient tends to be small away from edges) while the peak of P_{on} occurs at larger values of y (i.e. the image gradient is likely to be non-zero at edges), see figure (2).

2.2 Asymptotic Error Rates

Our task is 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 – such as a road [5].

Suppose we have a sequence of samples $\vec{y} = y_1, y_2, ..., y_N$ of the responses of the road detector. The optimal tests for determining whether the samples come from P_{on} or P_{off} will depend on the *log-likelihood ratio*[1] (see the Neyman-Pearson lemma [4]):

$$\log\{\frac{P_{on}(y_1, ...y_N)}{P_{off}(y_1, ...y_N)}\} = \sum_{i=1}^{N} \log\{\frac{P_{on}(y_i)}{P_{off}(y_i)}\}. \quad (2)$$

The larger the log-likelihood ratio then the more probable that the measurement sample $\vec{y} = (y_1, y_2, ..., y_N)$ 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 [4] that, for sufficiently large N, the expected error rate of this test decreases exponentially by $2^{-NC(P_{on}, P_{off})}$ where $C(P_{on}, P_{off})$ is the *Chernoff Information* [4] between P_{on} and P_{off} defined by:

$$C(P_{on}, P_{off}) = - \min_{0 \leq \lambda \leq 1} \log\{\sum_y P^{\lambda}_{on}(y) P^{1-\lambda}_{off}(y)\}. \quad (3)$$

[1]This can be thought of as the maximum likelihood test between two hypotheses which are equally likely a priori.

A closely related quantity is the Bhattacharyya bound:

$$B(P_{on}, P_{off}) = -\log\{\sum_y P^{1/2}_{on}(y) P^{1/2}_{off}(y)\}. \quad (4)$$

From these definitions it is straightforward to see that $C(P_{on}, P_{off}) \geq B(P_{on}, P_{off})$ for any P_{on}, P_{off} (because Chernoff selects λ to minimize $\log\{\sum_y P^{\lambda}_{on}(y) P^{1-\lambda}_{off}(y)\}$ with respect to λ while the Bhattacharyya bound just sets $\lambda = 1/2$). Our results, see section (5) show that Chernoff and Bhattacharyya give very similar values in our application domain.

The Chernoff Information and the Bhattacharyya bound are directly related to the order parameters determined by Yuille and Coughlan [9]. For the road tracking task, the order parameter K is given by $K = B(P_{on}, P_{off}) + B(P_{\Delta G}||U) - H(U)$, where $B(P_{on}, P_{off})$ is the Bhattacharyya bound providing a measure of effectiveness of the local edge detectors, and $B(P_{\Delta G}||U) - H(U)$ is a measure of the prior knowledge about the likely shapes of the road (as determined by $P_{\Delta G}$) compared to the number of possible distractor shapes (as measured by the uniform distribution U and its entropy $H(U)$).

The bigger the value of K the easier the road tracking task becomes (the task is effectively impossible if $K < 0$). Intuitively, the local edge detector, as measured by $B(P_{on}, P_{off})$, must be effective enough to overcome the number of possible distracting false paths.

It can be shown [4] that when two independent edge cues are combined then their Chernoff/Bhattacharyya bounds will simply add. In practice, few edge cues are completely independent and so the Chernoff/Bhattacharyya of two coupled cues is usually a lot less, see section (5), than the sum of their individual Chernoff/Bhattacharyya's.

3 Edge Detection Filters

Many edge detectors have been proposed in the literature and it is clearly impossible to test all of them. Most of them assume that the intensity profile across an edge can be represented by a step edge profile possibly with additive Gaussian noise – see [2] for a thorough and insightful analysis of this type of model. It is clear, however, that other types of edge profile exist. Psychophysicists show that human observers are indeed sensitive to such edges and can use shadow edges as cues for whether an object is touching the ground or is floating above it.

Therefore we first decided to evaluate the step edge model. To do so, we calculated edge profiles from the Sowerby dataset. We aligned them and performed principal component analysis. The results (see technical report) strongly supported the step edge model because the mean profile and the dominant eigenvectors corresponded to step edges. For a second test, we used Fisher's linear discriminant to determine the best linear projection for distinguishing between edge profiles and non-edge profiles. Our results showed that the best linear filter was the (discretized) first derivative operator indicating that the dominant feature of edges was their intensity discontinuities (see technical report for more details). From these "reality checks", and visual inspection of the images, we concluded that the vast majority of edges in the dataset were, at least roughly, of the step edge type. These results *do not say* that all edges are step-edges. Nor do they imply that other types of edge profile are unimportant.

