Attention-Focusing Artificial Neural Networks for Image Analysis

Stuart E. Barker

A thesis submitted in partial fulfilment of the requirements of The Nottingham Trent University for the degree Doctor of Philosophy

2000

Abstract

Some of the common operations humans take for granted, for example the human vision system, have been found very difficult to emulate. Although humans are able to perceive visual information almost instantly, this belies the complexity of this process.

This thesis describes a computer vision strategy that involves Artificial Neural Networks (ANNs) to perform accurate and efficient object identification. Face location is used as the primary test domain. This involves the processing of real world scenes to distinguish between faces of different shapes, sizes and different viewpoints. Object identification in a complex environment is an extremely difficult task and research into this area of computer vision is currently not being fully exploited. Many previous models for computer vision have applied techniques that only solve particular well-defined problems.

An efficient two-stage vision strategy is presented which removes the necessity to process an image at full resolution through the use of low resolution. The first stage uses a multi-resolution approach to identify areas of interest at an optimally low resolution. The focus areas are then passed to a classification stage to perform more accurate analysis to reject the area of interest or confirm the presence of the pre-determined object.

Thesis Abbreviations

AF – Attention-Focusing.

ANN – Artificial Neural Network.

ART – Adaptive Resonance Theory [Carpenter and Grossberg, 1985].

Backprop – Back-Error Propagation [Widrow and Hoff, 1960].

ED – Eye distance.

FA – Focus Area.

FNs – False Negatives.

FPs – False Positives.

MLP – Multi-Layered Perceptron.

TNs – True Negatives.

TPs – True Positives.

Table of Contents

Chapter 1

Introduction

Over the past thirty years the computing world has seen a continued growth in the technological improvement of computers in terms of their processing power, storage, and sophistication of applications performed on them. This increase in computing power naturally leads to the desire to solve ever more complex problems. An area that continues to receive much interest are the problems associated with the human vision process and the ability to extract semantic information from what is *seen* [Bischof, 1995].

In order to interact with our environment, the main source of sensory input is from our visual system. Processing this information allows humans to build an understanding of the world around them. This task is seen to be performed effortlessly without being even aware of the complexities involved. *Computer Vision* is an attempt to model the same information electronically, and to form internal representations that allow computers to build meaningful information about the external surroundings.

1.1 Why is Vision So Difficult?

Some of the common operations humans take for granted, for example the human vision system, have been found very difficult to emulate. This is in part due to the fact that the whole biological process is not completely understood. Although humans are able to perceive visual information almost instantly, this belies the complexity of this process.

For example, in order to process the visual information of any scene the brain deals with numerous problems. The first step is to segment the image into areas of interest and background information. Areas of interest may be defined from priori knowledge that the observer wishes/expects to identify. Areas of the background may be as complex as the objects that constitute the areas of interest, and may indeed comprise of objects themselves.

Classifying the areas of interest is also problematic and such issues include: different lighting conditions, partial occlusion, profiles, rotation, shape deformation, and size variance. It is not difficult to understand the limited progress made in developing robust models that can cope with all, or even the majority, of these problems. For this reason, algorithms have been developed to solve particular issues regarding the vision process.

1.2 Techniques for Computer Vision

The field of computer vision is a diverse discipline that encompasses many different techniques to solve particular computer related problems. These range from low-level image manipulation algorithms to higher level AI methods. To extract meaningful information from an image may require the combination of many different processes, which allows this understanding to take place.

A digital image at its most basic level can be regarded as a complex array of pixels each of which can be a discrete colour intensity. Image Processing is a particular area of computer vision that manipulates the image directly in order to improve or simplify the image. This is the first stage of the processing hierarchy, which then allows further representations to be built that model higher level information. Examples of typical image processing algorithms are: Thresholding, Histogram Equalisation, Convolution, Boundary Tracing, Edge Thinning, Region Merging, etc.

Pattern recognition, as its name suggests, is the process by which the structured composition of an image is identified and its form classified. This, generally, encompasses higher level techniques that uses the representations derived in the algorithms defined above. Processing methods that can be included at this stage are Matched Based Segmentation, Contextual Image Classification, Snake Growing, Texture Analysis, Artificial Neural Networks, etc.

The techniques presented are too numerous to be described in detail, but there are many texts (e.g. Sonka *et. al,* 1995) which give better explanations than can be provided here.

1.3 Motivations for a General Model of Computer Vision

Many previous models for computer vision have applied techniques that only solve particular well-defined problems. It is the purpose of this investigation to reduce the constraints on the problem to enable a novel and flexible model to be developed which considers more general aspects of computer vision.

A computer vision system capable of performing the tasks of human vision even to some small degree would be very useful. A framework that is flexible rather than specific to a particular application (e.g. recognising anomalies of a product on a conveyer belt) is more appealing because of its re-usability for similar problems. There are many application areas that could incorporate such technology in some way, e.g. automatic video surveillance; computer guided robots, etc. There is much research therefore, into solutions that attempt to tackle the wider aspect of general vision problems.

The problem of *face identification* is used to test the strategy. Face identification should not be confused with face recognition. Face recognition is the process by which a system has been trained to recognise particular people and is required to determine whether a novel instance of a face belongs to the known identity database. Typically the faces and constraints used for such systems are very limited; i.e. the faces are usually of a fixed size and orientation against a simple continuous background. Research on face recognition has received much more attention than the earlier vision problems described above. Many systems [Brunelli and Poggio, 1993, Edelman *et. al*, 1992, Fukuda *et. al*, 1992, Garrison *et. al*, 1990, Marsic and Micheli-Tzanakou, 1992, Turk and Pentland, 1990], exist that are able to provide a high degree of recognition accuracy. The identification of faces in an image is a more challenging task, as generally there is no prior knowledge about the image and the information is less constrained than that used for recognition.

Face identification is the process of identifying the locale of an object where there are as few constraints placed upon the problem domain as possible. The unrestrained nature of this problem in comparison to face recognition makes this task more difficult. Identification is also a general problem that is not necessarily specific to faces whereas face recognition has a more defined goal.

1.3.1 Artificial Neural Networks

As the problems undertaken have become ever more complex, the Artificial Intelligence field has looked for alternative methods to help provide solutions to difficult areas of research. A more recent approach in the field is to use Artificial Neural Networks (ANNs). ANNs are not specific to computer vision and they have also seen a greater increase in other application areas. Perhaps the reason why they have received great acceptance in many fields are that they are universal function approximators [Masters, 1993]; they do not require a large and complicated rule base; and have demonstrated a proven success of outperforming comparable techniques.

ANNs differ significantly to other techniques by their structure and method and can be regarded as general learning models. They are best described as biologically inspired models of the brain, and although they do not work in exactly the same way, they nevertheless use a number of simple, highly connected processing elements, as do their neurological counterparts. A more detailed explanation about the different types of ANN, their connectivity, and how learning takes place can be found in Appendix A.

It was concluded by Feldman and Ballard [Feldman and Ballard, 1982] that massively parallel models are the only biological plausible ones, as these are the only models which satisfy the A100 step rule[1]. For this reason, an ANN approach has been chosen as the most suitable method for developing a framework for a general computer vision system.

1.4 Aims and Objectives for the Vision System

In order for a general-purpose vision system to effectively cope with real world scenes, it will need to distinguish between objects of different shapes, of different sizes, and at different viewpoints. These are basic requirements if a system is to function successfully [Roth and Frisby, 1986]. The aim of the research is the design of a general model for vision that deals with position, size, viewpoint and shape tolerance against any type of background. No prior knowledge about the scene will be known, and the only input to a system is quantised digital information.

Faces can be described as a suitably complex object, but distinguishing them from all other objects is very problematic. The nature of the information required to enable efficient face recognition is complicated. All faces generally contain the same number of primary features (eyes, nose, mouth, etc.) but all of these vary to some degree for different face examples. For a general introduction to the problems of face identification and recognition, Samal and Iyengar [Samal and Iyengar, 1992] give a brief survey of the difficulties involved.

[1] The definition of the 100-step rule is that most neurons compute at maximum rate of 1000 Hz. Perception occurs within 100 milliseconds, and therefore biological models can require more than 100 steps.

The faces used for analysis will be directly facing the camera, and the system will not be expected to deal with the problem of extreme head rotation. The analysis is on static images only, and it is outside the bounds of this research to cover motion of objects in temporal sequences. A strategy is proposed to deal with the following:

- Position invariance. The system has no prior knowledge of where an object may be located, and thus must deal with objects occurring anywhere in the image.
- Tolerance to multiple occurrences. There is no prior knowledge of how many objects the scene contains.
- Background invariance. The system must be capable of locating objects with any type of background, however complex.
- Size invariance. There is no prior knowledge as to the size of each object contained in the scene, and the objects may even be of different sizes within the same image.

In order for the system to be as efficient as possible it is the aim of the design to use the input of the digital image directly for processing. This will involve the investigation of the necessary pre-processing required for input to the ANN model(s).

An area of the image that has been classified as the target object may then be used for further processing, that may include more detailed visual inspection or some high order manipulation on the information extracted. It may be useful to some applications to be able to make inferences based on the number, position, size of the objects etc. determined. In the case of face identification, further processing will most likely be the recognition of the identity of the person. The type of further processing is less general and is more specific to the target application. It is not the aim of this work to develop a recognition module but to develop strategies that address vision associated problems earlier on in the processing. However, recognition is the natural progression for a face identification system.

1.5 Outline of Thesis

This thesis describes a computer vision strategy that involves ANNs to perform accurate and efficient object identification. This also draws to the attention of the reader to the problems associated with the large quantities of data used for analysis and processing.

The author has highlighted a particular area within computer vision that is currently not being fully exploited. This chapter has defined the outline of the problem and discussed an approach, using ANNs, as to how this might be solved. The following chapters describe the issues considered to derive a general framework to this problem, and the logical manner by which they were addressed.

Systems for Computer Vision evaluates the field of computer vision and what other methods are being employed in this area. An outline method is presented to achieve the goals of object identification and this is compared with different approaches to determine the feasibility of this strategy.

The computer vision task is broken down into manageable problems, the first of which is to identify areas of interest within a complex scene. To simplify this difficult task, an initial restriction is imposed that areas of interest are of one size only. *High Speed Location of Fixed Sized Objects* describes a suitable solution to this initial problem.

The following chapter, *Size Invariant Object Location,* extends the problem to cover areas of interest of any size. The chapter discusses how the previous architecture is combined with suitable algorithms that enable the method to tackle this problem.

Now that areas of interest have been identified, reliable techniques are required to discriminate between correct and incorrect focus areas. *Focus Area*

Classification describes a combination of ANNs and algorithms that process the focus areas to give a robust means of verification.

The final chapter *Conclusions and Further Work* identifies what has been achieved by this framework and also its limitations. This naturally leads on to a discussion of how the work could be progressed further.

Chapter 2

Systems for Computer Vision

The previous chapter introduced an overview of the problems associated with computer vision and identified the need for a general vision model, which the author believes, is not being fully exploited. This chapter discusses some of the approaches that have been investigated and developed in this area. This involves a critique of these methods and considers the salient points of each. The conclusion of this review is to formalise a possible framework for the problem identified. This is then investigated further in the following chapters.

2.1 Image Processing and AI Techniques

With a traditional approach to image analysis, a collection of processing tasks are required. ANN systems in common with other approaches require pre-processing to produce a form that is more suitable for subsequent operations. This may involve functions that improve the clarity of the image, to counter the effect of uneven lighting, poor contrast levels and also to remove any noise that may have been included during image capture.

To identify areas of interest in the image, the scene must be processed to differentiate between the background and the objects contained within it. This process is generally referred to as scene segmentation [Haralick and Shapiro, 1985]. Different approaches can be used to achieve this. Some of these methods are discussed below.

One of the most basic approaches for image segmentation is Thresholding. After selecting an appropriate threshold value the image is transformed into a digital representation. The success of this technique is reliant upon the choice of threshold that can be determined by analysis of the intensity range of the original image. The algorithm can either enhance high or low intensities but does not necessarily improve the clarity of any particular objects within a scene, and may even make distinct objects become connected to the surrounding image.
Convolution is another segmentation method. This approach convolves a user-defined template with the image. The resulting effect to the transformed image is determined by the size of the matrix and the values within it. Such transformations include noise removal and edge enhancement. Typical template examples include the Roberts and Laplacian operators. Convolution may not enhance small detail within an image but distort it, and edge enhancements are generally directed towards linear edges. In the case of face identification, the image contains both detailed features and also non-linear edges.

After segmentation in a traditional image processing approach, the image is processed further for edge and boundary detection. This is to allow features to be extracted to aid object classification. Assumptions are made that the object can be distinguished using a contour due to some kind of contrast in intensity, colour or texture. In the problem defined, the objects to be identified can vary in size so any method must be capable of shape deformation. The fuzzy Hough transform [Philip, 1991] allows detection of objects whose exact shape is a little uncertain and finds matches that are closest to an approximate contour model. One of the other issues of this research is to process real world images. A characteristic of these types of images is that the objects and the background are not easily segmented. To use a contour approach is therefore problematic if no distinct boundaries can be found.

Statistical pattern recognition is an alternative to ANNs as a means to distinguish between different pattern classes. Objects are represented as numerical descriptors called feature vectors that are mapped onto pattern space. An initial problem with this method is selecting the best feature vectors, and the optimum number of vectors to describe the object. Too few descriptors and the object is insufficiently described, too many and the mapping to the feature space becomes more difficult.

A typical method of statistical analysis is probability density estimation [Fukanaga, 1990]. If the feature vectors are chosen appropriately, similar objects are represented as clusters in pattern space. This method produces a similar output to a self-organising ANN [Kohonen, 1989]. The same techniques are applied to discriminate the different classes in pattern space. Another particular problem is in the analysis of the output feature space. Ideally, the different classes are well-separated clusters. Alternatively, a non-linear decision boundary may be required to separate them.

The methods discussed previously have formed representations to extract features of information about an image. Further techniques can be used to combine this information and illicit more high level knowledge about the complex items and their interactions with each other within the image. The use of AI methods is very common at a higher processing level, and a good knowledge representation is required in order to complete full image understanding. Trees and graphs are methods [McHugh, 1990] used to construct a representation of a collection features, or primitives, that are related. Depending upon the position in the hierarchy of the tree determines the level of object complexity. Although the data structure assembles the information meaningfully, some further processing is required to parse and interrogate it. This type of methods tends to be more problem specific. This is only one particular approach and other methods exist to achieve similar goals.

ANNs are architecturally very different to other approaches. Because they are general models, they are suited to many different applications. There are two main approaches to using ANNs in computer vision.

The first and more biologically driven approach is modelling the human retina. Work in this area often uses artificial environments involving simple shapes for processing, rather than *real* images and the complex problems associated with them.

The second strategy involves systems which attempt to solve a computer vision problem in a way that achieves a specific goal, i.e. problem driven. This may mean that the system model is tightly constrained to a specific problem (e.g. anomaly analysis of components on a conveyer belt) but uses approaches that may achieve success in more general situations. Both strategies are of equal merit and are not mutually exclusive and there is possibly some overlap between them.

2.2 Biologically Motivated Low Level Vision

The following section discusses a small collection of ANN systems that attempt to produce biologically plausible methods for computer vision. Much research is being performed in this area and a whole new field of computation neuroscience has evolved to cater for the biological/computing overlap. A brief overview of prominent works in this field are given below.

The work of Marr [Marr, 1982] has been influential to many in the field of visual psychology. The visual process was decomposed into three sub-processes. These include the *Primal Sketch* which is a two dimensional representation of significant grey-level changes in the image. The *2.5D Sketch* which is a partial three-dimensional representation recording surface distances from the viewing point, and finally, the *Solid Model* based representation of objects in the scene. Watt [Watt, 1988] uses Marr's idea of vision and extends the primal sketch to develop an algorithm/model to process an image and extract and interpret the symbols generated from the scene.

Grossberg [Grossberg *et. al*, 1989] proposes a model for pre-attentive vision. This involves three sub-systems that include a Feature Contour System (FCS), Boundary Contour System (BCS), and an Object Recognition System (ORS). The FCS is designed to detect surface colours under variable lighting conditions. The BCS is designed to recognise invariantly object boundary structures, and the ORS is designed to recognise familiar objects in the environment. The general-purpose capabilities depend upon the decomposition into BCS, FCS, and ORS sub-systems. Both the BCS and FCS operate pre-attentively on images, even if they have been experienced before. The performance of both the FCS and BCS are interrelated and are thus combined to provide both parallel and hierarchical stages of neural processing. The feedback interactions between pre-attentive BCS and FCS and the attentive, adaptive ORS show that the systems are not independent modules. A definition for the whole adaptive system has been described as a series of algebraic equations and results are given for simple shape and texture analysis at a single scale.

A vision system [Sajda and Finkel, 1992] has been designed by integrating a top-down computation based approach with a bottom-up biologically motivated architecture. Its aim is to address occlusion-based object segmentation through the use of a hybrid ANN. It has been found to be capable of discriminating objects relative to their depth. To achieve this, feature extraction is applied, similar to that of Grossberg [Grossberg *et. al*, 1989] to perform simple activities such as edge detection, line orientation etc. The network architecture that has been used has included Programmable Generalised Neural units, which attempt to mimic properties found in the visual cortex.

Ahmad and Omohundro [Ahmad and Omohundro, 1990] discuss an implementation of attention-focusing. The work has been applied to simple geometric shapes such as equilateral triangles. To create a dynamic receptive field, a gating layer is constructed so that there is one *gate* unit per input unit. The action of the gate unit is to control the activity of the input unit depending upon whether it is within the focus of attention. Methods are also described to alter the size and position of the focus by determining the error of activations falling outside the focus of attention. Similarly, Anderson [Anderson, 1990] discusses the use of an attentional spotlight for visual attention. It is well known that selective attention mechanisms exist in the human vision system and that the operation is to some extent (a combination of parallel and serial processing) a sequential task.

The different models described briefly above are only a small selection of the many other authors that have developed biological representations of the biological process. All of the work discussed has provided different implementations based upon the understanding of the human visual process. The work has concentrated in developing coherent models to achieve a greater understanding of the biological process. To substantiate these theories the systems have been applied to problems using simple shapes. These models of human computer vision are currently inadequate to be applied to real world problems such as face identification. A common theme found in all these systems that can be extracted and used at a higher level is the idea of attention-focusing and selectively identifying areas of interest.

2.3 ANNs for Computer Vision

The systems described in this section have applied ANNs to solve particular problems in vision. Although the ANN models discussed have not necessarily addressed the specific problem of face identification, they do suggest suitable methods that may be applied to this problem.

Hutchinson and Welsh [Hutchinson and Welsh, 1989] describes a standard MLP trained on right eyes from full resolution images, in order to help detect moving features. A 16x16 window is passed across part of the input images to train the ANN. The ANN is trained with a response of 1 for the eye centre that decreases linearly away from this point. Using an alternative method for comparison, a Kohonen network is used to evaluate the performance of the first ANN. The Kohonen network is trained with the same input data as the first MLP, and has a 10x10 output layer. A further ANN is then trained to interpret the output produced from the Kohonen network. Neither ANN method shows a vast improvement over the other. Performance was generally good except for those images where glasses were worn.

No effective use of low resolution is employed, but Hines and Hutchinson [Hines and Hutchinson, 1989] use the same MLP model described above on reduced resolution images to reduce the quantity of data processing. An attempt at increasing the resolution did increase performance, but at the expense of more processing. The target output has been modified to a fixed output response.

The concept of using reduced resolution to reduce the amount of processing is investigated by Vincent [Vincent *et. al*, 1992]. Faces are also used for image analysis, but in a constrained manner, using head and shoulder images of a fixed size. A two-stage approach is used. Initially, a number of ANNs are trained to identify eyes, mouth etc. by scanning the image at coarse resolution. The output from these ANNs are then post-filtered and presented to further ANNs at high resolution. The post-filtering adopts knowledge about the relative positions of the micro features and is a good method of removing false activations. Using micro-features at coarse resolution however does not allow much scope to reduce the image significantly as some level of detail for these features is still required. [Vincent *et. al*, 1992] also does not propose strategies to deal with variabilities in the size of the head and shoulders and complexities of different backgrounds.

Allinson and Johnson [Allinson and Johnson, 1992] adopt a novel approach to attention-focusing by using a binary N-tuple sampling method which, although a very different architecture to the majority of ANN implementations, works similarly to the Kohonen self-organising network. Using head and shoulder images the system is system is trained to focus attention upon the right eye. Because the N-tuple method is a binary approach, the grey scale images have to be first converted to an appropriate form. This is achieved by using a rank ordering code. A supervised self-organising map is then used to produce a positional map of the attention window. The position is used to provide an *x, y* error of the maximum response relative to the centre of the window, and the amplitude of the response determines the size of the subsequent window (i.e. the resolution of the image). Using the N-tuple sub-space classification technique, although relatively efficient, does require a large amount of memory to process an image. The system also requires many saccadic jumps and resolution shifts before the correct position of the feature is found. The time taken to fully perform this operation is not given. As the system is dependant upon the processing of binary images this indicates that it is a less suitable method when considering the complexities of real world images.

A limited approach to locating right eyes is investigated by Evans [Evans *et. al*, 1991], which searches for eyes across three separate resolutions using three separate MLPs (one for each resolution). The area of interest with the most values above a threshold is determined to be the location of the eye. Using three separate MLPs restricts the variability in face size that can be used, and to cope with a greater range would require further trained MLPs at different resolutions to process the image. This is obviously impractical. Using this method does not significantly reduce the amount of information processed required, as each MLP is required to examine the image at its associated resolution. Furthermore, the system does not address the problem of false positives and how they may be reduced to limit the number of incorrect classifications.

There are some common problems that can be identified with the majority of the methods discussed. These include the problem of false activations and how these might be reduced. None of these methods have dealt particularly with varied backgrounds and this is likely to be the largest source of false activity. A further problem related to false activations is selecting an appropriate set of training example in terms of quantity and balance. Finally, although reduced resolution has been a central theme to the different architectures, optimally selecting the degree of reduced resolution has not been fully investigated.

2.4 ANNs for Face Identification

Among the literature concerning computer vision, there are a selected few authors that have addressed the particular problem of using ANN systems for face identification. The following sections review some of the key papers in this area.

2.4.1 Face Identification Using Receptive Fields

The work of Rowley [Rowley *et. al*, 1995] is closely related to this research as their aims are very similar. An ANN based system is used to detect frontal views of faces, and also allows for a slight degree of *xy* head rotation. An input window (of size 20x20 pixels) is used to scan across an image over multiple resolutions. The resolution of the input window used, ensures that only the eyes, nose and mouth are contained within it. The input frames are pre-processed before being presented to the ANN. The pre-processing includes *correctional lighting* and *histogram equalisation*. It is discussed by Rowley [Rowley *et. al*, 1995] that the combination of these techniques does visually enhance the clarity of the image being presented as input to the ANN. However, metrics about the increase in performance with these techniques are not given. Although the effect of these processing methods may help the performance of the ANN system, positive trade-off between increased recognition and computational overhead is unclear.

The ANN is a partially connected three layer paradigm. The input frame is connected to three separate receptive fields, which are designed to extract different features from the input image. The receptive fields can be regarded as a cluster of hidden units connected to selected areas of the input in a structured manner, where each hidden unit is connected to a limited number of inputs from the input frame. From the 20x20 input, the image is divided into 4 10x10 receptive fields, 25 4x4 receptive fields, and 10 20x2 receptive fields. 39 hidden units are then connected each to a single receptive field.

Using localised hidden units that are partially connected in this fashion is a way of looking for features without being specific to the problem domain, i.e. not being defined necessarily for extracting face features. This approach allows position dependent information about the input object to be extracted. There seems to be no rule as to how many receptive fields are required. Although the structure of the receptive fields adopted is not particularly specific to face feature extraction, different input configurations may allow for better features to be extracted. A further problem associated with the definition of the receptive fields is that the input is not overlapped and a single feature may lie in more than one receptive field. The manner by which the structure and number and of receptive fields are derived is not discussed or even whether it is an optimum configuration specific to the problem.

A technique has been applied to broaden the range of training examples by applying a transformation to the input faces. This produces slight variations in *xy* rotation, scaling, translation, and mirroring ensuring that the ANN system has better generalisation capabilities. This seems to be a good way of creating more examples, but care must be taken so as not to affect the characteristics of the face.

The facial information for a positive response is tightly constrained allowing for little, if no, tolerance to translation invariance. Therefore, the receptive fields will respond only to faces in a limited region. All of the hidden units are connected to a single output unit. This unit indicates a two-class state of either face or background.

After the ANN has been presented to the image, a number of techniques are then used to group and filter the multiple activations generated by the ANN. One technique adopted is to use multiple ANNs, trained on different data, and combine the output by means of *AND/OR* boolean operators. No benefit is found to be gained from using multiple expert ANNs rather than increasing the amount of training for a single ANN since the experts are extracting the same information from the image. Multiple ANNs also have the disadvantage of increasing the computation for each input frame. A further alternative discussed is to increase the number of receptive fields. This would then allow different information to be extracted from the input.

An approximate accuracy of up to a 93% has been quoted for size invariant identification of faces from a given test set provided by Sung and Poggio [Sung and Poggio, 1994]. Unfortunately, the criteria used to determine correct identification are not given. The figure for number of false positives per input frame is quoted to be 1 in 27,416. Again, it is not clear as to whether this figure relates to an incorrect identification of a face, or typically how many frames are examined when searching for a face. Several other performance figures are given which use slight variations in performance against the basic model. These variations have either increased or decreased the number of successful identifications and number of false positives.

2.4.2 Face Identification using a Shared Weight Network

The work of Vaillant, Viennet and Fogelman Soulie [Vaillant *et. al*, 1993, Vaillant *et. al*, 1994, Viennet and Fogelman Soulie, 1992] presents another method for face identification. A shared weight MLP architecture is adopted and used to train two ANNs. The shared weight architecture contains three hidden layers.

The input window is sub-sampled in 5x5 blocks being connected to a single unit in each first hidden layer feature detector. The first hidden layer consists of a group of four clusters, each of which are fully connected to the input layer, but not to each other. The second hidden layer also contains four clusters where each cluster is only connected to each cluster in the previous layer. The third and final hidden layer, which contains four hidden units, is fully connected to all the clusters in the previous layer. Finally, the ANN has a single output unit denoting the presence/absence of a face.

The first hidden layer has been designed to act as a low-level feature detector, with the second and third hidden layers extracting higher-order features. Using selected connectivity is similar in principle to the receptive fields paradigm investigated by Rowley [Rowley *et. al*, 1995]. Both models examine isolated areas of the image, and the same argument directed at Rowley [Rowley *et. al*, 1995] of how the connectivity is organised is also not answered.

The ANN described is very large in comparison to other models and contains many network weights which makes the model processor intensive per forward propagation. However, the authors indicate the network uses some form of weight sharing, which, does not reduce the amount of weights in the ANN but does reduce the number of weights that can be adapted whilst being trained. How this weight sharing is integrated within the ANN and the justification for using this type of architecture is not discussed.

The vision system comprises two ANNs (both using the architecture described above), a shift tolerant and a shift intolerant model. The shift tolerant ANN is trained to give a peak response for correctly centred faces, a distance function for translated faces and a negative response for background data. The shift intolerant ANN is trained to give a peak response only for correctly centred faces, and for all other cases a negative response. For the shift intolerant ANN the training consists of translated face examples only, and no background data. How the ANN responds to distracter images is not clear, but having none in the training data would suggest an unknown response. This may be a particular issue when examining images, which is explained below.

Pre-processing consists of convolving each resolution image. The input window of the shift tolerant ANN input window is scanned across an image at seven different resolutions. Why the system has been limited to this number of scales, and whether this is a restriction of the model is not explained. The size of the input window is 20x20 pixels, but unlike Rowley [Rowley *et. al*, 1995] a full sized face is contained within it.

Highly activated areas are then passed to the shift intolerant ANN for further processing. Activations produced above a threshold by shift intolerant ANN are put forward as possible face candidates. If the shift tolerant ANN passes distracter patches to the shift intolerant ANN may produce a peaked negative response but may also as likely produce an unknown response. The unpredictability of the output of this ANN when presented distracter images passed from the shift tolerant ANN is the most likely source of false alarms.

Although it is quoted that there are fewer *free parameters* which is useful for training purposes, the normal mode of operation uses two large ANN architectures. Using the ANNs to process images is therefore slower, because of the number of connections they contain. Whether this particular paradigm was chosen to aid training or for some other reason, this is not explained. Unfortunately, only the methodology is described and no performance figures are presented. Although the approach is valid there is no means with which to compare the accuracy of the system against other comparable models.

2.5 ANNs for Face Recognition

Recognition is the natural successor to identification. In the problem domain of face identification, the next task is to perform recognition of the face identified. There is certainly more published research in this area, not only for ANN approaches but also for other methods. Although the problem of face identification is a difficult one, especially when dealing with novel instances, and large data sets, the problem is much more tightly constrained than that of identification. As such, the author believes identification to be a more difficult problem. However, both problems address similar issues of characterising the data in some manner. For this reason, some ANN approaches to recognition are described below.

The shared weight architecture used by Vaillant, Viennet and Fogelman Soulie [Vaillant *et. al*, 1993, Vaillant *et. al*, 1994, Viennet and Fogelman Soulie, 1992] is also adopted by Bouattour [Bouattour *et. al*, 1992] for face recognition. The system consists of a two-stage model. The first stage uses the shared weight architecture to provide high level features that describe a face. These feature descriptors are then passed to a classifier MLP ANN to perform the face recognition. An LVQ network was also tried as the classifier but produced very similar results to the MLP.

A database was created containing ten people using slightly different profiles and scales. Various lighting conditions were used, as well as uniform and printed backgrounds. The ANN was then trained with this database. Recognition of a face, like so many other systems, is a result of closest match. This means that a person not contained in the database at all, will be given an output to the person most similar to it.

From a 100 test images a best performance of 96% accuracy has been recorded. Images containing stronger head movements, in profile and rotation, gave a lower recognition rate of 89%. The ANN is relatively large considering only ten different people are used for analysis, each requiring many training examples. Unfortunately, the extent to which the ANN can reliably store further faces and what the saturation level is has not been fully explored.

