EARLY STAGE OBJECT RECOGNITION USING NEURAL NETWORKS

Chris Bellman
Department of Land Information
Royal Melbourne Institute of Technology
Melbourne 3000
Australia

Chris.Bellman@rmit.edu.au

Assoc. Prof. Mark Shortis
Department of Geomatics
University of Melbourne
Parkville 3052
Australia

M.Shortis@unimelb.edu.au

Working Group III/4

KEY WORDS: Object recognition, Neural Networks, Wavelets, Support vector machines

ABSTRACT

Object recognition has been the focus of much research in the photogrammetric and image processing communities. The goal of this research is often the quantitative three-dimensional measurement of objects found in two-dimensional image space. Although a multitude of different approaches to the automation of this problem have been developed and thoroughly tested, few could claim that the process is more than semi-automated even in cases that involve a specialised application.

One reason for this could be the reliance of these techniques on recognition-by-reconstruction. In this approach, images are processed to extract edges or homogeneous regions. These edges are combined using geometric and/or perceptual rules to complete the object description. In some cases, the edges are matched to models of generic objects. The recognition of the object occurs as a result of the reconstruction phase. This suggests a cognitive approach to vision, where the recognition task is largely performed as a cognitive rather than a visual process.

This paper reports on an investigation into the use of neural network approaches for the initial recognition of objects within images. This research considers the initial identification of the objects as a visual rather than cognitive process. It is analogous to the classification problem in image processing and requires that characteristic image signatures are identified for particular object classes. The research focuses on the identification of image patches containing buildings as the first stage in the reconstruction of the building geometry. The general applicability of the method is still to be determined but the initial results are promising.

1 INTRODUCTION

Photogrammetry has long been concerned with the extraction of object dimensions from imagery. The advent of digital imagery has produced many techniques and algorithms for automating this task. Despite an extensive research effort, no dominant method or algorithm has emerged (Agouris et.al., 1998).

Of particular interest in digital photogrammetric applications has been the extraction of man-made features such as building and roads. 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.

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)

Despite the success of many of the algorithms developed, few if any, could claim to be fully automated. Most rely on some form of operator guidance in determining areas of interest or providing seed points on features.

Several trends emerge from previous research in computer vision that are relevant to photogrammetry. Early research into vision tried to solve the vision problem in total. That is, researchers worked towards a single, unifying theory of vision that could be replicated in a computer. Marr (Marr, 1982) proposed a representational framework for deriving shape information from images as a model of the human visual system.

Many photogrammetric applications have generally followed the view 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. A good example of processing at this level is the Canny edge detector (Canny, 1986). This is a linear detector that uses local maxima in the intensity gradient at each position in the image to identify edges. Many other feature detectors have been developed for edges, texture and homogeneous regions.

Intermediate-level processing involves the grouping and connection of these image primitives based on some measure of similarity or geometry. This forms what Marr calls 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. Most low level processing strategies create a large number of artefacts that the mid-level grouping strategies find difficult to resolve.

The problem can not be solved simply by segmentation, as this is difficult for an aerial image (Nevatia et.al., 1999). The image contains many objects, only some 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 trying to find 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.

Gulch et.al. (1998) describe a Semiautomatic 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.

Another approach to establishing candidate regions is the use of spatial information systems 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.

1.2 Other Processing strategies.

Another view of visual information processing is that a combination of symbolic and connectionist systems is required (Minsky, 1990). There are many potential advantages in combining the expressiveness and procedural versatility of symbolic systems with the fuzziness and adaptability of connectionist systems.

Recent work in image understanding (Draper, 1993;Draper et.al., 1999) suggests that biological vision consists of a number of visual processing strategies, each solving a visual sub-task. These processing strategies vary according the nature and source of the image and must be combined in the correct order to achieve the goal of many vision systems – object recognition.

The ADORE system (Draper et.al., 1999) uses symbolic and algorithmic approaches to undertake visual tasks and uses neural networks to learn control strategies for sequencing the visual processing procedures. As such, it combines symbolic and connectionist processing in the one system but the connectionist processing is used for decision making in the control sequencing rather than directly in the visual processing tasks.

The use of neural networks for visual processing tasks has largely been confined to low-level visual processes such as classification, edge detection, pattern recognition and segmentation. There has been little work done that tries to embody higher level visual processes in a neural network processing architecture, perhaps because others strategies have proven too successful or perhaps because it is perceived as impossible or irrelevant.

There is much evidence 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 probably not correct models of our biological vision. While this may not be significant in the quest for automated object extraction processes for aerial imagery, it is possible that neural processing strategies will offer some benefits in the early stages of object extraction.

