VICTORIA UNIVERSITY OF WELLINGTON

Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences

Computer Science

Pixel Based Neural Networks for Multiclass Object Detection

Mengjie Zhang

Technical Report CS-TR-01/13
September 2001

School of Mathematical and Computing Sciences
Victoria University
PO Box 600, Wellington
New Zealand

Tel: +64 4 463 5341
Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences
Computer Science

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341, Fax: +64 4 463 5045
Email: Tech.Reports@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

Pixel Based Neural Networks for Multiclass Object Detection

Mengjie Zhang

Technical Report CS-TR-01/13
September 2001

Abstract

We describe a pixel based approach to the use of neural networks for multiclass object detection problems in which the classes and locations of relatively small objects in large pictures must be determined. The networks use a squared input field which is large enough to contain every single object of interest and are trained by the back propagation algorithm on examples which have been cut out from the large pictures. The trained networks are then applied, in a moving window fashion, over the large pictures to detect the objects of interest. This approach has been examined on three object detection problems of increasing difficulty. The results suggest that this approach can be used to detect simple and regular objects with the translation and rotation invariance in large pictures against a relatively uniform background.

Keywords Object recognition, object classification, target detection, multiclass target detection, pixel based neural networks.

Author Information

The author is a lecturer in computer science, Victoria University of Wellington, New Zealand.

1 Introduction

As more and more images are captured in electronic form the need for programs which can find objects of interest in a database of images is increasing. For example, it may be necessary to find all tumors in a database of x-ray images, all cyclones in a database of satellite images or a particular face in a database of photographs. The common characteristic of such problems can be phrased as "Given $subpicture_1, subpicture_2 ... subpicture_n$ which are examples of the object of interest, find all pictures which contain this object and its location(s)". Figure 1 shows examples of problems of this kind. In the problem illustrated by Figure 1 (b), we want to find centers of all of the 5 cent and 20 cent coins and determine whether the head or the tail side is up. Examples of other problems of this kind include target detection problems [4, 13, 15] where the task is to find, say, all tanks, trucks or helicopters in a picture. Unlike most of the current work in the object detection area, where the task is to detect only objects of one class [4, 9, 10], our objective is to detect objects from a number of classes.

Several approaches have been applied to automatic object detection and recognition problems. Typically, they use multiple independent stages, such as preprocessing, edge detection, segmentation, feature extraction and object classification[2, 8], each of which uses one or more programs. The final results rely too much upon the results of each stage. If some objects are lost in one of the early stages, it would be very difficult or impossible to recover them in the later stage. To avoid these disadvantages, this paper introduces an approach with a single trained neural network for the multiclass object detection problems.

Two main kinds of approaches have been used in neural network systems for object classification/recognition problems – feature based and pixel based. In feature based approaches various features such as brightness, colour, size and perimeter are extracted from the sub-images of the objects of interest and used as inputs to the networks, as in [2, 8, 9, 14]. These features are usually different and specific for different problem domains. In pixel based approaches [3, 6, 12] the pixel values are used directly as inputs. To achieve the goal of domain independent object detection, this paper uses pixel based approach for multiclass object detection problems.

1.1 Multiclass Object Detection

Object detection here refers to the detection of small objects in large pictures. It consists of both object classification, which determines the classes of the objects of interest, and object localisation, which identifies the positions of all the objects in the large pictures.

In automatic object detection systems, in most cases, all the objects of interest are considered a *single class*. These are actually the *object* vs *non-object* or *target* vs *non-target* or *object* vs *background* detection problems. In contrast, *multiclass object detection* refers to the case where there is more than one class of objects and both their classes and locations must be determined. In general, multiclass object detection problems are much harder than single class detection problems, and multiclass detection using a single trained program (such as a neural network) is an even more difficult problem.