Shimada [Shimada, 1992] attempts face recognition that relies on symmetry operations to detect the eyes nose and mouth in an image. Using the location of these features, the face is normalised, using a 2D affine transformation, to a set size and then is presented to a set of Gaussian receptive fields. The use of the receptive fields aims to compress the dimensionality of the data as an aid prior to recognition. The activities of these fields are used by a radial based function classifier to interpolate the values. A different recogniser was created for each person.

Garrison [Garrison *et. al*, 1990] reduces the dimensionality of the input (512x512) by using an MLP trained for compression. This is performed by training an MLP with the same output vector as the input vector but using considerably fewer hidden units. The output from the hidden layer is then passed to a single layer classifier with two outputs signifying face and gender. The system has reported 100% accuracy in classifying the face against non-face images for the training set, but a 37% error for gender using novel faces. Although the compression network forces a representation of features, why this should perform any better than trained MLP classifier with the same input is doubtful. Admittedly, a different representation of features is likely, but the two networks should be comparable.

2.6 Summary

An overview of ANNs has been described and their applicability to vision type tasks. Biologically motivated paradigms have been explored which address more basic vision problems. In contrast, higher level oriented vision solutions have demonstrated similar approaches. ANN methods have been developed for eye/face location and recognition, but questions remain as to their generality or applicability to other domains. This project seeks to clarify these issues, and develop techniques that are robust, and transferable to other vision domains.

Chapter 3

Location of Fixed Sized Objects

The following chapter discusses the development of an ANN approach to object identification. The aims are to develop an identification system tolerant to size variation, position, background and multiple occurrences. This is applied to faces as defined in the first chapter. A simplified framework is considered initially which addresses the problem of location of objects of a similar size. Issues regarding size variance are addressed in the following chapter.

3.1 The Problem Domain

This research aims to produce a general approach to object identification that uses "real world data". The term *real world data* refers to images that are captured from a camera, or other digitising device, from everyday surroundings. The data used is not contrived in any manner and is equivalent to that which might be easily processed by the human visual system. This means that the lighting conditions may be variable, the object to be identified may appear in any kind of surroundings and be at any proximity from the camera. Figure 1 shows a typical example of the type of image that the system should be expected to process, i.e. an image with complex natural surroundings where there are multiple occurrences of faces all at various positions.

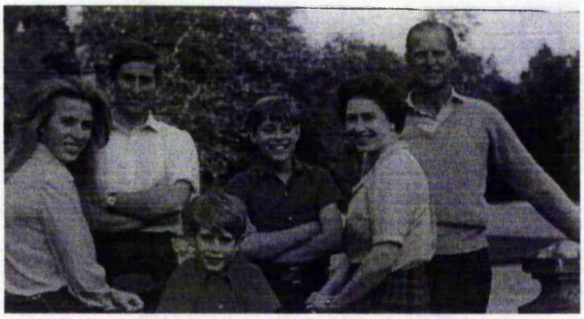


Figure 1 - Typical Real World Image

Faces have been chosen as a suitable object exemplar because of the variety and variability of faces. They can be regarded as having a *fuzzy*[2] definition and with no simple rules that can easily manage the degree of variety. Although faces have been selected and will be used to develop the techniques discussed in this thesis, the algorithms developed should applicable to any other real world problems that require similar image analysis.

The previous chapter discussed the merits for using ANNs for computer vision in comparison with other methods. The purpose of this research is to define a *general* framework for computer vision, which when applied to other problem domains (i.e. non-face identification) should only require different training data. The strategy will be the same.

An example where the method could be directly applied to another problem is a traffic monitoring system where there is a need to identify cars in a scene. Cars share similar properties to faces in that they contain variances in their attributes but also contain regularities. For example different models of cars are slightly different in size and shape, and are different colours, but they also contain characterising sub-features, i.e. wheels, wing-mirrors, lights and indicators, number plate etc. Obviously a car is a 3D object and one of the initial

[2] Although faces can be easily categorised as they all contain a set of primary features (e.g. eyes, nose and mouth), it is extremely difficult to construct a description with a formal set of rules that can cope with the variation in skin texture and colouration, gender, facial expression etc.

requirements for the traffic monitoring system would be to select the processing orientation of the car, i.e. side, front or back profile. This would determine which sub-features, as mentioned above, might aid identification.

The location and classification of faces in grey-level images are good exemplars because they contain many difficulties that are also present in many other domains. These include coping with large quantities of background distracter[3] information, object variation, position and size uncertainty. These problems must be overcome for computer vision strategy to be deemed effective.

3.2 Scope of the Problem

It is necessary to define the fundamental requirements and limitations of the model of what it can be expected to process. This allows suitable data to be selected for the problem and also allows suitable qualifiers to be selected that allow the performance of the system to be measured. Defining the scope of the problems also determines the generality of the system and also its suitability to other problem domains.

The principal aim of the system design is to perform accurate object identification without the need to process the entire image at full resolution. This has the obvious benefit compared to other systems such as Kwon and Lobo [Kwon and Lobo, 1994], that it is more efficient in terms of information processing. As well as the benefit of processing less information, reducing the resolution also reduces the variability in the data, and depending upon the reduction process applied can also remove noise. Processing at too low a resolution has the disadvantage of loss of information. The degree of reduced resolution is investigated such that processing is performed at the lowest possible resolution. This is discussed further in section 3.7.

[3] Distracters are parts of the image not containing the object to be identified, which may lead to false alarms being produced.

The collection of faces used for object location are those appearing approximately full on to the camera. Initially, discussion will be of faces of a similar size until the problem of size invariance is addressed in the next chapter. Faces used will *not be* occluded, but *may contain* glasses, beard or moustache. Slight profiles (up to approximately 15 degrees, and not excluding any of the primary features of the face, e.g. eyes, nose, mouth) and head rotation (again up to approximately 15 degrees) are allowed, but the system is not expected to perform in extreme cases of deviation from a "full-on, upright" position. The minimum size of face that can be successfully identified will be determined by the investigation into the minimal processing resolution studied in section 3.7.

The system should be able to cope with variability in the size of the objects to be identified without any prior knowledge about the input image. To simplify the initial problem faces of a fixed size are addressed initially in this chapter. The simplified framework is then extended in the following chapter to address the full problem of size invariant detection.

It is intended that the scope of this work will not address the issues defined below:

- ➢ Large degrees of object rotation (i.e. greater than 15 degrees in any plane).
- ➢ Partial occlusion of the object or its sub-features (except for natural partially occluding items such as beard and glasses).
- ➢ Object tracking in temporal sequences. The analysis is on static images only. Temporal analysis uses different strategies and cues not applicable to this computer vision approach.
- ➢ Recognition. This is a further stage in visual processing which is too large a task to be tackled as well as identification. Recognition is also a further task more appropriate to the problem of faces rather than a general task required by most computer vision problems.

These elements have been excluded as essential requirements for a successful computer vision strategy as they do not preclude the functioning of a basic, generic computer vision model. However, if the functionality of the system were to be extended then these problems would be logical choices to be considered.

3.3 Performance Measurement

The problem domain has described an object identification system, and the scope of the problem has been defined as to what it can be expected to process. However, some criteria are required to measure the success of the identification process. These two criteria are for the accuracy of identification for:

- The position of the object. The centre of the identifying object (i.e. a face) will be specified as the mid point between the centre of the eyes and the tip of the nose. This *xy* position will indicate the centre location of the face, and will be called the *focus point*. The position of the focus point is allowed to deviate from the target position for a limited number of pixels in any direction. This will be referred to generally as the *distance-error*, and specifically for face analysis, the *face distance-error*. The face distance-error is the distance between the focus point identified and the actual face location. This measure is less important for faces identified within the distance threshold, but more importantly determine the faces unsuccessfully identified (i.e. those focus points that lie further away than the distance threshold allows). Any face identified within this distance threshold is regarded as a successful identification (face or true positive). How the distance threshold figure is derived and the reason for its value is described later in this chapter. (Section 3.14).
- The size of the object. The area identified within the image by the system shall be called the focus area. The focus point is the centre of the focus area. The size of the focus area with reference to the size of the face will be within a defined *size tolerance error*. This will be defined such that the principal components of the object identified must be within the bounds of the focus area. This chapter deals only with faces of a fixed size, and will therefore not

use this metric until the topic of size variance has been introduced in the next chapter. This will also discuss a derived threshold value for the size tolerance error.

The following sections describe an approach to object identification. This chapter concludes with a performance assessment for fixed sized faces, and uses the metrics described above.

3.4 Approach to Object Identification

A particularly difficult problem that must be addressed is selecting a set of features that captures the information appropriate to the task. According to Vaillant [Vaillant *et. al*, 1994] this particular aspect has not been completely addressed. Many of the approaches discussed in chapter 2 use the sub-features of the face as the primary aid to identification. This is quite acceptable except that the lowest processing resolution remains high.

It can be argued that to identify possible faces does not initially require full detail of the sub-features, but a high level holistic view. The relationship between the sub-features is implicit as the object is processed as a whole. The research presented in the previous chapter did not seem to exploit this idea.

A novel approach taken by this research is to use a two stage approach to identification. This will comprise initially of low resolution holistic detection for the object(s). This method allows the useful information of the sub-features to be used without the necessity of a high resolution. To ensure correct and more robust object identification, selected focus areas are processed at a higher resolution as used in other approaches.

In the two-stage approach, the first stage will be referred to as attention-focusing. This stage is used to determine areas of interest. Only areas of interest identified as possible face candidates are processed by the second stage. The second stage of processing will be referred to as classification. Both stages comprise the identification process.

The benefit of a two-stage approach is that more processing can be done at lower resolutions, and also only selected areas identified by the attention-focusing stage are processed further. This enables less of the image to be processed at a higher resolution, which reduces the overall amount of processing required.

A lower resolution requires less information, which allows an ANN architecture to be smaller in size. Having less input being fed to an ANN also reduces the degree of variability of the pixels which benefits the training and generalisation capability of the ANN.

Using a holistic approach may require that other features are considered as an aid to image segmentation. An example of discriminating features not used in other systems are, the hairline or the nature of the shape of the head, where typically only the facial features of the head are used. It is of the opinion of the author that in a holistic approach to scene segmentation the features surrounding the face are as important as those contained within it. These ideas are investigated below.

3.5 ANN Architecture

It has been already chosen that the principal components of the computer vision model will consist of an ANNs. Therefore, a fundamental understanding regarding the concepts of ANNs are assumed. However for reference Appendix A provides a brief introduction to the key aspects of ANNs, how they are constructed, the methods of "learning", and the different ANN architectures available.

There are a number of questions that need to be answered in order to choose the appropriate ANN architecture, learning algorithm, training data etc. To achieve this, it is best to analyse the problem and identify the task placed upon the ANN, i.e. what are its inputs, and what are its outputs. From the answers to these questions it is possible to develop an informed selection of the appropriate ANN for the task.

In order to use an ANN for scene segmentation it is necessary to teach it what is a face and what is not. This is an extremely good example for requiring a supervised learning algorithm. A multi-layered Perceptron using the *backprop* learning algorithm is the most used and successful paradigm configuration for most problems. It has also been established that it has the ability of solving image processing problems [Dayhoff, 1990]. The *backprop* algorithm is a supervised learning algorithm and has the ability to learn any function and in particular non-linear problems makes it the most obvious choice as the foundation on which the attention-focusing ANN search strategy is developed.

The process of the attention-focusing stage is to search a reduced resolution image, and identify areas of interest on the original image where it is likely that an object of interest is located. A representation is therefore necessary to configure the ANN to receive reduced resolution input and output areas of interest in an easily interpretable form. Two approaches to this problem are investigated. These are a feature map representation, and a moving window ANN.

3.5.1 Feature Map Representation

With this method the input layer corresponds to the maximum size of the input image and the output layer corresponds to a *feature map* (where each output unit is a focus area). An example model size for this is:

- 100 x 100 (10,000) input units. This determines the maximum reduced image size that can be presented to the system. The size of the input is arbitrary, but should be a reasonable size given that it represents the reduced image. The greater the reduction the more of the real world is contained within the input frame.
- 100 hidden units. Having a large number of input and output units requires a suitably large number of units for the hidden layer. Unfortunately, there are no reliable formulas by which (given the size of the input and output) the optimum number of hidden units can be calculated. Other factors include the quantity of training data, noise, features in the data, etc. Using a hidden layer size of 100 units means there is a 100:1 input to hidden ratio. This ratio may be too large, but increasing the number of hidden units would produce (in neural network terms) a particularly large ANN model. (1 further hidden unit would produce 10,000 extra input weights).
- 25 x 25 (625) output units. As the input is image based, the size of the output must correspond to the input layer dimensions. However, the output does not necessarily need to be at the same reduced resolution. The size of the output is 1/16th of input size. Therefore, each output unit represents an area of 4 x 4 pixels of the reduced resolution input image.

This ANN topology has several disadvantages. The number of weight connections is relatively high (approximately 106,250) making it expensive in terms of computation and training. It has also been shown that models with many connections are poor at generalising [Atlas *et. al*, 1989] due to the massive amount of training required. A large training data set is needed in order for the network to activate the feature map appropriately. This is because several examples of faces in **every** location would be necessary for the network to generalise for *position only*. Many examples of faces and non-faces in every location would then be necessary to generalise for faces.

Initial experiments in training the example model were extremely poor. As the feature map was smaller than the input layer size, this also incorporated an increased positional error, including the error from the reduced resolution of the input image. Increasing the size of the feature map to a 1 to 1 correspondence would have also increased the burden on training and its generalisation capabilities. Due to these limitations an alternative approach to attention-focusing is considered.

3.5.2 Window Attention-Focusing ANN

An alternative method to the feature map representation for identifying focus areas is to use a sliding window approach. This has been inspired by Waite [Waite, 1991] for eye location within faces. The input layer for the ANN can be regarded as a single frame where only portions of the reduced resolution image are presented to it. The output for the ANN is a single output that indicates the presence or absence of the object.

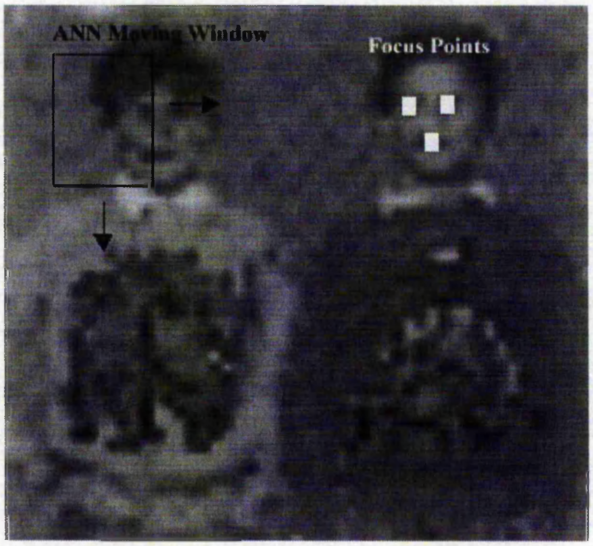


Figure 2 - Attention-Focusing ANN Window Method

To identify any faces within the image the ANN input window is scanned[4] across a low-resolution image and *focus points* are generated at positions where the *input frame* has been trained to indicate the presence of a face.

In contrast to the previous architecture, this kind of model adopted for face location is small:

- 10-15 x 15-25 input units. The input layer is proportional to the size and shape of the search object. This has the advantage that the image area that can be covered is unlimited.
- For an input layer of this size, only 10 to 30 hidden units are needed to give sufficient generalisation capability.
- There is only a single output unit, to indicate whether a face is present.

The architecture for this ANN model means that there are less connections (typically 4020 weights for a ANN with a window size of 10x20 pixels and a hidden layer of 20 hidden units) than the previous ANN which allows for easier training and better generalisation capacity. Although the size of the ANN is smaller than the feature map ANN, the sliding window ANN has to be presented to the whole area of the reduced resolution image. The accuracy required and the degree of reduced resolution determine whether it is necessary to present the ANN window to every location within the reduced resolution image.

3.6 Sampling the Image

The input images are sampled with an eight-bit monochrome intensity resolution. Although the original images are in colour it has been decided that classification does not require colour information, and the extra information required may make it more difficult to train an ANN. The majority of all research regarding image

[4] Although a process of scanning the input window across the image has been described, focus points can instead be found by processing the entire image in parallel. This would require the ANN model to be replicated and each ANN input window placed at a different position on the low-resolution image.

analysis, and particularly faces have used grey level images. None of the models discussed by the literature in chapter 2 have used colour information.

3.7 Optimising the Spatial Resolution

The idea of processing information at a reduced resolution in order to reduce the quantity of input data has been considered before by Evans, Vaillant, Marsic and Micheli-Tzanakou [Evans *et. al*, 1991, Marsic and Micheli-Tzanakou, 1992, Vaillant *et. al*, 1993]. There are several advantages to using reduced resolution input data. These include:

- A reduced amount of input data to the ANN means that finding face candidates in real-time is more feasible.
- Only selected areas of interest need to be processed at a higher resolution.
- Low resolution produces a decrease in the variability of input possibilities. This leads to easier training for an ANN as the variety of examples becomes less complex.

The input images have been digitised and sampled at 100 dpi. Faces used for training are all reduced to approximately the same size. The faces are reduced such that the entire head is included within the ANN input frame. The size of the faces digitised vary, but are typically 2" x 3" in size. This results in a 50 x 75 pixel area for the complete head and partial surroundings when sampled at this resolution.

Experiments to find the lowest possible processing resolution where reliable classification could be performed have been investigated. The approach taken is to use an incredibly high reduction that is unlikely to give any reasonable classification and then to gradually increase the resolution. It has been found from these experiments that for faces of size 2" x 3", the reduction in image size is optimal at 1/25th of the original area. Visually, this also corresponds to the threshold where faces can be reliably detected from human analysis. A typical face at this resolution will have a width of 8 and a height of 12 pixels. The size of the ANN window is set to slightly larger than this (11 x 16 pixels). This is because this method is based upon a holistic approach and it is believed that some of the surrounding detail helps to classify the face, e.g. hairline, shoulder tops etc.

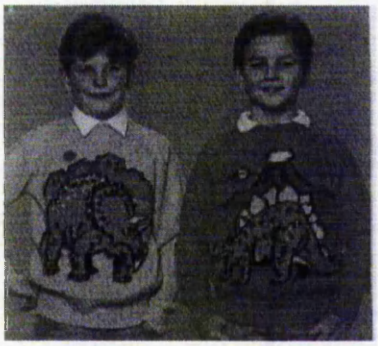


Figure 3 - Full Resolution Image

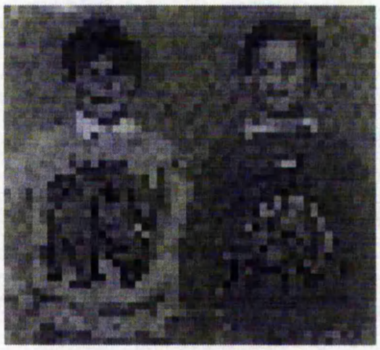


Figure 4 - Reduced Resolution Image

As can be seen from the reduced resolution image shown in Figure 3 and 4, the ANN is being presented information at a resolution close to the border of human recognition. In comparison with Rowley, Vaillant, Viennet and Fogelman Soulie [Rowley *et. al*, 1995, Vaillant *et. al*, 1993, Vaillant *et. al*, 1994, Viennet and Fogelman Soulie, 1992] the input frame dimensions are smaller, and also the frame contains the whole head. This suggests that faces at a much lower resolution can be successfully identified than previously recorded. This may be due to the holistic approach adopted. The benefit that an initial lower processing resolution has been found means that less processing in an image is required at the classifier stage than those comparable systems.

There are various means by which the image can be reduced. Two methods for modifying the spatial resolution were tried: *N-pixel* sampling and pixel averaging. For fixed sized faces used in the analysis above the method used to reduce the image is not significant in terms of ANN classification. However, for size invariant object analysis this becomes important. The effects of different sample methods are discussed in more detail in chapter 4.

3.8 Image Pre-Processing

It has already been identified that the first image pre-processing task performed by the attention-focusing module is image reduction. As well as this processing, there other methods which might be useful in terms of improving the quality of the data, especially at low resolutions, to help ANN classification. As the content of images vary significantly, unless particular pre-processing methods are found to improve the classification process, then ANN processing should use the reduced image directly. Other methods that may improve the image and impact on ANN classification are evaluated below. These include *normalisation*, *histogram equalisation* and *scaling*.

3.8.1 Histogram Equalisation

Histogram equalisation is a standard function to improve the quality of images captured with variable lighting. It has the effect of separating the intensity values more evenly throughout the intensity range, resulting in a general improvement in contrast.

The sample images used in training and testing all had a reasonable spread of levels in the intensity range. Histogram equalisation did not significantly improve the performance of the attention-focusing ANN. The extra computation could not therefore be justified. This method maybe particularly more pertinent for vision problems where image capture is not able to rely upon a good lighting environment.

3.8.2 Normalisation

Normalisation is a particular method of non-linear scaling. It is applied to each window frame to ensure the same contrast across the image, as described by [Hutchinson and Welsh, 1989]. A fundamental property of normalisation is that the total magnitude of each input frame is equal to one. However, even after normalisation the relative difference between inputs remains constant.

Applying normalisation should ensure that areas of different pixel intensity have no more significance than other areas. This may be particularly important when the ANN is presented with areas in the image of high intensity. The possible cumulative effect of high value inputs may cause the ANN to activate incorrectly.

The normalisation function is given in Equation 1.

$$a_{i\,new} = \frac{a_i}{\sqrt{\sum_{i=0}^{i=num\ inputs} a_i^2}}$$

Equation 1 - Normalisation

3.8.3 Linear and Non-Linear Scaling

Scaling is similar to normalisation, where the magnitude of the input values are reduced by some function. Reducing the size of the input values before they are presented to the ANN removes the burden of scaling via the learning algorithm. Two different scaling methods have been evaluated, a linear and a non-linear method.

The linear method simply divides all inputs by the maximum intensity value determined by the magnitude of the intensity range (Equation 2). Using this scaling method for pre-processing had no effect on the learning of the ANN.

$$new\, a_i = \frac{a_i}{\max intesity\ value}$$

Equation 2 - Linear Scaling

The non-linear scaling function scales the input values according to the magnitude of the sum of the squared input. This is shown in Equation 3.

$$new\, a_i = \frac{a_i^2}{\sum_{i=0}^{i=num\ inputs} a_i^2}$$

Equation 3 - Non-Linear Scaling

This function is similar to normalisation, except that the sum magnitude for all inputs is not equal to one. Figure 5 shows that applying normalisation to the input data does not in fact improve the representation of the input data, but rather has the effect of making some patterns, in terms of Euclidean distance, more similar to each other. The has the effect of making it more difficult for the ANN separate particular patterns which may belong to different categories, ie. faces and distracters.

Non-linear scaling does not force the input vector to lie closely to other dissimilar input vectors, which can happen with normalisation. The performance of applying normalisation to the ANN input compared less favourably than with the non-linear scaling method. Applying non-linear scaling to the input data as opposed to the other methods improves training. Training the attention-focusing ANN, the descent of the RMS was found to be more reliable compared with no pre-processing and simple linear scaling (Equation 2).

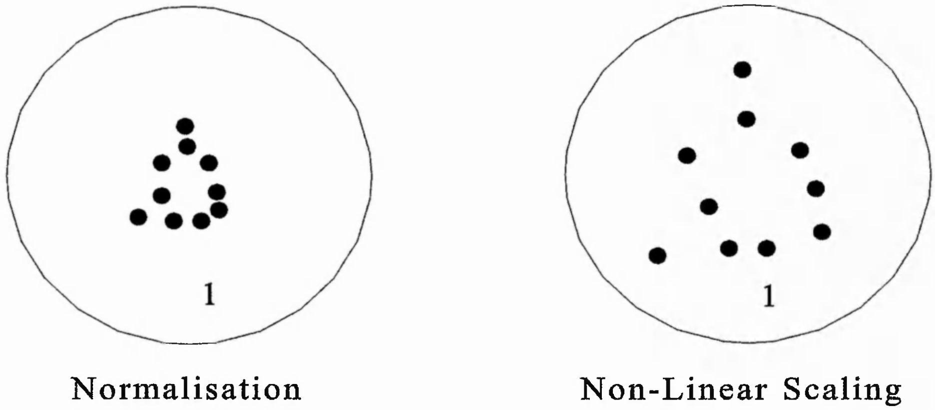


Figure 5 - Comparison of Transformed Input Data in 2D Space

3.9 Output Response

The previous section discussed how the image data is manipulated prior to being presented to the attention-focusing ANN. This section covers how the attention-focusing ANN is trained to respond to the target object. A supervised learning algorithm is adopted for the training, and ideally a response is required to indicate when the ANN input window passes over a face. In order to achieve this, particular issues need to be considered. These are described below:

3.9.1 Determining the Output Range

Generally, most target output activations lie in a defined range, typically between 0 and 1. The most common transfer function used is the Sigmoid function which squashes all input values to lie in this range, and is the one adopted within the backprop learning algorithm. Alternative functions such as the radial basis functions can be employed instead which produces an output between -1 to +1. Radial basis functions differ from the Sigmoid, as the output is cyclic.

The target outputs used for training are given the range 0.1 to 0.9, and these adjusted boundary values have been chosen because of the properties of the

Sigmoid function. The values have been chosen because it is less likely that ANN learning will fall into a local minimum. The reason for this is that the weight update is always non-zero, even at the boundaries (0.1 and 0.9), whereas at boundaries 0 and 1 this can produce near zero weight updates. Therefore, it is much more difficult to move weights from this response, and requires many more iterations to do so. Highly incorrect classifications are also helped because the boundary values are closer to the steepest slope of the function, and hence it is easier to move the weights away from the incorrect weight space. Figure 6 illustrates how the gradient at both ends of the Sigmoid curve tends to zero, whereas at the activations of 0.1 and 0.9 it is steeper.

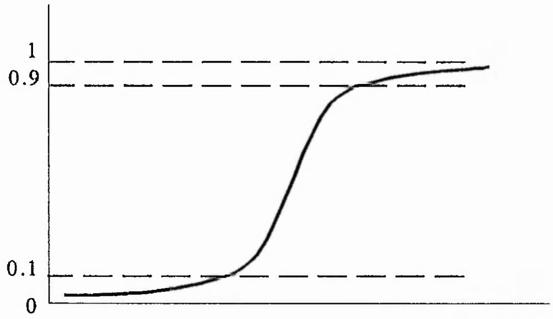


Figure 6 - The Sigmoid Function

3.9.2 Target Output Activation Function

The moving window attention-focusing paradigm contains only a single output unit. This is designed to indicate the current state of the input frame. Ideally, within an image there are two distinct categories; input frames that are positioned exactly centred on a face and those that cover entirely a selection of background. Unfortunately, with a moving window approach there is a fuzzy boundary where input frames contain a degree of both categories. For the two distinct cases the target response is simple, but for the other cases a different approach is required.

Some method or function is required given an input to determine the appropriate output response of the network. This will be referred to as the target output activation function, and should be confused with the transfer function discussed earlier. The following section describes different target output activation functions considered to for the different possible input scenarios.

3.9.2.1 Distance Activation

One possibility is a linear distance function Figure 7, where the value of the output is relative to the proportion of the face contained within the ANN input frame.

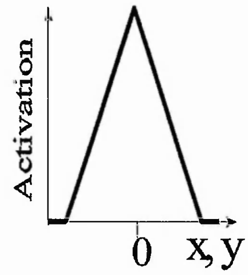


Figure 7 - Distance Activation Function

This is expressed by Equation 4:

$$t = 1 \; \frac{\sqrt{(xfacepos \;\; x\,)^2 \; + \; (yfacepos \;\; y\,)^2}}{window\ size}$$

Equation 4 - Distance Activation Function

This method not only trains the ANN to detect a face, but would also incorporate the percentage of a face in the input frame into the magnitude of the output value. A similar technique using a Gaussian function has been investigated by Vaillant [Vaillant *et. al*, 1993].

Applying the distance activation function the attention-focusing ANN is unable to learn the problem because of the ambiguity of representation in the output between positional information and recognition uncertainty. This is due to the function creating a spread of activation values, with no distinct boundary between the two categories. It is difficult to extract meaningful information from the output, as a high activation might indicate either a large percentage of a face within the input frame *or* a high uncertainty as to the class of input.

3.9.2.2 Modified Distance Activation

The distance activation can be modified to try to improve the distinction between categories by considering an input frame that contains less than half a face as a distracter. This removes the difficulty of classifying those input frames containing very small parts of faces. Any pattern that gives an output of above 0.5 is counted as a face frame.

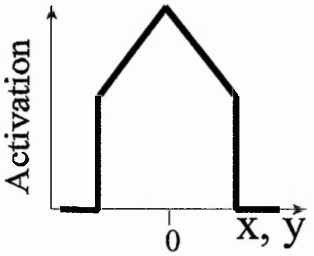


Figure 8 – Modified Distance Activation Function

Using the modified distance activation function is also unsuccessful at learning the problem producing a bias towards an output activation of 0.1. An explanation of failure to generalise may be because the *average* target value is biased towards distracters; i.e. the face category contains a linear range of values, whereas the background class contains a single value. This produces a larger number of 0.1 values in comparison to any other target response.

Although a function to determine the target output activation was successful in Vaillant's work [Vaillant *et. al*, 1993], the ANN architecture adopted here is not capable of generalising to this kind of representation. The relatively simple ANN architecture used makes it difficult to learn the function without increasing the size and complexity of the ANN as described by Vaillant [Vaillant *et. al*, 1993].

3.9.2.3 Simple Two Category Activation

Due to the difficulty in training the ANN to give an output that combines recognition with distance, a simple binary activation is evaluated, i.e. an activation that gives outputs of only either 0.1 or 0.9. This presents a problem. What should the target activation be when only part of a face is present? To overcome the learning problem suffered by the two target activations (Figure 7 and Figure 8), the area that contains a high proportion of both categories (i.e. partial faces) is ignored by the function illustrated by Figure 9.