For the rest of this paper, we will therefore restrict our attention to edge detectors which have been designed for step edges. Detectors for other edge types are left for a later study.

3.1 The Edge Detectors

Based on the results from the last section, we concentrate on three basic edge detectors. These are the intensity gradient, the Nitzberg edge detector [7], and the Laplacian of a Gaussian. These edge detectors are examined in both the intensity and colour regimes. We also examine their behaviours at a variety of different scales. It is straightforward to couple different edge detectors to obtain a vector valued output and to determine the additional information conveyed by combinations of edge detectors.

The modulus of the gradient and the Laplacian of a Gaussian operators are specified by the equations $\left|\vec{\nabla} G(x;\sigma) * I(x)\right|$ and $\nabla^2 G(x;\sigma) * I(x)$ where $*$ denotes convolution and $G(x;\sigma)$ is a Gaussian at different spatial scales parameterized by the standard deviation σ. The Nitzberg edge detector is non-linear and exists in several variants. It involves computing the matrix $\mathbf{N}(x;\sigma) = G(x;\sigma) * \{\vec{\nabla} I(x)\}\{\vec{\nabla} I(x)\}^T$ where ^{T} denotes transpose. We consider three possible edge measures: (i) the biggest eigenvalue $\lambda_1(x;\sigma)$ of $\mathbf{N}(x;\sigma)$, (ii) the trace of $\mathbf{N}(x;\sigma)$ – i.e $\lambda_1(x;\sigma) + \lambda_2(x;\sigma)$, and (iii) the two-dimensional vector consisting of both eigenvalues $(\lambda_1(x;\sigma), \lambda_2(x;\sigma))$. (Of course, the two eigenvalues of $\mathbf{N}(x;\sigma)$ are also used to detect corners in images.)

The use of colour information for edge detection has been debated (although its effectiveness for segmentation based on regional properties is not in doubt). We use a variant of the NTSC colour space where $Y = 0.299R + 0.587G + 0.114B, I = (0.596R - 0.274G - 0.322B)/Y, Q = (0.211R - 0.523G - 0.312B)/Y$. Here Y is interpreted to be the grey-scale image and I, Q are the *chromaticity vectors* representing *hue* and *saturation*. Observe, that *unlike NTSC*, we have normalized the chromaticity by the greyscale. We will examine the relative effectiveness of chromaticity and grey-scale either by themselves or when coupled.

The biology of human vision, combined with more pragmatic motives, strongly suggests that images should be processed at different scales. 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 the approach followed in this paper, *the optimal combination arises naturally* subject to the quantization procedure we use. Multiscale is performed by varying the parameters σ of the Gaussian convolutions. For linear filters, such as $\vec{\nabla}$ and ∇^2, this Gaussian convolution commutes with the derivative operation[2]. This is not true for the Niztberg detector because of its non-linearity.

4 Stability

An important practical issue of our approach is to develop an appropriate quantization for the distributions. There is a trade-off involved. If the number of quantization bins is too small then the results we obtain will be crude. It will also be important when comparing the performance of different edge operators (because crude binning may affect some operators more than others). By contrast, if we have too many quantization bins then the resulting Chernoff measures may be too sensitive to the "noise" in the data. In addition, the Chernoff Information can become unstable. More practically, the greater the number of bins the larger the amount of computations and memory requirements. At a more abstract level, we are faced with the danger of overfitting the data, which is a common problem inherent to all learning procedures [8].

After considerable experimentation and theoretical analysis (see technical report) we settled on an adaptive quantization scheme. It became clear that most of the reliable information could be extracted using only 6 adaptive bins for each dimension of the filter, see figure (2) (we emphasize that this adaption was performed over the entire dataset and *not* for each individual image). This enabled us to perform statistics on up to 6 coupled edge cues – which requires 6^6 quantized bins.

Theoretical analysis of the Chernoff information (see technical report) showed that it became increasingly unstable if the number of bins was too high

[2]Though, in practice, there can be small difference due to discretization.