1.2 Goals

The overall goal is to develop a domain independent approach to multiclass object detection problems and to determine the strengths and limitations of the technique. To achieve this goal, a sequence of multiclass object detection problems of increasing difficulty are applied. Specifically, we investigate

- Will this approach work for detecting simple objects in large pictures with a uniform background?
- Can it be used for regular object detection in large pictures with a relatively uniform background?
- Is it possible to detect complex objects in large pictures with a highly cluttered background using this approach?
- Will the performance deteriorate as the degree of difficulty of the detection problems increases?

1.3 Performance Evaluation

Performance in object detection is measured by detection rate and false alarm rate. The detection rate is the number of objects correctly reported as a percentage of the total number of real objects and false alarm rate is the number of objects incorrectly reported as a percentage of the total number of real objects. For example, a detection system looking for grey squares in Figure 1 (a) may report that there are 25. If 9 of these are correct the detection rate will be $(9/18) * 100 = 50\%$. The false alarm rate will be $(16/18) * 100 = 88.9\%$. It is important to note that detecting objects in pictures with very cluttered backgrounds is an extremely difficult problem and that false detection rates of 200-2,000% (that is the detection system suggests that there are 20 times as many objects as there really are) are common [10, 12]. Also note that most research which has been done in this area so far only presents the results of the classification stage and assumes that all other stages have been properly done. However, the results presented in this paper are the performance for the whole detection problem (including the localisation and classification). Thus, it is unfair to simply compare the results in this paper with those obtained by other systems which are developed only for the classification problem. It is not a fair comparison either if one compares the performance achieved in this paper with that only for one class detection problems.

2 The Image Databases

We used three different databases in the experiments. Example pictures and key characteristics are given in Figure 1. The pictures were selected to provide problems of increasing difficulty. Database 1 (Easy) was generated to give well defined objects against a uniform background. The pixels of the objects were generated using a Gaussian generator with different means and variances for each class. There are three classes of small objects of interest in this database: black circles ($class1$), grey squares ($class2$) and white circles ($class3$). The coin pictures (database 2) were intended to be somewhat harder and were taken with a CCD camera over a number of days with relatively similar illumination. In these pictures the background varies slightly in different areas of the image and between images and the objects to be detected are more complex, but still regular. There are four object classes of interest: the head side of 5 cent coins (class $head005$), the head side of 20 cent coins (class $head020$), the tail side of 5 cent coins (class $tail005$) and the tail side of 20 cent coins (class $tail020$). All the objects in each class have a similar size. They are located at arbitrary positions and with different rotations. The retina pictures (database 3) were taken by a professional photographer with special apparatus at a clinic and contain very irregular objects on a very cluttered background. The objective is to find two classes of retinal pathologies – haemorrhages (class $haem$) and micro aneurisms (class $micro$). To give a clear view of representative samples of the target objects in the retina pictures, one sample piece of these pictures is presented in

Number of Images: 7	Number of Images: 20	Number of Images: 15
No. of Classes: 4	No. of Classes: 5	No. of Classes: 5
Input Field Size: 14x14	Input Field Size: 14x14	Input Field Size: 14x14
Number of Objects: 240	Number of Objects: 600	Number of Objects: 164
Picture Size: 700x700	Picture Size: 640x480	Picture Size: 1024x1024
Easy (Circles and Squares)	Medium Difficulty (Coins)	Very Difficult (Retinas)
(a)	(b)	(c)

Figure 1: Object Detection Problems of Increasing Difficulty

Figure 2. In this figure, haemorrhage and micro-aneurism examples are labeled using white surrounding squares. Note that in each of the databases the background counts as a class.

