Automatic Interpretation of Outdoor Scenes

Neill W. Campbell†, William P.J. Mackeown†
Barry T. Thomas† and Tom Troscianko§

† Advanced Computing Research Centre
§ Department of Psychology
University of Bristol, Bristol, BS8 1TH
United Kingdom
Barry.Thomas@bristol.ac.uk

Abstract

This paper describes recent work on a neural network approach to outdoor scene interpretation. The results of evaluating a range of automatic region-based segmentation techniques based on a new *segmentation quality metric* are presented. The optimal technique is used to segment images of natural outdoor scenes. A powerful set of features designed for outdoor scene analysis is extracted from regions in the segmented images and used to train a neural network to recognise eleven different *classes* of objects, including sky, road, building and vegetation. The system is tested on a large number of images which have been hand-labelled to provide ground-truth segmentations and interpretations. This fully automated system achieves an impressive mean classification accuracy of 81.4% per image on unseen data, and runs in approximately 1 CPU minute on a Sun SPARCstation 20.

1 Introduction

Outdoor scene analysis is a challenging problem in computer vision. Several different approaches have been used in previous work. Model-based approaches have met with some success in domains where objects can be well described using geometric primitives [2]. However, this approach is not ideal since it can be difficult to find good models for complex objects in outdoor scenes, e.g. trees. Knowledge-based approaches use a set of hand-coded rules to describe the properties and geometric relationships of each object to be recognised, together with rules for object-specific recognition strategies which reduce the overall complexity of the interpretation process. Examples are the SCHEMA vision system [4] and the region-based scene analysis system of [13], which have both been demonstrated to give good results in interpreting typical outdoor scenes. A drawback of these approaches is the significant effort required to develop the knowledge base.

BMVC 1995 doi:10.5244/C.9.30

More recently, statistical approaches to outdoor scene analysis based on region classification have been demonstrated [16, 7, 11]. The method described in [11] used a set of features based on properties of a region such as colour, brightness, texture, shape, topological properties, etc. The system achieved a mean classification accuracy of 86.3% per image by area on a model of ideal segmentation. Although the performance of this scene analysis system is impressive, the use of an ideal segmentation model, whose implementation is largely manual, is unrealistic. With automatic segmentation techniques, the segments generated do not necessarily correspond to whole objects or even object components, thus increasing the difficulty of later semantic analysis. A key question for the scene analysis system is how use of an automatic segmentation technique in place of the ideal segmentation model would affect the accuracy of the image interpretation process. This paper describes an investigation extending this system to fully automatic outdoor scene interpretation.

2 The Labelled Image Database

The work presented here has used a sample of 80 colour images of outdoor road scenes from the Bristol Image Database [10]. The database consists of over 350 images of a wide range of urban and rural scenes. The images were digitised using a calibrated digitiser from small-grain 35mm colour transparency film to produce high-quality 36-bit colour images. The statistics of the image content and acquisition conditions have been carefully controlled. Briefly, of the 80 images in the sample, 40 are of urban scenes and 40 are of rural scenes spanning a wide range of viewpoints. Environmental conditions during image capture were dry, fully overcast and good atmospheric visibility (at least 1km). For all images, the camera was focussed at infinity and an aperture no larger than f/8 was used in order to ensure an adequate depth-of-field. The declination of the line of sight was approximately 4°. These conditions were uniformly and independently distributed over each other. The images have been hand-labelled to identify the objects in the scenes, thus providing a ground truth about the image contents. This database is a key resource in our work on quantifying vision system performance.

3 Machine Segmentation

The purpose of a segmentation process is to divide an image into a number of regions which typically correspond to objects and/or parts of objects. In this paper we are concerned with region-based segmentation. The ability to carry out such a process automatically has been a topic of much research in computer vision [5, 9, 1, 14]. One of the difficulties with all segmentation techniques has been to quantify their success with respect to a particular application. The best algorithm or parameter settings to use for a particular purpose is often left to the user to decide 'by-eye', using a small set of images which is in some sense 'typical'.

The database described above provides a ground-truth segmentation against which any other segmentation technique may be compared. We now discuss the segmentation quality metric we have developed for comparing any machine-

generated segmentation of an image with its ideal counterpart in the database, thus enabling objective evaluation of the quality of different segmentation techniques to be made.