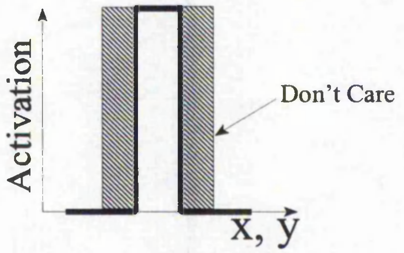


Figure 9 - Two Category Activation Function

Frames containing a mixture of face and background are excluded from training of the ANN as "don't care" states, so as not to lower the boundary between the two classifications. When an image is searched the ANN can generalise itself to the closest classification category. It is not crucial whether the ANN classifies such images as face or distracter because either is acceptable. This approach reduces the requirement for a large positional tolerance to be learnt by the ANN.

Frames within 3 pixels of the best face frame have also been included where they also contain full facial information and these have been given a target output of 0.9. This has a benefit of creating slight variations in training data but using the same face. It should also provide a slight positional tolerance to be learnt by the ANN.

3.10 The Data

Probably the most important issue after deciding on the type of ANN paradigm is the training data. Selecting appropriate data and how this is presented effects how successful the ANN generalises. The issues that must be considered are:

- The quantity of examples required for the network to distinguish between faces and the background.
- The ratio of different types of input in the training set.
- The ordering of examples in the training set.

3.10.1 Quantity and Quality of Examples

Initially, the training set consisted of faces from passport photographs, and miscellaneous distracter images. Detection was found to be in excess of 95% accuracy for novel passport face photographs, but less than 40% accuracy for faces within real world images. Classifying faces within passport photographs is easier as the size of the input window allows for a small portion of the surroundings to be included. For the passport photographs the surrounding area of the face is a simple constant background which produces a clear separation of object and background. The failure of the attention focusing ANN to respond to faces within a natural environment is not surprising given that complex texture types surround the face were included within the training of the attention-focussing ANN.

The number of false activations on real world images was also high. This was attributed to the ANN being presented with too little variation in background data.

In response to this, the training set was modified to comprise of only real world images. This allows faces to be situated against natural surroundings, and also allows a greater number of background examples to be included.

The type of possible backgrounds a face may appear against is almost infinite. Therefore it is necessary to select as much of a wide selection of varied input backgrounds as possible. Although faces vary, the possible variations compared to the background are much less. The number of examples required for the ANN to generalise is a difficult answer, given the number of possibilities. The larger the hidden layer the greater the number of patterns the ANN is able to generalise to. Even so, there must be a trade-off between the size of the ANN and the number of false alarms it produces. As the attention-focusing stage is only the precursor to classification it is only necessary to achieve a moderately high classification performance. However, the attention-focusing should achieve a reasonable success such that computation by the attention-focusing stage reduces the overall computation of object identification.

3.10.2 Balance of Input Examples

A further consideration was the balance in the number of faces to distracters. A balance is necessary in order for each input frame to compete equally without the output value having a bias to one type of image category [Evans *et. al*, 1991]. Most pictures contain a greater quantity of distracter information compared to face information. To achieve a balance, faces are repeated to give an equal face to distracter ratio.

3.10.3 Ordering of Examples within the Training Data

The ordering of faces and distracters within the training set was found to be very important to the success of the ANN learning the problem. With a highly unordered data set that contains large numbers of one category followed by another, training performance was found to be poor. This was due to the learning fluctuating between the two categories without fully learning either. In order to overcome this, the ordering of the examples presented to the ANN follows the sequence of a distracter followed by a face.

3.11 Choosing an Appropriate Step Value

Once the ANN has been trained, it can then be used to search an image to generate focus points. There is no prior knowledge about the position or number of any objects in the image, and therefore it is necessary for the ANN input window to search the entire image. The step size for the moving input window frame needs to be carefully chosen.

Having a moving input window frame that has a minimum step size of 1 produces multiple focus points when the frame is passed over a face. A large step value has the advantage of reducing the number of input frames to examine, and hence decreases the search time; but too large a value, and the frame may miss an object. A further disadvantage of any increase in step size is that the search will not produce clusters that can aid identification.

It is desirable to differentiate as much as possible between spurious activations and positive face identification before the focus points are used for classification. This can only be achieved by using a small step size and post processing the focus points created. This is because in the search, faces produce focus point clusters[5], whereas false positives are more likely to produce single unconnected activations.

3.12 Focus Point Post Processing

Once the image has been searched, it contains a map of focus points. With an equal balance of training examples, activation below 0.5 is determined to be a distracter, while any value above 0.5 is considered to be a positive candidate. A higher threshold value may be chosen, but may eliminate faces that produce lower activations.

[5] If the ANN is trained with a target activation function that is tolerant to positional shift, multiple activations will be generated in the area of the target object.

As one object generates several focus points, these need to be grouped before the classification stage. The aim of focus point post processing is to reduce the number of *focus points* to a reasonable level to create *focus areas*, without removing possible focus points corresponding to faces. The following sections discuss possible methods for transforming the focus points to focus areas and the advantages and disadvantages of each.

3.12.1 Distribution of Activation Values

There are two important components that can be used to describe a cluster of focus points. These are the number of focus points per cluster and the total activation of the focus area. As stated previously, focus points are those activations from the ANN that are over 0.5. Focus areas could be ranked in order of their activation, but it is unclear whether a focus area containing a few highly activated focus points is more significant than a larger number of reduced value focus points. For example, a novel texture may be a little similar to a face at low resolution to sufficiently activate the ANN at multiple positions of the face. However, the same novel texture may be dissimilar to a face in most positions of the texture except at particular positions and orientations where there is a strong response from the ANN.

The unpredictability of the response of the ANN for all novel textures has determined that the information regarding the number of focus points in a cluster, and the some activation of the focus area is not fully utilised. It would also seem reasonable that there should be a maximum number of focus points that can describe a face. For most cases this will hold true, except for instances where faces are very close together, which may produce a merged cluster. The grouping algorithm can quite easily become very complicated when considering all possible eventualities. However, two reasonably simple methods are discussed in the following sections:

3.12.2 Focus Area from the Mid-Point of Extremities

This is the most basic grouping algorithm. (See Figure 10). Contiguous focus points are grouped together. A record of all extremity positions of each group is kept. From these values, the centre point for the focus area is calculated by taking the mid-point in both the *x* and *y* directions.

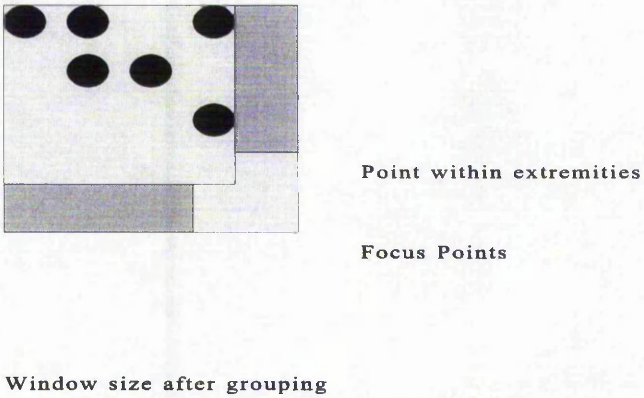


Figure 10 - Extremities Grouping Method

The advantage of this method is that it is easy to calculate and therefore relatively fast. Its main drawback however is that it is very crude in calculating the centre point accurately when taking into account the shape of the focus points within the group. It may also, in some circumstances, group together unconnected areas that are found within the growing window. This is very difficult to avoid without using the contour of the shape to determine which points should be included.

3.12.3 Focus Area using the Centre of Gravity

Assuming that each focus point is a binary value, the centre of gravity for a 2D shape can be calculated to find the mid-point of the focus area. This method should produce a more accurate centre point for non-uniform shapes, i.e. those that are non-rectangular. It less likely that the ANN trained with a position tolerant target activation function will generate shapes of this kind.

The above method does not take into account the weightings of the focus points. It can be extended to calculate the centre of gravity for three-dimensional shapes.

In this case, the activity of the focus points is used to describe the third dimension. Using this method is an even more accurate means with which to find the focus area centre point.

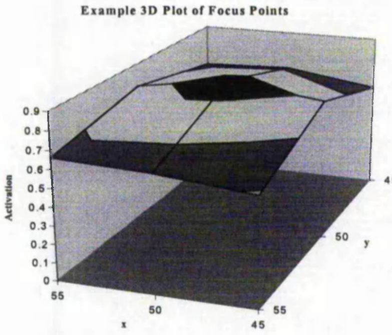


Figure 11 - 3D Plot of Focus Area used to Calculate the Centre of Gravity

Both centre of gravity methods add a degree of complexity to the calculation which may not be necessary if simpler methods produce satisfactory results. These methods only describe a means to calculate the centre of the cluster, and not how the focus points are grouped. Therefore, a method is required to obtain the cluster information prior to centre of gravity analysis. A technique such as the extremities grouping function or some other would be necessary to determine the points included in a cluster.

3.12.4 Focus Area using a Moving Window Grouping

A simple but alternative method to group the map of focus points is to use a moving window across the image. This is applied in the same manner as the ANN input window. However, the function of it is to simply count the number of focus points found within the grouping window. A window that contains more than a single activation are transformed to a focus area.

There are two parameters that define the accuracy and number of the focus areas produced. These are the grouping window size, and the grouping window step.

The values of these have no relationship with the ANN input window. To ensure there are no false negatives the parameters need to be chosen carefully. A moderately sized grouping window and small step value should ensure all possible face positives are found. A larger a grouping window size and step will reduce the number of focus areas produced, both false positives and false negatives. An analysis of different combinations of values is described later.

In comparison to the extremities grouping algorithm, this is even simpler but is guaranteed to catch more face positives assuming that the window size and step value are optimal. The main drawback of the grouping algorithm is that it produces a general increase in the number focus areas.

3.12.5Focus Point Filtering

Due to the large possible background pattern variation, the ANN will sometimes indicate the presence of an object when the input is just background. Filtering is a method to reduce the false activations of the ANN. A basic assumption is made that focus areas containing a single focus point should be removed. Filtering in this way is empirically justified as real objects produce clusters, and it decreases the number of false positives [Viennet and Fogelman Soulie, 1992].

It may be possible to increase the threshold to pairs of focus points for filtering out false positives, but to do this may result in objects being pruned. Because the emphasis is ensuring that there are *no* false negatives, there will be *some* false positives. As a single point can be the only guarantee of false activity[6] these are the only activations that are removed.

[6] An ANN that uses a position tolerant target activation function is trained to activate at multiple positions of the object. Therefore if an ANN is expected to produce a response at one position, then the ANN can also be expected to produce a response at another. An assumption is made that no target position is greater than another.

3.13 System Overview

The previous sections have described ideas and issues that have impacted on the resulting design of the attention-focusing ANN. This section brings together the solutions proposed for these problems and describes the subsequent architecture developed and operation of the attention-focusing ANN. The next chapter uses the paradigm developed, and extends it to consider the problem size invariant object recognition.

The system is capable of locating objects of a fixed size at a relatively high speed through the use of low resolution and minimal image pre-processing. An ANN input window is scanned across a low resolution image and *focus points* are generated at positions where the *input frame* has a high probability of containing a high percentage of the trained object (ie. a face).

The image is reduced to 1/25th of its original size and a non-linear scaling function is applied to each input frame. The input frame size is 11x16 pixels and there are 15 ANN hidden units. This reduced resolution and the appropriate ANN size have been found to give best performance from studies performed.

As the centre of the input frame nears the centre of an object, multiple focus points are created within the area of the face. Focus points are grouped together to form *focus areas*. Single ungrouped focus points are filtered out as likely errors. Focus areas generated are then available for further processing.

3.14 Performance Evaluation

To determine the effectiveness of the system described and the variations on this model, it is necessary to evaluate how well the system works on unseen data. Section 3.3 described the criteria by which the system, and in this particular instance the attention-focusing stage, is to be measured. However, appropriate values were not given to the error thresholds that include the face distance-error and size tolerance error. Numbers and justification for the values derived are discussed in the following section.

3.14.1 Criteria For Evaluation

The face distance-error is determined by whether the focus area encapsulates the main components of the object. For successful identification, the face distance-error should not be so great as to exclude the primary features (ie. eyes, nose and mouth) contained within the frame. From the author's visual examination, it was determined that the face is able to receive a shift of up to 4 pixels in any direction at the reduced resolution and still maintain that the primary features are contained within the focus area frame. Re-scaling to the full resolution relates this distance to be equal to 18 pixels in the original image. The threshold value of 18 pixels determines the boundary value for successful location. As this value is resolution independent (assuming that the same image capture method is used), it is also applicable to size invariant attention-focusing and classification analysis.

Although it is desirable for the face distance-error to be zero, using reduced resolution introduces a quantisation error. The degree of error is dependent upon the amount of image reduction determined by the scaling factor (in the case of the attention-focusing ANN the images are reduced to 1/25th of their original size) and will therefore usually position the focus area slightly away from the full resolution true centre (of up to 5 pixels in either the x or the y axis).

Using the threshold value described above the performance of the attention-focusing stage is measured by:

- The number of faces that have failed to be located within each scene (*false negatives*).
- The number of false focus areas (*false positives*) that are generated for each image. It is unrealistic to expect the system to have no false positives given the reduction in resolution and also the vast amount of background data that is inevitably unseen by the ANN.

3.14.2 Analysis of Results

All of the results discussed in this section have used an attention-focusing ANN trained on a set of 18 images that include a total of 30 faces. A set of 10 test images that include 10 faces have been used to analyse the performance of the trained attention-focusing ANN. The test set was chosen to be a similar representative of the type of images used to train the ANN, although it can be almost guaranteed that novel objects and textures will be subjected to the ANN due to the diversity of natural images. Particular novel textures can be easily identified as those which the attention-focusing ANN produces large numbers of false positives.

The images captured are all variable in size, and represent varying amounts of data to process per image. Using the reduction factor and input frame size discussed, this results in a total of 19,447 separate input frames available for analysis (from the 10 test images) by the attention-focusing ANN. Of these frames, 528 positions indicated a positive response. This shows a relatively high ratio of frames scanned to focus point produced. However, this figure does give a good measure of performance, as each image contains a face, which in turn comprises of several different frames. The most important aspect that is observed from the direct output of the attention-focusing ANN is that a focus point has been produced within the distance-error threshold for all the test cases. A more reliable measure of performance is to use the output from the grouping algorithm, as this is passed on to the next stage of processing.

The table below compares the performance of the *extremities* and *window* grouping methods. Both use the same focus point data produced by the attention-focusing ANN.

Window Grouping Method					
Box Size (Pixels)	**Step Size**	**FA**	**FP**	**FN**	**Average Distance-Error**
3	3	243	225	2	9.97
4	2	908	849	1	10.23
5	4	293	277	2	9.53
5	5	178	168	2	9.51
6	3	648	616	0	9.85
6	5	247	236	2	10.70
6	6	160	153	3	9.14
9	3	1015	881	0	10.57
9	6	265	257	2	8.68
10	5	393	377	0	10.73
12	6	306	296	1	9.72
Extremities Grouping Method					
N/A	N/A	102	91	2	8.50
Key: FA = Focus Areas, FP = False Positives, FN = False Negatives					

Table 1 - Performance Comparison of Window and Extremities Grouping Methods (1)

The table shows various different configurations for the *window* grouping method. In comparison, the *extremities* grouping method does not rely on any parameters that can alter the overall performance of the algorithm. From the different *window* grouping configurations, some basic rules can be extracted. An approximately equal box and step size leads to less focus areas being produced, as there is no window overlap. This leads to the possibility of focus points falling between box windows. The data given in the table shows that this does have an effect, and most false negatives are created from this arrangement.

Both grouping methods give approximately similar average distance-errors, which range between 8.5 to 10.73 full resolution pixels. This equates to of between 1.7 to 2.1 pixels in the reduced resolution image. All frames within ~3 pixels of the best frame were trained with an equal target activation value. Therefore a tolerance of ~4.7 reduced resolution pixels is built-in to the training, so the small positional error observed is better than expected. Although it is not apparent from the above results, faces that were in partial profile or were slightly tilted tended to produce larger although acceptable face distance-errors.

Window Grouping Method				
Box Size (pixels)	Step Size (pixels)	Face Positive : Face	ANN Frames : Focus Areas	ANN Frames : False Positives
3	3	2:1	80:1	86:1
4	2	6:1	21:1	23:1
5	4	2:1	66:1	70:1
5	5	1:1	109:1	116:1
6	3	3:1	30:1	32:1
6	5	1:1	79:1	82:1
6	6	1:1	122:1	127:1
9	3	13:1	0.79	22:1
9	6	1:1	73:1	76:1
10	5	2:1	50:1	52:1
12	6	1:1	64:1	66:1
Extremities Grouping				
N/A	N/A	1:1	191:1	214:1

Table 2 - Performance Comparison of Window and Extremities Grouping Methods (2)

Appendix B and C show a graphical output from the analysis of the attention-focusing ANN on the test images. Appendix B shows focus point grouping using the window grouping method. Appendix C shows focus point grouping using the extremities grouping method. Focus point activations (derived from the output of the attention-focusing ANN at a particular *x,y* position on the image) are depicted as linear grey scale squares. The intensity of each square represents the magnitude of the activation by the ANN, i.e. some value between 0.5-1.0. Black rectangles are invalid focus areas; i.e. the centre point of the focus area is beyond the distance-error threshold. Conversely, white rectangles represent valid focus areas; i.e. the centre point of the focus area is within the bounds of the distance-error threshold.

Examining the images shows that the ANN has produced focus points in every area where there is a face. The ANN has therefore achieved the main objective, which is to produce activations for all faces. However, test images 6 and 16 do not produce a cluster of connected points. Applying post-processing, i.e. the grouping algorithms remove the face candidate from further processing. Test images 7 and 17 contain a cluster of points but the position of the activations are off-centre which obviously affects the focus distance-error.

From the list of test images presented, some of them have generated a higher proportion of false activations. This is due to some images containing textures (for these cases polka dots and stripes) that had never been presented to the ANN before. It is unrealistic to expect a trained ANN to cope with all patterns that can be taken from real world data. In a real environment if particular test images contained textures that caused the attention-focusing ANN to falsely activate then the solution would be to add an appropriate selection of these test images to the training set and retrain. It is difficult to predict the performance of a trained ANN on novel images and the training of the attention-focusing ANN should be regarded as an iterative training process until a required performance level is achieved.

Appendix B shows the focus area output for the window grouping method, and Appendix C shows the output for the extremities grouping method. Focus areas are depicted as either white or black rectangles. A white focus area means that it is within the face distance-error, and is labelled a face positive. A black focus area means that it is outside the face distance-error, and is labelled a false positive.

Increasing the box overlap increases the focus areas produced, but not necessarily the number of false negatives. A configuration of box size 4 and step size 2 has an extremely high number of focus areas, but does not manage to remove all of the false negatives. The window grouping configurations that provide the best results are those with a larger window size. This has the ability to include unconnected focus points in a larger area. Because the grouping covers a wider area, the number of total focus area positions is less. The distance-error should also increase leading to the possibility of false negatives. A box size of 12 and step of 6 demonstrates the effect of having too large values. The optimal configurations from this set of results indicate that a box size of 6 and step size of 3, or a box size of 10 and step size of 5 give no false negatives and the fewest number of false positives.

The extremities grouping algorithm has produced considerably less focus areas than any window grouping configuration. The average face distance-error is also less than the other method. Although other grouping methods are available, of the two methods evaluated the extremities grouping method provides best performance using the data provided by the attention-focusing ANN. Although more time could be spent investigating whether a better grouping algorithm can be found, the aim is to provide a flexible computer vision framework that provides acceptable result which may not necessarily be optimal (if that can be achieved) but defines the particular components required for this approach.

3.15 ANN Weight Analysis

The main factor that governs the performance of the attention-focusing stage is the generalisation performance of the trained ANN. To analyse what has been learnt by the ANN one method is to examine the hidden unit weight connections leading to the input units.

Figure 12 shows graphically the weight connections of the network. The weight values can be unbounded in size and reflect the magnitude of the inputs. The weight values therefore, can contain a relatively large range of values (both positive and negative). To enable a visual representation of the weights, they are passed through the Sigmoid function to scale each weight to an appropriate grey level intensity. Low intensity pixels represent highly negative values and high intensity pixels represent large positive values.

Figure 12 - Graphical Representation of ANN Weights

Each two dimensional block represents a hidden unit of the size of the input frame. Figure 12 shows the ANN weights for the attention-focusing network used to produce the results in this chapter. Observing some of the units, some can be clearly seen as representing generalised low-resolution faces. These are not specific faces found in the training set, but rather composite faces. This illustrates that particular units have learnt features that are common to the trained object. Analysis of the output profile of these units shows a positive contribution to focus point activation. Conversely random (or noisy) looking units provide negative contribution to face type inputs. It is quite reasonable to expect this, but it is perhaps rather more surprising that most of the units have face like properties. It has been observed that none of the literature regarding the problem of image identification/classification/recognition has performed any visual analysis of the weight connections in regard to their ANN performance studies.

Examining the hidden units would seem to indicate that the quantity of training data has saturated the ANN. This is illustrated by the fact that the majority of weight values are at the extreme of the grey level range (i.e. black and white). Any improvement in terms of classification, i.e. increasing the training to cover an even broader range of faces and distracters, should require further hidden units to be added to the attention-focusing ANN architecture. Whether the number of distracters passed on to the next stage of processing is acceptable is discussed in the following chapters.

3.16 Summary

This chapter has examined the problem of face identification and investigated various means by which this problem might be tackled. Particular problems have been identified that has driven the paradigm presented. This has concluded with performance analysis of the selected model design, which highlights the benefit, and also the limitations of the first stage, the ANN attention-focusing system.

From the results provided, Test image 3 is a good representative of the performance of the ANN system. The face distance-error values are not zero, but they do illustrate that having such errors still allows the whole face to be contained within the input frame.

The false positives generated for the image clearly do not at full resolution look like faces. However, at low resolution some patterns may show some general face qualities, given that at this resolution faces may not appear to be particularly "face-like". Some other example pattern types the ANN has not been trained with also lead the ANN to produce a high activation. This will always be the case.

A two-stage computer vision strategy allows for classification error in compensation for speed, low resolution and reduced processing. A higher resolution classifier should be able to reject these patterns as false positives. Until the classification and then overall system performance has been evaluated it is difficult to specify whether more robust classification performance is required of the attention-focusing stage.

Firstly, the next stage in processing is to address the problem of size invariance and how this can be applied to the attention-focusing model already presented.

Chapter 4

Size Invariant Object Location

The system described in the previous chapter deals only with objects of an approximately similar size and at an optimal resolution. This chapter describes the development of the attention-focusing ANN to cope with size invariant object location.

4.1 Requirements for Size Invariance

The type of faces and background to be used for analysis is the same as identified previously for fixed size face analysis. The only difference in the data set are that face sizes are now unconstrained in and across images. This means that a face can be of any physical size and can occupy anything between a small area of the image up to the full image frame. Also, for images that contain more than one face in an image these may be of different sizes. Unlike the fixed sized object search, the size of faces is not known prior to the search.

The amount of image reduction determines the object size. Because the face identified may not be at the optimum resolution, a size error may be incurred with the size invariant attention-focusing ANN. The classification ANN must therefore be tolerant to this error.

4.2 Size Invariant Object Location

Now that an attention-focusing ANN has been established that can identify fixed sized faces in images efficiently and reasonably reliably, the system can be extended to cope with size variant faces. The following sections describe work that attempts to use the ANN trained for fixed sized faces to detect variable sized faces.

The previous chapter identified problems regarding the classification of faces at minimal reduced resolution. To redevelop a different architecture for size invariance classification without exploring means by which the fixed size attention-focusing ANN can be extended is wasteful. In fact, sub-sampling an image to reduce the object to the same resolution as expected by the fixed size attention-focusing ANN require no difference in ANN processing. The only difference is in the pre-processing necessary that enables the correct resolution to be presented to the ANN. Other authors [Anderson, 1990], [Evans *et. al*, 1991], [Marsic, 1992] have adopted a multi-resolution approach to their ANN architectures for multi-resolution processing.

4.2.1 ANN Architecture

Although the fixed sized attention-focusing ANN needs to be adapted to enable size invariant classification, it is intended to explore techniques that enable size invariant classification without significant modification to the fixed size attention-focusing ANN architecture. This can be achieved by processing the image across multiple resolutions to determine position and size of the object.

Having trained at a fixed resolution, one might assume that recognising objects of different sizes would require several ANNs to be trained at different resolutions to cope with the size variance. Indeed Evans [Evans *et. al*, 1991] adopts this approach to size invariant classification, but this severely limits the range of resolutions that can be processed and subsequently the extent of the objects that can be detected.

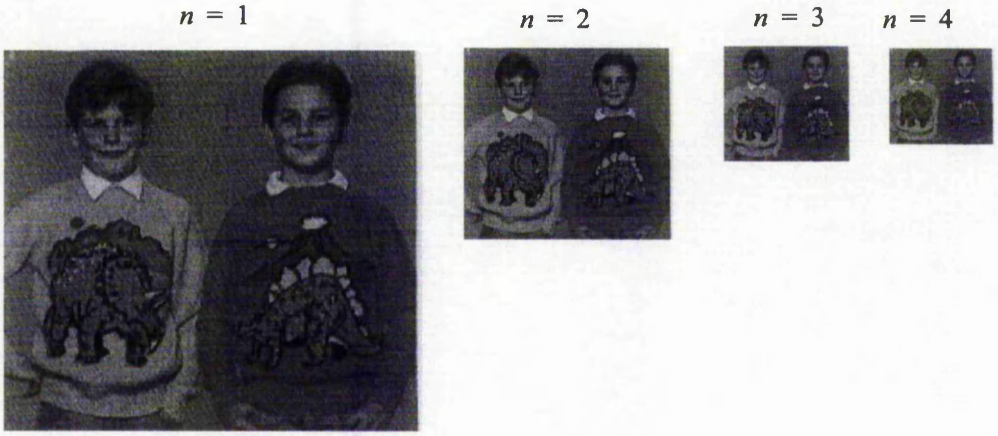
The variability in the images means that after image reduction there is a slight difference in size of faces being presented to the attention-focusing ANN for training. However, the size difference between objects is minimal. To incorporate any further size deformation into the training would burden the attention-focusing ANN with an added complexity requiring a larger input frame to ensure that the input object is not largely clipped and also a larger number units or connection strategy. The main objective of the attention-focusing stage is to identify possible areas of interest simply and allow further stages of processing to determine the accuracy of the focus areas presented to it.

4.2.2 Image Sampling

For single resolution analysis, the simple *n*-pixel method used by the attention-focusing ANN in the previous chapter is inadequate for multi-resolution analysis. The disadvantages of this method and techniques to overcome the problems associated with it are discussed below:

4.2.2.1 Simple *N*-Pixel Sampling

The initial approach to image reduction uses a simple *n*-pixel sampling method. This reduces the image by a scale factor *n*, every *n*th pixel in both the *x* and *y* direction. This is one of the fastest methods for shrinking an image, but the effect of scaling with consecutive scale factors produces a series of non-linear reductions.



Greater difference at higher resolutions than at lower resolutions

Figure 13 - Effect of *n* Pixel Scaling

The consequence of processing with the *n*-pixel method is that there is a bigger difference between adjacent resolutions closer to full resolution than at lower resolutions. If the image contains an object at low resolution, this method has the possibility of the attention-focusing ANN activating across several different resolutions, as the difference in sampling becomes increasingly insignificant. Conversely, the large differences in resolution change for small values of *n* may mean that the attention-focusing ANN fails to be presented with the appropriate object size expected (i.e. trained with) and therefore fails to activate at all.

4.2.2.2 Linear *N*-Pixel Sampling

An alternative method of image sampling is therefore required that allows for greater flexibility in the amount an image can be scaled. It also allows adjacent resolutions to be related and identified by a constant scaling factor.

To decide what pixels are extracted from the full resolution image to create the reduced resolution image is determined by the sample interval. This is described by Equation 5.

$$\textit{sample interval} = \frac{1}{\textit{reduction}^{n}}$$

Equation 5 – Determining the Sample Interval

Where 1 ≥ *reduction* ≥ 0; and *n* ≥ 0. The reduction value corresponds to an image scaling factor (where *%reduction* = (1 – *reduction) x 100*), and *n* determines the number of reductions carried out (i.e. the number of resolutions processed disregarding full resolution). A smaller reduction value corresponds to greater image reductions. If a single face covers the entire image, then the maximum number of reductions required to shrink the image to the size of the ANN input window is given by Equation 6.

$$\textit{maximum possible n} = \frac{\log_{10}\left(\dfrac{\textit{window size}}{\textit{picture size}}\right)}{\log_{10}\ (\textit{reduction})}$$

Equation 6 - Determining the Maximum Linear Steps

Where *window size* is the dimensions of the ANN input window, and *picture size* is the pixel dimensions of the image. The derivation of Equation 6 is given in Appendix G.

Although the size of the face is not known, *n* determines the number of resolutions the ANN must examine to guarantee that all possible face sizes have been explored. The value assigned to the *reduction* factor is user definable, and controls the degree of resolution reduction at each stage. An appropriate value for the *reduction* factor is explored later in this chapter.

The main advantage of applying this function to perform image reduction is that like simple *n*-pixel reduction it is a fast and effective means by which the image can be scaled. However, according to the *reduction* value selected this may cause bunching of the pixel values as the n^{th} pixel is rounded up or down. The simple alternative to this method is to use average sampling.

4.2.2.3 Average Sampling

Modal and *mean* sampling are two methods of average sampling. It is likely that both of these methods will produce a more accurate reduced pixel representation of the full resolution image because the surrounding pixel values are taken into account when determining the new pixel values.

A problem with both *n*-pixel and modal sampling is that important information may be lost in the reduction as the resulting image is not derived from all pixels. In comparison, mean sampling overcomes this problem but introduces blurring of well-defined edges.

To select the most appropriate algorithm amongst the three methods is almost an arbitrary decision. However, because of the amount of reduction the system has to perform to reach the optimum resolution (determined from experiments described in the previous chapter), the technique that provides the most accurate reduced representation has been chosen as the most suitable. Average sampling is the only method that takes into consideration every pixel in the reduction. This is the method adopted for the image reduction shown in the results presented later in this chapter.