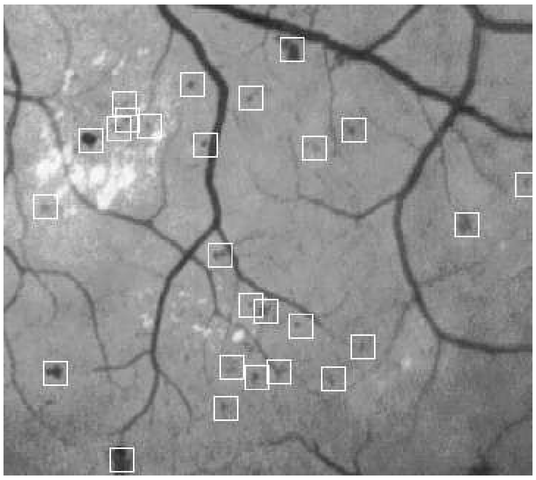


Figure 2: An enlarged view of one piece of the retina pictures

3 The Method

3.1 Overview

The data sets used in this paper and their relationships are illustrated in Figure 3 (a). An overview of the approach is shown in Figure 3 (b). A brief outline of the method is as follows:

1. Assemble a database of pictures in which the locations and classes of all the objects of interest are manually determined. Divide these full images into two sets: a *detection training set* and a *detection test set*.

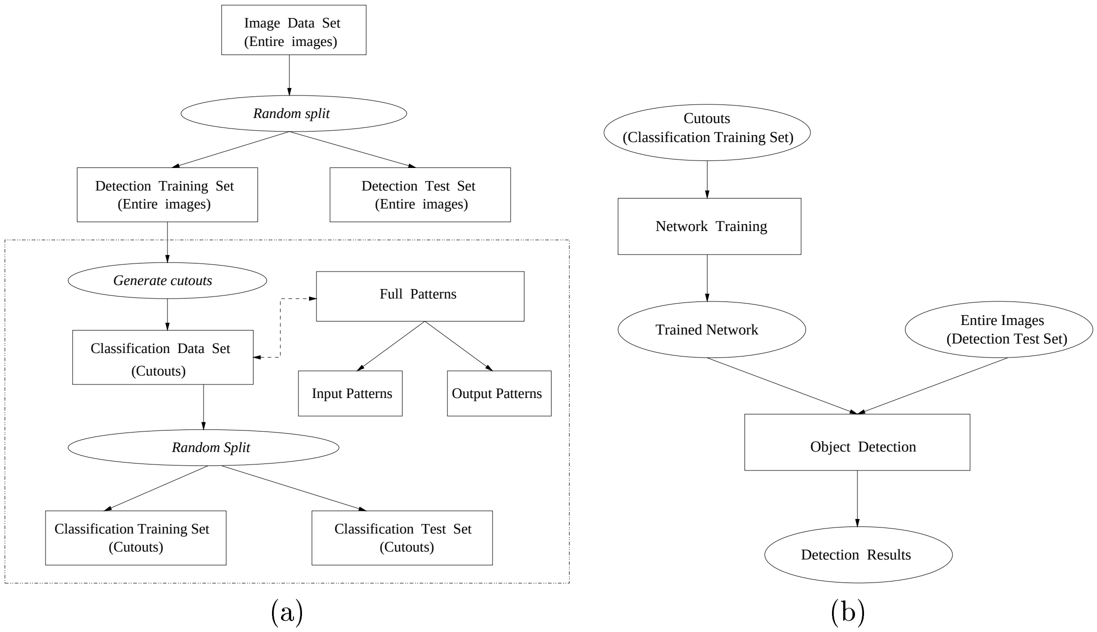


Figure 3: Overview of the approach

2. Determine an appropriate size (n) of a square which will cover all of single objects of interest and form the input field of the networks. Generate a classification data set by cutting out squares of size $n \times n$ from the detection training set. Each of the squares, called cutouts or sub-images, only contains a single object and/or a part of the background. Randomly split these cutouts into a classification training set and a classification test set. These classification data sets are used for object classification, which involves network training and network testing. Note that all of the backgrounds are also considered as one class, but not a class of interest.

3. Determine the network architecture. A three layer feed forward neural network is used in this approach. The $n \times n$ pixel values form the inputs of a training pattern and the classification is the output. The number of hidden nodes is empirically determined.

4. Train the network by the backward error propagation algorithm [5, 11] on the classification training data. The network training process will be stopped according to the termination strategy. The trained network is then tested on the classification test set to measure the object classification performance. The classification test data is also used to avoid overtraining the network. This step is designed for finding the best trained network for object detection.