3.1 The Segmentation Quality Metric

We will define the segmentation quality metric, $\mathcal{Q}(a, b)$ which quantifies the similarity between two different segmentations a and b of the same image, where b is normally the ground-truth segmentation from the database. This metric will be used to quantify the effectiveness of some common segmentation techniques on images in the database.

In designing the metric we require that the following criteria be satisfied:

1. The metric is bounded: $0 \leq \mathcal{Q}(a, b) \leq 1$.
2. The metric is symmetric in a and b, i.e. $\mathcal{Q}(a, b) = \mathcal{Q}(b, a)$.
3. The metric satisfies the triangle inequality, $\mathcal{Q}(a, b) \leq \mathcal{Q}(a, c) + \mathcal{Q}(c, b)$.
4. The maximum value is for an identical pair of segmentations: $\mathcal{Q}(a, a) = 1$.
5. A segmentation a consisting of over-segmented regions compared to the ideal segmentation b, has a score less than 1 and in the limit, where every region in a is a single pixel, $\mathcal{Q}(a, b) \simeq 0$.
6. A segmentation a consisting of under-segmented regions compared to the ideal segmentation b, has a score less than 1 and in the limit, where a consists of a single region (the whole image), $\mathcal{Q}(a, b) \simeq 0$.

The first step is to define a pairwise area-based comparison between regions $a_m$and b_n, where $b_n$is the region having the largest overlap with region a_m:

$$\mathcal{O}(a, b) = \sum_{m=1}^{N} \frac{(a_m \cap b_n)}{\mathcal{A}} \tag{1}$$

where N is the number of regions in segmentation a and $\mathcal{A}$ is the total area of the image. To provide the required symmetry, the metric is defined as:

$$\mathcal{Q}(a, b) = \mathcal{O}(a, b) \times \mathcal{O}(b, a) \tag{2}$$

Consider the synthetic segmentations shown in Figure 1. Here, segmentation a is under-segmented with respect to segmentation b. The calculation of $\mathcal{Q}(a, b)$ is shown in Figure 2.

3.2 Segmentation Techniques

The following low-level segmentation techniques, all of which guarantee closed regions, have been evaluated using the metric on monochromatic images.

Zero-Crossings : Zero-crossings of the Laplacian of the image [12]. This method is implemented by convolution of the image with a 15×15 Laplacian kernel.

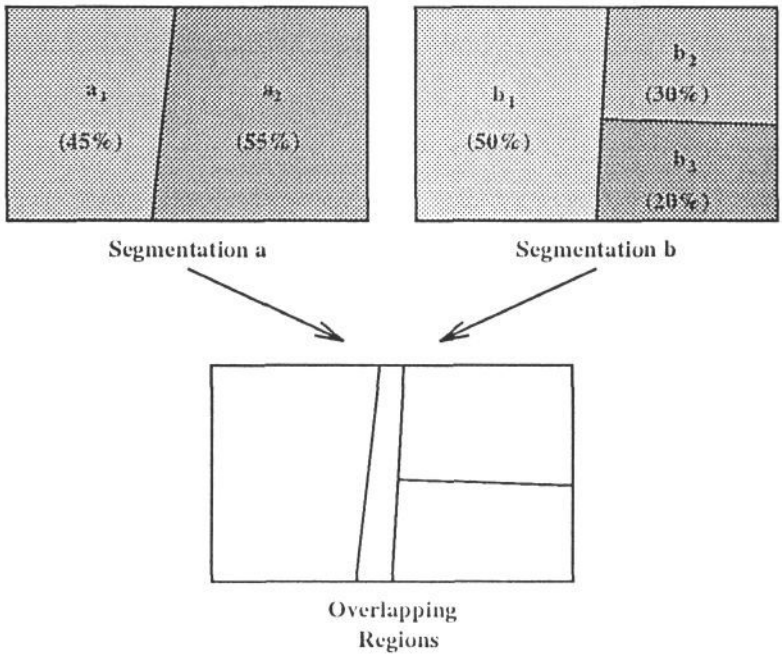

Figure 1: Overlap of Two Synthetic Segmentations

Region-Growing : A region growing algorithm as described in [6]. This method, beginning at the pixel level, iteratively merges pairs of regions if they have a similar grey-level distribution and a low average edge contrast along the shared region boundary.

