Neural Network Analysis of MINERVA Scene Analysis Benchmark

MARKOS MARKOU
MONA SHARMA
SAMEER SINGH

Department of Computer Science
University of Exeter
Exeter EX4 4PT
United Kingdom

Tel: +44-1392-264053
Fax: +44-1392-264067
Email: {m.markou, m.sharma, s.singh}@ex.ac.uk

Abstract

Scene analysis is an important area of research with the aim of identifying objects and their relationships in natural scenes. MINERVA benchmark has been recently introduced in this area for testing different image processing and classification schemes. In this paper we present results on the classification of eight natural objects in the complete set of 448 natural images using neural networks. An exhaustive set of experiments with this benchmark has been conducted using four different segmentation methods and five texture-based feature extraction methods. The results in this paper show the performance of a neural network classifier on a ten fold cross-validation task. On the basis of the results produced, we are able to rank how well different image segmentation algorithms are suited to the task of region of interest identification in these images, and we also see how well texture extraction algorithms rank on the basis of classification results.

1. Scene Analysis

Outdoor scene analysis is a complex problem. A number of different approaches have been used for recognising different objects in such scenes. In early experiments on scene analysis, simple problems were tackled. For example Brice and Fennema[5] defined the procedure to interpret simple objects in images such as wedges, cubes, wall and floor. Model based approaches, such as the one proposed by Brooks[6], have met with some success provided that objects can be defined with geometric primitives. These approaches have problems with recognising natural objects such as trees for example where such primitives are hard to define. Another approach is based on the use of a knowledge-based scheme where hand-coded rules are used for object recognition. These rules describe the characteristic properties of objects of different types. Some examples of such work include SCHEMA vision system by Draper et al.[10], region based scene analysis by Ohta[25], and VISOR connectionist system for scene analysis [21]. Although these approaches have shown reasonable results, a significant amount of computational effort is required even for simple scenes. More recently, far more complex problems are

being solved using texture based image analysis. The theses of Becalick[4] and MacKeown[22] provide an excellent review treatment of other studies in this area and summarise the progress made. In particular, Becalick reviews important studies based on knowledge based approaches in scene analysis, Autonomous Ground Vehicle (AGV) research, and image database and multimedia search.

Batlle et al. [3] review a number of studies in the area of scene analysis for urban and natural scene classification. Campani et. al.[7], and Parodi and Piccioli [26] have used colour information in the classification of natural objects including road boundaries, road signs, vehicles, buildings and trees. Colour information has also been used by the following authors: Draper et al.[10], and Hanson and Riseman[13] for the discrimination of objects in (road scenes) including sky, foliage, shoulder, trunk, sign-post, wire, warning sign, phonepole, road, roof, building, roadline, grass, unknown, and (house scenes) including sky, tree, grass, bush, shutter, wire, housewall, roof, roadline, road, and film border; Strat[31] for the discrimination of geometric horizon, complete sky, complete ground, skyline, sky, ground, raised object, foliage, bush, tree trunk, tree crown, tree, trall, and grass; Asada and Shirai[1], Hirata et al.[15] and Taniguchi et al.[32] for discriminating between road, roadlines, crosswalk, sky, trees, buildings, poles, cars, and trucks; Ohta et al.[153] for the classification of sky, tree, building, road, unknown, car, car shadow, and building window; Bajscy and Joshi[2] for the classification of ground, sky, horizon, skilliness and trees; Lavine[19], Laveen and Sheehan[20] and Nazif and Lavine[24] for the classification of bushes, car, fence, grass, road, roof, and ground; Douglass[9] for the classification of trees, house, grass, sky, car, street, window, brick-wall, concrete wall, roof and ground; Kim and Yang[17] for the classification of sky, foliage, road, grass, wall, roadline, window, footway, and trees; and Campbell et al.[8] for the classification of sky, vegetation, road markings, road, pavement, building, fence/wall, road sign, signs/poles, shadow, and mobile objects. In studies dealing with colour information, a range of segmentation techniques, features, and classification schemes have been used. Segmentation techniques have been based on edges, colour, vanishing point detection, colour histograms, texture operators, shape and contextual information. The use of colour undoubtedly helps the recognition process as different objects are often of different colours. Features used include those based on colour, texture, and shape. Classification strategies that have been used include rule based systems, bayesian classifier, nearest neighbour classifier, and neural networks.