5. Use the trained network as a moving window template [1] to detect the objects of interest in the detection test set. If the output of the network for a class exceeds a given threshold then report an object of that class at the current location. It is important to note that the thresholds for different classes are different.

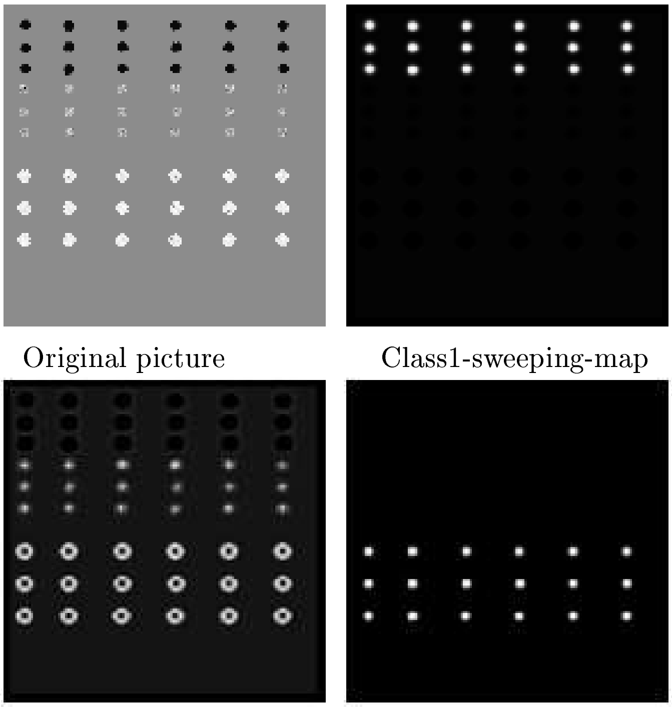
6. Evaluate the object detection performance of the network by calculating the detection rate and the false alarm rate.

3.2 Object Detection

While object classification corresponds to network training and network testing on the cutouts in the classification data sets, object detection corresponds to network sweeping on the entire images in the detection test set, which were not used in any form for network training. *Network sweeping* involves both object classification and localisation.

During *network sweeping*, the successfully trained neural network is used as a template matcher, and is applied, in a moving window fashion, over the large pictures to detect the objects of interest. The template is swept across and down these large pictures, pixel by pixel in every possible location.

After the sweeping process is done, an *object sweeping map* for each detected object class will be produced. An object sweeping map corresponds to a PGM image. Sample object sweeping maps for *class1*, *class2* and *class3* together with the original picture for the easy detection problems are shown in Figure 4. During the sweeping process, if there is no match between a square in a detected picture (an original picture in the detection test set) and the template, then the neural network output is 0, which corresponds to black in the sweeping maps; partial match corresponds to grey on the centre of the object, and best match is close to white in the sweeping maps.



Original picture Class1-sweeping-map

Class2-sweeping-map Class3-sweeping-map

Figure 4: Sample object sweeping maps in object detection.

4 Results

4.1 Object Classification Results

Network training and testing results for object classification are presented in Table 1. In all cases, the network training and testing procedure was repeated 15 times and the average results are presented. Line 1 shows that the best network for the easy pictures is 196-4-4 (input-hidden-output nodes), the average number of epochs used to train the network and get it converged is 199.40, and the trained network can achieve 100% accuracy on the cutouts of both the classification training set and test set. For the coin pictures, we can also achieve the ideal performance for object classification. However, this is not the case for the retina pictures, where only 71.83% accuracy was obtained on the test set of the cutouts.

Image Databases	Network	Training Epochs	Training Accuracy	Test Accuracy
Easy	196-4-4	199.40	100%	100%
Coin Pictures	576-3-5	234.6	100%	100%
Retina Pictures	256-6-5	475.8	81.62%	71.83%

Table 1: Object classification — Network training and testing results for the three databases.