4.2.3 Multi-Resolution Search

As the system has no predetermined knowledge about the objects contained within an image, a strategy is required that selects the resolution to process. Biological methods such as those discussed by Ahmad and Omohundro [Ahmad and Omohundro, 1990], Leow and Miikkulainen [Leow and Miikkulainen, 1991], and Sajda and Finkel [Sajda and Finkel, 1992] use a random approach to fixating upon the desired object to direct the next focus position and resolution. This approach is not feasible within the framework already developed, as the output from the attention-focusing ANN can only indicate the presence or absence of the trained object. A more formalised approach is therefore necessary.

Two strategies are therefore available; either process from high resolution to low resolution or vice versa. The first strategy offers no benefit in the attention-focusing approach of reducing the information to be processed and questions what a two-stage classification strategy would offer compared to processing the image at full resolution. Processing the lowest resolution first identifies large objects sooner and can remove the necessity to process the same area at higher resolutions. The strategy to manipulate how the different resolutions are processed is described as follows:

In order for the system to be flexible and cope with objects of any size, the image is reduced to the minimum resolution a face can occur. This is determined from the *maximum possible n* given in Equation 6. From this starting resolution the image is scanned to produce focus points, which are then grouped to provide single resolution focus areas. This is performed in exactly the same manner as fixed size face analysis. Any grouping method would be suitable, but the extremities grouping method (3.8.3) has been used initially. This is repeated for a number of higher resolutions using a fixed scaling factor, towards full resolution. It is necessary to perform this processing through the resolutions as there is no prior knowledge of the size of any object (face) in the scene, and it may be that multiple object occurrences will be of different sizes.

Large objects are located at low resolution (large reduction) and small objects at high resolution. The smallest possible face that can be accurately located has been determined previously by the experiments to determine the optimal resolution that are described in chapter 3.

Each separate resolution is scanned in the same manner as the fixed sized resolution attention-focusing ANN, producing focus areas for that particular resolution. These can at this point be passed on to the next stage for further processing. However, adjacent resolutions may be able to provide additional support to the validity of the focus areas. This is wholly dependant upon the reduction value employed, which is investigated in the following section.

4.2.4 Determining the Best Reduction Value

One of the most important aspects of the size invariant location method is the choice of the reduction value. Initial investigations have revealed that an object may be recognised across a number of contiguous resolutions, providing the image reduction is not too large. A reduction that is too large may miss the object entirely as the computed resolutions are too far from the resolution that the ANN has been trained with.

It is difficult to determine the most suitable value for the reduction term. A smaller reduction leads to a greater number of resolutions to search and therefore increases the time to process an image. This in turn increases the number of false positives generated as more input patterns are presented to the ANN. The number of false positives generated is proportional to the increase in number of patterns examined. The benefit to having a small reduction is that a closer resolution to the object is processed leading to a more robust detection of the focus area and also more accurate calculation of the objects' size. Analysis of different reduction values is given in the performance evaluation at the end of this chapter.

It is likely that an area of interest may produce focus point activations at the same location across different spatial resolutions provided that the difference between consecutive resolutions is not beyond the tolerance of the attention-focusing ANN. If this does occur, then there is the possibility of grouping these points to create a single multi-resolution focus area. Also, areas of interest receiving focus point activation at a single resolution only can be rejected as an unlikely area of interest.

4.2.5 Focus Area Post Processing

With multi-resolution analysis, a point in an image may be active across many different resolutions. Assuming that contiguous resolutions are close enough to each other, the presence of an object produces multi-resolution activation. For this reason, consecutive single resolution focus areas are grouped together to form multi-resolution focus areas.

An area of the image may be active across many resolutions and grouping all activations may lead to inaccurate object size calculations. Therefore, grouping all similarly positioned focus areas and averaging the size is restricted. This is to ensure that if any activations appear across a wide spectrum of resolutions they are *not* recognised as a single item. It is not feasible that an ANN tolerant to only a small change in size will activate correctly across many resolutions at the same point in the image. For this reason a limit on the maximum number of consecutive resolution focus areas that can be grouped has been set to a maximum of three.

Because the focus areas are of different sizes, it is necessary to determine which overlapping focus areas can be merged. Equation 7 is used to determine the maximum distance allowed between the extremity focus points of the two focus areas. Instead of examining the overlapping area that the different resolution focus areas cover, the focus areas are grouped on whether the cluster of focus points overlap. This allows overlapping focus areas to be merged that do not necessarily activate at the same point within the image. The equation determines the maximum distance (in integer pixels) that the cluster can be displaced to be considered a multi-resolution focus area. Focus areas within this distance and within two consecutive resolutions are grouped together to form a single multi-resolution focus area.

$$\mathit{maximum\ distance} = \mathrm{int}\left(\frac{1}{\mathit{reduction}^{\mathit{minimum\ focus\ n}}}\right)$$

Equation 7 - Maximum Distance between Focus Areas

The middle resolution of the grouped focus areas determines the estimated size of the object. As with single focus points at a fixed resolution, focus areas in just one of the multiple resolutions are also removed. Analysis of focus areas produced before post processing indicated that focus areas recognised at only a single resolution generally indicate a false positive at that spatial resolution. Filtering these it is possible to reduce the number of incorrect focus areas.

4.3 System Overview

The ANN system designed to locate objects of a fixed size has been extended to cope with size variant faces through the use of multi-resolution processing. The image is reduced to the smallest possible size a face can occur, and then the ANN input window is scanned across the low resolution image to generate any focus points. Any focus points produced are grouped to form single resolution focus areas. The resolution of the image is increased by a reduction factor, and the process repeated. This action is performed until a predetermined resolution is reached. (Currently full resolution).

Any similarly positioned consecutive single resolution focus areas are grouped, up to a maximum of three. Any non multi-resolution focus areas remaining after grouping are removed. The average size of the multi-resolution focus areas determines the size of the object. These are then passed to the classifier for further analysis.

The strategy of focus areas found at lower resolutions to eliminate processing parts of the image by the attention-focusing ANN at higher resolutions is explored later in Chapter 5.

4.4 Performance Evaluation

To determine how effective the multi-resolution search has been, similar experiments to that detailed in chapter 3 for single resolution have been performed on a collection of unseen images. The following describes the criteria for measuring the performance of the multi-resolution ANN system. Included in this chapter are results showing the success of location of size variant faces.

4.4.1 Criteria For Evaluation

The same performance criterion as for fixed resolution identification is applied for the multi-resolution identification. These include:

- The number of false negatives.
- The number of false positives.

To determine whether a multi-resolution focus area is a face positive, two measures are required. These are the face distance-error and the focus area size. A focus area is determined to be a face positive if the following two rules are met:

- The position of the focus area is within a maximum face distance-error.
- The size of the focus area is within the resolution of a single reduction step away from the optimum resolution for the face.

4.4.2 Analysis of Results

All of the results discussed in this section have used the same trained ANN as was used for analysis in the previous chapter. A similar set of 10 test images has been used to measure the performance of the multi-resolution search. A number of different reduction factors were tried, giving the total number of separate input frames examined by the ANN of between 284,503 and 2,323,316.

The test image outputs included in appendix C show the output of the ANN using a reduction value of 0.85. Focus points and focus areas are depicted in the same manner as previously except that different sized focus areas are shown which represent identification at different resolutions.

Table 3 shows the performance of the ANN across multiple resolutions using a range of different reduction values. As can be seen by the data and the associated graph, the number of false positives increases as the amount of image reduction decreases. This is expected as more information is being presented to the ANN.

Reduction	ANN Frames	FA	Frames :FA	FP	FN	Average Distance -Error	Average Size Error
.70	284,503	133	2139 : 1	126	4	9.79	-2.26
.75	453,214	172	2635 : 1	163	1	7.73	1.22
.80	733,120	312	2350 : 1	300	1	10.59	-2.35
.85	1,241,044	604	2055 : 1	590	0	8.51	-9.78
.90	2,323,316	1,031	2253 : 1	1,021	3	9.03	-5.08
Key: FA = Focus Areas, FP = False Positives, FN = False Negatives							

Table 3 - ANN performance across multiple resolutions and reduction values

As one might expect those images that generated a higher proportion of false activations at optimum resolution also generated false activity across multiple resolutions. Again, this is due to these images containing textures that have never been presented to the ANN before. This highlights the limitation of the number of varied examples in the training set, and how this inadequacy is propagated in the results for multi-resolution. A larger and more varied training set should overcome this problem to some extent. Although the number of possible background patterns is almost infinite, the number of distracters in the training set is relatively insignificant to the number of examples required to minimise false activation. For those test images that contained similar patterns to those that the ANN had been previously exposed to, the multi-resolution search did not create an inordinate number of false positives compared to the increase in the number of separate ANN window patches examined.

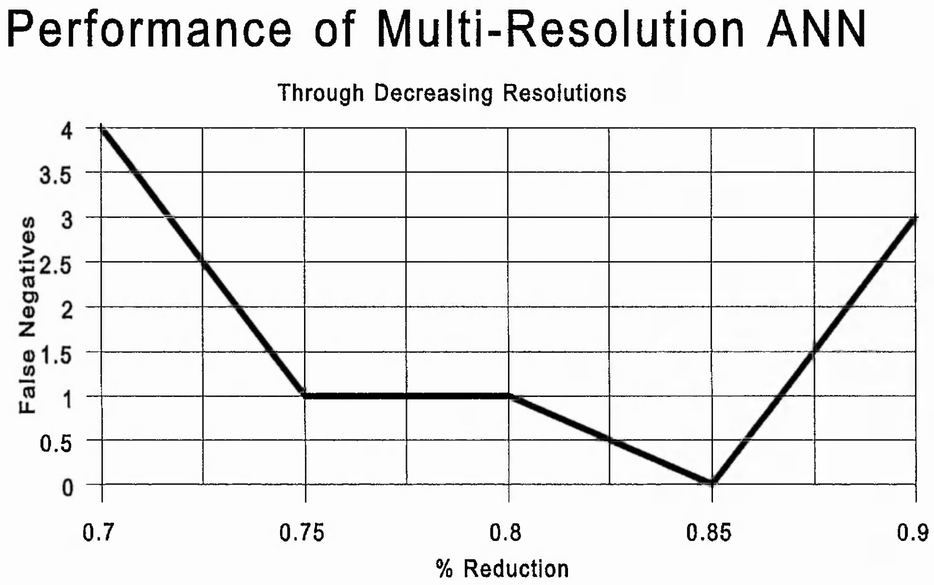


Figure 14 - Performance of Multi-Resolution ANN through Decreasing Resolutions

As the image reduction decreases, the number of false negatives also decreases, leading to the optimum reduction value for false negatives of 0.85. Image reductions greater than 0.75 produce resolution differences that are outside the bounds of the ANN size tolerance. Having a larger reduction may miss the optimum size of the object and therefore miss it entirely. Also, multi-resolution activations are required for successful identification, as single resolution activations are removed.

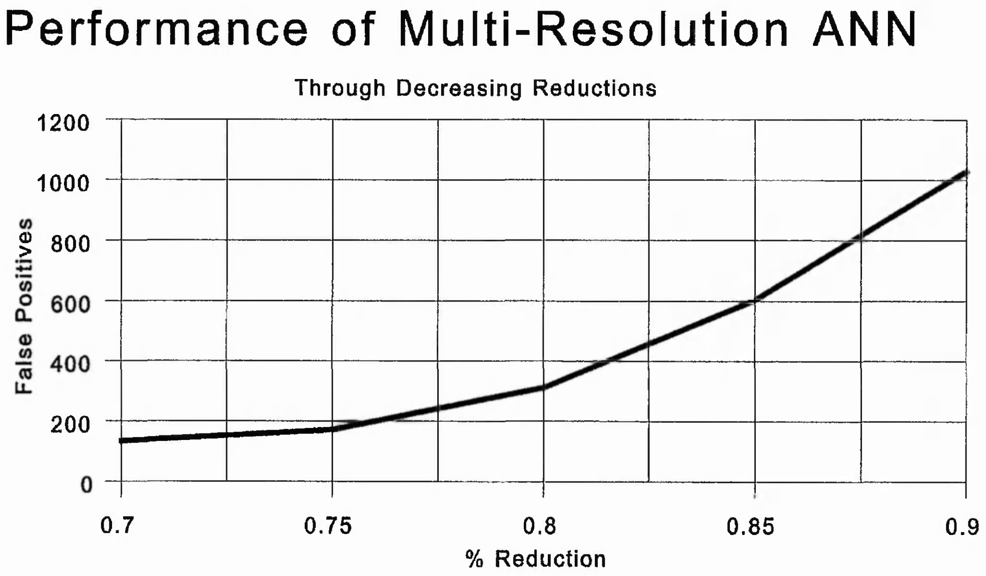


Figure 15 - Performance of Multi-Resolution ANN through Decreasing Reductions

The reason that there is a sharp increase of false negatives for reduction 0.9 is that the output becomes saturated with activity and subsequent grouping produces incorrect sized focus areas. Increasing the number of image reductions also increases the number of multi-resolution activations. This is because a maximum has been set of only three consecutive resolution focus areas that can be grouped. Extending the number of allowable resolutions to be grouped, and the reduction value tolerance, would reduce the number of false negatives. With a varying reduction factor, the grouping algorithm *should also* vary the number of focus areas that can be merged. However, a simple linear relationship between the reduction value and the number of consecutive resolutions to group is difficult to determine. Although the effect of different reduction values has been presented, varying the resolution grouping has not been fully explored to find the optimum reduction/grouping configuration.

Examining the face positive *size error* shows no discernible pattern through the different reduction factors. The results present a general bias towards smaller focus areas than the set size, but even with the largest size error this equates to an error of approximately two pixels at the reduced resolution. This amount of size error is negligible, and the results suggest on average the correct resolution for the object has been found.

4.5 Summary

The results presented show that a technique of multi-resolution analysis incorporating the ANN trained at an optimum resolution is a feasible method for size invariant object location. A reduction factor of 0.85 has been found that has enabled all faces to be found in the test data with minimum generation of false positives.

Even with multi-resolution analysis, a relatively small number of false positives have been produced in comparison to the amount of processing per image. The exception to this has been with pattern types unfamiliar to the trained ANN. As stated previously in the single resolution analysis, this highlights deficiencies in the amount of varied training examples.

This chapter has extended the attention-focusing model to consider objects of different sizes. The classification performance of the fixed sized resolution ANN has managed to identify all faces within the images without the need for retraining or a change to the ANN architecture. The following chapter investigates an appropriate architecture for the next stage in processing and how the output produced by the multi-resolution attention-focusing ANN can be classified more accurately.

Chapter 5

Focus Area Classification

The previous two chapters have dealt with the first stage of a two-stage identification process. This first stage, the attention-focusing ANN has generated areas of interest, which need a more accurate classification. This chapter presents a model to achieve this goal.

5.1 Identification - A Two Stage Approach

Focus Area Classification establishes whether the output from multi-resolution analysis contains the trained object in any of the focus areas presented. The attention-focusing ANN uses minimal resolution to identify possible areas of interest. It is only a quick identifier of *possible* areas of interest (faces), and some errors are incurred as a compromise for the low resolution used. It is the function of the classifier to perform a more accurate analysis of the focus areas to resolve the uncertainty about which category the areas of interest belongs to. More information is therefore necessary to perform reliable classification.

Increasing the resolution to provide more information with which to perform classification provides its own problems. The initial studies for the attention-focusing ANN architecture investigated the minimum resolution that an ANN could perform reasonable classification. However, the study also identified that increasing the resolution and also the relative size of the ANN produced poor classification performance. This is because the increase in information used for analysis requires that the ANN architecture is also larger and more complicated. An increase in the amount of data being presented also produces a greater variation that requires more examples to train the ANN.

Therefore, although more detailed information is required to perform a more accurate classification this has to be balanced with the necessity to keep the complexity of the ANN classifier architecture as simple as possible.

5.2 Requirements for Classification

The main aim of the attention-focusing stage is to determine areas of interest as effectively as possible with likely candidates of the desired object. However, the requirements of the classification stage are to:

- Reject incorrect focus areas presented to it by the attention-focusing ANN.
- Able to correctly classify the true areas of interest as being of the pre-determined object, e.g. a face.

5.2.1 Removing False Positives

It has been accepted that a certain degree of incorrect focus areas will be passed to the classification stage. As well as being able to re-affirm the presence of the desired object, classification must also be able to reject the false positives presented to it. It is possible that similar types of patterns giving rise to false positives at the attention-focusing stage may also cause classification to fail for the same reasons; i.e. the focus area has some type of pattern components which might confuse both ANN paradigms to activate incorrectly. The ANN classifier must therefore cope for this.

5.2.2 Greater Classification Accuracy

As well as removing the false positives the second aim of classification is to classify the true positives with a greater degree of certainty. The classification stage may also be able to adjust some of the original measures given by the attention-focusing stage in regards to exact position and size adjustment. Although a focus area may have been correctly identified by the attention-focusing ANN, the low resolution used incurs a tolerance for error. Increasing the resolution for the classification process enables some of the original metrics to be adjusted to improve the information known about the objects contained in the image presented.

5.3 Increasing the Amount of Information

In order to achieve a greater accuracy in classification more information is required than that provided by the attention-focusing stage, otherwise a one pass identification process would be all that is required.

Classification is required to provide a more accurate assessment of the focus areas. The degree of how much extra information required is difficult to establish. To simply increase the amount information and then train an ANN classifier is not going to produce a classifier that is able to meet the requirements stated above, for the reasons already given.

A problem that may impact on the usability of the identification system is the degree of increased resolution the classification stage requires. This imposes the size, and detail necessary for object identification. A system that requires too high resolution makes it less flexible to generic problems not related to face identification, and typically for problems where the area of interest is small in relationship to rest of the scene. Therefore, the increase in resolution for classification does not want to be too high, in terms of functionality and also for ANN training.

An increase in resolution does not necessarily demand that more information is required. As the focus areas have been determined to contain certain information, this can be used to the advantage of the design of the ANN classifier to impose justifiable constraints. In the case of face classification, and probably for most other problem scenarios, the object contains identifiable sub-features (e.g. for faces sub-features include eyes, nose, mouth etc.) that support the classification process. This method of using sub-features to support identification is also found in the human vision system [Bruce, 1988], [Treisman, 1982]. Therefore it seems a logical approach to exploit the use of sub-features as cues for classification.

Using sub-features is a good method to reducing the amount of information. It has already been established that to offset the increase in resolution the amount of information the classification ANN has to be reduced in order to keep the ANN architecture relatively simple and also have the chance of generalising a solution. Using sub-features assumes that areas within the area of interest have particular well-defined classifiable attributes and it should only be necessary to process these to perform an accurate classification.

The positional relationship of the sub-features can aid in the discrimination of background distracters. Areas of interest not containing these sub-features can be rejected. This means that it is unnecessary to examine the whole focus area, and allow extraneous areas of the focus area to affect the classification.

Now that it has been established that sub-feature classification is a suitable mechanism for the classifier stage, the next step is to identify how, and which, sub-features should be selected to aid classification. How this is achieved impacts on the design of the ANN architecture and search strategy employed. This is investigated further below:

5.4 ANN Models for Classification

Two different approaches to higher resolution classification are investigated. Both use the concept of sub-features to support classification. This reduces the quantity of information (i.e. only part of the focus area needs to be examined) and thus making it potentially easier for ANN classification. Although the two strategies examined use sub-features, how this is achieved is very different.

A detailed description of the functionality of each method is described in detail below. This also includes a comparison of the performance of each method discussing the advantages/disadvantages of each approach.

5.4.1 Sub-Feature Model

One classification method is to search for a sub-feature or sub-features belonging to the object within the focus area [Adams *et. al*, 1992]. The first step is to identify possible characterising features belonging to the object. For the chosen exemplar, faces, eyes are recognisable strong indicators that enable true positive classification to be asserted. Many authors, [Hutchinson and Welsh, 1989], [Vincent *et. al*, 1992], [Waite, 1991] have used ANN models to detect eyes successfully. Other sub-features that can be used are the nose and mouth.

The number of sub-features that may be applied to support classification is wholly dependent upon the type of object being identified. Unfortunately, selecting the appropriate sub-features is arbitrary. However, an object being distinct in its own right should always include at least two characterising sub-features that should be fairly obvious indicators that belong only to the object being identified.

The greater the number of sub-features used to aid classification, the greater the flexibility and potentially the more accurate the final classification. Multi sub-feature classification offers greater flexibility in that there is not a single reliance on a single sub-feature, and failure to classify a sub-feature does not necessarily result in the misclassification of the focus area. However, the logic required for the data fusion of this information is complicated and the more sub-features used increases the amount of processing required to perform focus area classification.

Having chosen a sub-feature such as the eyes, An ANN is trained on left eyes (from the viewers perspective) in a similar fashion to the attention-focusing ANN, except that the sampling resolution is higher. Only one eye is used to train with as both eyes can be regarded as separate features. The right eye is as significant a sub-feature as the left, and is possibly unique in that it is essentially the vertical mirror of the left. To reduce the burden of training, a right eye sub-feature classifier can be simply created by mirroring the input weights along the vertical axis of an ANN trained on left eyes. As with the ANN for faces, the dimensions of the ANN for left eyes have been chosen to fit the measurements of the feature including some peripheral information such as the eyebrow. The same training method as the attention-focusing ANN is used to train the ANN for left eyes.

Having trained the sub-feature ANNs, the classification procedure is as follows:

1. Each focus area determined from the attention-focusing stage is increased in resolution to match the resolution at which the ANN is trained for eyes. Having prior knowledge about the geometric position of the sub-features, this can be used to reduce the area of the search as necessary. The search area must also allow for the possible error of the attention-focusing stage to accurately position the focus area. The search must also allow for positional shift according to some head movement.
2. As previously discussed, the multi-resolution search may have incorrectly estimated the actual size of the object. Therefore, a single resolution search may be insufficient to locate the sub-feature successfully. To compensate for a size error, a multi-resolution search for the sub-feature is required. Because an approximate object size has been determined, the number of consecutive resolutions that the classifier ANN has to search is limited to within a single resolution step either side of the focus area.

The identification of the sub-feature within the focus area is more evidence that the area of interest is the designated object (face). Those focus areas not containing the sub-feature can be rejected, and the others can be selected for further processing if necessary. This may involve searching for other sub-features to progressively increase the level of certainty. Alternatively, if a sub-feature is misclassified, as long as there are sufficient other sub-features, the classifier may be robust enough to cope with this and not necessarily reject the focus area.

5.4.2 Holistic Feature Extraction Model

In contrast with the sub-feature model, the holistic feature extraction model attempts to determine the salient features for an object automatically rather than choosing these manually. This is achieved by training an ART (Adaptive Resonance Theory) network [Carpenter and Grossberg, 1985], which is very suited to feature decomposition/compression problems. The trained ART contains a selection of nodes that comprise features common to a varying number of examples. Each node is then connected to particular features that it characterises as being significant for the number of examples associated with that cluster. These feature clusters are then used within an MLP paradigm to specify the connectivity to the focus area, using only the ART determined features to determine the inputs to the focus area. Using the ART to predetermine the ANN connectivity is a completely novel approach. The MLP is then trained in a similar fashion. The method to finding the holistic features in the ART style network is described in the sections that follow.

5.4.2.1 The ART Network

To identify distinct sub-features automatically requires a model that uses some method of unsupervised learning. Chapter 2 has already identified the Kohonen and ART networks as being two of the most widely used models of this type. The ART paradigm has been chosen in preference to the Kohonen for the following reasons: it has an extremely fast learning period, converging in only three epochs; and more importantly a self evident cluster representation.

There are several advantages for using the ART to determine the weight connections. These are:

- Connectivity is less than full connectivity. This reduces the amount of processing that would be required by full connectivity.
- The ANN is only connected to the predetermined important areas of the input. Unimportant areas of the image cannot therefore influence the final classification.

Two slightly different approaches to training an ART type network have been investigated[7]. Both approaches produce feature clusters that can be used as a connectivity matrix for an MLP classifier. The first approach is based upon the original ART network design as developed by [Carpenter and Grossberg, 1985], and uses a binary representation as input (Binary ART). The second approach differs mainly in that it uses grey-scaled values as inputs to the system. Although the development of ART2 [Carpenter, 1987] attempts to overcome the limitations of binary input, the second ART training method presented (grey level ART) uses an alternative and more sophisticated approach to achieve this.

5.4.2.2 Binary ART

The ART network deals only with binary input patterns, and therefore requires the 8-bit grey-level input to be passed through a threshold function before being presented to it. A mid-point value within the grey scale range has been selected as the most appropriate threshold as it is not specific to any particular image type.

It is unknown which features of the object are important, and these may either be areas of low or high intensity. In order to produce clusters for both of these

[7] For the remaining text the name >ART network= will refer to both of the ART networks investigated unless otherwise explicitly specified.

feature types it is necessary to produce positive binary inputs for patterns above and below the threshold. This is easiest to achieve by training two separate ART networks, with the second having inverted inputs of the first. To increase the number of examples and to produce more robust feature clusters, each face can be mirrored along the *y*-axis. This is perfectly valid, as faces are never exact mirrors [Bruce, 1988]. The feature clusters can then be combined together to produce the full set of nodes and weights that can be used as a connectivity matrix for a MLP type ANN. It is this partially connected ANN that is then trained for actual classification.

To train the ART network, requires two parameters to be set. These determine the number of clusters[8] and size of the features. These shall be referred to as *ADEQUACY* and *MARGIN*.

ADEQUACY represents the uniqueness requirement for a cluster. In training mode, for any cluster to win it must be similar to the input pattern by a certain degree as set by the *ADEQUACY*. This parameter also sets the minimum size of the feature (i.e. the number of connections that describe it). Unlike the sub-feature method the cluster feature may not be localised to a set region in the focus area.

To inhibit clusters converging to represent the same features, the *MARGIN* parameter requires a specified difference to be met between clusters, otherwise a new cluster is created. This is useful when two clusters are similar to the input pattern but represent different features of the input. If no single cluster is significantly closest than any other then a new cluster is created.

To fully train the binary ART classifier and determine the clusters used within the classifier MLP requires only three epochs. The speed by which the ART network can derive a solution enables different parameter configurations to be investigated

[8] A cluster from the ART net is equivalent to a hidden unit in an MLP and exhibits partial connectivity only.

prior to training the classifier MLP. This ensures that the clusters represent reasonable sized features, and also that there are a sufficient number of clusters to represent the number of features for an object type.

The training data to generate the feature clusters for the ART is slightly different to that required by the attention-focusing ANN. As the ART is concerned only with feature extraction, it is unnecessary to present background distracter information. For the attention-focusing ANN some position invariance was also encoded into the training, but for the ART training it is intolerant shift invariance. Therefore it is necessary that the object of interest be presented as near to the centre of the input frame as the resolution allows.

5.4.2.3 Grey Level ART

The binary ART method presented above, is quite basic in the method of dealing with the varying grey levels that constitute the object. The identification of "good" features is paramount if the classifier ANN is to produce robust classification. The grey level ART is an exploration in a slightly different approach to feature extraction, where the aim is to provide more intelligent processing of the grey level data.

An obvious choice to achieve this would be to examine the suitability of the ART2 paradigm [Carpenter and Grossberg, 1987]. This is a development upon the ART model to allow non-binary vectors. The ART2 processes grey-level data by essentially pre-processing into a binary representation. A similar training method is then employed to the original ART model. Although this is a move towards a method of processing continuous values, the underlying architecture is the same and does not offer any significant advantages to the binary ART implementation.

For the grey-level ART method adopted, the approach by which grey-level data is handled is quite different to the method used by ART2. A brief description of the algorithm is as follows:

1. The first step is to load the grey level data into the input layer. Next, histogram equalisation is applied to the input values. This helps improve the coverage of intensities across the grey level range, which allows similar patterns to adopt the same grey level bands. This is useful when determining similarity of features in an image.
2. The next step is to calculate the sum activation of the input values for all clusters (Equation 8). The winning cluster is the one that has the most active connections (Equation 9). An active connection is determined by whether the connected input is within a specified grey level range. This is represented by the constant *GREY_LEVEL_DIFF*. The number of connections for the winning cluster must also meet the requirements as specified by the constants *ADEQUACY* and *MARGIN*.

$$ac = \sum_{i=0}^{i<ni} |\ (1 - ia_i \, x \, w_i) \leq gld\ |$$

$$\textit{where } ac \equiv \textit{active connections; } ia \equiv \textit{input activation; } w \equiv \textit{weight}$$
$$gld \equiv \textit{grey level diff; } ni \equiv \textit{num inputs}$$

Equation 8 - Calculating the Active Connections

$$\textit{winning cluster} \rightarrow \textit{cluster}_n = \max\left(\frac{\Sigma ac_i}{ni_n}\right)$$

Equation 9 - Determining the Winning Cluster

3. If there is no winning unit and a cluster is to be added, then all weight values are set to the reciprocal of the input (Equation 10). If weights are to be updated for a cluster, any connections falling outside the bounds of the *GREY_LEVEL_DIFF* parameter are removed. The values for the new weights are set to the sum of old weights and the reciprocal of the input divided by two (Equation 11). Although the structure of the feature is determined by its connections, its weight value determines the grey level range. The weight is updated in this manner to best represent both the old and new inputs.

$$\begin{aligned} new\ cluster_{n+1} &\rightarrow w_i\ of\,active\ input_i = \frac{1}{ia_i} \\ new\ cluster_{n+1} &\rightarrow w_i\ of\,inactive\ input_i = 0 \end{aligned}$$

Equation 10 - Determining the Weights of the New Cluster

$$\begin{aligned} updated\ cluster_n &\rightarrow w_i\ of\,active\ input_i = \frac{w_i + \frac{1}{ia_i}}{2} \\ updated\ cluster_n &\rightarrow w_i\ of\,inactive\ input_i = 0 \end{aligned}$$

Equation 11 - Determining the New Weight Values for the Winning Cluster

5.4.2.4 Connectivity Matrix for the MLP Classifier

Once either the binary ART or the grey level ART has been trained, the derived model becomes the hidden layer of the classifier ANN. Each cluster from the ART network becomes a hidden node in the MLP. The features represent the connectivity of each hidden node. Even though there may be many hidden nodes (typically more than that required for a fully connected MLP), the reduced connectivity still produces an architecture with fewer connections in total than a fully connected network using less hidden nodes.

