A MACHINE LEARNING APPROACH TO BUILDING RECOGNITION IN AERIAL PHOTOGRAPHS

C.J. Bellman[1,*] **, M.R. Shortis**[2]

[1]Department of Geospatial Science, RMIT University, Melbourne 3000, Australia - Chris.Bellman@rmit.edu.au
[2]Department of Geomatics, University of Melbourne, Parkville 3052, Australia - M.Shortis@unimelb.edu.au

Commission III, Working Group III/4

KEY WORDS: Building detection, Learning, Classification, Multiresolution

ABSTRACT:

Object recognition and extraction have been of considerable research interest in digital photogrammetry for many years. As a result, many conventional tasks have been successfully automated but, despite some advances, the automatic extraction of buildings remains an open research question. Machine learning techniques have received little attention from the photogrammetric community in their search for methods of object extraction. While these techniques cannot provide all the answers, they do offer some potential benefits in the early stages of visual processing. This paper presents the results of an investigation into the use of machine learning in the form of a support vector machine. The images are characterized using wavelet analysis to provide multi-resolution data for the machine learning phase.

1. INTRODUCTION

The advent of digital imagery has resulted in the automation of many traditional photogrammetric tasks.

However, the automatic extraction of man-made features such as building and roads is far from solved. These objects are attractive for automatic extraction, as they have distinct characteristics such as parallelism and orthogonality that can be used in the processing of symbolic image descriptions. Despite an extensive research effort, the problem remains poorly understood (Schenk, 2000).

Object extraction from digital images consists of two main tasks:

- identification of a feature, which involves image interpretation and feature classification and,
- tracking the feature precisely by determining its outline or centreline.

(Agouris *et. al.*, 1998)

Although many algorithms have been developed, none could claim to be fully automated. Most rely on some form of operator guidance to determine areas of interest or providing seed points on features.

This paper addresses the issue of determining areas of interest (candidate patches) using machine learning techniques.

Most photogrammetric applications for building recognition have followed the principle, established by Marr (Marr, 1982), that there are three levels of visual information processing. The first, low-level processing, involves the extraction of features in the image such as edges, points, and blobs that appear as some form of discontinuity in the image.

Intermediate-level processing involves the grouping and connection of these image primitives based on some measure of similarity or geometry. This forms the primal sketch (Marr, 1982) and is the basis for testing object hypotheses against rules that describe object characteristics. Many approaches are possible for establishing these rules such as semantic modelling (Stilla & Michaelsen, 1997), similarity measures (Henricsson, 1996), perceptual organisation (Sarkar & Boyer, 1993) or topology (Gruen & Dan, 1997).

High-level processing usually involves extracting information associated with an object that is not directly apparent in the image (Ullman, 1996,pg 4). This could be determining what the object is (recognition), or establishing its exact shape and size (reconstruction). In computer vision, recognition is the most common problem pursued. In photogrammetry, reconstruction of the geometry of features is more typically required.

1.1 Candidate regions

Despite the advances that have occurred in automated object extraction, most photogrammetric applications require some form of operator assistance to establish candidate image regions for potential object extraction. This is usually necessary to reduce the search space and make the problem tractable. Low-level processing strategies such as edge detection create a large number of artefacts that the mid-level grouping strategies find difficult to resolve.

This problem cannot be solved simply by segmentation, as this is difficult for an aerial image (Nevatia *et. al.*, 1998). An image contains many objects, only some of which should be modelled. The objects of interest may be partially occluded, poorly illuminated or have significant variations in texture.

In the case of building extraction, Henricsson (1996) solves the candidate problem in a pragmatic way. Rather than finding candidate regions using a computational process, the operator identifies candidate regions of the same building in multiple images. The computer system then extracts the edge features, groups these based on several measures of similarity and computes a 3-dimensional reconstruction of the building.

* Corresponding author

Gulch *et. al.* (1998) describe a Semi-automatic Building Extraction System that has undergone extensive development over a number of years. In this system, an operator interprets the image contents and automated tools assist the operator in the acquisition of 3-D shape data describing a building. In another system (Michel *et. al.*, 1998), the operator need only provide a seed point within the building roof-line. The building is then extracted automatically using a pair of epipolar images.