4.2 Object Detection Results

This section describes the detection performance of this approach on the three detection problems. In all cases, the 15 trained networks described in Table 1 are performed for object detection and the average results are presented. During the object detection procedure, various thresholds result in different detection results. The higher the threshold, the fewer the objects that can be detected by the trained network, which results in a lower detection rate but also a lower false alarm rate. Similarly, the lower the threshold is selected, the higher the detection rate and the higher the false alarm rate are produced. Thus there is generally a trade-off between the detection rate and the corresponding false alarm rate.

The best detection rates and the corresponding false alarm rates for the three classes in the easy pictures are presented in Table 2. The best detection rates achieved for all the three classes are 100%, showing that this approach can successfully detect all the objects of interest in this database. At this point, detecting *class1* (black circles) and *class 3* (white circles) did not produce any false positives (false alarm rates are zero), while detecting *class2* (grey squares) resulted in a 91.2% false alarm rate on average.

Easy Pictures	Object Classes		
	Class1	Class2	Class3
Detection Rate(%)	100	100	100
False Alarm Rate(%)	0	91.2	0

Table 2: Object detection results for the easy pictures.

The detection results for the coin pictures are described in Table 3. As in the easy pictures, the trained neural networks achieved 100% detection rates for all the classes, showing that this approach correctly detect all the objects of interest from different classes in the coin pictures. In each run, it was always possible to find a threshold for the network output for class *head005* and *tail005* which resulted in detecting all of the objects of these classes with no false alarms. However, detecting classes *head020* and *tail020* was a relatively difficult problem. The average false alarm rates for the two classes at a 100% detection rate were 182% and 37.5% respectively.

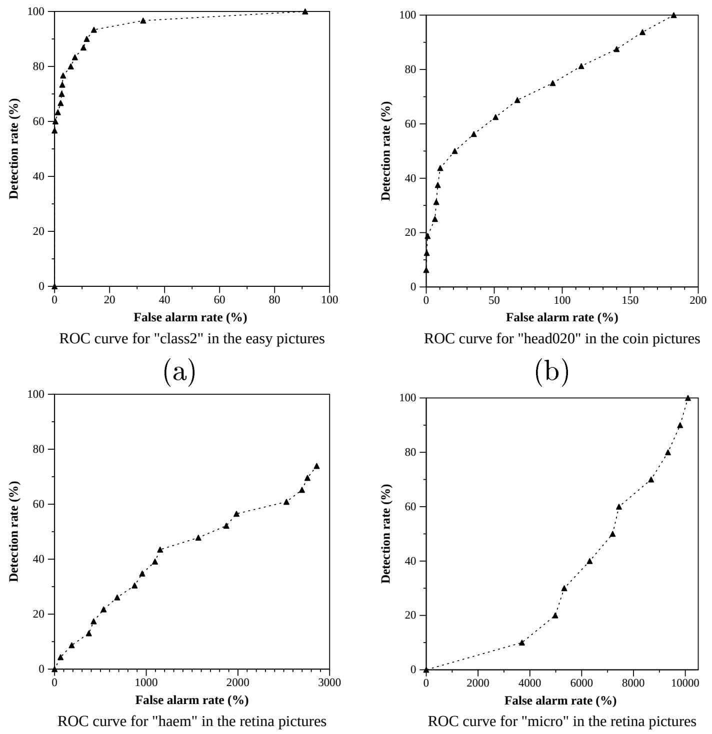
Compared with the performance of the easy and coin pictures, the results of the two classes *haem* and *micro* in the very difficult retina pictures are disappointing. All the objects of class *micro* were correctly detected (a detection rate of 100%), however, with a false alarm rate of 10104%. The best detection rate for class *haem* was 73.91% with a corresponding

Coin Pictures	Object Classes			
	head005	tail005	head020	tail020
Detection Rate (%)	100	100	100	100
False Alarm Rate (%)	0	0	182	37.5

Table 3: Object detection results for the coin pictures.

false alarm rate of 2859%. Even at a detection rate of 50%, the false alarm rate was still quite high (about 1800%).

