Statistical cues for Domain Specific Image Segmentation with Performance Analysis. Proc. CVPR'2000.

Scott Konishi A.L. Yuille

Smith-Kettlewell Eye Research Institute
2318 Fillmore Street
San Francisco, CA 94115

Abstract

This paper investigates the use of colour and texture cues for segmentation of images within two specified domains. The first is the Sowerby dataset, which contains one hundred colour photographs of country roads in England that have been interactively segmented and classified into six classes – edge, vegetation, air, road, building, and other. The second domain is a set of thirty five images, taken in San Francisco, which have been interactively segmented into similar classes. In each domain we learn the joint probability distributions of filter responses, based on colour and texture, for each class. These distributions are then used for classification. We restrict ourselves to a limited number of filters in order to ensure that the learnt filter responses do not overfit the training data (our region classes are chosen so as to ensure that there is enough data to avoid overfitting). We do performance analysis on the two datasets by evaluating the false positive and false negative error rates for the classification. This shows that the learnt models achieve high accuracy in classifying individual pixels into those classes for which the filter responses are approximately spatially homogeneous (i.e. road, vegetation, and air but not edge and building). A more sensitive performance measure, the Chernoff information, is calculated in order to quantify how well the cues for edge and building are doing. This demonstrates that statistical knowledge of the domain is a powerful tool for segmentation.

Proceedings Computer Vision and Pattern Recognition CVPR'2000. Hilton Head, South Carolina. 2000.

1 Introduction

Although there has been recent progress in general purpose image segmentation, for example [2], [10], [12], it remains an extremely difficult problem. In this paper we examine the effectiveness of segmentation using domain specific cues which are learnt from image databases. The basic idea is that most images can be grouped into domains. Within these domains the statistical properties of images are likely to be very similar, and such knowledge has been exploited for the segmentation of aerial images. Will the same approach work for images of city and country scenes photographed from ground level? How effective are such cues for the segmentation of such images?

The goal of this paper is to learn simple filter cues for segmentation, based on texture and colour, within two image domains (one containing 100 images, the second containing 35). We then evaluate the performance of the cues for these databases. This part of our work is in the spirit of performance analysis [1].

Our first image domain is the Sowerby image database which consists of one hundred presegmented images of road scenes in the English countryside. The second image domain are street scenes in San Francisco. In both domains, we applied a set of filters which were sensitive to colour, texture, and edges. We looked at the empirical joint probability distributions of these filter re-

sponses at multiple scales. Then we constructed a probabilistic model for the domain using these empirical distributions and prior knowledge about the typical number of each class per image. We can then apply Bayesian classification for each set of filters and evaluate their performance on these datasets.

We are careful to avoid the dangers of overfitting when estimating the joint probability distributions. Due to the so-called "curse of dimensionality" the amount of data required to learn joint distributions increases exponentially with the dimension of the distribution. This means that we can only use a limited number of filters and also our classes must be restricted to those for which there is sufficient training data.

Our experiments show that, with suitable choice of filters, the Bayesian classification scheme is successful. We measure this in two ways: (i) the false positive and false negative classification rates into the six classes (evaluated over the entire dataset), and (ii) the *Chernoff information* between the probability distributions for each class, which gives a measure of the asymptotic error rate of classification. For classes for which the filters statistics are approximately homogeneous spatially – e.g. road, air, and vegetation – the classification rates are very accurate using either texture or colour, or their combination. The Chernoff information is a more sensitive measure which is used to evaluate how well the filters do on the more difficult classes (i.e. edge and building).

We stress that our classifications are based on local filter properties only and hence is *fast*. We *do not use any knowledge about the likely shape of region boundaries or even that neighbouring pixels are likely to belong to the same image class*. Such knowledge, even in the simplest forms of regional grouping by a boundary smoothing constraint such as snakes – see [12], would definitely improve the quality of the segmentation. The goal of this paper, however, is to demonstrate how much information is available in local filter cues only.

2 Background

There has, of course, been extensive work on image segmentation using colour, texture, and other cues – see [2], [10], [12] and references therein. Much of this work is orthogonal to the goals of our paper as it does not attempt to learn segmentation cues within a domain. Our work can complement approaches of this type by providing prior models for the image properties of regions.