The initialisation of the weights for the classifier ANN can differ for the two types of ART model. For the binary ART network, each connected input is given a random weight. The grey level ART can adopt either the weights already derived or use random values. The benefit of the ART derived values are that they already represent positive discrimination of the object, and this may reduce the learning time when training the classifier. Further analysis is required to determine the actual effect or benefit of setting the initial weight configuration to any other setting than the standard random values.

In case the object exhibits some other properties not found by ART feature extraction, an extra hidden node is added to the architecture. This also helps balance the ANN so that, if required, the node can be used to represent the distracters. Initial experiments in training the MLP indicated an improvement in the speed of generalisation when an extra fully connected hidden unit was included in the MLP architecture.

Although the ART has determined composite features, the nature of the learning algorithm will have the effect that not all the feature clusters will be used for positive discrimination. In fact, learning may turn some clusters into negative clusters. The *backprop* learning algorithm is quite independent of the ART clustering and although the aim of the aim of the ART is to, in some way, force the learning of the MLP to the features detected the nature of the algorithm will determine its own importance of the ART connectivity.

5.4.2.5 Training the ART-MLP Classifier

The ANN network is trained in the same fashion as the attention-focusing ANN, with both positive and negative examples. Having trained the ART networks at a fixed size and position, the ART-MLP classifier is trained in the same manner. The reason for this is that the connectivity of the clusters represents templates and should be only expected to fire when the template is in the trained position. Therefore, the MLP classifier must be trained in the same way. Having a more rigid size and position along with the higher resolution allows for a greater accuracy to be determined about the object.

The architecture of the ART-MLP consists of hidden units with varying numbers of weight connections as well as a final fully connected unit. Initial experiments in training the ART-MLP found that the learning was biased against nodes with fewer connections. It was therefore necessary to modify the learning algorithm slightly to ensure that all nodes had an equal chance of generalisation. Equation 12 shows the change made to the calculation that determines the new weight adjustment. This slight change equalises all hidden units, with the same chance of contributing to the classification.

$$\Delta w_{ji} = \eta \, \delta_j \, a_i x \left(\frac{iu}{ic}\right)$$

where iu ≡ input units; ic ≡ input connections

Equation 12 - Learning Rate Adjustment

5.5 Texture Analysis

Chapter 3 showed that the attention-focusing model is reasonably simple but suffers from having to process a large number of different pattern types. Although it is reasonably successful at classifying the majority of input patterns, a certain number of false positives are passed on to the classification stage. Classification must be very robust and cannot continue to support false positives as possible true positives.

To address the problem of the large feature space (i.e. the degree of variation in the input patterns), a method is required to, if possible, partition the feature space into more manageable "chunks". This is to reduce the burden upon training and classification. This is especially important since an increase in resolution also increases the possible variation in the input patterns. The technique of texture analysis is explored below to determine whether the categorisation of the input patterns can help improve the robustness of classification. This is applicable to both classification (sub-feature and ART-MLP) models previously described.

The idea of applying texture analysis prior to classification is to determine which background category a pattern may fall into. Several classifiers are then trained to distinguish between the desired object and the background type. This helps reduce the problem and split it into more manageable parts. Training an ANN on a limited set of background data should improve performance in terms of being able to disambiguate between background and the desired object. With a small number of background examples, the performance of the ANN generates a high proportion of false positives. This is due to the limited variety of examples exposed to the ANN. Increasing the sample size has a beneficial effect on the ratio of number of false positives but can lead to the introduction of false negatives. This problem should be reduced with a limited domain of background data.

Using texture analysis allows the burden of classification to be reduced in two ways:

- As a filter before classification, e.g. Determine the type of texture of the focus area, and if not typical of a sub-feature type, i.e. background distracter texture type, reject frame. Although it is not possible to filter all input patterns by texture, a significant proportion should be able to be rejected.
- To segment the training, i.e. Select only background distracters similar to a sub-feature to help improve the discrimination of the classifier.

Texture analysis may be performed prior to classification or as a means of post processing ANN activations to remove potential false positives. To incorporate texture analysis as a further means of discrimination was determined undesirable as this may lead to more techniques to constantly improve identification and may lead to a process of diminishing returns.

5.5.1 The Co-Occurrence Matrix

The texture analysis implemented is based upon well-defined statistical techniques. It consists of a first stage of building a probability matrix based upon; the image data; and a pixel template to compare structure in the pattern. Using this matrix several different measures can be used to extract different key values that describe the pattern texture in some way. The number of vectors used is arbitrary. Seven different measures were described by Sonka [Sonka *et. al*, 1993] and these have been used to form a seven-dimensional vector to represent the texture. As long as the measures used are able to adequately represent different general background types, the number used is adequate. For a greater separation of the data more measures may be needed.

Using the *k*-means clustering algorithm [Sonka *et. al*, 1993], the background data can be separated to form a number of different clusters. The value of *k* can be predefined to represent some small number in which *k* classifier ANNs can be trained. Each ANN is then trained only with background distracters that are closest to the particular *k* cluster, and all other patterns that belong to the target object class.

As discussed above, texture analysis comprises of a number of measures that form a characterising vector of the pattern being examined. To enable the metrics to be calculated, an intermediate stage is required. Firstly, a co-occurrence matrix is created which allows the statistical measures to be taken which describe the texture.

The co-occurrence matrix is created by defining a matrix that records the occurrence of some grey level configuration. With varying textures the configuration varies accordingly. To calculate the co-occurrence matrix requires a series of steps:

1. The first step is to determine the size of the co-occurrence matrix. This is defined by Equation 13.

$$msize = constxconst;\ mindex = \frac{256}{const};\ const = 16$$

Equation 13 - Determining the Size of the Matrix

msize represents the size of the co-occurrence matrix, and *mindex* represents a quantising factor. This reduces the size of the co-occurrence matrix, and therefore the computation required deriving it. To represent a full 8-bit grey scale matrix would require 65,536 elements, which is realistically too large to be used extensively in the classification process. Therefore, the value of *const* is set to 16 as a computational compromise.

2. The next step is to initialise all values in the matrix to zero. This is defined by Equation 14.

$$co\text{-}occurrence_{v1,\,v2} = 0$$

Equation 14 - Initialising the Matrix

3. To represent texture, a simple measure is used to define the type of texture. A diagonal pairing has been used in this case. The next step is to insert appropriate pairings into the matrix represented by Equation 15.

$$v_1 = \frac{activation_{xy}}{mindex};\ v_2 = \frac{activation_{x+1,\,y+1}}{mindex}$$
$$co\text{-}occurrence_{v1,\,v2} = co\text{-}occurrence_{v1,\,v2} + 1$$

Equation 15 - Inserting values into the Matrix

activation$_{xy}$ is the grey level value at the *x,y* co-ordinate of the image. This value is scaled to the size of co-occurrence matrix using the quantising factor. Each cell within the co-occurrence matrix relates to a particular grey-level pairing. After determining the values for v_1 and v_2, the associated cell within the co-occurrence matrix is then incremented.

4. The final step is to convert the values contained within the matrix into normalised probabilities as shown in Equation 16.

$$tot\text{-}val = \sum_{x=0}^{x<16} \sum_{y=0}^{y<16} co\text{-}occurrence_{x,y}$$

$$co\text{-}occurrence_{v1,\,v2} = \frac{co\text{-}occurrence_{v1,\,v2}}{tot\text{-}val}$$

Equation 16 - Converting the Values to Normalised Probabilities

tot-val determines the sum of all values within the co-occurrence matrix. Each value within the co-occurrence matrix is then scaled by *tot-val* converting the values to lie between 0 and 1. After this conversion, the sum of all values within the matrix is equal to 1.

Using the co-occurrence matrix, measures can then be applied to it to characterise the input patterns.

5.5.2 Texture Measures

To characterise the textures for this classification problem, seven different measures have been chosen. These measures are then combined together to form a seven valued vector. The distribution of sub-features and distracter textures is then represented in a seven-vector feature space.

Any number of measures can be taken. For the purposes of classification, a fairly coarse method of categorisation is all that is required. Therefore, there are only a small number of measures applied. The measures chosen are standard texture analysis functions. These comprise the following:

$$\begin{gathered} moment \; = \; 0; \; k \; = \; 2 \\ moment \; = \; moment \; + \; (y - x^{k} \; x \; co - occurrence_{x,y}) \end{gathered}$$

Equation 19 - 2nd Moment

$$\text{max-}val \; = \; \max \; (co - occurrence_{x,y})$$

Equation 17 - Maximum Probability

$$\begin{gathered} moment \; = \; 0; \; k \; = \; 1 \\ moment \; = \; moment \; + \; (y - x^{k} \; x \; co - occurrence_{x,y}) \end{gathered}$$

Equation 18 - 1st Moment

$$\begin{gathered} entropy \; = \; 0 \\ if \, co - occurrence_{x,y} \; > \; 0 \; entropy \; = \; entropy \; + \\ |(co - occurrence_{x,y} \; x \; \log 10 \; (co - occurrence_{x,y}))| \end{gathered}$$

Equation 20 - Entropy

$$\begin{gathered} uniformity \; = \; 0 \\ uniformity \; = \; uniformity \; + \; (co - occurrence_{x,y} \; x \; co - occurrence_{x,y}) \end{gathered}$$

Equation 21 - Uniformity

$$\begin{gathered} imverse - moment = 0; k = 2 \\ where \; x \neq y \; inverse - moment = inverse - moment + \frac{co - occurrence}{y - x^{k}} \end{gathered}$$

Equation 22 - Inverse 2nd Moment

$$\begin{gathered} imverse - moment \; = \; 0; \; k \; = \; 1 \\ where \; x \neq y \; inverse - moment \; = \; inverse - moment \; + \; \frac{co - occurrence}{y \; - \; x^{k}} \end{gathered}$$

Equation 23 - Inverse 1st Moment

Having created a vector for each input pattern, a method is required to place divisions upon the vector in order to help categorise the texture and thus aid classification.

5.5.3 Texture Clustering

Clustering is a means of categorising vectors that are similar. The method chosen to partition the texture vectors has been based upon the *k* means clustering algorithm [MacQueen, 1967]. An alternative clustering method is to use a Kohonen feature map [Kohonen, 1989], which performs a similar but different approach to the problem, i.e. dimensionality reduction and clustering. The *k* means clustering algorithm has been chosen in preference to the Kohonen feature map because of its simplicity and speed. Having selected an appropriate number of clusters, the cluster centre points are adjusted over time to represent the distribution of vectors presented to them.

There are several initial decisions that need to be made prior to clustering. These are:

- The number of clusters that are to be used to represent the textures. The fewer the clusters the more crude the partitioning. The greater the number of clusters the more complicated, and possibly less meaningful, the partitioning. There is no simple rule of thumb as to how this value is chosen. If the number of classes is known prior to clustering, then this can sometimes be a good value. Ideally, in this case the number of clusters required might be 2, i.e. object and anything else. Unfortunately, it has already been identified by the performance of the attention-focusing ANN that discrimination between features and distracters is extremely difficult and some amount of overlap does occur.
- How the data is to be presented to the clustering algorithm. This influences how the vector centres are modified, and thus how representative of the data the clusters become.

➢ The initial cluster starting points. This may determine the meaningful migration of the cluster centres and the final cluster positions.

To produce the representative clusters, the following steps are performed:

1. Select the number of clusters and assign an exemplar vector to each.
2. Present a vector from the data set to all clusters. Associate the vector with a cluster according to the cluster with the minimum Euclidean distance.
3. Repeat from 2 for all vectors in the data set.
4. Present a vector from the data set to its associated cluster and adjust the cluster centre point as an average function of all vectors previously assigned to that cluster including the current vector.
5. Repeat from 4 for all vectors in the data set.
6. Repeat from 2 until cluster centres converge.

To examine the similarities between texture vectors, and the advantages/disadvantages of varying the numbers of clusters, clustering has been performed for a number of different configurations. A discussion follows on the analysis of the clustering method and what information this tells us about the data. Since classification is base upon sub-features, the texture analysis has used left eyes and a wide variety of background distracters.

5.5.4 Defining the Number of Clusters

The first problem to be addressed is what is the ideal number of clusters required to best represent the data? The merits of different cluster configurations are discussed below.

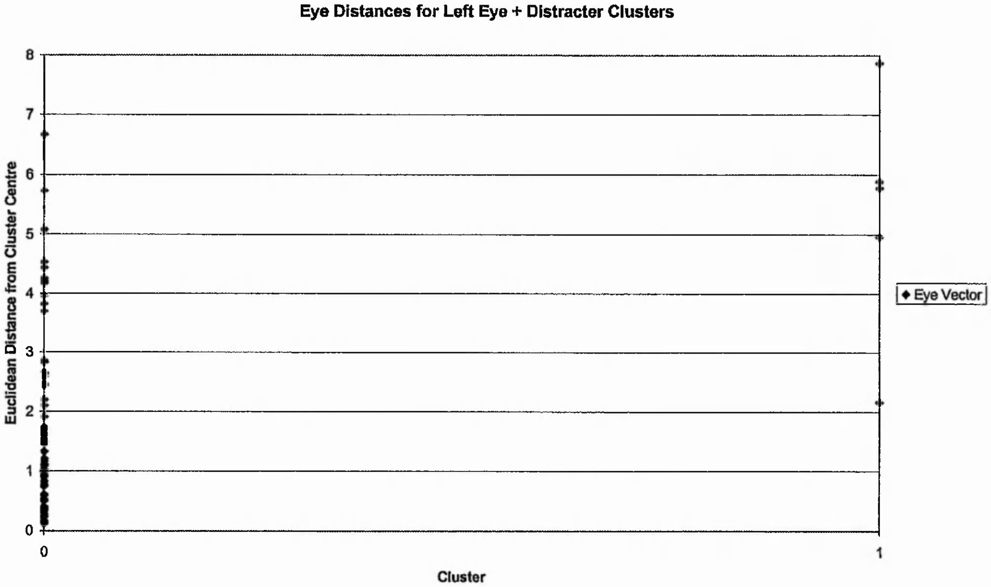


Figure 16 - Distribution of Eye Texture Vectors across Two Clusters

The first step is to identify whether there is a natural separation between eye and texture vectors. This is illustrated in Figure 16. Although, eye vectors predominantly occupy cluster zero, a number of eye vectors however are closer to cluster one. As well as fewer eye vectors being associated with cluster one, the distribution of these vectors is more widespread.

The diversity of the textures shows that there is not a simple and natural segregation of the eye and background distracter textures. This is not surprising given the difficulty of the attention-focusing ANN being able to discriminate between similar patterns.

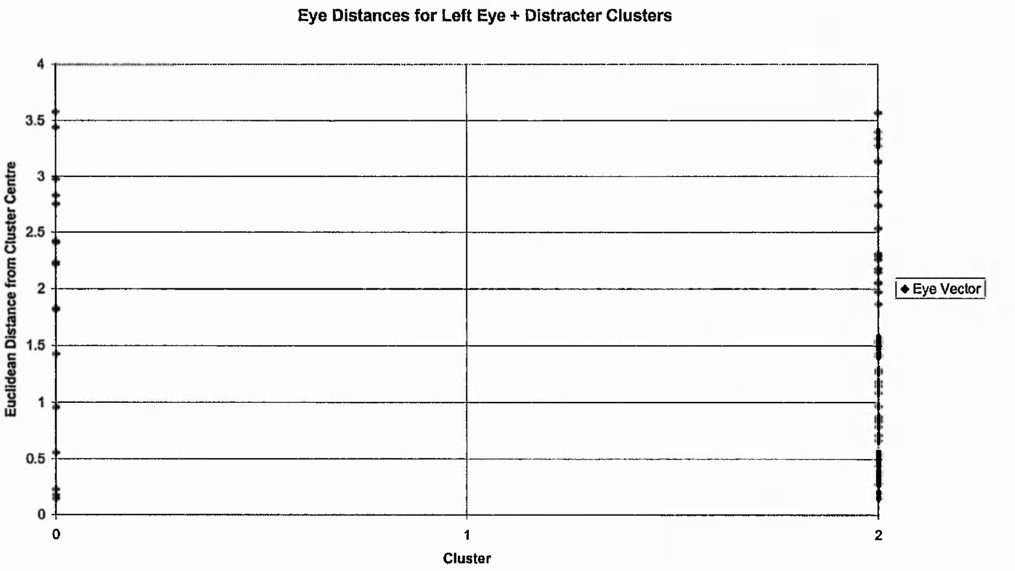


Figure 17 - Distribution of Eye Texture Vectors across Three Clusters

Figure 17 shows the effect of increasing the number of clusters to three. Again the eye vectors are distributed across two clusters. The remaining cluster can be regarded as a distracter cluster as no eye vectors are closer to this cluster centre than the other clusters. Increasing the number of clusters has made a noticeable difference in that the eye vectors are much more concentrated around the cluster centre, and the distance of the outlier vectors is half than previously.

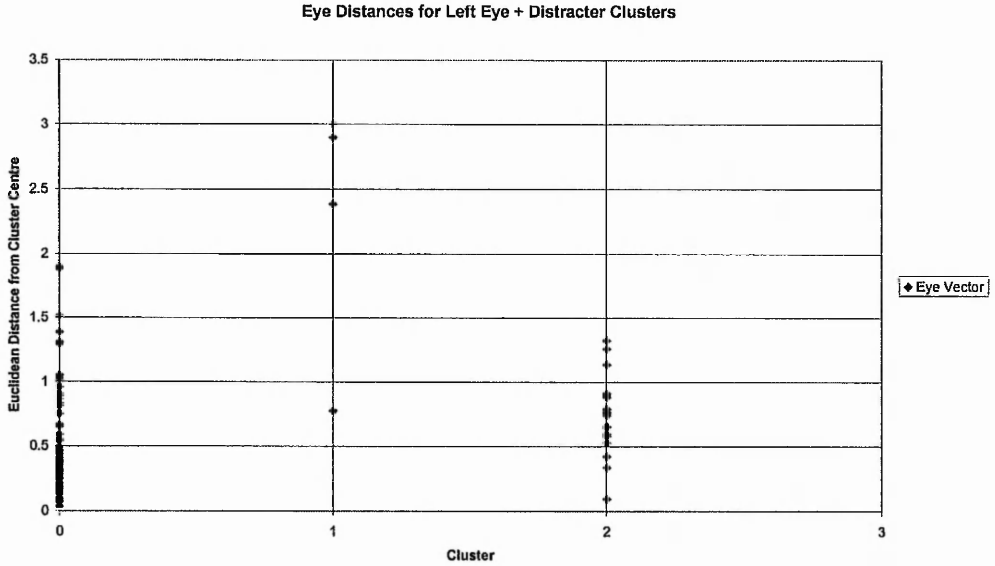


Figure 18 - Distribution of Eye Texture Vectors across Four Vectors

Similar to the properties of three clusters, clustering with four (Figure 17) still produces a distracter cluster. The distribution of eye vectors is now spread across three vectors. The two clusters with the majority of eye vectors have become even more concentrated around the cluster centre. The other cluster has a small distribution of relatively widely spaced vectors.

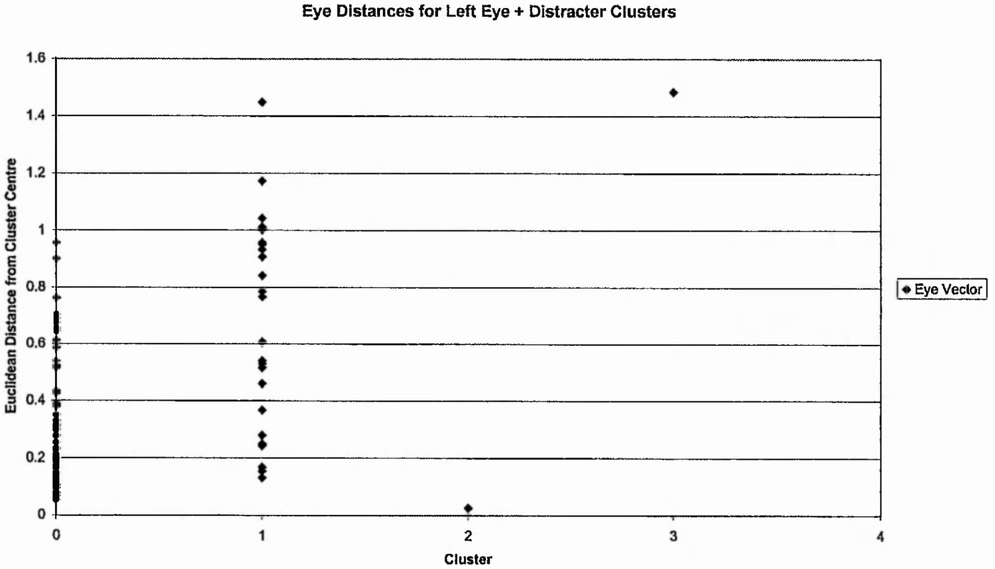


Figure 19 - Distribution of Eye Texture Vectors across Five Clusters

Increasing the number of clusters shows an emerging pattern. As more clusters are introduced (see Figure 18), the distance of the of eye vectors from a cluster centre reduces. Also, outliers migrate away towards other closer clusters, e.g. 2 and 3. A single cluster still attracts distracters only, as this is sufficiently distant from other clusters to attract any eye textures, and is central to most of the distracter patterns. The other clusters contain distracter patterns, but the number of distracters varies with the number and proximity to the cluster centre of the eye texture vectors.

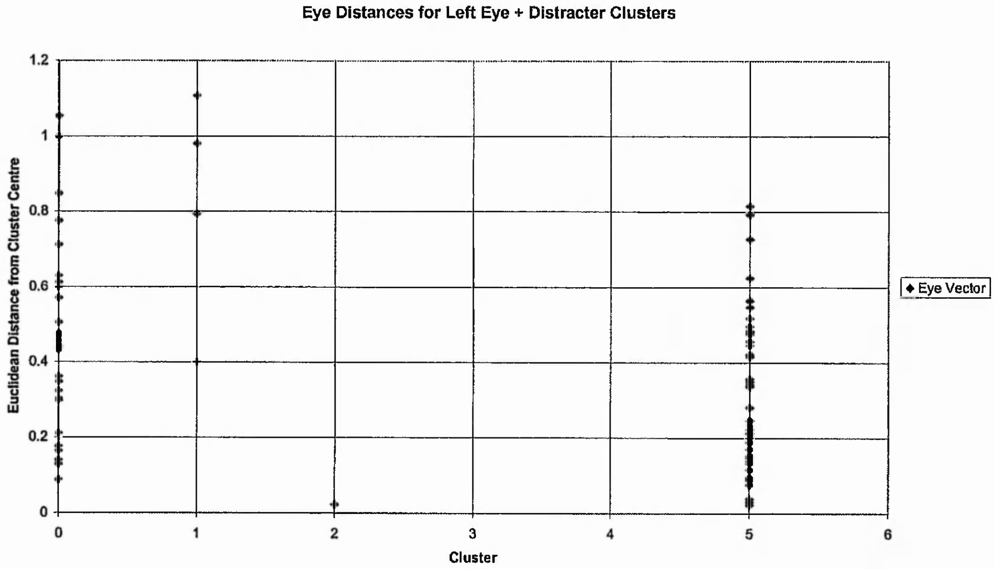


Figure 20 - Distribution of Eye Texture Vectors across Six Clusters

Figure 20 shows a continuation of the trends described previously, particularly greater eye vector fragmentation. The previous distracter cluster has now separated into two clusters, and has passed the optimal number of clusters. A further increase of the number of clusters leads to greater fragmentation and dispersement (not shown here).

Selecting the most appropriate number of clusters from the distributions given, is an arbitrary decision. Observations from varying the number of clusters is summarised below:

- Two Clusters. Loose clustering that contains both class categories.
- Three Clusters. Contains a distracter only cluster, but eye clusters still widely dispersed.
- Four Clusters. Identifies outlier eye vectors that form new a cluster.
- Five Clusters. Greater concentrations of vectors surrounding cluster centre, but to the expense of eye vectors becoming more disperse.
- Six Clusters. Clusters losing their logical groupings.

This investigation has shown that there are problems associated with having either too few or too many clusters. Unfortunately, there appears to be no obvious optimum number of clusters, and therefore five clusters has been chosen by the author as the most suitable compromise.

5.5.5 Characterisation of the Clusters

The previous section illustrated how the object feature vectors (eye texture patterns) was distributed amongst the clusters. Although this provides useful information to guide how texture clustering should be performed, it does not describe the size of the cluster, what the distribution of distracters per cluster, and how separate the clusters are from one another. This section intends to provide an answer to these questions in order to provide a more informed decision in the type of clustering adopted.

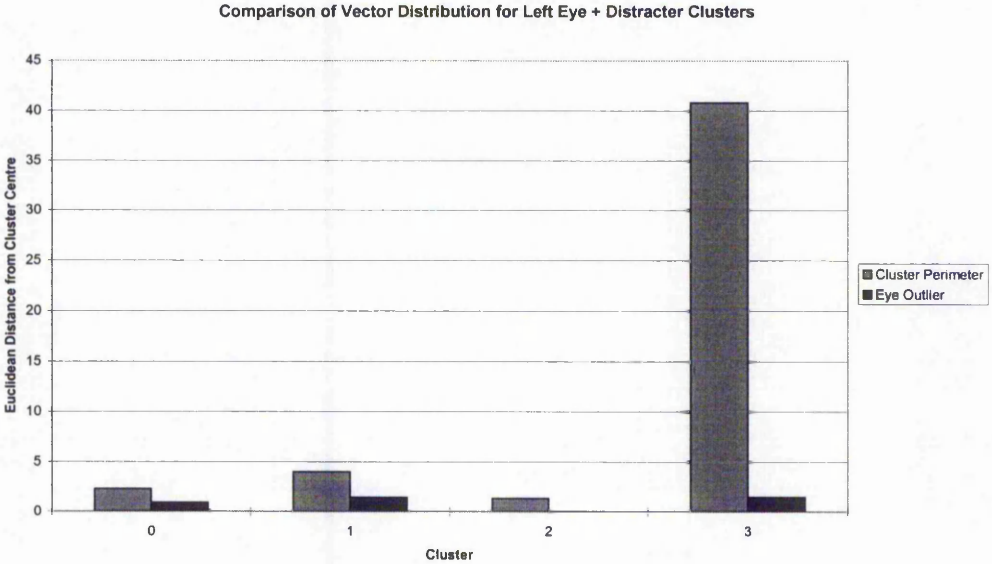


Figure 21 - Perimeter Distance for Clusters

Figure 21 shows the distribution of the perimeters and the furthest eye outlier of each cluster. The vector determines the cluster perimeter with the largest Euclidean distance assigned to that cluster. Unfortunately, this does not fully describe the distribution of vectors, primarily distracters, but only a single vector that may or may not be typical of the next closest vector.

The distribution is based upon 5 clusters. The fifth cluster, and not illustrated here, is the largest cluster and contains distracter only textures. Apart from cluster three, the remaining clusters are approximately the same size. A useful attribute of cluster 3 is that the spread of eye vectors is relatively close to the cluster centre as opposed to the perimeter. The distribution of eye vectors is known from Figure 17, which allows sensible thresholds to be applied to all clusters. This enables the filtering of vectors with large Euclidean distances which may be closest to a particular cluster but is dissimilar enough to be rejected, e.g. if the winning cluster is cluster 3, and the Euclidean distance is, say, greater than five, then this is clearly not an eye texture.

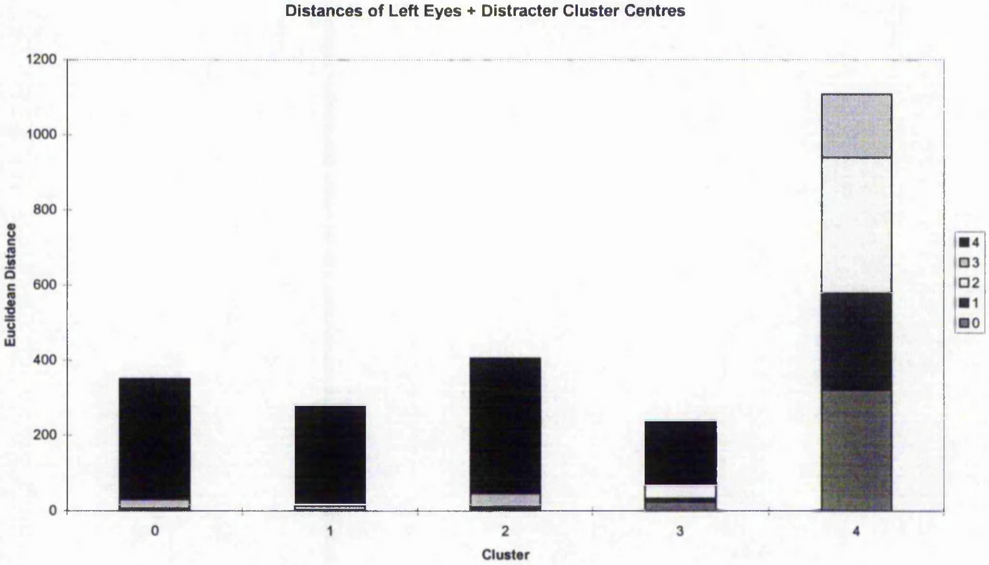


Figure 22 - Euclidean Distances from Cluster Centres

To understand the similarity of the data and how the clusters are represented in the figure above feature space shows the distances of the clusters from each centre point (using five clusters). It can be seen that cluster four, the distracter cluster, is much more distant from all other clusters, implying that the vast majority of distracter textures are unlike eye textures. Figure 21 also shows from the magnitude of the perimeter that the cluster encompasses a much greater volume of the feature space compared to any of the other clusters.