The identification of objects in greyscale images is more complicated, though computationally simpler than processing colour images. The recognition ability of the system is limited as colour information is not used, however, for a number of real-time scene analysis applications, greyscale analysis is a better choice. A number of studies on scene analysis have used greyscale images for analysis. For example, Efenberger & Graefe[11], and Regenberger and Graefe[27] tried to identify road, tree trunk, tree, rock, barrels and cars in natural images using edge and grey level information; Ide et al.[16] trained a classifier to recognise pole images on the basis of diameter, height and layout information; Gamba et al.[12] and Mecocci et al.[23] classified natural objects, lateral vertical surfaces, frontal vertical

surfaces, horizontal surfaces and vanishing points; and Kumar and Desai[18] classified sky, tree, sidewalk and road. In greyscale images, the shape and texture information is useful for characterising objects. For natural objects such as trees, snow, etc., fractal features can also be used for characterising them. For objects in urban scenes, shape feature play a more important role than they do otherwise.

In this paper we investigate the problem of scene analysis on greyscale images. MINERVA benchmark for scene analysis includes a collection of 448 images in both colour and greyscale format. In this paper we investigate the classification success on eight natural objects using neural networks. The features sets have been generated using a range of segmentation and texture extraction methods. This study gives us an important insight into the utility of image processing techniques with regards to their utility in defining how well they can be used in scene analysis. The paper is laid out as follows. First we describe the characteristics of the MINERVA benchmark. Next, we describe the generation of feature sets on this data and give detailed results based on using a neural network for a ten fold cross-validation study.

2. MINERVA Benchmark

Outdoor data was collected from the University of Exeter campus using a Panasonic digital video camera (Model NV-DS77B) having 720x576 pixels resolution. The data was collected in the form of 448 coloured digital stills. Images of different natural scenes containing grass, trees, sky, clouds, pebbles, road and bricks were collected. The camera was stabilised using a tripod. All of the images were taken during daytime to give a realistic view of the real environment. Environmental conditions during capture were dry, fully overcast, and good atmospheric visibility. For all images the camera was focussed at infinity. The images contain nine different natural objects trees, grass, sky, clouds, bricks, pebbles, road, water, leaves and water. Water is not included for classification, as it is present only in four images, and therefore it is not possible to have enough samples for classification. All of the images collected are stored in bitmap (.bmp) format with 16-bit colour depth and a resolution of 720x576 pixels. The images have been reduced to a size of 512x512 pixel resolution using the ‘convert’ command in Linux for further analysis. At the same time, the images have been converted in .pgm format with 256 grey levels.

The benchmark data is available as both colour and greyscale images at http://www.project-minerva.ex.ac.uk or http://www.dcs.ex.ac.uk/minerva. Some sample images are shown in Figure 1. The results that we present here have only been produced on greyscale images.

3. Methodology and Feature Extraction

The main steps involved in feature extraction include image segmentation, texture feature extraction and classification using neural networks. We have used for segmentation methods including fuzzy c-means clustering (FCM), histogram thresholding, region growing and split and merge. We use five

texture extraction methods including autocorrelation, co-occurrence matrices, edge frequency, Law's, and run length. The images are first segmented into separate regions on the basis of homogeneity criterion. This is a particularly difficult process for images of natural objects. Also variable lighting conditions make it difficult for an otherwise competent image segmentation technique to give a reasonable result. We do not apply any form of image enhancement as it is not clear how this may have an impact on the quality of texture measures extracted. Next, the segmented images are labelled. The tagged images form the ground truth data. Once the image regions have been tagged, a unique texture feature vector is extracted from each region that describes its texture properties. These feature vectors are then used for teaching a classifier to learn characteristic patterns associated with corresponding classes. In our case, the true performance of the taught system has been evaluated using cross-validation approach. We use ten fold cross-validation for testing data with the backpropagation based multi-layer perceptron neural classifier.