There has been previous work on using colour cues to detect structures such as roads. A successful example was demonstrated by Crisman for road tracking [4]. Her work, however, continually estimated colour models for roads interactively and did not attempt to do statistics of road, or non-road, properties over a large dataset. Other work on the use of colour cues for recognizing specific objects includes Swain and Ballard [11]. In addition, there have been successful models of texture obtained using the Minimax Entropy learning theory [13]. These works, however, have not explored the use of domain specific statistical knowledge for segmentation. A recent learning method [8] is very different from our approach and make use of reinforcement learning with high-level feedback.

Our recent work studied the effectiveness of different edge cues for segmenting the Sowerby dataset [6]. This study measured the Chernoff information provided by specific edge cues and demonstrated that the effectiveness of these cues was approximately constant over the entire database. The work in this paper is based on a similar methodology but is more general. Instead of two classes – edge versus non-edge – there are six. Moreover, in addition to Chernoff information we also evaluate the false positive and false negative rates of classification.

3 Statistical Basics

This section provides the statistical basis of our approach. It describes how, for any set of filters, we can obtain an empirical joint probability distribution for their responses to the six classes. From these *learned* distributions we apply Bayesian probability theory to determine which class a filter response is likely to be a member of. This is evaluated in terms of the false positives and false negatives for each class (expressed in terms of a confusion matrix).

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

In addition, we use a more sensitive measure to address the related question of which class a *set of samples* is most likely to be in. This is evaluated using the Chernoff information, which determines the asymptotic error rates, see [3].

3.1 Determining empirical probability distributions.

Any class cue (or combination of cues) is represented by *a filter* $\phi(.)$, which can be evaluated at each position in the image. $\phi(.)$ can be a linear, or non-linear filter, and can have a scalar or vector valued output. For example, one choice 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 choice 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 a filter $\phi(.)$ 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 filter will depend on this quantization scheme so care must be taken to determine that the quantization is robust and close to optimal, see section (5). The filter is run over the image and its empirical statistics (histograms) are evaluated for the operator's responses to the six different classes. These histograms are then normalized to give six *conditional distributions* $P(y_j|\alpha)$, where α denotes the six classes $\{edge, vegetation, air, road, building, other\}$.

For example, for the filter $\phi(.) = \left|\vec{\nabla}\right|$ we would anticipate that the probability distribution for $P(y_j|\alpha = air)$ is strongly peaked near $y_i = 0$ (i.e. the sky tends to have small image gradients), while the peak of $P(y_j|\alpha = edge)$ occurs at larger values of y (i.e. the image gradient is likely to be large at edges of objects).

It is important to ensure that we have enough training data so that we do not overlearn the data, see [9]. This restricts us to using a limited number of filters (because the amount of data required grows exponentially with the number of filters used in our joint histograms). We *stress that use standard procedures to ensure this such as learning the distributions on half the dataset and evaluating them on the other half.*

We must also determine prior probability distributions $\{P(\alpha)\}$ for the six classes. These are estimated by the empirical number of image pixels in each class computed over the entire dataset. From these two types of distributions (conditional and priors) we construct the Bayesian

decision rule: label a pixel x as lying in class $\alpha^*(x)$, where $\alpha^*(x) = \arg_{\alpha=} \max P(y_j(x)|\alpha)P(\alpha)$. Classifications, confusion matrices, and false positive and false negative error rates for this rule will be given in section (6) for a set of different filters. In addition, we will consider the classification when we set the prior class probabilities to be uniform, and hence classify the pixels by the maximum likelihood estimator $\alpha^*(x) = \arg_{\alpha} \max P(y_j(x)|\alpha)$ (this may be a useful strategy when some classes are very rare and hence the "data driven prior" biases strongly against them).

3.2 Asymptotic Error Rates

We may also want to determine whether a *set of samples* is more likely to be in one class or another (i.e. all members of this set are assumed to be in a single class). This issue is important if we intend to group a set of pixels using spatial information. It is a more sensitive measure than the false positive and false negative classification rates. It is useful for those region classes for which the false positive and false negative rates are poor.

The optimal test for determining whether a set of samples $\vec{y} = y_1, y_2, ..., y_N$ comes from classes α or β is given by the log-likelihood test (see the maximum likelihood-Pearson lemma [3]). It can be shown [3] that, for sufficiently large N, the expected error rate of this test decreases exponentially by $e^{-NC(P(y|\alpha),P(y|\beta))}$ where $C(P(y|\alpha), P(y|\beta))$ is the *Chernoff Information* [3] between class α and β defined by $C(P(y|\alpha), P(y|\beta)) = -\min_{0\leq\lambda\leq 1} \log\{\sum_y P(y|\alpha)^\lambda P(y|\beta)^{1-\lambda}\}$.