To give an intuitive view, the extended ROC curves (Receiver/Relative Operating Characteristic curve [7, 16]) for class *class2* in the easy pictures, class *head020* in the coin pictures and classes *haem* and *micro* in the retina pictures are presented in Figure 5 (a), (b), (c) and (d) respectively.



(a) (b) (c) (d)

Figure 5: Detailed results (extended ROC curves) for some "difficult" classes in the three databases

5 Analyses and Discussions

5.1 Analysis of Results

As can be seen from the detection results obtained here, it was always possible to detect all objects of interest in the easy and the coin pictures. This reflects the fact that the objects in the two databases are simple or regular and the background is uniform or relatively uniform. While detecting the easy pictures only resulted in a few false alarms, detecting objects in the coin pictures resulted in a relatively higher false alarm rate. This is mainly because the detection problems in the coin pictures are more difficult than in the easy pictures.

Due to the difficulty of the detection problems, the results for the retina pictures are not good. For class *micro*, while all objects were correctly detected, a very high number of false alarms were produced. This is mainly because these objects are irregular and complex and the background is highly cluttered. For class *haem*, it was not possible to detect all objects of interest (the best detection rate was 73.91%). This is mainly due to the size variance of these objects (from 7×7 to 14×14 pixels). Another reason that we did not obtain good results on the retina pictures might be the insufficient number of training object examples (only 164, see Figure 1).

5.2 Weight Analysis

To analyse why this approach can be used for multiclass object detection, this section interprets the network internal behaviour through visual analysis of the weights in the trained networks. For presentation convenience, we use the trained networks for regular object detection in the coin pictures. Most other networks contained similar patterns.

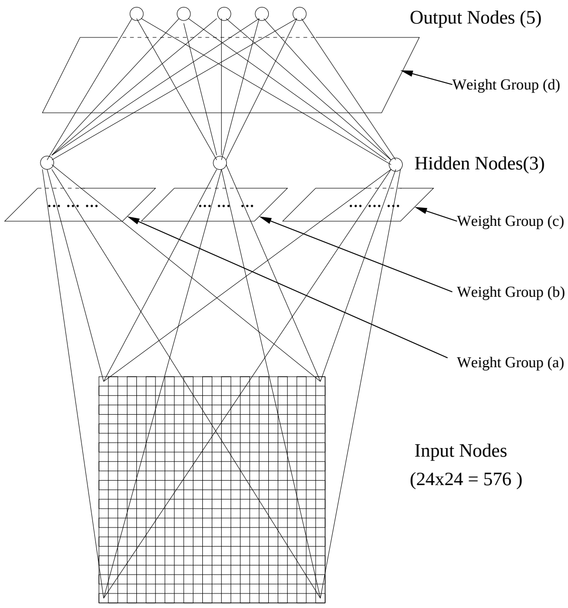
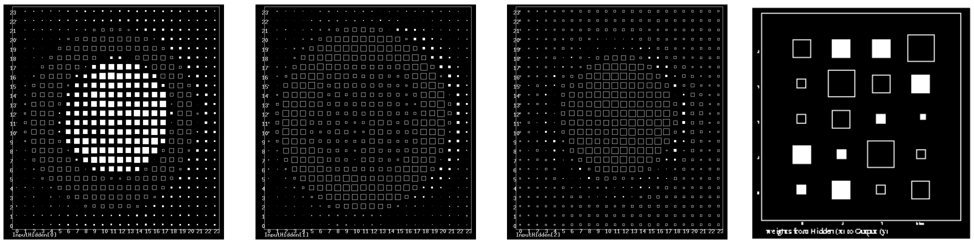


Figure 6: Network architecture with four weight groups for object detection in the coin pictures.

Figure 6 shows the network architecture (576-3-5) used in the coin pictures. The weight groups between the input nodes and the hidden nodes and between the hidden nodes and the output nodes are also presented. The weight matrices (a), (b), (c) and (d) shown in Figures 7 correspond to weight groups (a), (b), (c) and (d) in the network architecture in this figure.