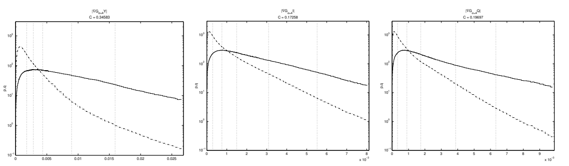

Figure 2: The marginals over all the dataset. The vertical axis gives the probabilities and the horizontal axis is the filter response. The adaptive bins are shown by the vertical dotted lines. The P_{on} and P_{off} are represented by the heavy and the dashed line respectively.

(without a corresponding increase in the number of samples). In particular, it also showed some sensitivity to cases where several bins were empty. However, we developed robust techniques to avoid these dangers (see technical report).

When computing the Chernoff for an ensemble of images, we also detected an instability which we called the "rotten apple". It became clear that a single outlier image could corrupt the Chernoff information of the entire dataset. This could be seen because the Chernoff information for the ensemble became significantly worse than the average of the Chernoff information for each image. This led us to reject 5 images from our set of 103 because of the poorness of their edge maps (see technical report).

5 Results

We now show the results on a range of filters $\left|\vec{\nabla}\right|, \nabla^2, \mathbf{N}$ applied at a range of scales, to colour/greyscale, and with various coupling of the cues (Gaussian filters are used to get multiscale). For reasons of space, we restrict ourselves to showing the following filters: (1) $\nabla^2 Y(\sigma = 1)$, (2) $\nabla^2 Y(\sigma = 1,2,4)$, (3) $\left|\vec{\nabla} I\right|, \left|\vec{\nabla} Q\right|, (\sigma = 1)$, (4) $\left|\vec{\nabla} Y\right| (\sigma = 1)$, (5) $\lambda_1(Y), (\sigma = 1)$ (6) $\lambda_1(Y), \lambda_2(Y)(\sigma = 1)$ (7) $\left|\vec{\nabla} Y\right|, \left|\vec{\nabla} I\right|, \left|\vec{\nabla} Q\right|, (\sigma = 1)$, (8) $\left|\vec{\nabla} I\right|, \left|\vec{\nabla} Q\right| (\sigma = 1,2,4)$, (9) $\left|\vec{\nabla} Y\right| (\sigma = 1,2,4)$, (10) $\lambda_1(Y)(\sigma = 1,2,4)$, (11)$\lambda_1(Y), \lambda_2(Y)(\sigma = 1,2,4)$, (12) $\left|\vec{\nabla} Y\right|, \left|\vec{\nabla} I\right|, \left|\vec{\nabla} Q\right|, (\sigma = 1,2)$.

To calibrate the Chernoff measures, we calculated them for the Geman and Jedynak road tracking application [5] (from the plots in their paper). This gave a Chernoff of 0.22. This sets a baseline and, as we will show, we can obtain Chernoff's quite significantly higher by combining cues. (Though we stress that Geman and Jedynak were tackling a different problem

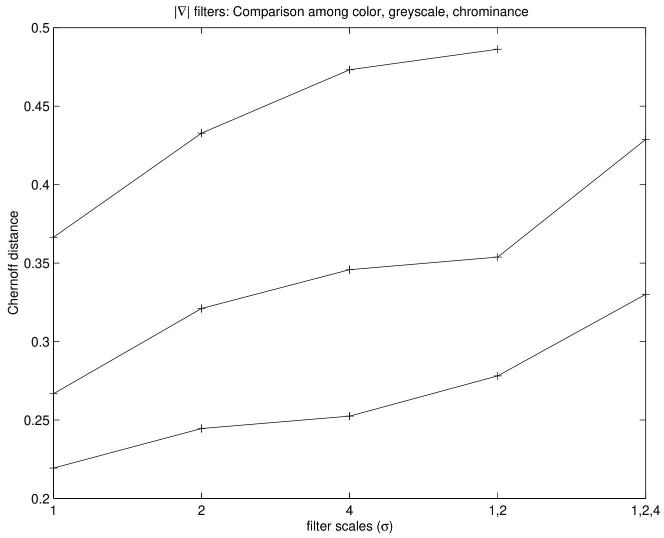

Figure 3: Multiscale results of the gradient filter. On colour (top), greyscale (centre), and chrominance (bottom). The vertical axis gives the Chernoff distance. The horizontal axis labels the filter scales, and joint scales (e.g. $1,2$). The graph demonstrates the advantages of using multiscale and full colour. Chromaticity is weaker than greyscale which, in turn, is significantly weaker than full colour.

and so our results are not directly comparable.)