K-means : The set of K centres in the grey-level histogram which minimises the sum of the squared distances of pixel grey-levels to their nearest centre [15]. Every pixel is then marked with its nearest centre.

Gabor Filters : Pixels are labelled with the index of the multi-scale Gabor filter [3] which gives the greatest response magnitude regardless of orientation.

3.3 Results

The results of scoring the segmentation techniques on the database are shown in Figure 3.

It is interesting to note that K-means, which is the best method, requires the least parameter tuning to optimise it for a large sample of images. The optimum value of K was found to be 3, but as shown in Figure 4, the performance varies little over a small range of values of K.

Region a_m	Region b_n	$\frac{(a_m \cap b_n)}{a_m}$	$\frac{a_m}{\mathcal{A}}$	$\frac{(a_m \cap b_n)}{\mathcal{A}}$
a_1	b_1	100%	45%	0.450
a_2	b_2	50%	55%	0.275
			Sub-Total $\mathcal{O}(a,b)$	0.725

Region b_m	Region a_n	$\frac{(b_m \cap a_n)}{b_m}$	$\frac{b_m}{\mathcal{A}}$	$\frac{(b_m \cap a_n)}{\mathcal{A}}$
b_1	a_1	80%	50%	0.400
b_2	a_2	100%	30%	0.300
b_3	a_2	100%	20%	0.200
			Sub-Total $\mathcal{O}(b,a)$	0.900
			Total $\mathcal{Q}(a,b) = \mathcal{O}(a,b) \times \mathcal{O}(b,a)$	**0.653**

Figure 2: Calculation of the Segmentation Quality Metric

Segmentation Technique	Metric Score
Human	1.000
K-means	0.589
Isotropic Gabor	0.443
Region-Growing	0.411
Laplacian Z-C	0.212

Figure 3: Comparison of Segmentation Techniques

4 Automatic Outdoor Scene Interpretation by Region Classification

This section discusses a neural network approach to outdoor scene interpretation based on region classification. In this context, it is more relevant to consider the mean accuracy of region classification, rather than the quality of the region segmentations themselves as measured by the metric defined above.

4.1 Method

Previous work [11] has demonstrated a neural network approach to region classification. The work used a set of features based on internal properties of a region: colour, brightness, texture, topological properties (e.g., Euler number and hollowness), shape information (e.g., compactness and multi-scale boundary descriptors based on curvature extrema), and a set of local contextual features defined in terms of properties of pairwise combinations of a region and each of its four largest

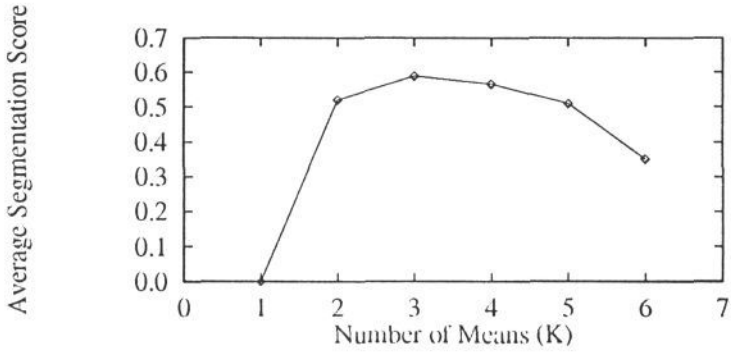

Figure 4: Average Score for K-means Segmentation for Several Values of K

nearest-neighbours. Although the use of viewpoint-dependent features such as absolute position and size would have increased the recognition accuracy for restricted viewpoints, as in [4, 13], such features were avoided in order to provide a more general solution. The feature set is described in more detail in [10].

A neural network was trained to classify the regions in segmented images from the image database using the labels of 11 different object classes as shown in Figure 5. The network architecture was a Multi-Layer Perceptron with a single hidden-layer and 11 output units. The labels were represented on the output units as a 1-out-of-11 binary code, thus enabling an approximation to Bayesian classification to be obtained [8]. The system achieved a mean classification accuracy of 86.3% by area on the ideal segmentations from the database.

	Label
1	Bounding_Object
2	Building
3	Cloud/Mist
4	Illumination Shadow
5	Mobile_Object Car
6	Road_Sign
7	Road_Surface Marking
8	Road_Surface Other
9	Road_Border
10	Telegraph_Pole
11	Vegetation

Figure 5: Labels used for the Object Classes

For direct comparison with the earlier work, an identical network architecture

and feature set are used here. The network consists of 29 input units and 11 output units. The hidden layer consists of 16 units which was optimised to give the best results in the previous study. The network is re-trained using conjugate gradients optimisation with batch presentation of examples. Test results are obtained with unseen examples.