In some situations, spatial information systems can be used to provide existing semantic and positional data about objects in an image (Agouris *et. al.*, 1998). A set of fuzzy operators is used to select the relevant data and control the flow of information from image to spatial database. The system offers the potential of fully automatic updating of spatial database but the relies on the existence of the database in the first place. It does not use image data to determine regions of interest.

The use of auxiliary data such as digital surface models (Zimmermann, 2000), multi-sensor and multi-spectral data (Schenk, 2000), provides another means of determining regions of interest in an image but issues of data fusion add complexity to the task.

There is much evidence from cognitive science that human processes for shape recognition are both rapid and approximate in many cases. Intuitively, this suggests that complicated and lengthy visual processing strategies are not complete models of our biological vision, particularly in the early stages of visual processing.

2. A MACHINE LEARNING APPROACH

Machine learning approaches, such as those based on neural networks and support vector machines, are popular strategies for image analysis and object recognition in many imaging applications (Osuna *et. al.*, 1997; Li *et. al.*, 1998). In photogrammetry, machine learning techniques have been applied to road extraction (Sing and Sowmya 1998), knowledge acquisition for building extraction (Englert 1998) and for landuse classification (Sester 1992). Neural techniques have been used in feature extraction (Li et al. 1998, Zhang 1996), stereo matching (Loung and Tan 1992) and image classification (Israel and Kasabov 1997).

The recognition task is generally treated as a problem of classification, with the correct classifications being learnt on the basis of a number training examples. Where the images are small (i.e. have few pixels), a direct connection approach is employed, where each image pixel is directly connected to a node in the connectionist architecture. For typical aerial digital imagery, such an approach is not feasible due to the combinatorial explosion that would result. Some preprocessing stage is required to extract key characteristics from the image domain. Many of the strategies for preprocessing are available, such as edge detection (Canny, 1986), log-polar-forms (Grossberg, 1988) and texture segmentation (Lee & Schenk, 1998).

Wavelet analysis is often associated with image compression (Rabbani & Joshi, 2002) but also has useful properties for the characterization of images. Of particular interest are the multi-resolution representations that can be generated (Mallat, 1989). Such an approach has been used successfully in system to recognize the presence of a pedestrian in a video image (Papageorgiou *et. al.*, 1998); (Poggio & Shelton, 1999) and for face recognition (Osuna *et. al.*, 1997). There are strong suggestions from psycho-physical experiments that mammalian vision systems incorporate many of the characteristics of wavelet transforms (Field, 1994).

2.1 Wavelet Processing

Wavelet processing allows a signal to be described by its overall shape plus a range of details from coarse to fine (Stollnitz *et. al.*, 1995). In the case of image data, wavelets provide an elegant means of describing the image content at varying levels of resolution.

The Haar wavelet is the simplest of the wavelet functions. It is a step function in the range of 0-1 where the wavelet function $\psi(x)$ is expressed as:

$$\psi(x) := \begin{cases} 1 & \text{for } 0 \le x < 1/2 \\ -1 & \text{for } 1/2 \le x < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The wavelet transform is computed by recursively averaging and differencing the wavelet coefficients at each resolution. An excellent practical illustration of the use of wavelets is provided by Stollnitz *et. al.*(1995).

As a discrete wavelet transform (DWT), the Haar basis does not produce a dense representation of the image and is not sufficiently sensitive to translations of the image content. An extension of the Haar wavelet can be applied that introduces a quadruple density transform (Papageorgiou *et. al.*, 1998; Poggio & Shelton, 1999). In a conventional application of the discrete wavelet transform, the width of the support for the wavelet at level n is 2^n and adjacent wavelets are separated by this distance. In the quadruple density transform, this separation is reduced to ¼ 2^n (Figure 1(c)). This oversamples the image to create a rich set of basis functions that can be used to define object patterns. An efficient method of computing the transform is given in Oren *et. al.*, (1999).