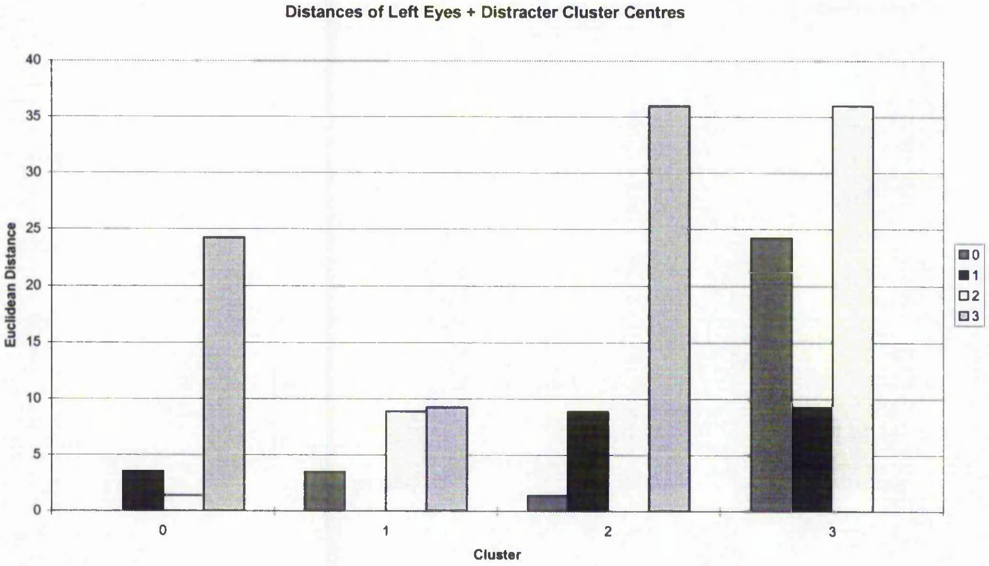


Figure 23 - Euclidean Distances from Cluster Centres

As the distance is so large for cluster four, it is difficult to examine the distances of the other clusters that contain the eye textures. Figure 23 plots the same information as Figure 20 but excluding cluster four.

Figure 20 shows the distance of the cluster centres from one another but not the distance of the cluster perimeters and whether each cluster is distinct or if overlapping occurs, i.e. the outlier vector determines the size of the volume of the cluster in all *n* dimensions. This is shown in Figure 24.

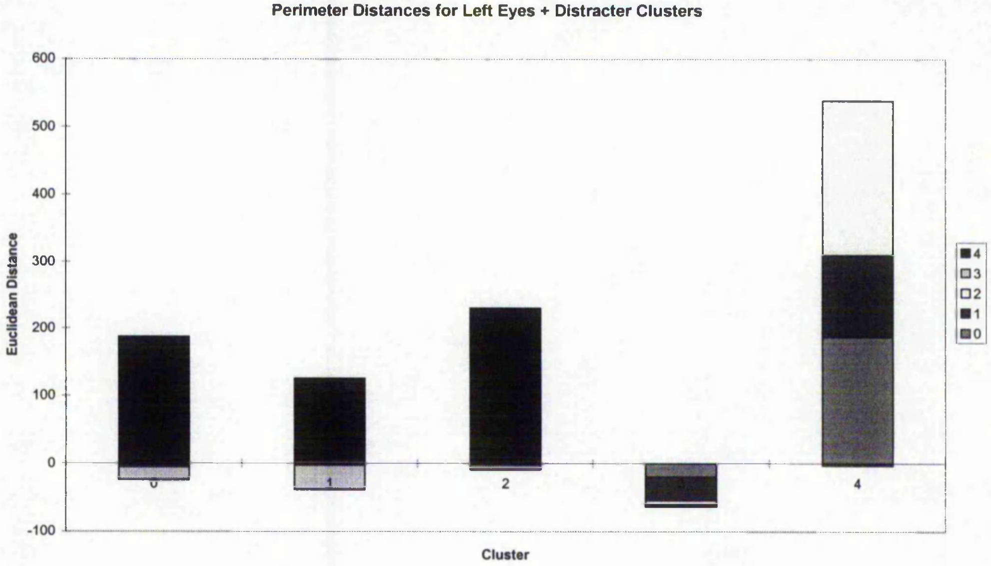


Figure 24 - Euclidean Distance from Cluster Perimeters

Like Figure 20, this illustrates further the distinct separation from all other clusters the distracter cluster occupies in the feature space. The negative distances shown in Figure 24 indicate that the perimeters overlap, which is based upon an equi-distant seven dimensional boundary. A vector which lies on the far side of a cluster may by its Euclidean distance set the perimeter of the cluster such that two or more clusters intersect.

Another aspect of the data for this problem is disproportionate number of examples for both categories. The large ratio of distracters has the possibility of affecting the position of the clusters to be biased towards these vectors. In the training of the attention-focusing ANN, replication of the class with fewer examples was successfully explored as a means to overcome this problem. This technique has been applied to clustering in order to determine the effect, if any, that replication has on the clustering

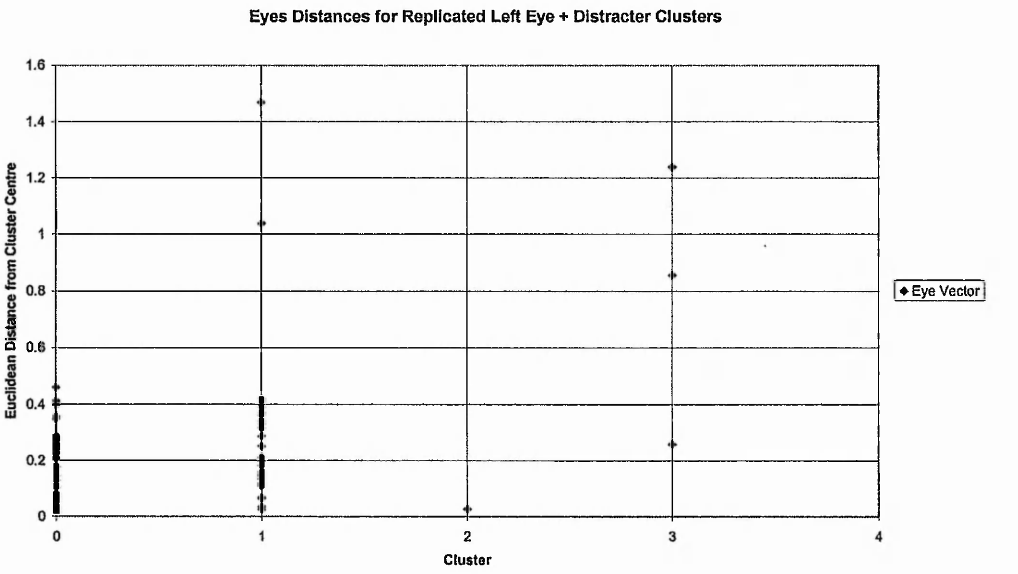


Figure 25 - Distribution of Eye Vectors using Replication

Figure 25 shows the distribution of eye vectors when these are replicated to the same ratio as the distracters. Comparing this with Figure 17, shows a slight improvement for all eye vectors, i.e. all eye vectors are closer to their respective cluster centre. The difference in the eye vectors has meant that the distribution of clustering has not altered. This fact is given further support when comparing Figure 21 with Figure 26, which shows a very similar distribution.

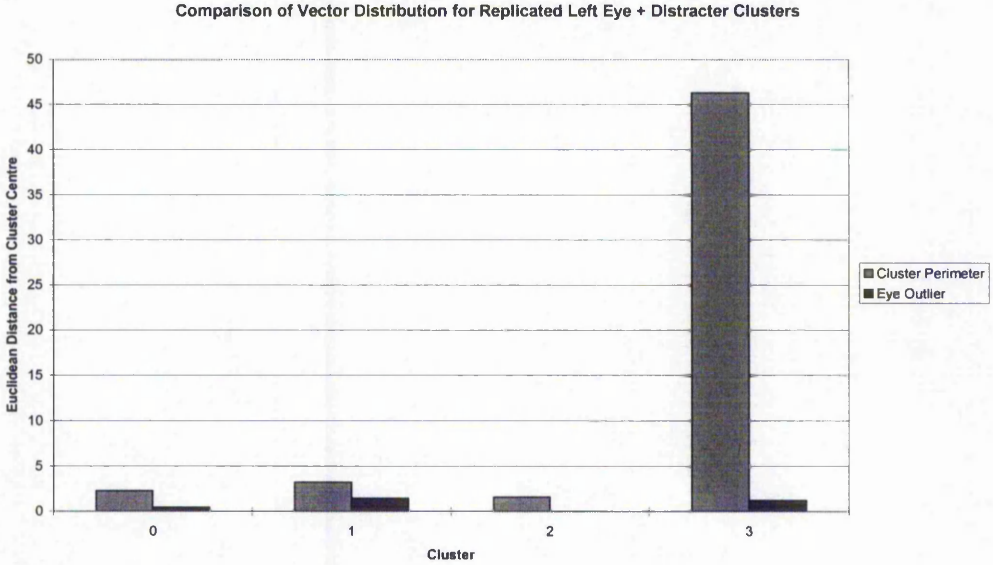


Figure 26 - Perimeter Distances for Clusters Using Replication

A different clustering approach is to generate clusters specific to eyes; i.e. to ignore distracter texture vectors when generating the clusters. Each cluster then represents a sub-set of eye texture vectors. The perimeter distance (incorporating a small margin for similar outlier eye vectors) can then be used as a threshold to reject distracter vectors that go to a particular cluster but lie outside the threshold distance. Using this approach allows the similarity/dissimilarity of the eye textures to be observed more closely and whether there are true anomaly outliers. These would be represented as lone vector clusters. This method also ensures a bias towards the eye vectors since they are the only vectors used to create the clusters. This results in the eye vectors being closer to the cluster centre than the previous cluster creation method.

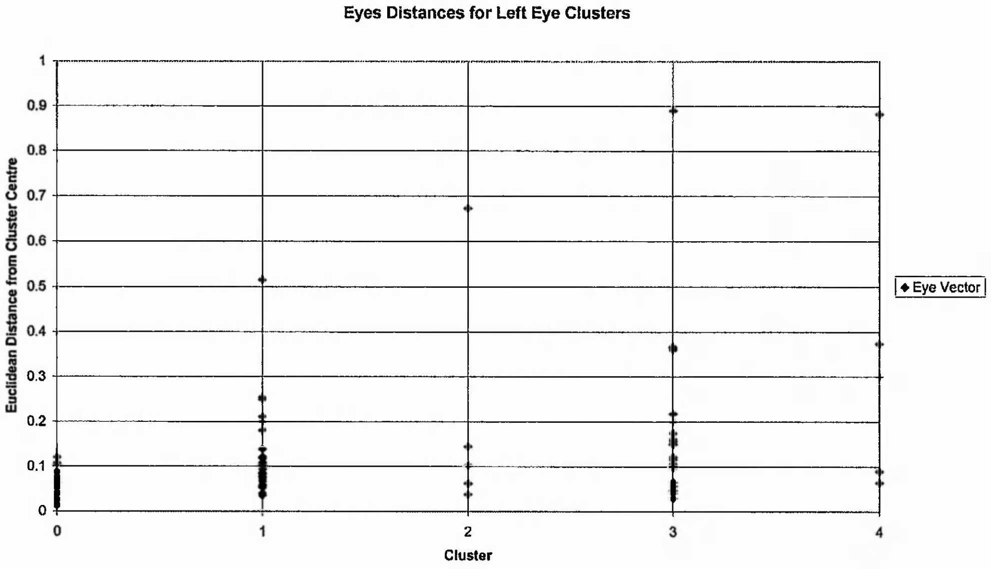


Figure 27 - Clustering with Eyes Only Textures

In comparison with Figure 25, Figure 27 shows the maximum distance of any eye vector is less than 1.0 compared with 1.4. Figure 27 also shows that the majority of eye vectors for each cluster are very similar. Typically, each cluster contains one or two eye outlier vectors, which on average double the perimeter distance. Even so, a smaller threshold (to positively reject distracter textures) can be used for this clustering method than for the combined eye/distracter clustering approach.

There are strengths to both cluster methods. Assuming the number of clusters has been set to five the advantages of both methods are as follows. For balanced object/distracter clustering, a single cluster is created that represents only distracter textures. This is particularly useful for rejecting distracter patterns. For object only clustering, more compact clusters are created which allows more refined discrimination between possible objects and distracters. Performance of both clustering methods is included in the following section.

There are two possible ways that texture analysis can be used within the classification process. Having trained a classifier ANN, texture analysis can be performed which pre-filters some of the distracter patterns. The remaining patterns are presented to the classifier ANN. Using texture analysis in this way is a simpler and alternative method of categorisation. It is also a faster method of categorisation, and only patterns that fall within the cluster perimeters are required by the ANN to classify. However, using texture analysis as a pre-filter does not reduce the complexity of the problem of the classifier ANN to generalise between the object and the distracters.

Since it has been shown that although clustering can determine the similarity of some distracter patterns, it can also provide an alternative to patterns that are clearly not the desired object. It would seem therefore a logical approach to use texture analysis to also help define the training for the classifier ANN as well as using it for categorisation in the classification process.

A method to achieve this is to direct the ANN to distinguish like object/distracter patterns and also provide greater coverage of the pattern space is to train multiple ANN classifiers. Using the object only cluster generation method allows directed training of the classifier ANNs so that distracters are chosen that are most similar to a particular subset of eye textures. This method is then able to concentrate on those patterns where there is greater uncertainty as to the correct classification. The number of ANN classifiers is equivalent to the number of clusters. For novel patterns presented for classification, the first step is to categorise the input pattern as a texture type. If the texture is sufficiently dissimilar from and of the defined texture clusters it is removed from further processing, i.e. the texture vector lies outside all of the defined cluster perimeters. If the input texture falls within the boundary of a particular cluster, that associated classifier ANN performs classification upon it.

This approach attempts to tackle two problems previously identified. By using multiple classifiers, greater coverage of the distracters can be included in training without excessively burdening the generalisation capacity of each ANN. The

ANNs are also trained more specifically with patterns that are more similar with the object they are trained to distinguish. Texture analysis also has the benefit of being able to be used with any type of ANN classifier approach adopted.

5.6 Performance Evaluation

The previous sections in this chapter have described various possible methods for classification. All of these methods use differing approaches to solving the problem of classification leading to full identification. The following sections attempt to discuss, and where appropriate, present a comparative study of these different approaches.

To evaluate the performance of classification is more difficult than the results presented in other chapters. This is because the performance of the classifier stage is generally dependent, to a certain degree, upon the success of the attention-focusing stage. If the attention-focusing stage fails to find a face, classification is unable to recover from this. Alternatively, testing the performance of the classifier in isolation is difficult as a representative of input samples need to be collected that relate to what the attention-focusing stage is likely to pass on to it. This is particularly pertinent for the texture analysis method where the training of the classification ANNs have been directed to learning specific pattern types. Therefore, the most sensible way to present the performance of the classification stage is to use the output from the most successful attention-focusing ANN and use this as the basis of the input to the classifier[9]. The performance of the classifier will therefore also show the overall performance of the identification system.

5.6.1 The Test Set

To ensure a more complete analysis of the overall performance of the ANN identification system the test set used for the analysis comprises of 30 test images.

[9] The favoured attention-focusing ANN finds all faces.

These are completely separate and different from the training set used for training both the attention-focusing ANN and the classifier ANNs. Included within these images are 43 faces that range in size, age and also differ in gender. Each image contains at least one face with some images containing two or more.

5.6.2 Criteria For Evaluation

The same performance criterion is applied to classification/identification as that used for other studies presented in previous chapters. These include:

- The number of false negatives.
- The number of false positives.

The same two measures (face distance-error and the focus area size) used in the attention-focusing analysis are also applied to determine whether an identified focus area is a true positive or false positive. It is possible for the classification stage to determine new values for the size and position of focus areas. In fact focus areas identified at the attention-focusing stage may be marginal false positives until classification re-adjusts the size or position of the focus area to become a true positive. Conversely, classification may perform the opposite. However, to highlight the performance of classification, none of the focus areas presented to it from the attention-focusing stage have been altered in any way other than to re-apply a new classification category.

Several different approaches to classification have been discussed in this chapter. The following sections describe a summary of the results from these different methods.

5.6.3 Binary and Grey-Level ART Classification

The binary ART method is relatively well defined and only a few modifications have been made to the original ART algorithm in order to present images as the input to the ANN. These are notably, presenting positive and negative images to define the ANN connectivity, adding an extra a fully connected hidden unit and adjusting the learning algorithm in the *backprop* stage to ensure all hidden units have an equal weighting.

However, the grey-level ART paradigm has been more of an investigation into the feasibility of adapting the binary ART model to process the grey-level images directly without the need to pre-process the image, and therefore potentially lose characterising attributes.

The first decision in training the ART ANNs is to determine the classification resolution. Three separate resolutions have been investigated. These include eye distances of 8, 12 and 16. (The attention focusing stage employing an eye distance of 4). An eye distance of 16 is close to the maximum resolution the images can be increased without going to full resolution. The size of the input layer is scaled according to the eye distance based upon the original attention-focusing ANN input size. For the eye distances mentioned the input layer sizes are 22x32, 33x48, and 44x64.

To determine appropriate values for the adequacy and margin is dependant upon a number of factors:

- The adequacy determines the minimum number of connections that can be used to describe a feature. A reasonable value would relate to, say, approximately 15% of the total input area. Any value less than this threshold is more likely not to indicate any meaningful holistic feature. Having too high an adequacy value would force a representation to include further connections which are more specific to particular face instances than being general. Obviously a high adequacy value would increase the connectivity for each hidden unit but also increase the total the number of hidden nodes in the whole classifier ANN.
- The margin determines how distinct the clusters have to be with one another. To ensure that the clusters' connections differ by a reasonable margin, the margin threshold should be between 40-80% of the adequacy value. Too low

a margin value and there is a possibility that non-similar input patterns may be merged together. Too high a margin value will reduce the generalisation of the hidden nodes and increase the overall number of hidden nodes in the ANN.

- The grey-level difference is a threshold particular to the grey-level ART paradigm. This value clusters the grey level values into bands, so that small variations in intensity can be grouped together as part of the same feature. Whereas the margin and adequacy compare connectivity, the grey level difference determines how close a grey level pixel is to another. The parameter can be best described as acting like a moving binary threshold. The larger the grey level difference, the greater the range of grey level values that are sufficiently similar, and vice versa. To determine an appropriate value is more a case of trial and error. A value of 0.5 (since all grey-level values have a value between 0 and 1) should configure the grey-level ANN to perform in a similar fashion to the binary ART paradigm.

The intention for the grey-level ART has been to provide greater flexibility in determing the holistic features through the use of a more sophisticated algorithm that improves on the cluster grouping selection criteria. Unfortunately, although several parameter variations have been investigated the grey-level ART does not perform as well as expected. Meaningful clusters appear to be extracted, as can be seen in Figure 28, and also the ANN generalises to some extent to distinguish between the face and background distracters. Unfortunately, initial studies using the grey-level ART classifier showed that the ANN did not distinguish between faces and distracters successfully and showed far worse performance than the attention-focusing ANN classifier. In comparison, approximately 70% of faces were classified correctly with the grey-level ART ANN classifier compared to 100% classification with the attention-focusing ANN.

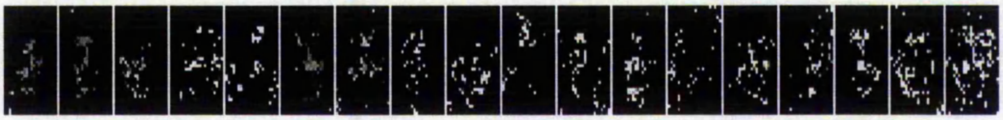


Figure 28 - Grey-Level Art Hidden Unit Clusters

The advantage of using a holistic approach to classification is that the level of detail required should not be as great as that required by the sub-feature classifier method as the whole feature. i.e. In this case areas throughout the whole face are being used rather than just a selected part of the focus area. Analysis of the ANN structure for the resolutions using an eye distance of 12 and 16 created large ANN topologies, even taking into account partial connectivity. It has already been established by Atlas [Atlas *et. al*, 1989] that large ANNs have poor generalisation capability and this was found in the attempts to train various ANNs with a large number of connections. Consequently, ANNs trained of this size could not learn to distinguish between faces and distracters with an accuracy for both pattern types of above approximately 60%.

The resolution chosen for the analysis of the binary ART ANN method uses an eye distance of 8 which relates to double the resolution of the attention-focusing ANN. This should be sufficient for accurate object classification. Using an eye distance of 8 provides a good balance between an increase in the amount of data and is still less than a classification system using full resolution analysis.

Binary Art Classifier Performance

ANN	Mean FPs Per Image	Mean Classifier FNs Per Image	Total No. of Classifier FNs	Classifier Topology	Classifier RMS	Test Description
1	23.1	0.0	0	22x32x56	0.0483784	Standard 1 feature face only classification using 3 resolution search
2	3.8	0.6	17	22x32x56	0.0431267	Higher Resolution focus areas filtered beneath classified low res FA
3	28.5	0.0	0	22x32x56	0.0328473	Standard 1 feature face only classification using 3 resolution search
4	20.8	0.0	0	22x32x56	0.0328403	Standard 1 feature face only classification using 3 resolution search
5	12.4	0.0	1	22x32x41	0.0585977	Standard 1 feature face only classification using 3 resolution search
6	20.8	0.0	0	22x32x41	0.0463396	Standard 1 feature face only classification using 3 resolution search
7	20.0	0.0	0	22x32x41	0.0383208	Standard 1 feature face only classification using 3 resolution search
8	5.0	0.1	4	22x32x41	0.0339621	Standard 1 feature face only classification using 3 resolution search
9	6.8	0.5	16	22x32x41	0.0339621	Higher Resolution focus areas filtered beneath classified low res FA
10	19.1	0.0	1	22x32x41	0.0339621	Standard 1 feature face only classification using 3 resolution search
11	13.9	0.0	0	22x32x41	0.0271449	Standard 1 feature face only classification using 3 resolution search
12	18.9	0.0	0	22x32x41	0.025171	Standard 1 feature face only classification using 3 resolution search
13	5.6	1.0	31	22x32x41	0.0234918	1 feature face only classification with hidden unit profile discrimination
14	8.4	0.1	3	22x32x41	0.0234918	Standard 1 feature face only classification using 3 resolution search
15	16.8	0.0	1	22x32x41	0.0219071	Standard 1 feature face only classification using 3 resolution search
16	4.8	0.2	5	22x32x41	0.0207085	Standard 1 feature face only classification using 3 resolution search
17	15.8	0.0	1	22x32x41	0.0200979	Standard 1 feature face only classification using 3 resolution search
18	10.0	0.1	3	22x32x41	0.0200979	1 feature face only classification using 3 resolution search. Using 0.6 threshold
19	46.6	0.0	1	22x32x41	0.0196245	Standard 1 feature face only classification using 3 resolution search
20	15.9	0.0	1	22x32x41	0.0194414	Standard 1 feature face only classification using 3 resolution search
21	1.7	0.5	13	22x32x37	0.0113408	Higher Resolution focus areas filtered beneath classified low res FA
22	14.0	0.4	11	22x32x37	0.0113408	1 feature face only classification using half resolution
23	5.0	0.3	8	22x32x37	0.0113408	Standard 1 feature face only classification using 3 resolution search
24	3.3	0.5	13	22x32x37	0.0113408	1 feature face only classification using single resolution search
25	9.9	0.0	1	22x32x37	0.0109504	Standard 1 feature face only classification using 3 resolution search
26	1.9	0.5	15	22x32x37	0.0109504	Higher Resolution focus areas filtered beneath classified low res FA

Table 4 - Binary ART ANN Classifier Performance

Table 4 shows a number of ART ANN performance studies to determine the best classification approach and model. The table presents three different ART derived architectures, all at an eye distance resolution of eight pixels. The difference in ANN size is dependent upon the ART parameters used. A number of different results are presented that relates to a number of different changes. These include:

- The performance of the ANN through training, i.e. the classifier performance at different stages of RMS.
- The number of resolution searches performed at classification, e.g. single or three-resolution search.
- Using selected true positive hidden unit profiles to influence final classification.
- Using positive classification to filter remaining unclassified focus areas.

The results have not been ordered in terms of ANN classifier performance but rather in ANN connectivity and decreasing RMS. Typically the fewer connections the ANN contains the smaller the RMS value. However, this does not necessarily provide better performance. The RMS of the ANN in fact does not seem to correlate at all with overall performance. This is perhaps due to the supporting algorithms and techniques that manipulate the focus points and areas that provide the final performance.

The default method is to perform a three resolution search around the determined resolution, grouping firstly at a single resolution and then across the three resolutions. Other methods investigated include an attempt to classify at only a single increased resolution; increasing the threshold of the ANN output for what constitutes a focus point; and finally to use the ordering of classification to automatically remove focus areas that lie under a larger classified focus area.

The final method (positive discrimination of focus areas) is desired because this would remove the need to process many focus areas at higher resolution. Many focus areas overlap spatially and this method shows that a significant number of focus areas are removed in this way. Unfortunately, the results also show that this inevitably removes true positive classifications (if at a higher resolution) and retains a false positive focus area.

The best overall performance is ANN 11. Firstly, it identifies all true positives correctly and is the ANN that identifies the fewest number of false positives. Unfortunately, the mean number of false positives per image is still relatively high. Other variations improve on the false positive performance but then misclassify some of the true positives.

This highlights a number of issues already discussed throughout this thesis. The performance of the ANN is erratic through training, but the general trend is that after a certain amount of training although the true negative performance will generally improve the true positive performance does not.

Although the holistic approach to classification, for some results presented, has made a general improvement on classification for the focus areas passed to it by the attention-focusing stage the performance is still disappointing. Post classification still presents an unacceptable number of false positives.

The binary ART method is only one approach to classification and the following sections discuss the merits of the sub-feature approach as well as investigating the complimentary technique of texture analysis.

5.6.4 Sub-Feature Classification

The main difference between sub-feature classification and the holistic approach is that sub-feature classification requires more detail to perform classification. However, the sub-features are limited in size, which means that the ANN topology is not excessively large. In fact, the size of the classifier ANN is no larger than that required for the attention-focusing ANN. The performance results presented in **Table 5** use an eye distance of 12 and 16.

Sub-Feature Classifier Performance								
ANN	Mean FPs Per Image	Mean Classifier FNs Per Image	Total No. of Classifier FNs	Eye Distance	Classifier Topology	Classifier Rms	No. of Classifiers	Test Description
1	2.3	1.0	30	12	12x8x10	0.0441249	1	Standard 1 feature left eye only classification using 3 resolution search
2	12.8	0.1	3	12	12x10x15	0.0219234	1	Standard 1 feature left eye only classification using 3 resolution search
3	13.7	0.2	5	12	12x10x15	0.0219234	1	Standard 1 feature right eye only classification using 3 resolution search
4	8.7	0.3	8	12	12x10x15	0.0219234	2	Standard 2 feature left AND right eye classification using 3 resolution search
5	7.7	0.4	10	12	12x10x15	0.0180799	1	Standard 1 feature left eye only classification using 3 resolution search
6	9.0	0.4	11	12	12x10x15	0.0180799	2	Standard 2 feature left AND right eye classification using 3 resolution search
7	5.0	0.1	2	16	16x13x15	0.0074871	1	Standard 1 feature left eye only classification using 3 resolution search
8	4.4	0.1	3	16	16x13x15	0.0074871	1	Standard 1 feature right eye only classification using 3 resolution search
9	3.6	0.3	8	16	16x13x15	0.0074871	1	Standard 1 feature left eye only classification using single resolution search. Single resolution focus areas removed.
10	4.8	0.0	1	16	16x13x15	0.0074871	2	Standard 2 feature left OR right eye classification using 3 resolution search

Table 5 - Sub-Feature ANN Classifier Performance

Classification of the true positives at an eye distance of 12 is poor throughout. Classification of the false positives at this resolution is better than the holistic approach. ANN 1 shows good false positive classification, but extremely poor true positive classification and can be rejected on this basis.

Increasing the classification resolution does improve the performance for both true positives and false positives. However, none of the results presented in

Table 5 have successfully classified all of the true positives. ANN method 10 comes closest but does not classify all true positives correctly. Although false positive classification is better than the holistic approach none of the methods satisfy the primary objective that all true positives shall be classified correctly. This is a difficult decision since only one face focus area has been classified incorrectly for ANN 10.

Turning the attention to particular aspects of the performance studies highlights some interesting analysis. ANNs 2 and 3, and ANNs 7 and 8 show the comparative classification performance between two identical ANNs looking for the same type of feature. In this case one ANN classifying left eyes and the other classifying right eyes. Interestingly the classification performance is worse for right eye classification for both ANN studies. Examining the images, there are instances where hair obscures the feature of interest. This would suggest that classification is very much driven by the images used. Although classification is shown to better for left eyes, expanding the test set might provide examples where this is not the case.

ANN 10 combines both sub-features into the classification relying on either to affirm a positive classification. This improves the true positive performance, and false positive performance for this arrangement will always relate to the worst performing sub-feature classifier. Other results show when ANNs must both produce a positive classification. As can be seen, true positive classification suffers because both sub-features need to be identified correctly in order to affirm the focus area is a true positive. Obviously, this method does help to reduce the number of false positives.

5.6.5 Sub-Feature Classification using Textures

As described in previous sections, texture analysis is an attempt to help segment the feature space in order to help remove the burden upon classification. Sub-feature classification has been used to investigate this method but there is no reason why it cannot also be applied to holistic classification also.