Each step of our analysis can now be explained in more detail. A total of four image segmentation methods have been applied to the image set including fuzzy c-means clustering (FCM), histogram thresholding (HT), region growing (RG), and split and merge (SM). In order to generate ground truth data, we needed to develop a technique whereby using a point and click system, an expert can label segmented regions with appropriate classes. One simple, but inefficient, way of doing this is to present both the original and the segmented images to the user and ask them to tag the segmented regions by looking at corresponding regions that are identifiable in the original. Unfortunately, this strategy does not work when an image contains several regions. A better interface can be designed that presents a single image that contains the highlighted boundaries of different regions. By clicking within a region, its tag can be supplied and the system can automatically colour the region to denote that it has been tagged. In order to develop this method, we first need to identify region boundaries. Region boundaries are not smooth on a segmented image due to its noisy nature. Median filtering is applied to remove noise in segmented images. The 5x5 filter is found to the best for our analysis. Once smoother region boundaries are available on segmented images, the boundary information is extracted using Susan algorithm [29]. Susan automatically extracts the region boundaries and these are overlaid on the original image. On the boundary highlighted original images, we now tag the segmented regions. The output of this analysis appears as a completely tagged image. The output of this process is a text file, called map file that shows the image number, region number and appropriate class tag. Texture is extracted from all pixels in the segmented regions. The five different texture extraction methods generate five different feature sets for the same set of images. When the features are extracted from the original image, they are mapped with this file to generate the feature set containing each region as a row vector appended at the end with its class name. Finally, the feature sets are presented to the classifier for analysis.

4. Experimental results

In our images, objects from eight classes are present including sky, clouds, bricks, pebbles, road, trees, leaves and grass. We first separate the regions on the basis of colour information as belonging to class "vegetation" or "natural objects". Vegetation objects consist of trees, grass and leaves. Natural objects include sky, clouds, pebbles, bricks, and road. We train separate neural network classifiers for classifying data. In order to ensure that the full data set is used and a meaningful comparison can be drawn when comparing feature sets, no outlier removal or other preprocessing techniques are used. The neural networks used for experiments are shown in Figures 2 and 3. The neural network for vegetation classification consists of three output nodes corresponding to the classes grass, trees and leaves. The neural network for natural object classification has five output nodes corresponding to the classes sky, clouds, bricks, pebbles and road. For each feature extraction method, different number of features have been extracted. This information is displayed in these figures showing that for Auto-Correlation Function (ACF) we have 99 features, for Co-occurrence Matrices (CM) we have 20 features and so on. Since the neural networks would take a considerable amount of time training on these large number of features, the first five principal components for each feature set have been used for training the networks. The feature extraction for these is described in detail by Sharma [28]. Methods except for co-occurrence matrix feature selection are also described in Sonka et al.[30]. Co-occurrence matrix measures are based on Haralick's suggested features [14]. The neural networks have been optimised for each of the ten cross-validation folds individually. The architecture of the networks has been optimised by varying the number of hidden nodes and monitoring the generalisation test error of the network. The network yielding the least generation error is selected for further analysis.

The number of samples used for analysis is shown in Table 1 for vegetation data and in Table 2 for natural object data. The co-occurrence matrix method fails to work with very small regions generated by the segmentation process. Hence the number of samples for this method are slightly less than the others in these tables. The number of samples generated by the segmentation process is fairly balanced. For example, using *fuzzy c-means clustering* (FCM) we get the following rough proportion: trees (21.3%), grass (14.8%), sky (16.2%), clouds (13.6%), bricks (7.5%), pebbles (6.7%), road (8.4%) and leaves (11.4%). For *histogram thresholding* we get the following rough proportion: trees (15.4%), grass (11.8%), sky (20.5%), clouds (14.8%), bricks (13.4%), pebbles (5.6%), road (9.5%) and leaves (9.0%). For *region growing* we get the following rough proportion: trees (17.7%), grass (6.8%), sky (18.5%), clouds (17.4%), bricks (14.1%), pebbles (6.4%), road (5.8%) and leaves (13.2%). Finally, for *split and merge* we get the following rough proportion: trees (12.7%), grass (15.2%), sky (16.1%), clouds (12.7%), bricks (12.2%), pebbles (12.5%), road (8.6%), and leaves (10.0%). It is therefore possible to have a reasonable training for chosen classifiers on such data.