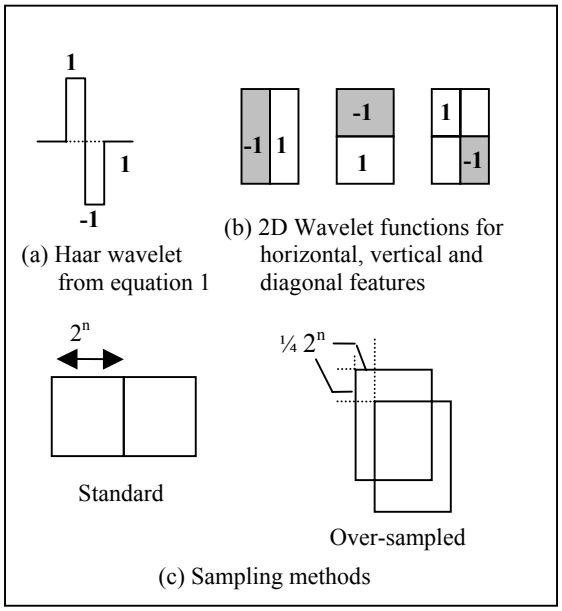
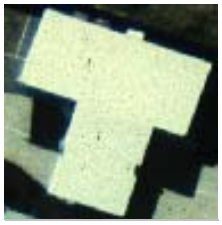
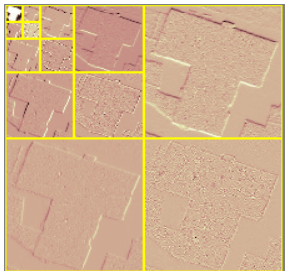


Figure 1: The Haar wavelet characteristics (after (Papageorgiou *et. al.*, 1998)).



(a) Building image patch



(b) Wavelet decomposition to level 4 of building patch in (a)

Figure 2. (a) Sample image and (b) corresponding wavelet representation (not over-sampled)

3. SUPPORT VECTOR MACHINES

The Support Vector Machine (SVM) is based on the principles of structural risk minimization (Vapnik, 1995). It has the attractive property that it minimizes a bound on the generalisation error and is therefore not subject to problems of local minima that may occur with other classifiers such as multilayer perceptrons (MLP).

Another property of the SVM is that its decision surface depends only on the inner product of the feature vectors. As a result, the inner product can be replaced by any symmetric positive-definite kernel (Cristianini & Shawe-Taylor, 2000). The use of a kernel function means that the mapping of the data into a higher dimensional feature space does not need to be determined as part of the solution, enabling the use of high dimensional space without addressing the mathematical complexity of such spaces. SVM's have been used successfully in applications for face detection (Osuna *et. al.*, 1997), character recognition (Schölkopf, 1997; Boser *et. al.*, 1992) and pedestrian detection (Papageorgiou *et. al.*, 1998).

4. TEST DATA

A set of classification test data, in the form of square image patches, was extracted from three colour digitized aerial photographs. The photographs were originally acquired for a project over the city of Ballarat in Victoria. The photographs were recorded at a scale of 1:4000 and had been scanned on a photogrammetric scanner at a resolution of 15 microns. Each image patch was 256 by 256 pixels and contained either a single building or a non-building area of the image. Although some care was taken to centre the building within the image patch, the exact location of the building in the image patch varied. The orientation of the building within the image patches also varied. This lead to a broader representation of the building class than if the buildings were carefully aligned in each image patch but created a more difficult classification problem.

The classification test was based on a balanced test set of 100 building images and 100 non-building images. Image coefficients were extracted using the wavelet process described in 2.1 above. A public domain Support Vector Machine (Joachims, 1998) was used to classify the image patches into building or non-building categories.

5. RESULTS

The image patches used to train the SVM classifier using several different kernels including polynomial and sigmoidal kernels. However, the best results were obtained with a simple linear kernel with no bias. Of the two hundred image patches, only one patch was classified incorrectly. This was an image of a large swimming pool that was classified as a building. Although this result appeared to be very good, the confidence measures produced by the SVM training suggested a reliability of only 55%. This could be due to overfitting of the decision surface to the data. However, the reliability measures produced by the SVM are also known to be pessimistic (Joachims, 1998), due to the unbounded nature of the problem. To see if that was the case here, an extensive leave-one-out test was undertaken. This produced a revised reliability measure of 73%. Although the reliability estimate improved, this indicates there may still be some overfitting of the data. To test this further, 20 building image patches were withheld from the training data and the SVM was re-trained. The withheld patches were then classified by the new SVM. Of the 20 building patches, only 8 were classified as buildings. This result is similar to the original reliability estimate. In this case, the decision surface of the SVM is unlikely to generalize well to a broader set of data. This could be due to the small size of the training set. Further work is required to expand the size and scope of the training set to determine if a more generalized decision surface can be established.