Figure 7 shows the weights from a trained 576-3-5 network which has been successfully used in the coin pictures. In this figure the full squares represent positive weights and the outline squares represent negative weights, while the size of the square is proportional to the magnitude of the weight. Matrices (a), (b) and (c) shows the weights from the input nodes to the first, the second and the third hidden nodes. The weights are shown in a 24×24 matrix (corresponding to the 24×24 input field) to facilitate visualisation. Figure 7 (d) shows the weights from the hidden nodes to the output nodes and the biases of these output nodes. The five rows in this matrix, in bottom to top order, correspond to the classes *head005, tail005, head020, tail020* and *other*. In left to right order, the first three of the four columns correspond to weights from the first hidden node (associated with weight matrix (a)), the second hidden node (associated with weight matrix (b)) and the third hidden node (associated with weight matrix (c)) to the five output nodes. The last column corresponds to the biases of the five output nodes.



(a) (b) (c) (d)

Figure 7: Weights in a trained network for object detection in the coin pictures.

Inspection of the first column of Figure 7 (d) reveals that weight matrix (a) has a positive influence on 5 cent coins and a negative influence on 20 cent coins. It has a strong influence on class *tail005* but a week influence on class *head005*. The same matrix has a strong negative effect on class *other* (background). Inspection of the second column reveals weight matrix (b) has a positive effect on the 5 cent coins and a negative effect on the 20 cent coins. Moreover, it has a strong influence on class *head005* but a week influence on class tail005. Also it has a very strong positive influence on class *other*. This indicates that the combination of matrices (a) and (b) not only can separate the 5 cent coins from the 20 cent coins and the background, but also can discriminate the 5 cent tails from the 5 cent heads. Inspection of the third column reveals that the weight matrix (c) has a strong negative influence on the tails of both 5 cent and 20 cent coins and a week influence on the heads of both 5 cent and 20 cent coins. It also strongly supports the background and has a strong negative influence on the tails of 5 cent coins. The fourth column suggests that the biases also play a complementary role for detecting objects, particularly for class *tail020*, class *other* and class *head005*. If we regard the nodes of the hidden layers as representing feature detectors learnt by the network, then Figures 7 (a)-(c) are a visual representation of these features. Visually these features 'make sense' as there are regions corresponding to the 5 cent coins, the annulus remaining to the background when a 5 cent coin is 'removed' from the centre of a 20 cent coin.

6 Conclusions

This paper presented a pixel based approach for detecting multiple class objects in large pictures using multilayer feed forward neural networks. The backward error propagation algorithm was used for the network training on the sub-images which had been cut out from the large pictures. The trained network was then applied, in a moving window fashion, over the entire images to detect the objects of interest. Object detection performance was measured on the large pictures in the detection test set. The experimental results showed that this approach performed very well for detecting a number of simple and regular objects against a relatively uniform background. It did not performed well on the difficult detection problems in the retina pictures, which indicates that it can not be well suited to detecting complex and irregular objects against a highly cluttered background. As expected, the performance degrades when the approach is applied to detection problems of increasing difficulty.

Visalisation of the weights in trained neural networks resulting from this approach revealed that trained networks contained feature detectors which "made sense" for the problem domain and could discriminate objects from different classes. In addition, the approach has the following advantages:

- The raw image pixel data are used as inputs to neural networks, and accordingly traditional specific feature extraction and selection is avoided.
- It is a domain independent approach and can be directly applied to multiclass object detection problems in different areas.
- Multiple class objects can be detected (classified and localised) from large pictures with a single trained neural network.

The approach also has a number of disadvantages, which need to be addressed in the future:

- This approach cannot successfully detect objects with the size invariance, such as class *haem* in the retina pictures.
- In this approach, the network was trained based on the object cutouts but the trained network was directly applied to the entire images. This might be one of the main reasons which result in many false positives in some difficult detection problems. We are investigating ways of refining the trained networks based on the entire images to improve the detection performance.
- The approach needs a relatively large number of training examples. For instance, the number of training examples in the retina pictures is not sufficient, which might be one of the main reasons resulting in poor performance.