Thus to determine the asymptotic error rates, as well as the individual pixel error rates, we compute the Chernoff information between different classes (as functions of the choice of filters).

4 The Filters

We concentrated on combinations of four basic filters. These are the intensity itself (for colour segmentation), the gradient, the Nitzberg edge detector [7], and the Laplacian of a Gaussian. These filters are examined in both the intensity and colour regimes and at a variety of different scales. Multiscale is performed by varying the parameters σ of the Gaussian convolutions and combining single scale responses into a vector filter. (In the approach followed in this paper, *the optimal combination arises naturally*, subject to the quantization procedure we use.) It is straightforward to couple different filters to obtain a vector valued filter and to determine the additional information conveyed by combinations of elementary filters.

Our results showed that the most effective filters were the intensity (i.e. colour) and the Nitzberg operator [7]. Colour is, not surprisingly, a very effective cue for distinguishing between different regions. The Nitzberg operator was originally designed as a corner detector and it turns out to be an effective operator for distinguishing between regions of different textures. More precisely, the Nitzberg operator involves computing the matrix $\mathbf{N}(x;\sigma) = G(x;\sigma) * \{\vec{\nabla} I(x)\}\{\vec{\nabla} I(x)\}^T$ where ^{T} denotes transpose. The output is the two-dimensional vector consisting of both eigenvalues $(\lambda_1(x;\sigma), \lambda_2(x;\sigma))$.

The gradient and Laplacian of Gaussian filters were less effective. They might be more effective if sufficient data were available to enable us to train them at a larger number of different scales. Similarly, we lacked sufficient data to reliably train filterbanks of Gabor filters (see technical report for details).

5 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. By contrast, if we have too many quantization bins, then the resulting probability distributions (and measures derived from them such as Chernoff information) may overfit the data. 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 [9]. (As a practical concern, the bigger the number of bins the larger the amount of computations required and the greater the memory requirements).

After 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

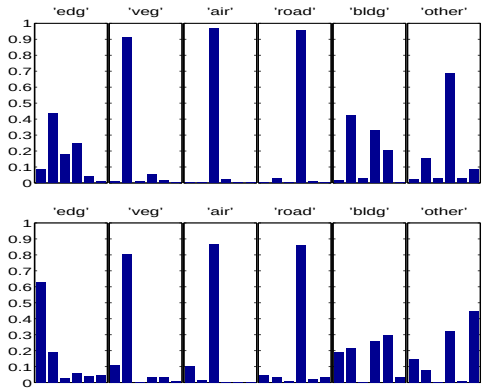

Figure 2: Colour at combined scales 2 and 4: Confusion plot with data driven prior. The top row shows the probabilities, $P(\alpha^*|\alpha)$, of the classification α^* when the true class is α. I.e. the top leftmost panel shows $P(\alpha^*|edge)$ for α^* being $\{edge, vegetation, air, road, building, other\}$. Observe that the classifications of *veg, air, road* are over 90%. The bottom row shows $P(\alpha|\alpha^*)$ with similar conventions. Observe that if a pixel is classified as *vegetation* then it has an over 80% chance of really being vegetation.

using only 6 adaptive bins for each dimension of the filter (we emphasize that this adaptation was performed over the entire dataset and *not* for each individual image). This enabled us to perform statistics on up to 6 coupled filters – which requires 6^6 quantized bins.

6 Results on the Sowerby Database

6.1 Classification and Confusion Matrices

We demonstrate the confusion matrix and the false positive and false negative rates. This confusion matrix, like our other statistics, is calculated over the entire database of 100 images. Not surprisingly, the best classes are road, air and vegetation. We see that colour by itself, see figure (2), is successful except for edge detection and such non-homogeneous classes as buildings (see section 6.3 for discussion), see figure (2) (colour is able to detect edges by combining filter responses at multiple scales – recall, for example, that the Laplacian of a Gaussian can be approximated by the difference of two Gaussians at different scales). Texture is perhaps surprisingly successful using only the Nitzberg operator to measure it, see figure (3). Moreover, texture (using Nitzberg) is significantly more effective than colour for detecting edges.