This section will study the effect of using the same network architecture, feature set, and labels with an *automatic* segmentation process.

4.2 Results and Discussion

Figure 6 shows the results of the classification process for the ideal segmentation

Classification Accuracy	Segmentation Method		
	3-means	4-means	Ideal
% Correct By Region	70.4%	72.4%	70.0%
% Correct By Area	73.2%	81.4%	86.3%

Figure 6: Region Classification Results for Machine-Generated and Ideal Segmentations

and for K-means segmentation. Since the quality metric gave a similar score for several values of K, we have provided results for both $K = 3$ and $K = 4$. As we see, the 4-means segmentation gives a better classification result. This is due to the fact that the scoring metric treats under- and over-segmentation equally, whereas the region classifier performs better with moderately over-segmented images.

The 4-means automatic segmentation leads to the impressive result of 81.4% of pixels being assigned the correct label. This is not much less than the performance obtained with the ideal segmentations, and is significantly greater than the chance recognition level (9%). The results expressed by area (row 2) are better than the corresponding results by region frequency (row 1) because large regions are more often correctly classified.

Figure 7 shows a scene typical of the database, and Figure 8 shows the labelled version of it. The 11 different object classes have been assigned different colours (shown here as grey levels). The largest regions in the image are correctly labelled, i.e. the sky, the road, the trees, the houses, the car and the pavements. Some smaller regions are mis-classified, e.g. the white label of the car bumper signifies a road marking, a small region of the road is classified as pavement, and distant sections of pavement have been classified as building.

Implementation Performance

A sequential implementation of our neural network approach to outdoor scene interpretation takes approximately 1 minute of CPU time per image for images of 768 × 512 resolution on a Sun SPARCstation 20. This is equivalent to $O(10^9)$ floating point operations per image. To the best of our knowledge, our system is the first implementation to reach this level of performance. Furthermore, the

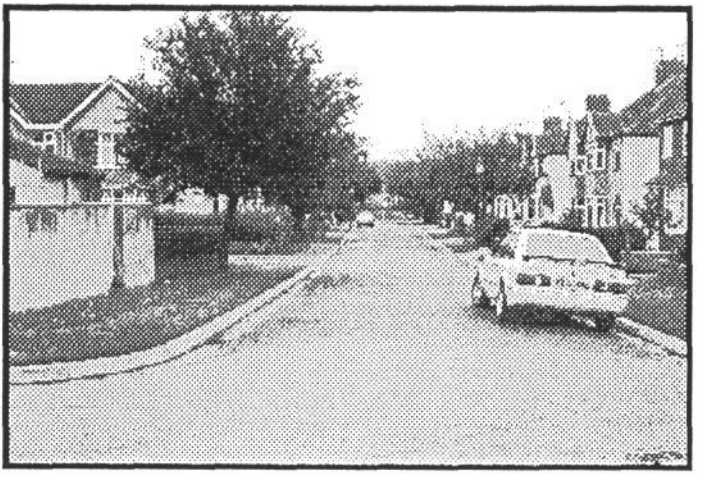

Figure 7: Example Image : "Film19a23"

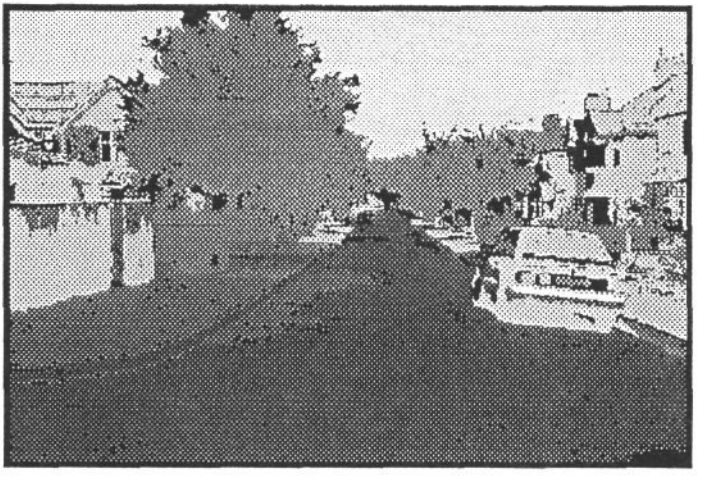

Figure 8: Region Classification for Figure 7

neural network approach has substantial implicit parallelism since features can be extracted in parallel and regions can be labelled in parallel. Spatial parallelism in the segmentation could be achieved by using a local K-means method. Real-time (5 frames per second) outdoor scene interpretation based on this system is therefore possible with present-generation parallel supercomputers.