Sub-Feature Texture Classifier Performance

ANN	Mean FPs Per Image	Mean Classifier FNs Per Image	Total No. of Classifier FNs	Eye Dist.	Classifier Topology	ANN Rms	No. of Classifiers	Test Description
1	3.8	0.0	0	16	16x13x15	0.027526	1x5	Standard 1 feature left eye only classification using 3 resolution search. Texture clusters derived from left eyes only.
2	28.4	0.0	0	16	16x13x15	0.0180428	1x5	Standard 1 feature left eye only classification using 3 resolution search. Texture clusters derived from both left eyes and distracters.
3	5.6	0.0	0	16	16x13x15	0.0159132	1x5	Standard 1 feature left eye only classification using 3 resolution search. Texture clusters derived from left eyes only.
4	5.0	0.1	2	16	16x13x15	0.0139436	1x5	Standard 1 feature right eye only classification using 3 resolution search. Texture clusters derived from both right eyes and distracters.
5	4.9	0.0	1	16	16x13x15	0.0139436	1x5	Standard 1 feature left eye only classification using 3 resolution search. Texture clusters derived from both left eyes and distracters.
6	4.0	0.1	3	16	16x13x15	0.0139436	2x5	Standard 2 feature left AND right eye classification using 3 resolution search. Texture clusters derived from both left eyes and distracters.
7	5.0	0.1	2	16	16x13x15	0.0139436	2x5	Standard 2 feature left OR right eye classification using 3 resolution search. Texture clusters derived from both left eyes and distracters.
8	2.4	0.2	4	16	16x13x15	0.0125595	1x5	Standard 1 feature left eye only classification using 3 resolution search. Texture clusters derived from both left eyes and distracters.
9	3.9	0.2	5	16	16x13x15	0.0125595	1x5	Standard 1 feature right eye only classification using 3 resolution search. Texture clusters derived from both right eyes and distracters.
10	9.5	0.4	11	16	16x13x15	0.0125595	1x5	Standard 1 feature left eye only classification using 3 resolution search. Texture clusters derived from both left eyes and distracters. Higher Resolution focus areas filtered beneath classified low res FA
11	2.6	0.2	6	16	16x13x15	0.0125595	2x5	Standard 2 feature left AND right eye classification using 3 resolution search. Texture clusters derived from both left eyes and distracters.

Table 6 - Sub-Feature Classifier Performance using Texture Clustering

The picture that is obtained immediately from the results shown in Table 6 is that texture analysis is a worthwhile method and does improve the overall classification. Two approaches to create the initial texture clusters have been examined. The first method is to use only the item of interest to categorise the clusters, and the second method uses all pattern types. ANNs are then trained for each object cluster using the usual method of training. The only differences are that patterns are pre-selected according to their texture type. For the first method this means an ANN is created for each texture cluster. For the second approach one cluster is a distracter only cluster and there is no ANN for this. Any textures that are assigned to this cluster are instantly rejected as distracters.

Of the two clustering methods, ANN 1 provides the best performance of the two clustering methods. The ANN 1 method uses object only clustering. The results for this ANN method also show the best overall performance of all the performance studies presented.

Other similar trends are evident that are present in sub-feature classification. These include the performance of left and right eye classification where right eye classification performs slightly poorer than its counterpart. Also, although the combination of sub-features helps reduce false positives it also impacts on true positive performance.

The identification results for the best approach, i.e. ANN method 1 in Table 6 are included in appendix E. This includes the initial output from the attention-focusing stage and the final output after classification. Similar to other results presented, focus areas highlighted in white represent true positive classification and focus areas highlighted in black represent false positive classification.

5.7 Summary

This chapter has investigated the classification stage of the identification process. Various approaches have been presented. Of the different models, the classification system that has shown best performance is the sub-feature classifier trained on left eyes using texture analysis as a pre-filter. Texture analysis has been proven to compliment the classification process and allows the number of ANNs performing classification to be extended and thus provide greater coverage in the feature space.

The sub-feature classification approach with texture analysis is able to classify all true positives correctly and also significantly reduce the number of false positives. Even so the mean number of false positives per image is approximately four. Other variations (such as multiple sub-feature classification, filtering of positively classified focus areas) that have been investigated to reduce the number of false positives have also affected true positive classification. The next chapter re-evaluates the approach used for object identification and discusses problems associated with developing a generic solution.

Chapter 6

Conclusions and Further Work

6.1 Summary

In the first chapter a set of requirements were outlined that it was hoped the research would fulfil. These are basic requirements if the intended Computer Vision system is to be used in a meaningful way. However, the implication of providing a solution to all of these problems is not trivial.

A model has been presented that does address these requirements, and does go a long way in succeeding to provide this functionality. However, the ultimate criteria for being a robust solution has not been met. The model, even though it is able to identify all faces it is incapable of removing all false positives. This is perhaps not unexpected given that comparable solutions by Sung and Poggio [Sung and Poggio, 1994], and Rowley [Rowley *et. al*, 1995] also present a degree of error.

The possible variation in background distracter is almost infinite and to train an ANN that can cope with this variation is possibly an ill posed problem. A method to improve the classification of particular pattern types is to include representative samples into the training data. However, the number of these examples can rapidly increase, such that it is unfeasible to continue to do so. Also, a point is reached where the capacity of the ANN to cope with such a large number of background distracters in the training set can only have a detrimental effect on the true positive classification performance.

Even though the system copes remarkably well, and processes the image efficiently, a more focused problem domain may have provided better overall identification performance. It is not unreasonable that, as the object of interest is pre-selected at the start, appropriate environments where the object may exist could also be selected. This is not to say the background information is any way less complicated but the scope of the variety that the identification system may examine is more constrained.

The architecture presented is flexible enough to be applied to similar problem domains. Even though faces have been used throughout this thesis as an example image analysis problem, the techniques presented are still applicable to all other image identification related problems. For example, an automated cancer cell screening system could be developed using this architecture. The object of interest would be an abnormal cell and distracter information would be all other tissue on the sample image. Similarly to faces, cancer cells are hard to describe yet provide identifiable characteristics that identify them as being such.

An identification system that uses a two-stage strategy to identification has proven to be a valid approach. It is able to perform more of its processing at lower resolutions than any comparable system. The outcome of processing less information is that it is extremely efficient and allows real-time processing to be realised.

The following sections will concentrate upon identifying the various limitations of the approach and possible ways to improve the model presented.

6.2 The Attention-Focusing ANN

The attention-focusing ANN is the first stage of the identification system and the resulting output is fundamental to the final identification performance. Finding the optimum resolution is key to the ANN strategy so that it is unnecessary to process an image at full resolution. The degree of reduced resolution has a large impact on the amount of processing required. A balance must be found between the advantage of using a lower reduction in comparison to the error that this introduces. The advantage of using ANNs that comprise any processing system is the ability to define an architecture without being restricted by set rules and methods. On the downside this puts great reliance upon the quality of the ANN training. This was shown to be a problem for both the attention-focusing ANN and also the classification ANN.

6.2.1 Selecting the Training Data

The objective to deal with any real world image presents a problem for ANN training as the quality and number of training examples has to be carefully selected. Conclusive studies were made to determine the optimum resolution for the attention-focusing ANN, and the size of the ANN reflected the resolution required. Although great thought was given to the ANN training, it was in some respects very simplistic. Throughout this research, it has become apparent that analysis of real world data requires a substantial quantity of training data. Although many examples of different faces and varied background are needed, it is very difficult to select distracter information that is representative of all possible patterns.

As the variety of test images are increased some patterns are better generalised to than others. However, for some particular novel patterns it also does not generalise well to and produces a high proportion of false activity. For these types of patterns these were re-introduced into the training which reduced the false activation.

This creates a dependency upon the ANN to continue to retrain with input patterns that it does not generalise to. In this respect, the ANN may never be able to generalise to all new patterns. This is fine for a specified domain where knowledge about the environment can direct the training, but less useful for generalised solutions that this model attempts to provide a solution to.

Organisation of a large number of examples has a definite effect on the learning of the ANN. As well as collecting the training data, it needs to be ordered in a specific manner to enable the ANN to generalise. Replication and pattern ordering was found to help training of both ANNs. Before this was introduced, using a large number of training patterns created instability in generalising to the two pattern categories and, also because more background distracters were available, created a bias towards these pattern types. Unless this is configured appropriately the ANN has a tendency to fall into local minima.

Even supplying an equal number of positive examples it does not really solve the eventual bias that the ANN training leads to, i.e. towards background distracters. This could be a problem with re-enforced replication. The overall large number of patterns should remove any possibility of the ANN memorizing any patterns no matter how frequent the replication. However, to gather the number of example face patterns required to balance the number of distracters is unfeasibly large. This is an obvious problem in itself that addresses the method by which simple capture of the relevant input patterns is achieved, but it is pure conjecture whether significant gains would be seen over using replication.

6.2.2 Over-Training

In order to cope with the large variety of pattern types, the number of patterns presented for learning is well into the tens of thousands. The background distracters are reinforced with positive object pattern types. Obviously the large volume of training examples ensures that the ANN cannot memorize the pattern types. A problem however does occur that after a certain period of training, the ANN performance continues to improve background distracter classification at the expense of the positive object patterns. Unfortunately, the RMS is not a good metric to use as the RMS value continues to decrease. The most reliable way to ensure that training has not generalised at the expense of the positive classification is to constantly test the ANN whilst training. This obviously creates an added burden upon training times. Unfortunately, no alternatives have been found to determine the best cut-off point for ANN training. Although, rules of thumb are available to guide training [Swingler, 1996] no techniques were found in the literature survey that deals with the issue of two category problems encompassing extremely large data sets. None of the other methods that use ANNs for pattern classification discuss training issues for their ANN paradigms. However, it is evident from the performance results that all systems exhibit similar generalisation problems, i.e. general failure to classify some test pattern types.

6.2.3 Single Resolution Grouping of ANN Output

Although two different approaches to focus point grouping were presented and investigated, the extremities method was the method adopted to present the attention-focusing and classification/identification results. However, there are limitations for both grouping methods. The window grouping method tends to produce many more focus areas than the extremities grouping algorithm. The extremities grouping algorithm is unable to remove all false negatives. A limitation of the extremities grouping algorithm could be addressed by modifying the criteria on how it currently creates clusters. A better method should take into account of:

- ➢ The contour of the cluster. Only focus points that are contiguous to another should be grouped.
- ➢ Limit the size of the cluster. This should help separate multiple focus clusters that are close together.

This is a more intelligent grouping algorithm, which should produce a slight increase in focus areas but to the advantage of fewer false negatives. The relatively high failure rate of both methods is due to the unpredictable nature of ANN misclassification.

6.2.4 Multi-Resolution Grouping of ANN Output

Multi-resolution analysis is used to perform two functions. The first is to identify the size and position of possible faces in the input image. The second is to provide support for this by analysing the focus point distribution across other resolutions. Examination of focus point output across different spatial resolutions does show that generally the ANN does activate for the trained object across small differences in spatial resolution. Single resolution activation that might be passed on to Classification can also be filtered. Smaller image reductions have been chosen exactly for this reason. Grouping of focus points are made at one resolution and then grouped across resolutions. This is fairly crude although any alternative would soon become very complicated. Although various other techniques were looked into, e.g. centre of gravity to obtain a better mid point, it was deemed that a lot of effort might be expended upon a better solution without any significant gains. This due in part to the unpredictable nature of one trained ANN against another. A small amount of effort has been made into the distribution of focus points and their values but no trends were identified that would the grouping and potential classification.

6.3 ANN Classification

The previous section critically analysed various aspects of the attention-focusing system. Some of the methods discussed are also applicable to classification, e.g. the ANN training and also the clustering techniques. Although there are slight differences to classification, these techniques are applied in essentially the same manner and thus the same problems are also pertinent.

6.3.1 ANN Models for Classification

Using an ART ANN to determine holistic features use this to specify the connectivity of the classifier ANN is a completely novel approach. Because of the way the features are determined automatically, this was initially the preferred method for the definition of the classifier. The paradigm lends itself to being a more generic solution than the sub-feature ANN method and could be more readily applied to other computer vision related problems. However, a combination of other techniques provided better means of classification.

Both ART ANN approaches described in this thesis differ from the original definition [Carpenter and Grossberg, 1985] and even the later revision [Carpenter and Grossberg, 1987]. The binary ART ANN presented deviates only in the manner in which input patterns are presented; i.e. both positive and negative examples are presented for training. The grey-level ART however is an attempt to process continuous values based upon the original ART algorithm. Even though the ART2 paradigm also attempts to do this, it is done in a much more crude fashion. Classification performance was disappointing from the grey-level ART ANN. The reason for this is unknown, but could be attributable to the configurability of the parameters. Although different configurations have been examined, the range of values that the parameters may hold has not been completely investigated, and as such there may be scope to investigate this further.

Using an unsupervised ANN architecture to determine connectivity is also an interesting idea that could be exploited further. There is no reason why the technique cannot be applied to other areas. It can be argued that this method is applicable to any problem requiring feature extraction. Either of the two topics discussed could form the basis of a research project in their own right.

6.3.2 Texture Analysis

To reduce the complexity of the problem and the burden of an ANN to learn many different background types, texture analysis has been suggested as a means to improve classification. This technique may have been applied before the attention-focusing stage but there are several reasons why this has not been implemented. The attention-focusing stage works reasonably well for the resolution the ANN examines. Applying texture analysis would require a significant amount of further processing. The aim of the classifier is to perform a more accurate analysis of possible faces. Therefore, in order to achieve this, analysis would be also required before classification. To perform this technique twice was therefore considered not to be computationally viable or promise a significant improvement on identification accuracy.

6.4 Further Work

This thesis has only dealt with a specific set of vision problems. Other aspects that the model could be adapted to consider include:

- Object profiles. This thesis only considered objects that were shown face-on to the camera. In true real world images it is likely that some objects would be shown in various degrees of profile. A full classification system would need to be able to identify an object from any angle.
- Object rotation. As for object profiles, objects in real world images make contain a degree of rotation, and therefore this would also need to be addressed by an identification system.
- Partial object occlusion. A particular issue that was not truly discussed in this thesis was if an image contained features that were not found in every example of the object. For example, glasses or facial hair. Although, these items can be said to occlude parts of the image, the face is still humanly recognisable.
- Moving images. This thesis has only considered static images. A natural progression for image analysis is to extend the problem to consider moving images. This is generally achieved by comparing consecutive time window frames and examining the difference between them to identify any changes that have occurred. Having identified the object of interest in one frame of the image, it would then be possible to track it throughout subsequent images. Processing temporal sequences however is computationally more expensive and would require sophisticated algorithms to cope with the considerable increase in the amount of information.
- Colour images. Grey-level images have been used within this research.

These were chosen to simplify the amount of information presented to an ANN for processing. The model could be extended to consider colour images that might provide additional information to aid classification. The human visual system uses colour along with shape and texture to identify objects, and it would seem logical to explore the effect of using colour images with regard to identification performance.

- Recognition. It has already been discussed in Chapter 1 that a face identification system could be a useful precursor to face recognition. There are many applications where identification could be used along with face recognition. An example of this could be a crowd surveillance system. Recognition however, is more applicable to faces.
- Depth. The identification system developed has performed image analysis with what could be argued as a software emulation of a single camera. Employing the technique of stereoscopic imaging would allow 3D information to be inferred from an image. The use of this would be more applicable for problem domains such as robotics in which 3Dinformation is required for navigation.

The goal to develop a system that is capable of processing any real world images is extremely challenging and the literature provides many examples that attempt to tackle a single area of this problem, as well as some object identification systems that have a similar objective.

It can be said that we are still a long way off from realising a generic object identification system that is reliably robust, and is able to perform accurately using a diverse set of real world image input. However, the potential benefits that could be gained from automated systems that are able to perform human vision activities are enormous.

References

[Adams *et. al*, 92], Adams, F.W., Jr.; Nguyen, H.T.; Raghaven, R.; Slawny, J. "A parallel network for visual cognition", IEEE Transactions on Neural Networks, 1992, Vol: 3 Iss: 6 p. 906-22.

[Ahmad and Omohundro, 1990], Ahmad, S.; Omohundro, S. "Equilateral triangles: A challenge for connectionist vision", Proceedings of 12th Annual Conference of Cognitive Science, 1990, p. 629-36.

[Allinson and Johnson, 1992],Allinson, N.M.; Johnson, M.J. "Application of self-organising digital neural networks to attentive vision systems", IEE International Joint Conference on Image Processing and its Applications, 1992, p. 193-6.

[Anderson, 1990], Anderson, G.J. "Focused attention in three-dimensional space", Perception and Psychophysics, 1990, vol: 47 p. 112-20.

[Atlas *et. al*, 1989], Atlas, L.; Marks II, R.; Donnel, M.; Taylor, J. "Multi-scale dynamic neural net architectures", IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, 1989, p. 509-12.

[Barker *et. al*, 1995], Barker, S.E.; Powell, H.M.; Palmer-Brown, D. "High speed face location at optimal resolution", World Congress on Neural Networks, 1995 Annual Meeting of the International Neural Networks Society, 1995.

[Barker *et. al*, 1996], Barker, S.E.; Powell, H.M.; Palmer-Brown, D. "Size invariant attention focusing (with ANNs)", Proceedings of the International Symposium on Multi-Technology Information Processing, Taiwan, 1996.

[Beasley *et. al*, 1993], Beasley, D.; Bull, D. R.; Martin, R. R. "An overview of genetic algorithms: part 1, fundamentals", Inter-University Committee on Computing, USA, 1993, vol: 15 no. 2, p. 58-69.

[Bischof, 1995], Bischof, H. "Pyramidal neural networks", Lawrence Erlbaum Associates, New Jersey, USA, 1995.

[Bouattour *et. al*, 1992], Bouattour, H.; Fogelman Soulie, F.; Viennet, E. "Neural nets for human face recognition", IEEE International Joint Conference on Neural Networks, IEEE New York, USA, 1992, p. 700-704 vol. 4.

[Bruce, 1988], Bruce, V. "Essays in cognitive psychology : recognising faces", Lawrence Erlbaum Associates, Hove, UK, 1988

[Brunelli and Poggio, 1993], Brunelli, R.; Poggio, T. "Face recognition: features versus templates", IEEE Transactions on Pattern Analysis and Machine Intelligence, 1993, vol: 15 no. 10 p. 1042-52.

[Carpenter and Grossberg, 1985], Carpenter, G. A.; Grossberg, S. "Category learning and adaptive pattern recognition, a neural network model", Proceedings of the Third Army Conference on Applied Mathematics and Computation, ARO Report 86-1, 1985, p. 27-56.

[Carpenter and Grossberg, 1987], Carpenter, G. A.; Grossberg, S. "ART2: self-organization of stable category recognition codes for analog input patterns", Applied Optics,, 1987, vol: 26 p. 4919-30.

[Dayhoff, 1990], Dayhoff, J.E. "Neural network architectures : an introduction", Van Nostrand Reinhold, New York, USA, 1990

[Edelman *et. al*, 1992], Edelman, S.; Reisfeld, D.; Yeshurun, Y. "Learning to recognise faces from examples", Proceedings of Second European Conference on Computer Vision, Springer-Verlag, 1992, vol: 1 p. 787-791.

[Evans *et. al*, 1991], Evans, M.R.; Ellacott, S.W.; Hand, C.C. "A multi resolution neural network classifier for machine vision", IEEE International Joint Conference on Neural Networks, IEEE New York, USA, 1991, p. 2594-9 vol. 3.

[Feldman and Ballard, 1982], Feldman, J.A.; Ballard, D.H. “Connectionist models and their properties”, Cognitive Science, vol. 6 p. 205-254.

[Fukanaga, 1990], Fukanaga, K. “Introduction to statistical pattern recognition”, Academic Process, Boston, 2nd Edition, 1990.

[Fukuda *et. al*, 1992], Fukuda, T.; Itou, S.; Arai, F. "Recognition of human face using fuzzy inference and neural network", IEEE International Workshop on Robot and Human Communication, 1992, p. 375-80.

[Garrison *et. al*, 1990], Garrison, W.; Cottrell W.; Fleming, M. "Face recognition using unsupervised feature extraction", Proceedings of International Network Conference, 1990, p. 322-5.

[Gouhara *et. al*, 1991], Gouhara, K.; Watanabe, T.; Uchikawa, Y. “Learning process of recurrent neural networks”, IEEE, 1991, p. 746-51.

[Grossberg *et. al*, 1989], Grossberg, S.; Mingolla, E.; Todorovic, D. "A neural network architecture for pre-attentive vision", IEE Transactions on Biomedical Engineering, 1989, vol: 36 no. 1 p. 65-84.

[Haralick and Shapiro, 1985], Haralick, R.M.; Shapiro, L.G. "Image segmentation techniques", CVGIP 29, Academic Press, New York, 1985, p. 100-132.

[Hines and Hutchinson, 1989], Hines, E.L.; Hutchinson, R.A. "Application of multi-layer perceptrons to facial feature location", Third International Conference on Image Processing and its Applications, IEE London, UK, 1989, no. 307 p. 39-43.

[Hopfield, 1982], Hopfield, J.J. "Neural networks and physical systems with emergent collective computational abilities", Proceedings of the National Academy of Scientists, 1982, vol: 79 p. 2554-2558. Reprinted in Anderson and Rosenfeld 1988, p. 460-64.

[Hutchinson and Welsh, 1989], Hutchinson, R.A.; Welsh, W.J. "Comparison of neural networks and conventional techniques for feature location in facial images", IEE London, UK, 1989, no. 313 p. 205-5.

[Jervis, 1994], Jervis, T.T. "Optimisation methods for neural networks", Neural Networks Summer School, University of Cambridge, 1994, p. 1-39.

[Kohonen, 1988], Kohonen, T. "Self-organisation and associative memory", Springer-Verlag, NY, USA, 1988, 2nd Ed.

[Kohonen, 1989], Kohonen, T. "Self-organization and associative memory", Springer Verlag, 3rd Edition, 1989.

[Leow and Miikkulainen, 1991], Leow, W.K.; Miikkulainen, R. "A neural network for attentional spotlight", IEEE International Joint Conference on Neural Networks, IEEE New York, USA, 1991, p. 436-41 vol. 1.

[MacQueen, 1967], MacQueen, J. “Some methods for classification and analysis of multivariate observations”, Proceeding of 5th Berkeley Symposium, p. 281-97, 1967.

[Marr, 1982], Marr, D. “Vision: a computational investigation into the human representation and processing of visual information”, W.H. Freeman, San Francisco, 1982.

[Marsic, 1992], Marsic, I.; Micheli-Tzanakou, E. "A framework for object representation and recognition", International Joint Conference on Neural Networks, IEEE NY, USA, 1992, vol: 3 p. 272-7.

[Masters, 1993], Masters, T. “Practical Neural Network Recipes in C++”, Academic Press, 1993.

[McHugh, 1990], McHugh, J.A. “Algorithmic graph theory”, Prentice-Hall, Engelwoood Cliffs, NJ, USA, 1990.

[Philip, 1991], Philip, K. P. “Automatic detection of myocardial contours in cine computed tomographic images”, PhD Thesis, University of Iowa, 1991.

[Roth and Frisby, 1986], Roth, I.; Frisby, J. P. “Perception and Representation : A Cognitive Approach”, Open University Press, Milton Keynes, 1986.

[Rowley *et. al*, 1995], Rowley, H.A.; Baluja, S.; Kanade, T. “Human face detection in visual scenes”, CMU-CS-95-158, 1995.

[Sajda and Finkel, 1992], Sajda, P.; Finkel, L.H. "Simulating biological vision with hybrid neural networks", Simulation, USA, 1992, vol: 59 p. 47-55.

[Samal and Iyengar, 1992], Samal, A.; Iyengar, P.A. "Automatic recognition and analysis of human faces and facial expressions: a survey", Pattern Recognition, Automatic Recognition and Analysis, 1992, vol: 25 part: 1 p. 65-75.

[Schaffer *et. al*, 1992], Schaffer, J.D.; Whitley, D.; Eshelman, L.J. "Combinations of genetic algorithms and neural networks: a survey of the state of the art", IEEE, 1992, p. 1-37.

[Shimada, 1992], Shimada, S. "Extraction of scenes containing a specific person from image sequences of a real world scene", IEEE Region 10 Conference, 1992, p. 568-72.

[Sonka *et. al*, 1993], Sonka, M.; Hlavac, V.; Boyle, R. "Image processing, analysis and machine vision", Chapman and Hall, Cambridge University Press, UK, 1993.

[Sung and Poggio, 1994], Sung, K-K; Poggio, T. "Example-based learning for view-based human face detection", A.I. Memo 1521, CBCL Paper 112, MIT, 1994.

[Swingler, 1996], Swingler, K. "Applying nearal networks - a practical guide", Academic Press, UK, 1996.

[Treisman, 1982], Treisman, A. "Perceptual grouping and attention in visual search for features and for objects", Journal of Experimental Psychology: Human Perception and Performance, American Psychological Association, 1982, vol: 8 no. 2 p. 194-214.

[Turk and Pentland, 1990], Turk, M.; Pentland, A. "Face processing: models for recognition", Proceedings of SPIE, Intelligent Robots and Computer Vision VIII: Algorithms and Techniques, 1990, vol: 1192 p. 22-32.

[Vaillant *et. al*, 1993], Vaillant, R.; Monrocq, C.; Le Chun, Y. "An original approach for the localisation of objects in images", Third International Conference on Neural Networks, IEE London, UK, 1993, vol: 372 p. 26-30.

[Vaillant *et. al*, 1994], Vaillant, R.; Monrocq, C.; Le Cun, Y. "Original approach for the localisation of objects in images", IEE Proceedings-Vision, Image and Signal Processing, IEE London, UK, 1994, vol: 141 Iss: 4 p. 245-50.

[Viennet and Fogelman Soulie, 1992], Viennet, E.; Fogelman Soulie, F. "Multi resolution scene segmentation using MLPs", IJCNN, IEEE New York, USA, 1992, vol: 4 p. 55-9.

[Vincent *et. al*, 1992], Vincent, J.M.; Waite, J.B.; Myers, D.J. "Automatic location of visual features by a system of multilayered perceptrons", IEE Proceedings F, 1992, Vol: 139 Iss: 6 p. 405-12.

[Waite, 1991], Waite, J. "Facial feature location using multilayer perceptrons and micro-features", International Joint Conference on Neural Networks, 1991, vol: 1 p. 292-299.

[Watt, 1988], Watt, R.J. "Visual processing: computational, psychophysical, and cognitive research", Lawrence Erlbaum Associates, UK, 1988.

[Widrow and Hoff, 1960], Widrow, B.; Hoff, M. "Adaptive switching circuits", August IRE Wescon Convention Record, 1960, Part 4. p. 96-104.

[Widrow *et. al*, 1976], Widrow, B.; McCool, J.M.; Larimore, M.G.; Johnson, C.R. "Stationary and non-stationary learning characteristics of the LMS adaptive filter", Proceedings of IEEE, 1976, Vol 64. Iss. 8.

[Widrow and Stearns, 1985], Widrow, B.; Stearns, S.D. "Adaptive signal processing", Prentice-Hall, 1985.

Appendix A – Overview of ANNs

This appendix is intended to provide a brief introduction to Artificial Neural Networks (ANNs). A brief outline on the workings of an ANN is given and discusses some of the main differences between ANN models.

ANN Overview

Artificial Neural Networks, or connectionist architectures, describe systems that use a model that is composed of a collection of simple processing elements (generally referred to as units), Figure 29, joined together usually by adaptive connections referred to as weights. Each processing element performs some function, commonly on the weighted sum of inputs leading to the unit. This kind of architecture is typical of most ANN paradigms. The Multi Layered Perceptron (MLP), Figure 30, is one of the most common ANN paradigms and is a popular example that comprises this structure.

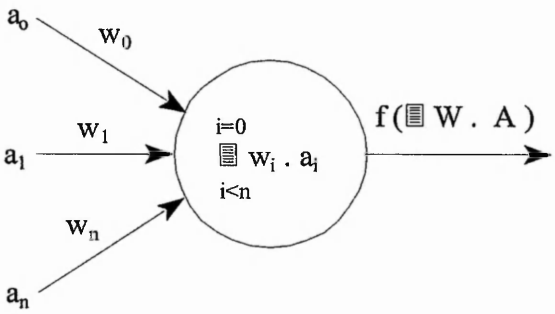


Figure 29 - Simple Processing Unit

Of all the different types of ANN models that exist, all can be described by two main characteristics:

- Learning Algorithm. This describes how the connections within the model are adapted. The modification of these values is termed *learning* and this can be either unsupervised or supervised.
- Connectivity. This describes the topological mapping of units and connections in the model, whether there is a partial connection between layers of units, lateral connections, or a fully interconnected model.

Learning Algorithm

The learning algorithm determines how, give an input vector, weights are to be adapted. For the problem of face identification the input vector will be some representation of the input image. There are two different types of learning, unsupervised and supervised.

Unsupervised Learning

Unsupervised learning describes the means with which the network adapts its connections, from the input patterns presented in an autonomous fashion. Typical examples of network models in this category include the Kohonen [Kohonen, 1988] and ART networks [Carpenter, 1985]. In comparison to supervised learning algorithms they do not require an input to control their output.

Being unsupervised leads to the network making its own generalisations on the importance of particular aspects of the input data. To derive any meaning from the output requires interpretation of the output values to determine what has been learnt, i.e. how the patterns have been grouped or transformed. Similar input patterns are likely to produce clusters in the output vector space.

Supervised Learning

In contrast to unsupervised learning, the type of weight modification that occurs is dependent upon the teacher signal used whilst training. For every training input example that is presented to the network, there is an appropriate desired output. The network generates its own output, and the difference between the two is often used to change the weights of connections. This learning method is repeated until all, or a required number of the input examples, have generated the correct answer to a pre-set precision.

The advantage of supervising the network is that the output representation is in a form that requires no interpretation, and the generalisations made on the network have been made to the requirements of the supervisory input. The main disadvantage of this learning method is that it is generally slower than unsupervised learning, and that it forces a solution to a complete problem, rather than allowing the system to find regularities in the data that may be useful later. One way to overcome this latter disadvantage is to decompose the problem in stages and use one network per stage. For example, in OCR, one network can split the alphabet into six subsets, and other networks (one for each subset) can discriminate between four or five characters.

Connectivity

The connectivity of the ANN is generally chosen to suit the problem or type of learning algorithm used. It may be that full connectivity is unnecessary as the ANN may contain redundant weight connections. Alternatively, recurrent connections are usually applied to problems requiring storage of temporal information. There are three main general types of network connectivity found common in ANN applications. These are:

- Fully connectivity. This generally describes feed forward only ANNs. Each unit in a fully connected network is connected to all other units in the layer below. The representations made in the ANN are distributed across connections.
- Partial connectivity. This can be used to form localist representations of the data. There is no rule as to when this should be applied, and is only chosen if partial connectivity has a higher probability of generalising to the data.