The results of the analysis are shown in Tables 3 and 4. It should be noted that these results are for the complete MINERVA benchmark of 448 images. The classification rates for vegetation data are better than natural object data. The following key observations can be drawn from these:

- For vegetation analysis, the best texture features appear to be edge frequency and autocorrelation. A best recognition rate of 78.7% correct classification is obtained using split and merge segmentation and edge frequency features.
- For natural object analysis, the best texture features appear to be edge frequency. The best recognition rate of 72.3% correct is observed with edge frequency and split and merge.
- For each texture extraction method, the use of different preceding segmentation method has an important bearing on the final classification results.
- Segmentation methods can be ranked on the basis of how well they perform. It appears that split and merge is by far the most superior and there are minor differences between others. The only disadvantage of split and merge is that it gives too many regions for analysis that makes it more computationally expensive.

5. Conclusions

In this paper we have provided the results of a comprehensive analysis of MINERVA scene analysis benchmark. As more studies are published with this benchmark, it will become possible to put our results into perspective. The results are quite good considering that only greyscale images are being used and texture is the only feature. In order to make our study comprehensive, we have used more than one segmentation method and more than one texture extraction method. This allows us to understand the amount of interaction between these two stages. In other words, by using different combinations of segmentation and texture algorithms, how different are the recognition performances. We find that the difference between the best and the worst recognition rates can be as high as 28% for vegetation data and 21.5% for natural object data. On the whole neural networks perform reasonably well with good recognition performances of 78.6% and 72.3% correct classifications on the two experiments.

References

1. M. Asada and Y. Shirai, Building a world model for a mobile robot using dynamic semantic constraints, Proc. 11th International Joint Conference on Artificial Intelligence, pp. 1629-1634, vol. II, Detroit, 1989.
2. R. Bajcsy and A.K. Joshi, A partially ordered world model and natural outdoor scenes, in *Computer Vision Systems*, A.R. Hanson and E.M. Riseman (eds.), pp. 263-270, Academic Press, New York, 1978.
3. J. Batlle, A. Casals, J. Freixenet and J. Marti, A review on strategies for recognising natural objects in colour images of outdoor scenes, *Image and Vision Computing*, vol. 18, pp. 515-530, 2000.
4. D.C. Becalick, Natural scene classification using a weightless neural network, *PhD Thesis*, Department of Electrical and Electronic Engineering, Imperial College, London, 1996.
5. C.R. Brice and C.L. Fennema, Scene analysis using regions, *Artificial Intelligence*, vol. 1, pp. 205-226, 1970.
6. R.A. Brooks, Model based 3D interpretation of 2D images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 5, no. 2, pp. 140-150, 1983.
7. M. Campani, M. Capello, G. Piccioli, E. Reggi, Visual routines for outdoor navigation, Proc. of Intelligent Vehicles Symposium, pp. 107-112, Tokyo, Japan, 1993.
8. N.W. Campbell, W.P.J. Mackeown, B.T. Thomas, and T. Troscianko, Interpreting image databases by region classification. *Pattern Recognition*, vol. 30, no. 4, pp. 555-563, 1997.
9. R.J. Douglass, Interpreting three dimensional scenes: a model building approach, *Computer Graphics, Vision and Image Processing*, vol. 17, pp. 91-113, 1981.
10. B.A. Draper, R.T. Collins, J. Brolio, A.R. Hanson and E.M. Riseman, The scheme system, *International Journal of Computer Vision*, vol. 2, pp. 209-250, 1989.
11. W. Efenberger and V Graefe, Distance invariant object recognition in natural scenes, Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1433-1439, Osaka, Japan, 1996.
12. P. Gamba, R. Lodola and A. Mecocci, Scene interpretation by fusion of segment and region information, *Image and Vision Computing*, vol. 15, pp. 499-509, 1997.
13. A.R. Hanson and E.M. Riseman, Visions: a computer system for interpreting scenes, in *Computer Vision Systems*, A.R. Hanson and E.M. Wiseman (eds.), pp. 303-333, Academic Press, New York, 1978.
14. R.M. Haralick, K. Shanmugam and I. Dinstein, Textural features for image classification, *IEEE Transactions on Systems, Man and Cybernetics*, vol. 3, no. 6, pp 610-621, 1973.
15. S. Hirata, Y. Shirai and M. Asada, Scene interpretation using 3D information extracted from monocular colour images, Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1603-1610, Raleigh, NC, 1992.