When colour and texture are combined with the data driven prior, see figure (4), we get the best results. Observe what happens if we use the uniform prior, see figure (5). This is better at finding buildings but worse at everything else. It also finds more true edges in the image but also has more false positives.

6.2 Chernoff Measures

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 nats, which was perfectly adequate for their task of tracking roads in aerial images. In our dataset, see figure (6,7,8), observe that we attain Chernoffs which are higher by *almost an order of magnitude*. This suggests that classifying a set of pixels into classes will be highly successful. Observe, again, that colour is relatively ineffective at detecting edges.

6.3 Classification Errors

On the whole, the Bayesian classification using joint texture and colour statistics is remarkably successful, particularly considering we are using no spatial grouping at all. However, we did detect some systematic biases.

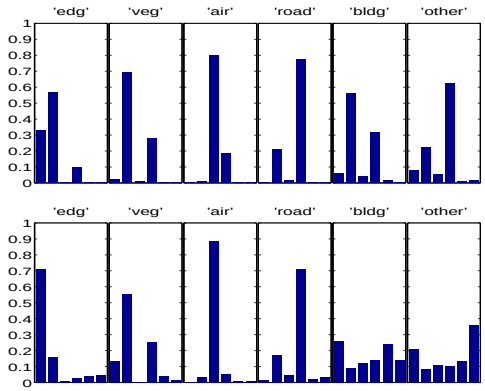

Figure 3: Texture at multi-scale using Nitzberg detectors: Confusion plot with data driven prior. Same conventions as previous figure.

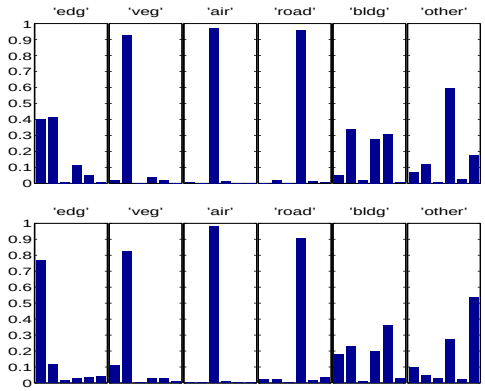

Figure 4: Colour and Texture: Confusion plot with data driven prior. Same conventions as previous figure.

Firstly, the *second most frequent error* occurred for *edge*s. This is hardly surprising since edges are notoriously hard to detect reliably. Moreover, most of the errors consisted of confusing *edge*s with *vegetation* (and vice versa). This is to be expected because one characteristic of *vegetation* is a high density of small scale edges. Overall, the classification rates and the Chernoff information for edges are good and suggest that a limited amount of spatial grouping will be sufficient to detect most of them, see also [6]. We observe also that edge-corners were often misclassified as *vegetation*. This is also not surprising because texture regions like *vegetation* will also have a high density of corners. Again, we expect that spatial grouping will be required to distinguish between corners due to *vegetation* and those due to *edge*s.

It does well on *vegetation*. There is some confusion with edges and corners, see previous paragraph. Also *vegetation* at a distance tends to get smoothed out and turns blue-grey. This affects its texture and colour properties and can cause it to be misclassified as *road*. Some texture on *building*s can get misclassified as *vegetation*.

Air is the most easily classified class. Smooth bright objects, as in *building*s and water on the road are sometimes seen as *air*. Very smooth road areas (more obvious with equal prior!) are sometimes seen as *air*. Dark thunderclouds are sometimes seen as *road*.

Road can also be reliably classified (with data driven prior). But *road*s in the far distance, where they are very smooth, can be classified as *air*. The *most common error* is that *building*s

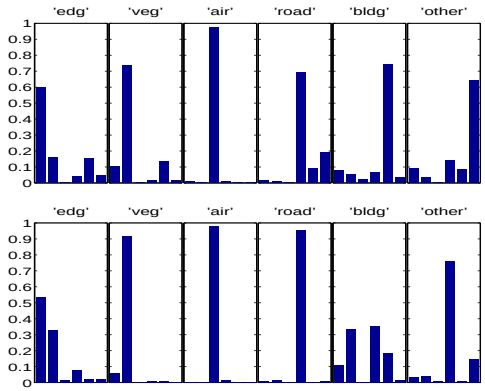

Figure 5: Colour and Texture: Confusion Plot with uniform prior. Same conventions as previous figure.