Acknowledgement

We would like to thank Dr. Victor Ciesielski for useful discussions and Chris Kamusinski who provided and labelled the retina pictures.

References

[1] Dana H. Ballard and Christopher M. Brown. *Computer Vision*. Englewood Cliffs, N.J: Prentice-Hall, Inc., 1982.

[2] David P. Casasent and Leonard M. Neiberg. Classifier and shift-invariant automatic target recognition neural networks. *Neural Networks*, 8(7/8):1117–1129, 1995.

[3] V. Ciesielski and J. Zhu. A very reliable method for detecting bacterial growths using neural networks. In *Proceedings of the International Joint Conference on Neural Networks*, pages 62–67, Beijing, November 1992.

[4] Paul D. Gader, Joseph R. Miramonti, Yonggwan Won, and Patrick Coffield. Segmentation free shared weight neural networks for automatic vehicle detection. *Neural Networks*, 8(9):1457–1473, 1995.

[5] Steven Hayes and Victor Ciesielski. A neural network approach to positional and rotational invariant recognition of 3d objects. Technical report, Royal Melbourne Institute of Technology, Melbourne, Victoria, Australia, 1997.

[6] Jack S. N. Jean and Jin Wang. Weight smoothing to improve network generalisation. *IEEE Transactions on neural networks*, 5(5):752–763, September 1994.

[7] Charles E. Metz. ROC methodology in radiologic imaging. *Investigative Radiology*, 21(9):720–732, September 1986.

[8] Steven K. Rogers, John M. Colombi, Curtis E. Martin, James C. Gainey, Ken H. Fielding, Tom J. Burns, Dennis W. Ruck, Matthew Kabrisky, and Mark Ocley. Neural networks for automatic target recognition. *Neural Networks*, 8(7/8):1153–1184, 1995.

[9] H. L. Roitblat, W. W. L. Au, P. E. Nachtigall, R. Shizumura, and G. Moons. Sonar recognition of targets embedded in sediment. *Neural Networks*, 8(7/8):1263–1273, 1995.

[10] Michael W. Roth. Survey of neural network technology for automatic target recognition. *IEEE Transactions on neural networks*, 1(1):28–43, March 1990.

[11] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, and the PDP research group, editors, *Parallel distributed Processing, Explorations in the Microstructure of Cognition, Volume 1: Foundations*, chapter 8. The MIT Press, Cambridge, Massachusetts, London, England, 1986.

[12] Mukul V. Shirvaikar and MMohan M. Trivedi. A network filter to detect small targets in high clutter backgrounds. *IEEE Transactions on Neural Networks*, 6(1):252–257, Jan 1995.

[13] Allen M. Waxman, Michael C. Seibert, Alan Gove, David A. Fay, Ann Marie Bernandon, Carol Lazott, William R. Steele, and Robert K. Cunningham. Neural processing of targets in visible, multispectral ir and sar imagery. *Neural Networks*, 8(7/8):1029–1051, 1995.

[14] Philip Winter, Shahab Sokhansanj, Hugh C.Wood, and William Crerar. Quality assessment and grading of lentils using machine vision. In *Agricultural Institute of Canada Annual Conference*, Saskatoon, SK S7N 5A9, Canada, July 1996. Canadian Society of Agricultural Engineering. CASE paper No. 96-310.

[15] Yonggwan Won, Paul D. Gader, and Patrick C. Coffield. Morphological shared-weight networks with applications to automatic target recognition. *IEEE Transactions on neural networks*, 8(5):1195–1203, September 1997. ISSN 1045-9227.

[16] Mengjie Zhang. *A Domain Independent Approach to 2D Object Detection Based on the Neural and Genetic Paradigms.* PhD thesis, Department of Computer Science, RMIT University, Melbourne, Australia, August 2000.