16. A. Ide, M. Tateda, H. Naruse, A. Nobiki and T. Yabuta, Automatic recognition and stereo correspondence of target objects in natural scenes, Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1597-1602, Raleigh, NC, 1992.
17. I.Y. Kim and H.S. Yang, Efficient image labelling based on markov random field and error backpropagation network, *Pattern Recognition*, vol. 26, no. 2, pp. 1695-1707, 1995.
18. K.S. Kumar and U.B. Desai, Joint segmentation and image interpretation, Proc. 3rd IEEE International Conference on Image Processing, vol. 1, pp. 853-856, 1996.
19. M.D. Lavine, A knowledge based computer vision system, in *Computer Vision Systems*, A.R. Hanson and E.M. Riseman (eds.), pp. 335-352, Academic Press, New York, 1978.
20. M.D. Lavine and S.I. Shaheen, A modular computer vision system for picture segmentation and interpretation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 3, no. 5, pp. 540-556, 1981.
21. W.K. Leow and R. Miikkulainen, Visual schemas in neural networks for object recognition, *Connection Science*, vol. 9, pp.161-200, 1997.
22. W. Mackeown, A labelled image database and its application to outdoor scene analysis, *PhD Thesis*, University of Bristol, UK, 1994.
23. A. Mecocci, R. Lodola and U. Salvatore, Outdoor scene interpretation for blind people navigation, Proc. 5th International Conference on Image Processing and its Applications, pp. 256-260, Edinburgh, UK, 1995.
24. A.M. Nazif and M.D. Lavine, Low level image segmentation: an expert system, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 5, pp. 555-577, 1984.
25. Y. Ohta, Region oriented image analysis system by computer, *PhD Thesis*, Kyoto University, March 1985.
26. P. Parodi and G. Piccioli, A feature based recognition scheme for traffic scenes, Proc. Intelligent Vehicles Symposium, pp. 229-234, Tokyo, Japan, 1995.
27. U. Regenberger and V. Graefe, Visual recognition of obstacles on roads, in Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 980-987, Munich, Germany, 1994.
28. M. Sharma, Performance evaluation of image segmentation and texture extraction methods in scene analysis, MPhil Thesis, University of Exeter, UK, 2000.
29. S.M. Smith and J.M. Brady, SUSAN - a new approach to low level image processing, *International Journal of Computer Vision*, vol. 23, no. 1, pp. 45-78, 1997.
30. M. Sonka, V. Hlavac and R. Boyle, *Image processing, analysis and machine vision*, PWS press, 1998.
31. T.M. Strat, *Natural object recognition*, Springer, Berlin, 1992.
32. Y. Taniguchi, Y. Shirai and M. Asada, Scene interpretation by fusing intermediate results of multiple visual sensory information processing, Proc. IEEE International Conference on Multisensor Fusion and Integration for Intelligence, pp. 699-706, Las Vegas, 1994.