The principal results are the effectiveness of colour and multiscale, see figure (3) for the gradient (we get similar trends for the Laplacian and the Nitzberg detectors). The effectiveness of the Nitzberg filter over the gradient and both (substantially) over the Laplacian of a Gaussian, see figure (3). A summary plot to illustrate the main effects is given in figure (5) (using the filter numbering given above). We give a relief map, see figure (6), showing that although the Chernoff's vary from image to image the relative effectiveness of the filters is approximately the same. Figure (7) also examines the variations between images by plotting the Chernoff's of the P_{on}, P_{off} of different images compared to the $P_{on}^{avg}, P_{off}^{avg}$ obtained for the entire dataset.

6 Conclusion

Our results give a quantitative measure of the effectiveness of different edge cues taking into account multi-scale and colour. The results are encouraging in that, by careful combination of cues, we are able to get very discriminability, as measured by Chernoff, and demonstrate that certain edge detectors are significantly better than others. Our measures of effec-

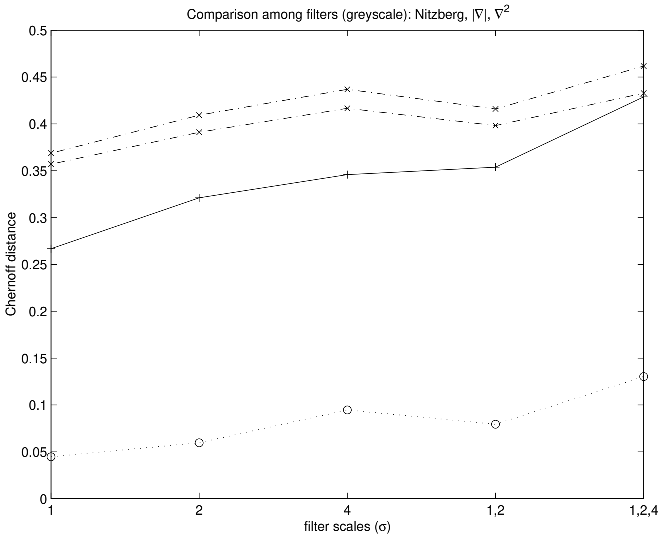

Figure 4: Multiscale results on greyscale images with similar conventions to the previous figure. The two dashed lines (top) are the Nitzberg filter with joint (λ_1, λ_2) and λ_1 respectively. The dark line (center) is the gradient filter. The dotted line (bottom) is the Laplacian of a Gaussian. Observe that Nitzberg is the most effective but it does not improve as rapidly as the gradient for multi-scale. The Laplacian of a Gaussian comes in a very poor third.

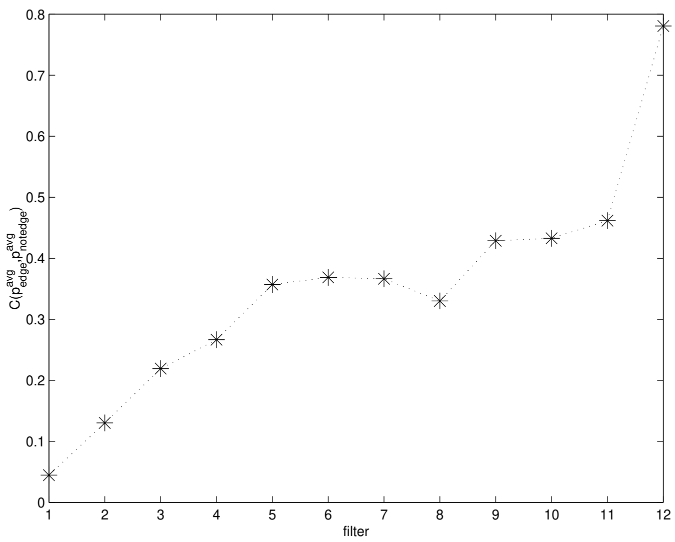

Figure 5: Summary graph of the performance of different filters over the entire dataset. Observe the poor performance of the Laplacian of a Gaussian (1,2) and chromaticity (3). The gradient is good (4) and all variants of Nitzberg (5,6). Colour and multiscale give big improvements (7,8,9,10,11,12).