2 A NEURAL NETWORK APPROACH

The high resolution of digital aerial imagery makes the use of connectionist approaches more problematic than in other image domains. A direct connection model, where each pixel is connected in the neural architecture, cannot be employed 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 used in other domains are also not suitable. The use of log-polar-forms (Grossberg et.al., 1988) is effective with binarised radar images but is difficult to apply to the grey scale images typical in photogrammetry.

One approach that offers some promise is the use of wavelets to characterise images (Papageorgiou et.al., 1998; Poggio et.al., 1999). In a pedestrian detection system, a wavelet function is used to extract significant image details at various image resolutions from coarse to fine. The wavelet coefficients represent a skeleton of the significant features in the image space. A neural network based around support vector machines is then used to train the network to recognise these image characteristics and associate them with a feature class in the image. The method does not extract the object in question, it merely recognises the presence or absence of the object. This technique has some appeal for the selection of regions of interest in digital images that contain buildings.

2.1 Wavelet Processing

Wavelet processing allows a function 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:

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

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 given in Stollnitz et. al.(1995).

Papageorgiou et.al. (1998) and Poggio et.al. (1999) use an extension of the Haar wavelet that introduces a quadruple density transform. In a conventional application of the 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 effectively 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., 1997).

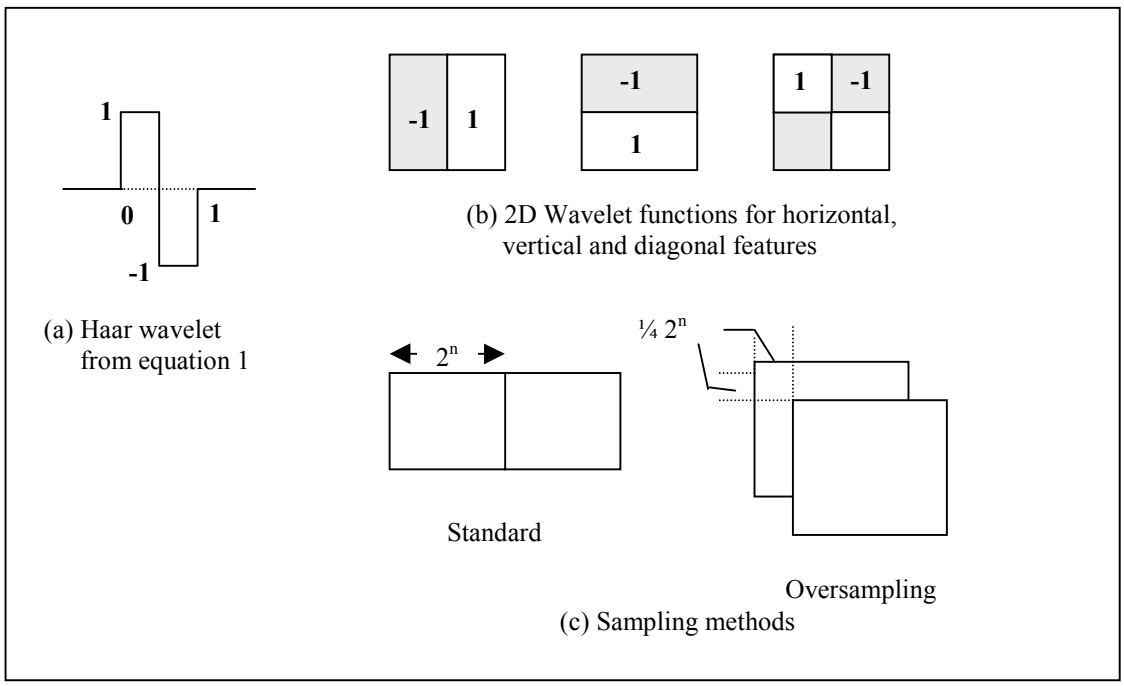

Figure 1: The Haar wavelet characteristics (after Papageorgiou, 1998).

3 PRELIMNARY RESULTS

A full implementation of the approach described above is yet to be applied to the detection of buildings in aerial images. Preliminary work has focussed on applying the wavelet transform to some typical aerial imagery. The Avenches data set (Mason et.al., 1994) was used to test the implementation. Figure 2 shows an example of a building patch from this data set and the resultant wavelet image. The wavelet transform is generated for successively coarser levels of the image (level 1 : 2x2; level 2 : 4x4; level 3 : 8x8) and is not oversampled. The wavelet coefficients at each scale are shown as

grey scale components of the image for each level. These are computed by applying an offset of 128 to the calculated coefficients to ensure that negative coefficient values can be shown as grey scale values in the image.