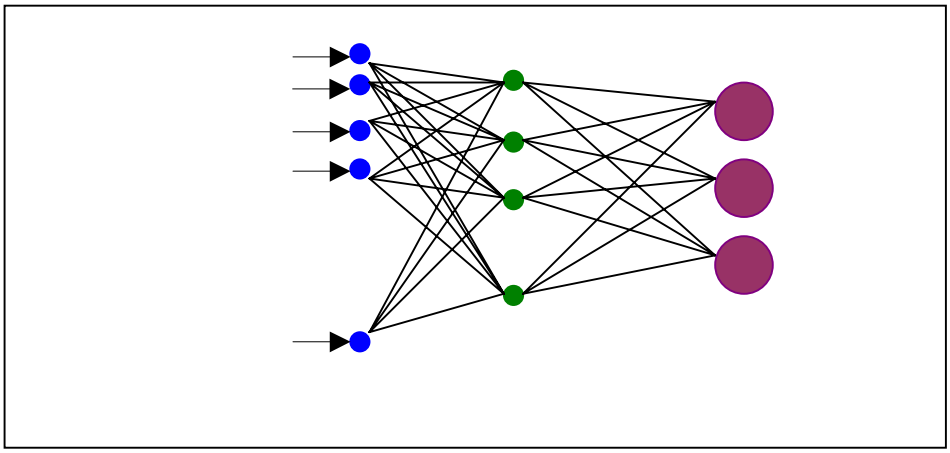


Figure 2. Neural networks for vegetation classification. ACF (autocorrelation function); CM (co-occurrence matrices), EF (edge frequency); LM (Law's measures); RL (run length)

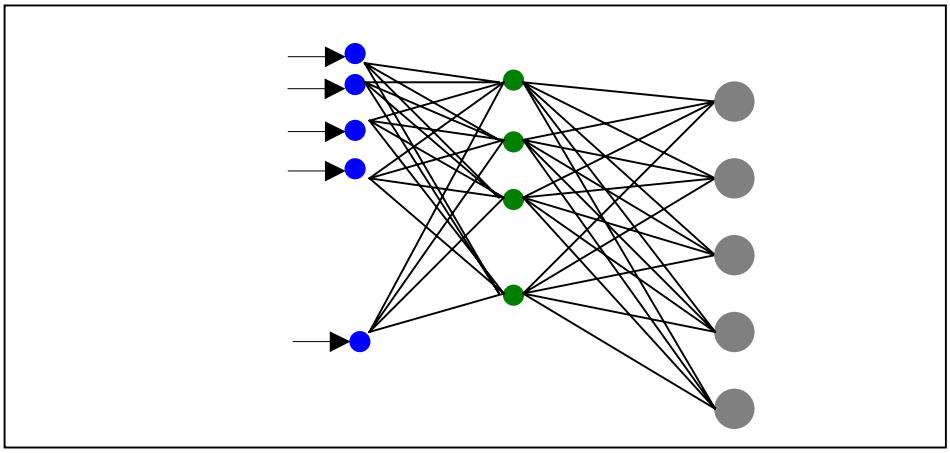


Figure 3. Neural networks for natural object classification. ACF (autocorrelation function); CM (co-occurrence matrices), EF (edge frequency); LM (Law's measures); RL (run length)

Feature extraction	Segmentation method			
	FCM	HT	RG	SM
Autocorrelation	861	739	383	943
Co-occurrence	822	729	346	937
Edge frequency	861	739	383	943
Laws	861	739	383	943
Run length	861	739	383	943

Table 1. The number of samples in feature data files for vegetation data

Feature extraction	Segmentation method			
	FCM	HT	RG	SM
Autocorrelation	951	1306	630	1542
Cooccurrence	828	1266	569	1520
Edge frequency	951	1306	630	1542
Laws	951	1306	630	1542
Run length	951	1306	630	1542

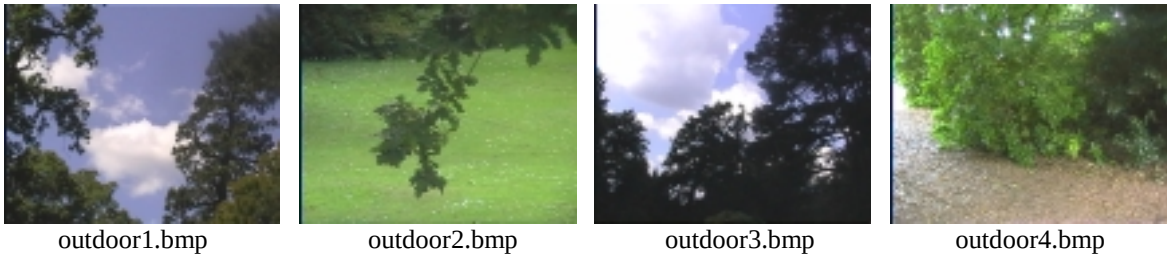
Table 2. The number of samples in feature data files for natural object data