5.1 Other Considerations

In applying multi-resolution analysis, an appropriate set of resolutions must be chosen for the task at hand. In this case, a choice must be made between minimizing the amount of data that is fed to the classifier and retaining enough information about the original image that a sensible classification is possible. In this example, wavelets with supports of 16 and 32 pixels were used, resulting in image coefficients of 16 x 16 and 8 x 8 for each of the horizontal, vertical and diagonal wavelet functions.

Other resolutions were tried but at higher resolutions, the number of coefficients expands rapidly and there was no significant gain in the accuracy of the classification. At lower resolutions, too much information about the image was lost to enable definite classification.

One limitation of this implementation is the size of the coefficient data set produced from the wavelet transform. As the wavelet transform is over-sampled, each image patch generates 960 coefficients. The current algorithm makes no attempt to optimize the storage of these coefficients.

6. CONCLUSION

Machine learning methods have been used successfully in several image processing and machine vision domains but there has been little research into their potential for photogrammetric applications. While these techniques often cannot satisfy the metric requirements of photogrammetry, they can provide useful starting points and heuristic filters in the area of automated object extraction.

The Support Vector Machine is well suited to this application, as it does not suffer from the problem of local minima and produces a statistically robust decision surface. The SVM recasts the problem into high dimensional feature space, where problems that are not linearly separable in lower-dimensional feature space may become separable.

An important aspect of machine learning in vision applications is to extract a representative set of characteristics from the image. The multi-resolution approach of wavelets does this quite nicely and is well supported by psycho-physical evidence suggesting that mammalian vision systems operate in a similar manner.

Although some refinement and further testing are required, the machine learning approach outlined in this paper could be used to identify image patches that are likely to contain a building. As such, it would act as a heuristic filter, providing only image patches that had a high probability of containing a building to the functions that perform the building extraction processes.

7. REFERENCES

Agouris, P., Gyftakis, S. & Stefanidis, A., 1998. Using A Fuzzy Supervisor for Object Extraction within an Integrated Geospatial Environment. In: *International Archives of Photogrammetry and Remote Sensing,* Columbus, Ohio, Vol. XXXII, Part III/1, pp. 191-195.

Boser, B. E., Guyon, I. M. & Vapnik, V. N., 1992. A training algorithm for optimal margin classifiers. In: The 5th Annual ACM Workshop on Computational Learning Theory, Pittsburgh, ACM Press.

Canny, J. F., 1986. A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6): pp. 679-686.

Cristianini, N. & Shawe-Taylor, J., 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge, UK, Cambridge University Press.

Field, D., 1994. What is the Goal of Sensory Coding? *Neural Computation* 6(4), pp. 559-601.

Grossberg, S., 1988. Nonlinear Neural Networks. Principles, Mechanisms, And Architectures. *Neural Networks* 1: pp. 17-61.

Gruen, A. & Dan, H., 1997. TOBAGO- a topology builder for the automated generation of building models. In : *Automatic Extraction of Man-Made Objects from Aerial and Space Images*. Eds. A. Gruen, Baltsavias, E.P. & Henricsson, O. Basel, Switzerland, Birkhauser,: 393.

Henricsson, O.,1996. Analysis of image structures using color attributes and similarity relations. Unpublished PhD Thesis, Department of Geodesy and Photogrammetry. Zurich, Swish Federal Institute of Technology: pp. 124.

Israel, S. and Kasabov, N., 1997. Statistical, connectionist and fuzzy inference techniques for image classification. Journal of Electronic Imaging, 6(3): 1-11.

Joachims, T., 1998. Making Large-Scale SVM Learning Practical. In: *Advances in Kernel Methods - Support Vector Learning*. Eds. B. Scholkopf, Burges, C.J. & Smola, A,J. Cambridge, USA, MIT Press.

Lee, D. & Schenk, T., 1998. An Adaptive Approach for Extracting Texture Information and Segmentation. In: *International Archives of Photogrammetry and Remote Sensing,*Columbus, Ohio, Vol. XXXII, Part III/1.