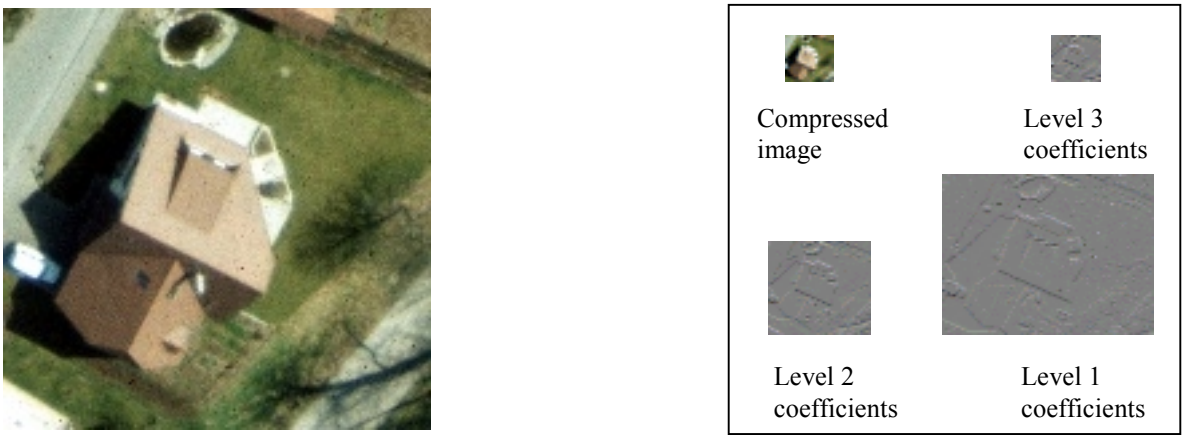

Figure 2. (a) An image patch containing a single building, (b) the resultant Wavelet coefficient images at each level.

When processing a colour image, each colour band generates its own set of wavelet coefficients. These can be combined in a variety of ways. In figure 2, the coefficients are simply combined to form a new RGB image. In figure 3, only the maximum coefficient value for each pixel is retained in the coefficient image. The coefficients can also be normalised over a specified range to reduce the effect of varying image intensities (Papageorgiou et.al., 1998).