5 Conclusions

This paper has presented the results of an investigation into a fully automatic system for outdoor scene analysis. This system is capable of classifying regions of images with a mean accuracy 81.4% by area. This is an extremely high level of performance for such a complex task.

A neural network has been trained on a set of features extracted from regions in the segmented images to classify the regions into one of eleven different classes of object. This image interpretation system has been tested on a large number of images, and takes approximately 1 CPU minute per image on a Sun SPARCstation 20, i.e. $O(10^9)$ FLOPs. Real-time (5 frames per second) outdoor scene interpretation based on this system is possible with present-generation parallel supercomputers.

A metric has been designed for comparing the quality of a segmentation with its ideal human-segmented counterpart from a labelled image database containing ground-truth segmentations and interpretations. A range of automatic segmentation methods have been evaluated using this metric, and the optimal method has been used to segment the images in our database.

Acknowledgements

This project is supported by EPSRC (grant GR/J36662). We thank British Aerospace PLC, Sowerby Research Centre, Bristol for their support with the development of the database. Additional thanks are due to Phil Greenway and Andy Wright for useful discussions.

References

[1] Bhanu, B. and Parvin, B.A. Segmentation of Natural Scenes. *Pattern Recognition*, 20(5):487–496, 1987.

[2] Brooks, R.A. Model-based 3D Interpretation of 2D Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5(2):140–150, 1983.

[3] Daugman, J.G. Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1169–1179, Jul 1988.

[4] Draper, B.A., Collins, R.T., Brolio, J., Hanson, A.R., and Riseman, E.M. The Schema System. *The International Journal of Computer Vision*, 2:209–250, 1989.

[5] Fu, K.S. and Mui, J.K. A Survey on Image Segmentation. *Pattern Recognition*, 13(1):3–16, 1981.

[6] Gay, M. Segmentation Using Region Merging with Edges. *Procs.* 5th *Alvey Vision Conference*, pages 115–119, 1989.

[7] Haddon, J.F., Boyce, J.F., Protheroe, S. and Hesketh, S., 1993, "Neural Networks for the Texture Classification of Segmented Regions in Forward Looking Infrared Images", J. Illingworth, editor, Procs. 4th British Machine Vision Conference, 197–206, Guildford, UK.

[8] Hampshire, J.B. and Pearlmutter, B.A. Equivalence proofs for multi-layer perceptron classifiers and the Bayesian discriminant function. In *Procs. 1990 Connectionist Models Summer School*, volume 1, pages 159–172, San Mateo, CA, 1990. Morgan Kaufmann.

[9] Haralick, R.M. and Shapiro, L.G. Image Segmentation Techniques. *Computer Vision, Graphics, and Image Processing*, 29(1):100–132, 1985.

[10] Mackeown, W.P.J. *A Labelled Image Database and its Application to Outdoor Scene Analysis*. PhD thesis, University of Bristol, 1994.

[11] Mackeown, W.P.J., Greenway, P., Thomas, B.T., and Wright, W.A. Road Recognition with a Neural Network. *Engineering Applications of Artificial Intelligence*, 7(2):169–176, 1994.

[12] Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, W.H., New York, 1982.

[13] Ohta, Y. *A Region-Oriented Image-Analysis System by Computer*. PhD thesis, Kyoto University, March 1985.

[14] Pal, N.R. and Pal, S.K. A Review on Image Segmentation Techniques. *Pattern Recognition*, 26(9):1277–1294, 1993.

[15] Tou, J.T. and Gonzalez, R.C. *Pattern Recognition Principles*. Addison Wesley, Reading, MA, 1977.

[16] Wright, W.A. Image Labelling with a Neural Network. *In Procs.* 5th *Alvey Vision Conference* Reading, UK, 227–232, 1989