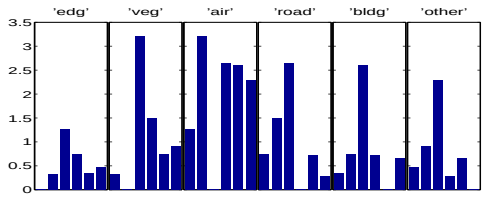

Figure 6: Colour at combined scales 2 and 4 : Chernoff Information in nats. For example, leftmost panel shows $C(P(y|edge), P(y|\alpha))$ for $\alpha = \{edge, vegetation, air, road, building, other\}$. Note that $C(P(y|edge), P(y|edge)) = 0$.

with flat surfaces are often seen as *road*s, see figure (9) (about 1/2 – unless the *building* prior is boosted).

Building is not a good class. A subclass – like stonework – seems to be a better class since they have homogeneous regular texture which differs from the less-structured texture in *vegetation*. *Building*s come in many styles and should be split into subclasses. Moreover, some specifically building-like features – such as straight-lines and right-angles – can not be easily detected by the filters we are applying. Our filters are local and only well suited for extracting homogeneous image properties.

The *Other* class is also poorly defined. It should ideally be split up into subclasses such as cars. At present, cars are partially classified as *other*. Parts like the windshield and bright smooth areas are classified as *air*. Areas near tail-lights, license plate, with high density of corners are often labelled *vegetation*. Noticeably, long straight thin objects – such as thin towers and, very occasionally, road curb boundaries – are labelled *other*.

7 Results on the San Francisco Database

The results for the San Francisco database are broadly similar. We investigated the classes *road, air, vegetation, car, building, and other*. We applied similar filters to those used for the Sowerby database.

As before, we calculated the confusion plots with the data driven priors and uniform priors. It should be stressed that *for this dataset the data driven priors can be misleading*. This is because the images are not fully segmented and the statistics for each class are obtained by samples which are interactively obtained. This means that the default class, other, contains many pixels which might best be assigned to other classes and hence its data driven prior is far larger than it should be. Moreover, certain classes such as vegetation are only sparsely represented in the images and

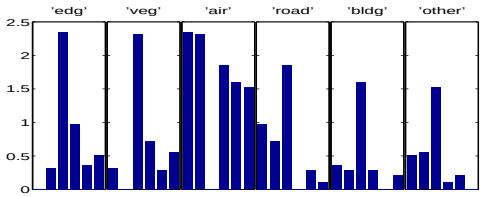

Figure 7: Texture with Nitzberg at scale 2: Chernoff Information. Same convention as previous figure.

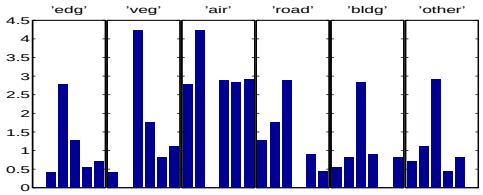

Figure 8: Colour combined with texture: Chernoff Information. Same convention as previous figure.

hence their data driven priors bias against them. This effect can be seen in figures (11,12); it seems that the uniform prior yields better results. Overall, the confusion plots show that the cues are very effective, yielding roughly similar success rates as the Sowerby dataset. The cues are most effective for the road, air, and vegetation classes. There is a slight decline due perhaps to the lack of an "edge class".

We show some typical San Francisco images in figure (13) with their segmentations, using the uniform prior, in figure (14).

8 Summary and Conclusions

Overall, we were surprised at how effective simple features could be in both domains. Using simple colour and texture filters we were able to get high classification rates into three of the six main classes (road, air, and vegetation). The classification is very fast because it is done by a simple loop through the image. This suggests that domain specific statistics are powerful for segmentation even without requiring spatial grouping. These domain statistics complement existing segmentation techniques and, when augmented by spatial grouping, should yield highly effective segmentations.

We have started comparing the statistics of the six classes between the two domains. Our preliminary results show some broad similarities but also some significant differences. Certain texture features, for example, seem to be surprisingly similar between the domains. On the other hand, not surprisingly, the air in the San Francisco database has different colours than the air in the Sowerby images (blue versus grey). In figure (15) we train on half the dataset and evaluate on the other half to ensure that we are not overfitting the data.

Finally, we encourage the development of similar segmented databases which can be used for statistical performance analysis of visual algorithms [1].

Acknowledgments

We want to acknowledge funding from NSF with award number IRI-9700446, from the Center for Imaging Sciences funded by ARO DAAH049510494, and from the Smith-Kettlewell core grant. We gratefully acknowledge the use of the Sowerby image dataset and thank Andy Wright for bringing it to our attention.