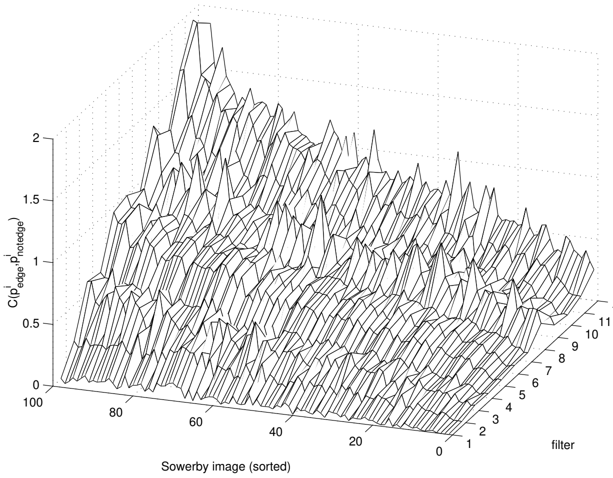

Figure 6: A relief figure demonstrating the relative effectiveness of different filters. Observe that although some images (on left of figure) have easier edges than others, the relative effectiveness of the detectors is roughly constant from image to image. The differences in performance between images is closely related to the amount of vegetation (i.e. background texture) in each image. Observe than even the most difficult images, far right, have Chernoff's approaching 0.5 for the best cue combinations.

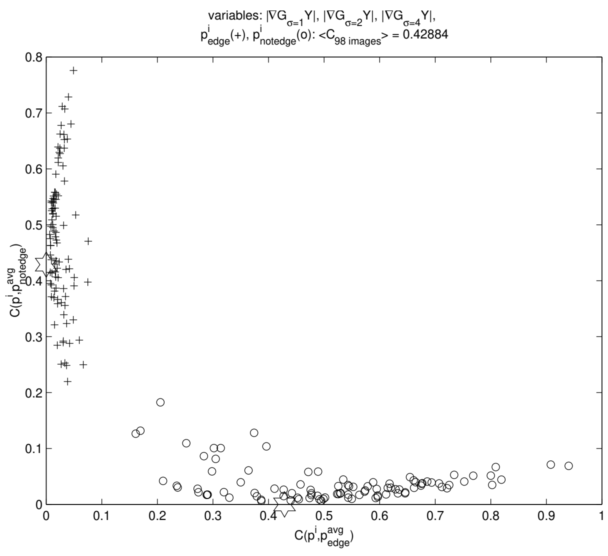

Figure 7: Chernoff figure. We plot $C(P, P_{on}^{ave}), C(P, P_{off}^{ave})$ for the P_{off} (circles) and the P_{on} (pluses). This demonstrates that although there is variation in the P_{on}, P_{off}'s between images they nevertheless have standardized behaviour. Positions of the P_{on}, P_{off}'s in this plot are determined by the relative amounts of sky, road, buildings, and vegetation in the images.

Figure 8: The typical image and its $\log P_{on}/P_{off}$ obtained using the gradient operator at three combined scales on the greyscale image.

tiveness are in a form directly suitable for combination with geometric grouping cues [9].

Future work involves investigating these, and other, edge cues on a range of domains. We stress that the results in this paper are in agreement with earlier studies on the gradient and Nitzberg filters (single scale and no colour) reported in [3] on a small set of five images (taken under very different circumstances). Combined with the results of [1], this suggests that our results will apply directly to other databases (though some variations are to be expected).

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 noname core grant, and the AFOSR grant F49620-98-1-0197 to anonymous. We gratefully acknowledge the use of the Sowerby image dataset and thank Andy Wright for bringing it to our attention.

References

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

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

[3] 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.

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

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

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

[9] A.L. Yuille and J.M. Coughlan. "Convergence Rates of Algorithms for Visual Search: Detecting Visual Contours". In *Proceedings NIPS'98.* 1998.

[10] 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.

[11] S.C. Zhu and D. Mumford. "Prior Learning and Gibbs Reaction-Diffusion". IEEE Trans. on PAMI vol. 19, no. 11. Nov. 1997.