Li, R., Wang, W. & Tseng, H.-Z., 1998. Object Recognition and Measurement from Mobile Mapping Image Sequences using Hopfield Neural Networks: Part 1. ASPRS Annual Conference, Tampa, Florida, USA, American Society of Photogrammetry and Remote Sensing.

Loung, G. and Tan, Z., 1992. Stereo matching using artificial neural networks. International Archives of Photogrammetry and Remote Sensing, XXIX(B3/III): 417-422.

Mallat, S. G. 1989., A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* II(7): pp. 674-693.

Marr, D., 1982. *Vision : A computational investigation into the human representation and processing of visual information*. New York, W.H. Freeman and Company.

Michel, A., Oriot, H. & Goretta, O., 1998. Extraction of Rectangular Roofs on Stereoscopic Images - An Interactive Approach. In: *International Archives of Photogrammetry and Remote Sensing,*Columbus, Ohio, Vol. XXXII, Part III/1.

Nevatia, R., Huertas, A. & Kim, Z., 1998. The MURI Project for Rapid Feature Extraction in Urban Areas. In: *International Archives of Photogrammetry and Remote Sensing,*Columbus, Ohio, Vol. XXXII, Part III/1.

Oren, M., Papageorgiou, C., Sinha, P., Osuna, E. & Poggio, T., 1997. Pedestrian Detection Using Wavelet Templates. In: Proceedings of Computer Vision and Pattern Recognition, Puerto Rico.

Osuna, E., Freund, R. & Girosi, F., 1997. Training Support Vector Machines: An Application to Face Detection. In: Proceedings IEEE Conference on Computer Vision and Pattern Recognition, San Juan, pp. 130-136.

Papageorgiou, C. P., Evgeniou, T. & Poggio, T., 1998. A Trainable Pedestrian Detection System. In: Intelligent Vehicles, Stuttgart, Germany.

Papageorgiou, C. P., Oren, M. & Poggio, T., 1998. A General Framework for Object Detection. International Conference on Computer Vision, Bombay, India.

Poggio, T. & Shelton, C. R.,1999. Machine Learning, Machine Vision, and the Brain. *AI Magazine* 20: pp. 37-55.

Rabbani, M. & Joshi, R., 2002. An overview of the JPEG 2000 still image compression standard. *Signal Processing: Image Communication* 17: pp. 3-48.

Sarkar, S. & Boyer, K., 1993. Perceptual Organisation in Computer Vision: A Review and a Proposal for a Classificatory Structure. *IEEE Transactions on Systems, Man and Cybernetics* 23(2): pp. 382-399.

Schenk, T., 2000. Object Recognition in Digital Photogrammetry. *The Photogrammetric Record* XVI(95): pp. 743-759.

Schölkopf, B., 1997. Support Vector Learning, Unpublished PhD Thesis, TU Berlin, Berlin pp. 173.

Sester, M., 1992. Automatic model acquisition by learning, International Archives of Photogrammetry and Remote Sensing, XXIX(B3/III) : 856-863.

Sing, S. and Sowmya, A., 1998. Rail: Road recognition from aerial images using inductive learning. International Archives of Photogrammetry and Remote Sensing, XXXII(3/1): 367-378.

Stilla, U. & Michaelsen, E., 1997. Semantic Modelling of Man-Made Objects by Production Nets. In: *Automatic Extraction of Man-Made Objects from Aerial and Space Images*. Eds. Gruen A., Baltsavias, E.P. & Henricsson, O., Basel, Switzerland, Birkhauser, pp. 393.

Stollnitz, E. J., DeRose, T.D. & Salesin, D.H., 1995, Wavelets for Computer Graphics: A Primer (Part 1). *IEEE Computer Graphics and Applications*, 15(3), pp.76-84.

Ullman, S., 1996. *High-Level Vision*. Cambridge, Massachusetts, Massachusetts Institute of Technology.

Vapnik, V., 1995. *The Nature of Statistical Learning Theory*. New York, Springer Verlag.

Zhang, Y.-S., 1996. A hierarchical neural network approach to three-dimensional object recognition. International Archives of Photogrammetry and Remote Sensing, XXXI(B3): 1010-1017.

Zimmermann, P., 2000. A New Framework for Building Detection Analysing Multiple Cue Data. *International Archives of Photogrammetry and Remote Sensing*, Vol. XXXIII, Part B3: pp.1063-1070.