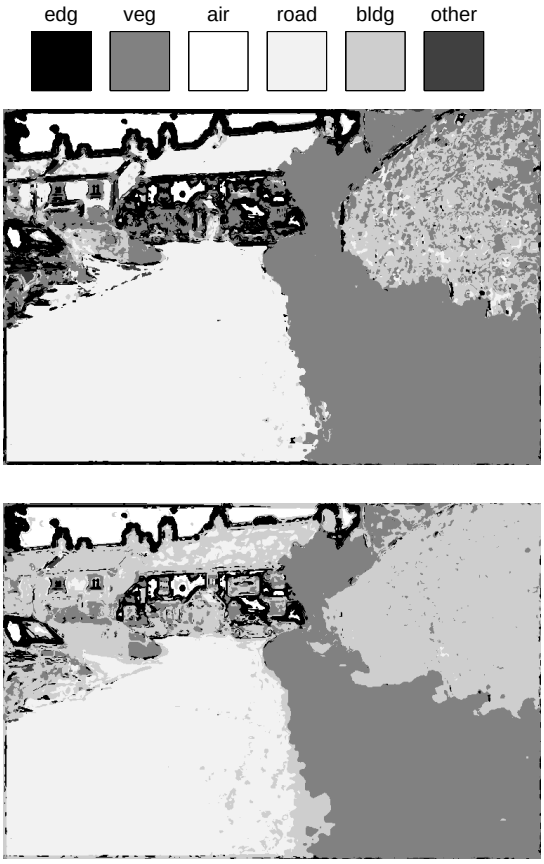

Figure 9: Observe (upper panel) that the buildings can be interpreted as roads when the data driven prior is used. If the prior for buildings is increased so as to equal that of road, then buildings are located more effectively (lower panel). (Stone walls, as in the right of the figure, are classified as "building".)See key, at top, for greyscale labelling conventions.

References

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

[2] D. Comaniciu and P. Meer. "Robust Analysis of Feature Spaces: Color Image Segmentation". In *Proceedings Computer Vision and Pattern Recognition*. CVPR'97. Puerto Rico. 1997.

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

[4] J. D. Crisman. "Color Region Tracking for Vehicle Guidance". In **Active Vision**. Eds. A. Blake and A.L. Yuille. MIT Press. 1992.

[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] S. Konishi, A. L. Yuille, J. M. Coughlan, and S. C. Zhu. "Fundamental Bounds on Edge Detection: An Information Theoretic Evaluation of Different Edge Cues." *Proc. Int'l conf. on Computer Vision and Pattern Recognition*, 1999.

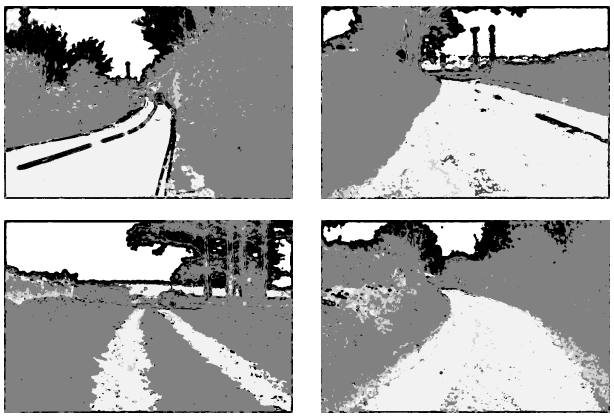

Figure 10: Examples of segmentations on typical Sowerby images. Same greyscale conventions as previous figure. The original images are shown in figure (1). Observe the effectiveness of the segmentations.

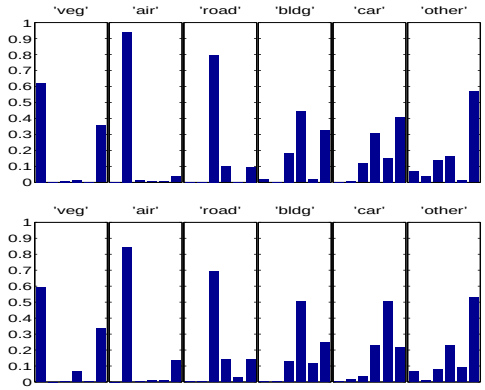

Figure 11: Colour and Texture: Confusion plot for San Francisco dataset with data driven priors (this is misleading). Same conventions as figures 2-6. Observe that the data driven prior biases against "vegetation" and in favour of the default class "other".