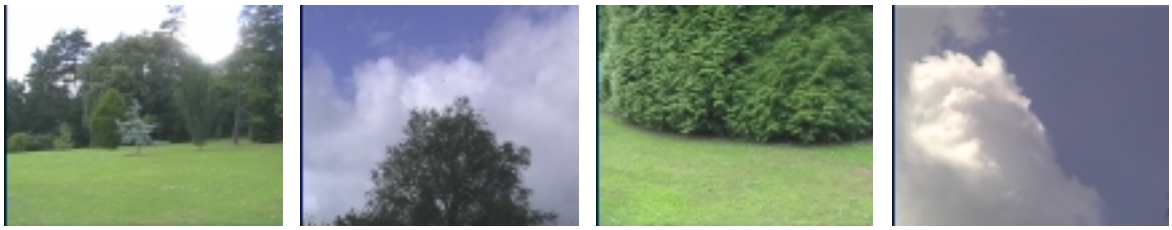
Features *Segmentation*	Auto-correlation	Co-occurrence	Edge frequency	Laws	Run length
FCM	71.3	54.4	73.1	65.5	52.5
Histogram	70.2	64.3	74.0	58.0	50.8
Region growing	61.4	53.4	68.3	58.8	59.6
Split and merge	64.3	69.1	78.7	71.8	61.9

Table 3. Neural network recognition rates in percentage as an average of 10 fold cross-validation for *vegetation* data classification

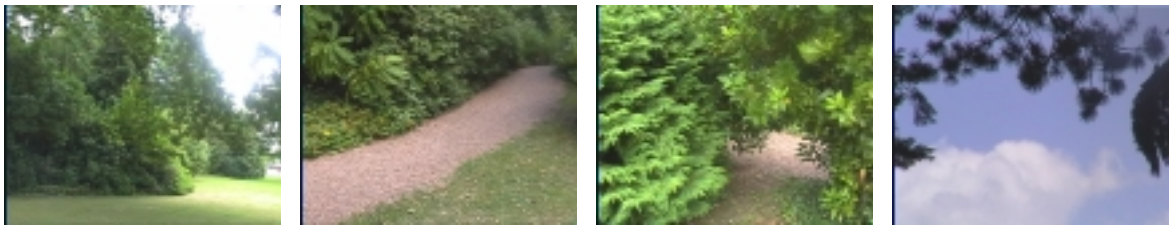
Features *Segmentation*	Auto-correlation	Co-occurrence	Edge frequency	Laws	Run length
FCM	68.4	51.6	71.5	56.5	54.2
Histogram	64.6	54.1	68.7	56.8	50.8
Region growing	67.2	51.8	67.7	58.0	55.6
Split and merge	67.7	55.6	72.3	53.7	56.8

Table 4. Neural network recognition rates in percentage as an average of 10 fold cross-validation for *natural object* data classification


outdoor1.bmp outdoor2.bmp outdoor3.bmp outdoor4.bmp


outdoor5.bmp outdoor6.bmp outdoor7.bmp outdoor8.bmp


outdoor9.bmp outdoor10.bmp outdoor11.bmp outdoor12.bmp


outdoor13.bmp outdoor14.bmp outdoor15.bmp outdoor16.bmp


outdoor17.bmp outdoor18.bmp outdoor19.bmp outdoor20.bmp


outdoor21.bmp outdoor22.bmp outdoor23.bmp outdoor24.bmp


outdoor25.bmp outdoor26.bmp outdoor27.bmp outdoor28.bmp

Figure 1. Sample images from the MINERVA benchmark