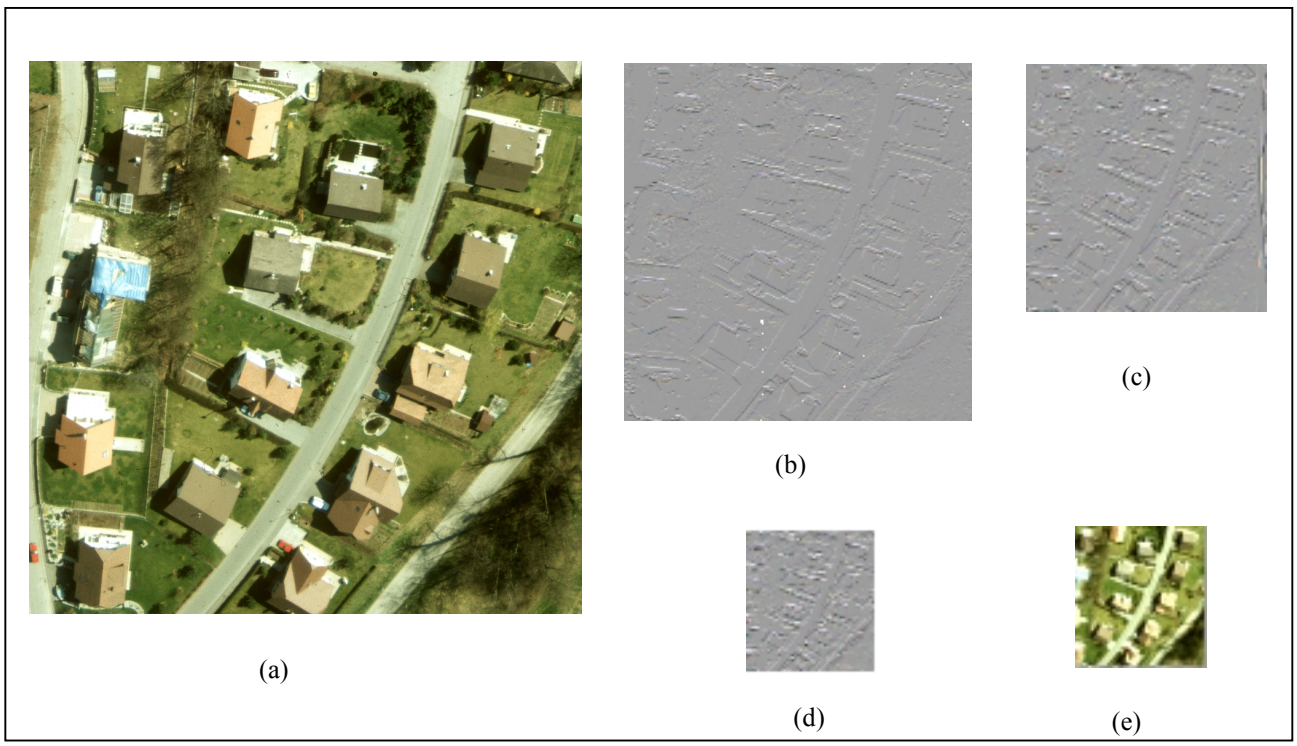

Figure 3. The wavelet transform applied to a larger image patch (a) of the Avenches dataset. The resultant coefficient images are (b) level 2, (c) level 3 and (d) level 4. (e) shows the resultant compressed image.

The coefficient images show that considerable detail can be extracted from an image using the wavelet coefficients. The next stage of processing will involve developing an oversampling technique and building a large library of images and

coefficient data sets that can be used to train the learning system. Although many connectionist systems could be used for the training, the Support Vector Machine (Vapnik, 1995) is capable of learning quickly in high-dimensional spaces and has been applied successfully to several other image domains (Poggio et.al., 1999).

The rationale for using wavelets to characterise images is the reduction of the data dimensionality without significant loss of information. Using the coefficient images as an initial classifier will allow candidate regions to be selected automatically for further processing and analysis by more computationally demanding solutions.

4 CONCLUSION

The use of wavelet transforms to extract significant characteristics of image features appears to offer some promise. In conjunction with a neural network based classifier, the technique has been applied successfully in the image domains of pedestrian, face and vehicle detection. Its suitability for building detection has yet to be shown but initial investigations suggest that the wavelet transform can extract characteristics of buildings. Employing these characteristics to train a detection system must be explored further before the usefulness of this technique can be fully evaluated. This will be the focus of future work.

ACKNOWLEDGEMENTS

The support of Dr. Stuart Robson is gratefully acknowledged for allowing us to use the IVW software for reading and displaying images.

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, Vol. XXXII, Part III/1, Columbus, USA. pp. 191-195.

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

Draper, B. A., 1993, Learning Object Recognition Strategies. Unpublished PhD Thesis, Department of Computer Science, University of Massachusetts, pp. 152.

Draper, B. A., Bins, J. & Baek, K. ,1999, ADORE: Adaptive Object Recognition. In :Proceedings of International Conference on Vision Systems (ICVS 99), Las Palmas de Gran Canaria, Spain.

Grossberg, S., 1988, Nonlinear Neural Networks: Principles, Mechanisms, And Architectures. Neural Networks, 1, 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 (Ed., Gruen, A., Baltsavias, E.P. & Henricsson, O.) Birkhauser, Basel, Switzerland, pp. 393.

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

Marr, D., 1982, Vision : A computational investigation into the human representation and processing of visual information. W.H. Freeman and Company, New York, pp. 397.

Mason, S., Baltsavias, M., & Stallmann, D., 1994, High Precision Photogrammetric Data Set for Building Reconstruction and Terrain Modelling. Institute for Geodesy and Photogrammetry, Swiss Federal Institute of Technology (ETH), Zurich, Switzerland.

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, Vol. XXXII, Part III/1, Columbus, USA.

Minsky, M., 1990, Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy. In: Artificial Intelligence at MIT, Expanding Frontiers. Vol. 1 (Ed, Winston, P. H.) MIT Press, Massachusetts.

Nevatia, R., Huertas, A., Kim, Z., 1999, The MURI Project for Rapid Feature Extraction in Urban Areas. In: International Archives of Photogrammetry and Remote Sensing, Vol. XXXII, Part III/2&3, Munich.

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

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

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

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, pp. 382-399.

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 (Ed., Gruen, A., Baltsavias, E.P. & Henricsson, O.) Birkhauser, Basel, Switzerland, 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, pp.76-84.

Ullman, S., 1996, High-level vision : object recognition and visual cognition. Massachusetts Institute of Technology, Cambridge,Massachusetts, pp. 412.

Vapnik, V., 1995, The Nature of Statistical Learning Theory. Springer Verlag, New York, pp. 193.