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

[8] 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.

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

[10] J. Shi and J. Malik. "Normalized Cuts and Image Segmentation". In *Proceedings Computer Vision and Pattern Recognition.* CVPR'97. Puerto Rico. 1997.

[11] M. Swain and D. Ballard. "Color Indexing". *International Journal of Computer Vision.* 7(1). pp 11-32. 1991.

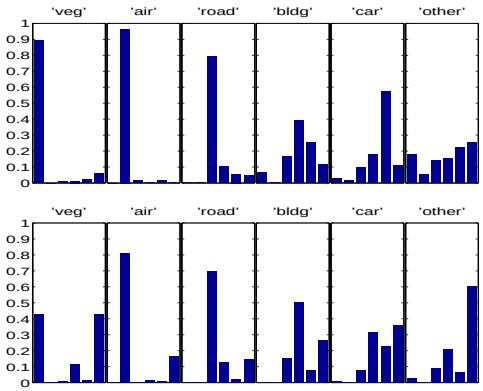

Figure 12: Colour and Texture: Confusion Plot for San Francisco dataset with uniform prior. Same conventions as previous figure. The uniform prior improves the assignment to classes such as "vegetation".

Figure 13: Examples of San Francisco images. See following figure for their segmentations.

[12] S. C. Zhu 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.

[13] 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.

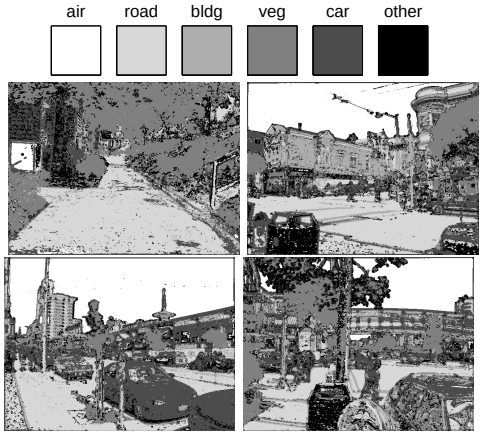

Figure 14: Examples of segmentation of the San Francisco images in the previous figure using the uniform prior. See key, at top, for greyscale labelling conventions.

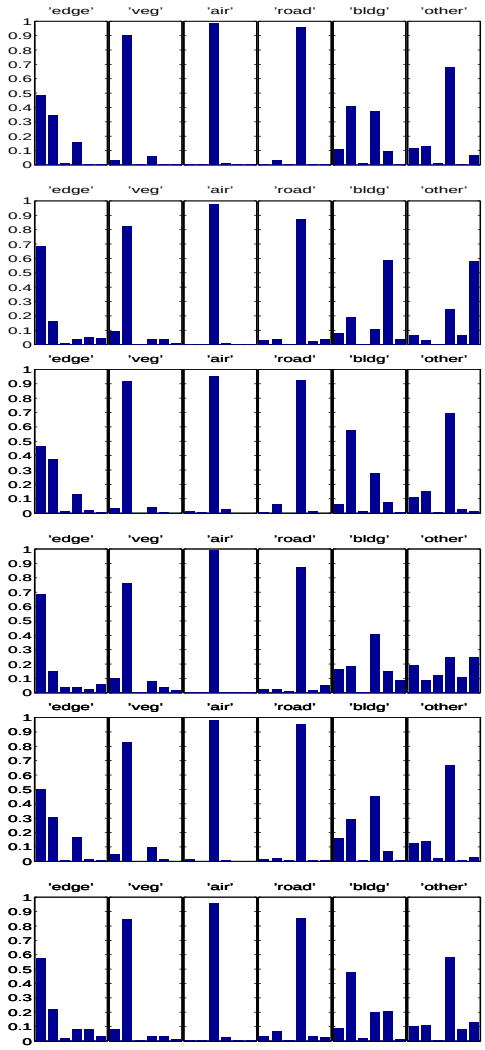

Figure 15: Checking for overfitting on Sowerby. The top two panels show the confusion plots trained and evaluated over the entire dataset. The middle two and bottom two panels show the confusion matrices where the conditional distributions are learnt on half the dataset (randomly chosen) and evaluated on the other half. Colour and texture filters used.