- Recurrent network architectures. These not only provide full feed forward connectivity, but also outputs leading from a unit back into the unit as input. An example of a recurrent ANN is the Hopfield network [Hopfield, 1982]. Alternatively, [Gouhara *et. al*, 1991] discusses recurrent network learning for MLPs. These kind of ANNs are suited towards problems that consist of temporal structure.

MLPs and Back-Error Propagation

Historically it was the advent of the Adeline and Madeline [Widrow and Hoff, 1960] that sparked the initial interest in ANNs. These early networks showed it was possible to use an adaptive system, consisting of only simple processing elements, that was able to learn solve a surprising range of problems [Widrow *et. al*, 1976], [Widrow and Sterns, 1985].

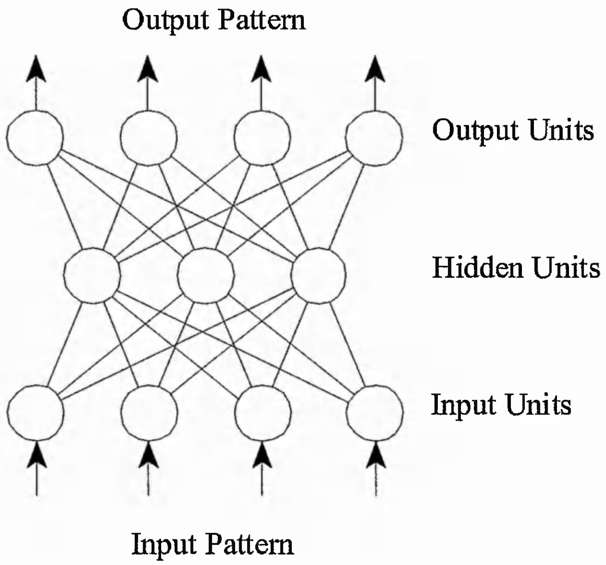


Figure 30 - Structure of a Multi Layer Perceptron

The Adeline is a two-layer network with a single set of adaptive weights. The Madeline is an extension that allows for multiple hidden layers. Only a single layer of weights can be modified. Unfortunately, these networks were limited in that they only dealt with binary input. More importantly, which caused them to be ignored for many years, it was proven that the Widrow-Hoff learning algorithm (Equation 24), which corrects the errors generated at the output, was unable to solve linearly inseparable problems. e.g. *xor*.

$$w_{ji} = w_{ji} + \eta(t_j - o_j)a_i$$

Equation 24 - Widrow-Hoff Weight Udate Rule

It was only with the development of the back-error propagation[10] learning algorithm that this limitation was eventually overcome. The algorithm allows for *multiple adaptive* layers to be used, and a means by which the weights leading to each unit are modified according to the error it generates. The learning can be described as a two stage process that consists of a forward pass and then a backwards pass.

Learning

The forward pass consists of the propagation of the unit activations of all units from the first hidden layer to the output layer. This calculation consists of the weighted sum. This is then passed through a function to compute the unit's activation as shown in Figure 29. This activation function is generally the Sigmoid function, and could be described as acting like a soft threshold.

The backward pass consists of two actions, to calculate the full error of the network, and to update all weight corrections using this error. To calculate the error for both the output and hidden layer units two equations are needed. Firstly, the error at the output, which is determined from the target values (Equation 25). Secondly, to calculate the error for all hidden units (Equation 26). The error at an output unit is determined by all the hidden units leading to it. Therefore, each hidden unit has a contributory factor to the final error. All of the weights are then updated using the same weight update rule (Equation 27).

[10] This learning algorithm is simply referred to as *backprop*, and this name will be used throughout the rest of this report.

$$\delta_j = (t_j - o_j) f'(S_j)$$

Equation 25 - Output-Hidden Unit Error

$$\delta_j = [\sum_{k=0}^{k<n} \delta_k w_{kj}] f'(S_j)$$

Equation 26 - Hidden-Input Unit Error

$$\Delta w_{ji} = \eta \; \delta_j \; a_i$$

Equation 27 – Unit Weight Change

The value of the learning rate determines the amount the weights are adapted. Generally, this lies in the range between zero and one. A low value signifies a small weight change, and a high value greater modification. Too large a value for the learning rate may cause instabilities during the progress of the learning. Ultimately, the choice of learning rate often determines how quickly the network converges to the solution but also the reliability of the learning.

Key
Δw is the weight change
η is the learning rate
δ_j is the error for unit j
a is the input activation
f= is the first derivative of the Sigmoid function
S_j is the weighted sum for unit j
t_j is the target output of the jth unit
o_j is the actual output of the jth unit
w_{ji} is the ith weight leading to the jth unit
a_i is the activation of the ith input

A way to determine the progress of the learning is to compute the Root Mean Squared (RMS) error using all patterns in an epoch. The RMS value is always in the range of between zero and one. The closer to zero, the more accurate the general performance of the network. Judging the performance of an ANN on the RMS alone can be misleading, e.g. A large error for one pattern will produce the same RMS as several small errors on a number of patterns.

Optimisations

Much research has gone into modification of the basic *backprop* algorithm in order to improve the convergence times of the network. These range from simple adjustments to complex algorithm extensions. Although the optimisations require more memory and computation per epoch, the aim is to significantly reduce the total number of epochs required.

The most basic addition to the *backprop* learning algorithm is to add a momentum term to the weight adjustment. This helps increase the weight change if the modification is proceeding in the same direction as the last update. An offline modification to *backprop* is to accumulate the error generated by all the patterns in the training set and adjust the weights after each epoch only.

The methods discussed so far are first order, and gradient descent is performed by making a linear approximation. Second order methods make no assumption that the problem space has a simple form and determine the weight adjustment by the use of a local quadratic model [Jervis, 1994]. Conjugate gradient descent and scaled conjugate gradient descent are second order methods. From the quadratic model the methods attempt to select the best direction without undoing the minimisation of previous iterations. The disadvantage of these algorithms is the time taken to perform the search for the optimum direction and how appropriate the quadratic model is.

Local methods for learning optimisation do not use global gradient descent to determine the weight adjustment. These methods consider local changes for each weight only. *Delta-bar-delta*, *Rprop*, and *Quickprop* are local optimisation methods.

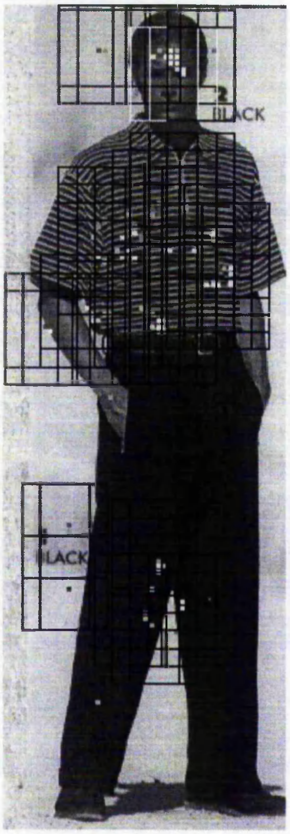
In general, either second order or local optimisation methods will find a solution in fewer epochs than *backprop* using momentum only. Unfortunately, all of the optimisation methods mentioned have a limitation of some kind. Different methods will perform better on different data, and there is no set criteria for selecting the most appropriate method for the problem. A more complete analysis of the different methods discussed is included in [Jervis, 1994].

Appendix B – Results for Window Grouping Method

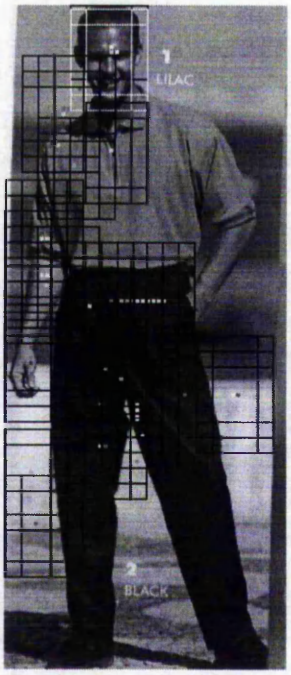
Test results from single resolution search using window grouping method.

- Window size: 6x6 reduced resolution pixels.
- Window step: 3 reduced resolution pixels.

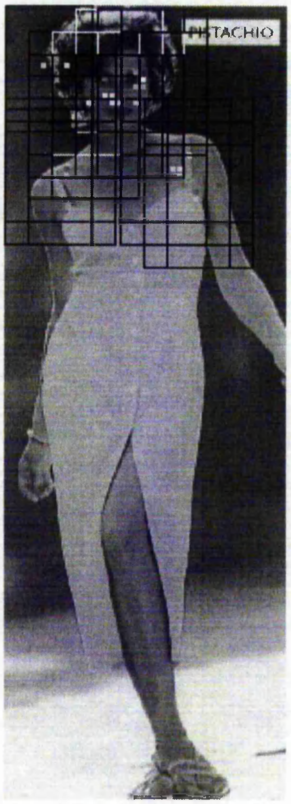
Test Image 1



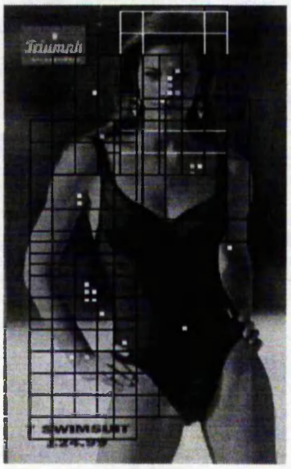
Test Image 2



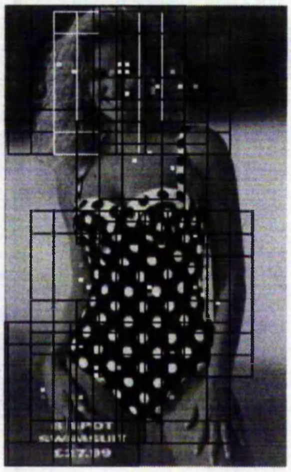
Test Image 3



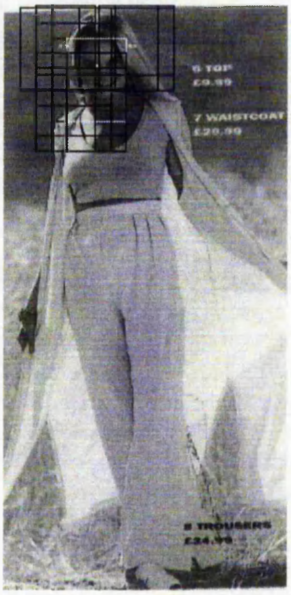
Test Image 4



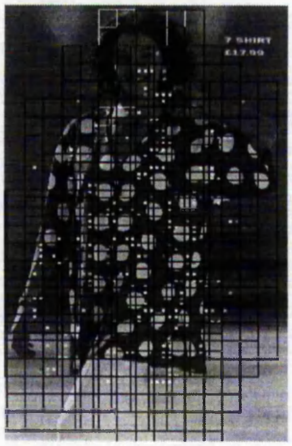
Test Image 5



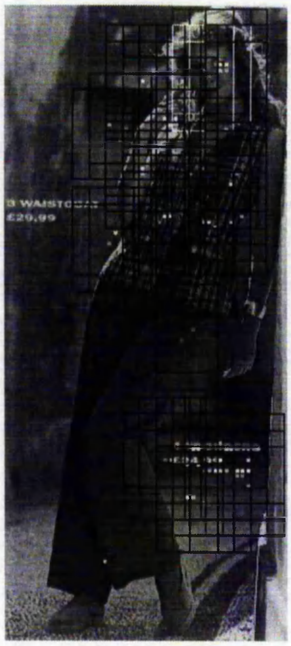
Test Image 6



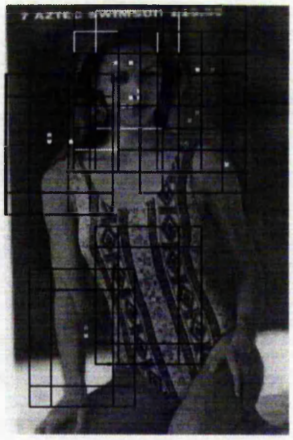
Test Image 7



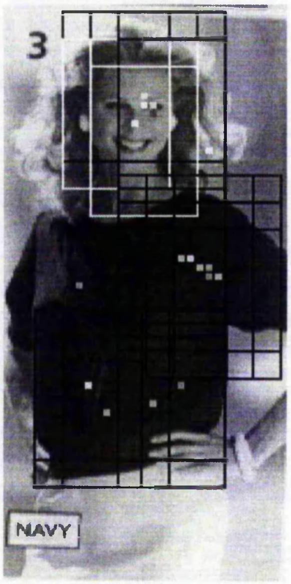
Test Image 8



Test Image 9



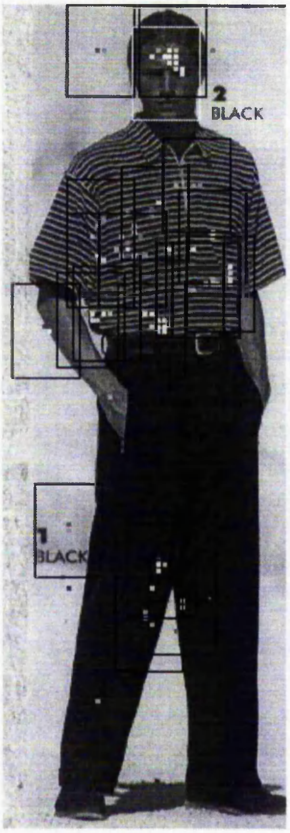
Test Image 10



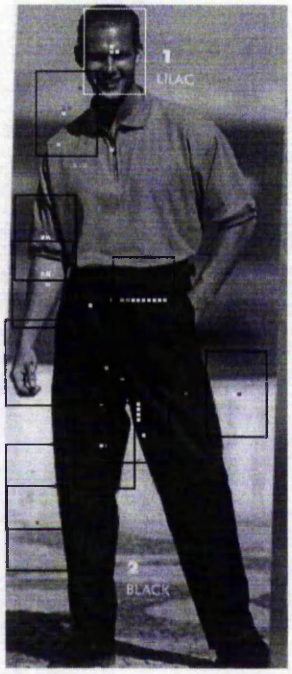
Appendix C - Results for Extremities Grouping Method

Test results from a single resolution search using extremities grouping method.

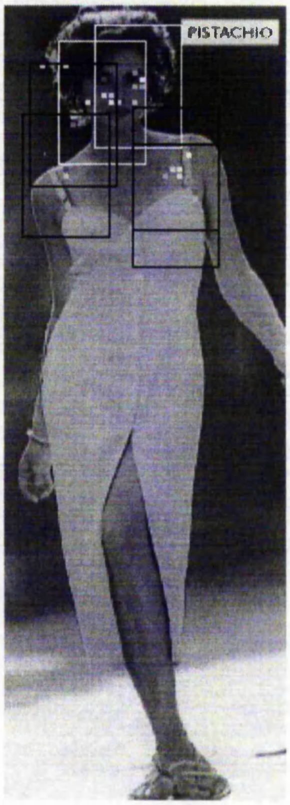
Test Image 11



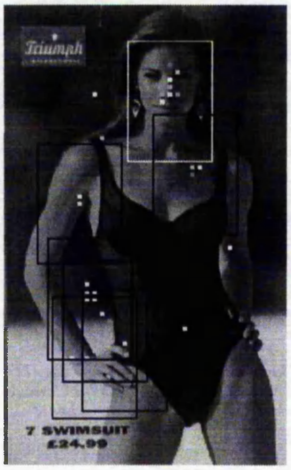
Test Image 12



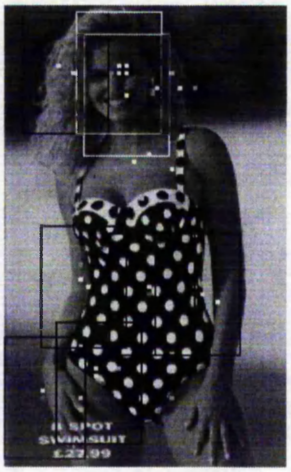
Test Image 13



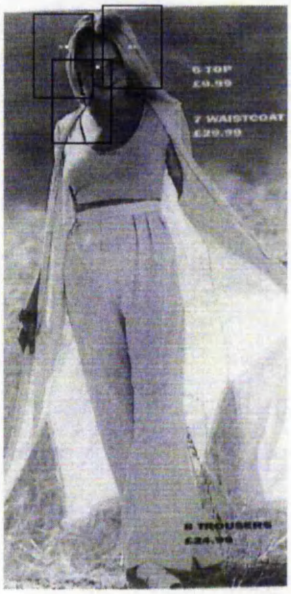
Test Image 14



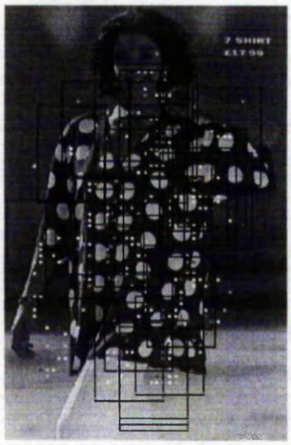
Test Image 15



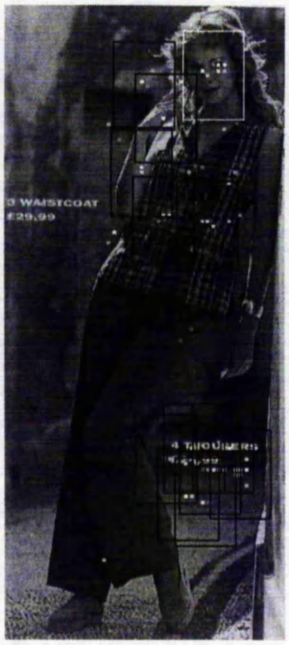
Test Image 16



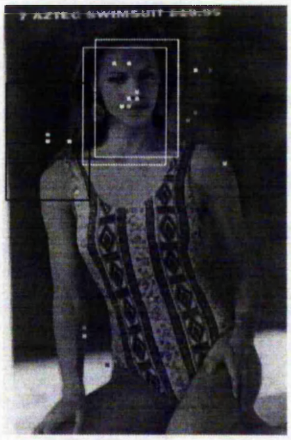
Test Image 17



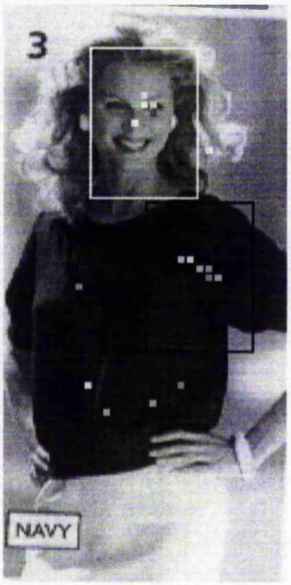
Test Image 18



Test Image 19



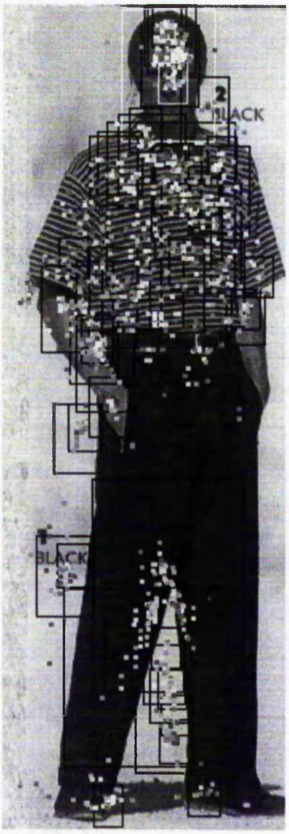
Test Image 20



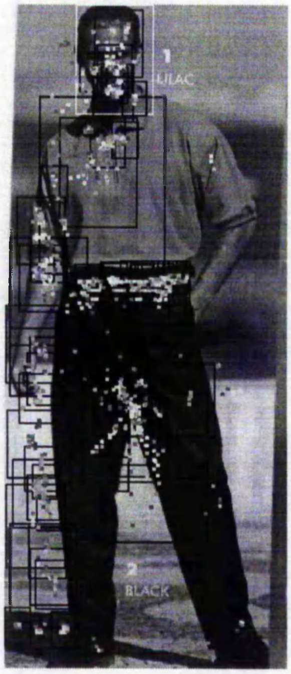
Appendix D - Results for Multi-Resolution Analysis

Test results from a multi-resolution search using a reduction factor of 0.85.

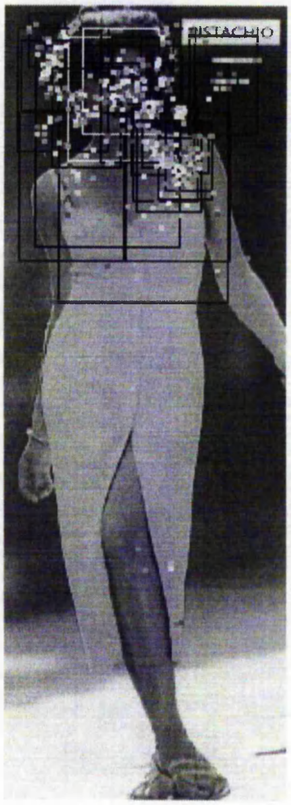
Test Image 21



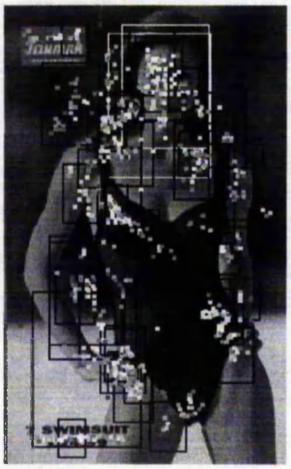
Test Image 22



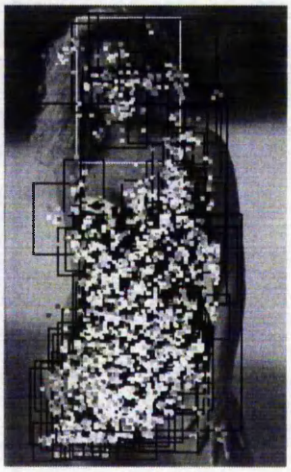
Test Image 23



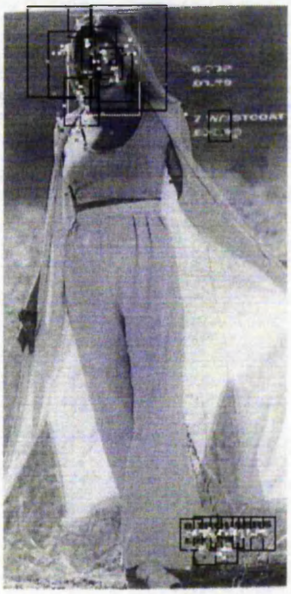
Test Image 24



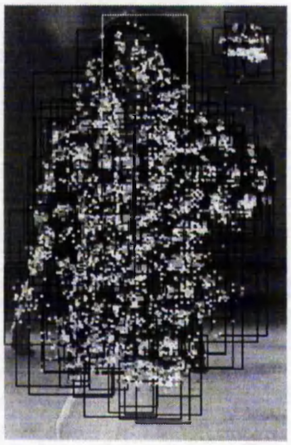
Test Image 25



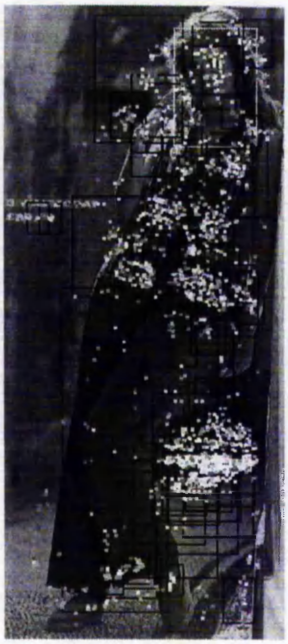
Test Image 26



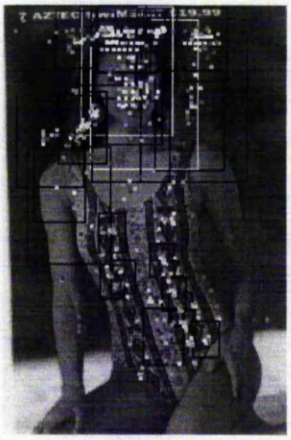
Test Image 27



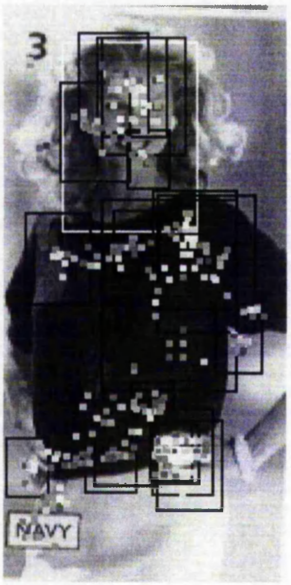
Test Image 28



Test Image 29



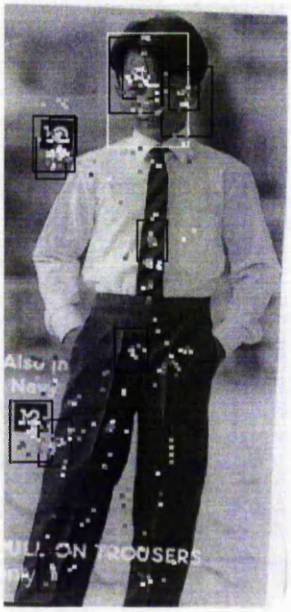
Test Image 30



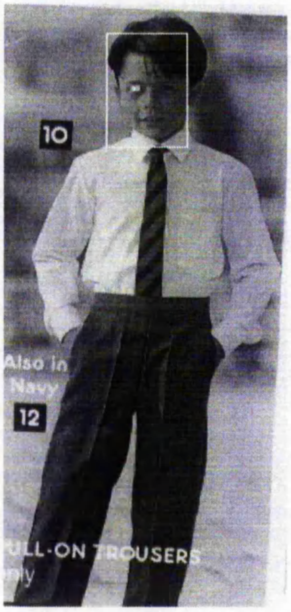
Appendix E - Results for Identification Analysis

Final identification results using left eye classification and texture segmentation and filtering. Left side output from attention-focusing; right side classification.

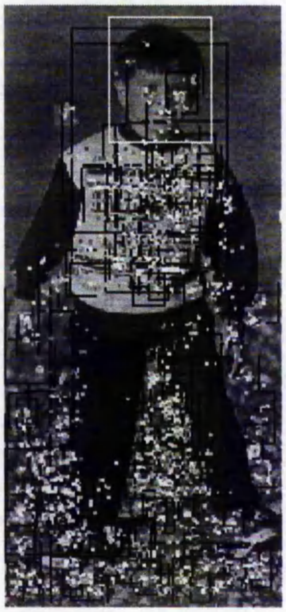
Test Image 31



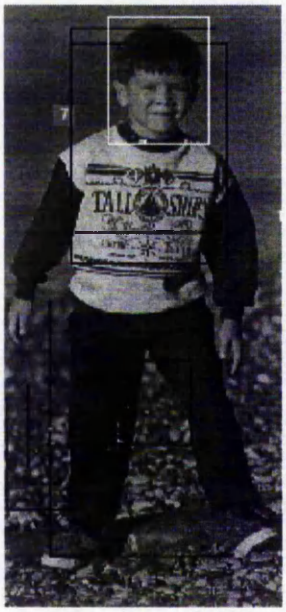
Test Image 32



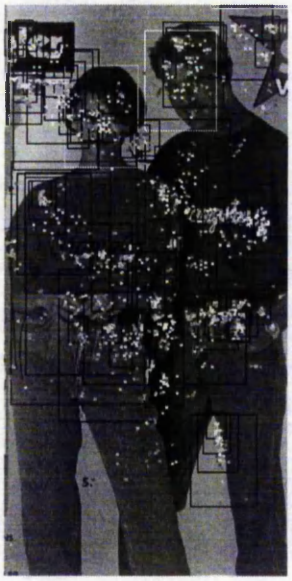
Test Image 33



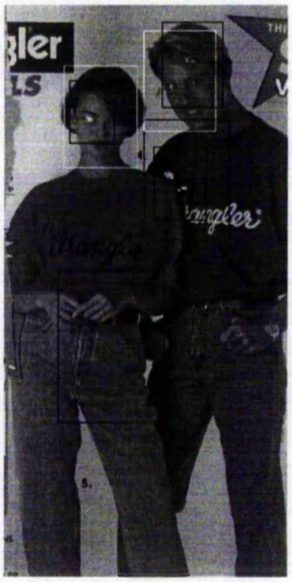
Test Image 34



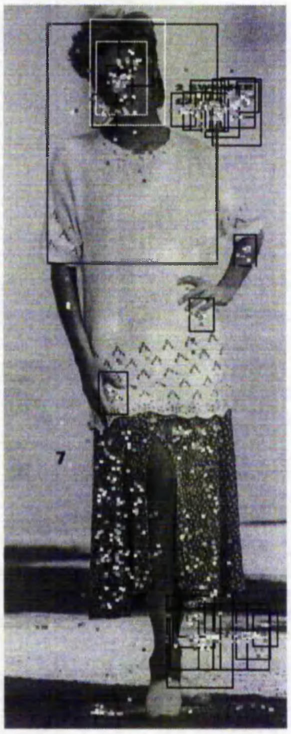
Test Image 35



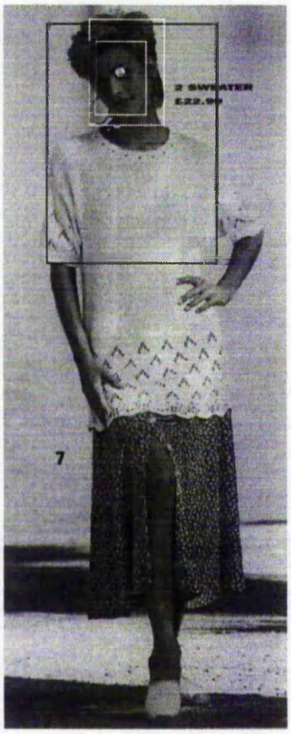
Test Image 36



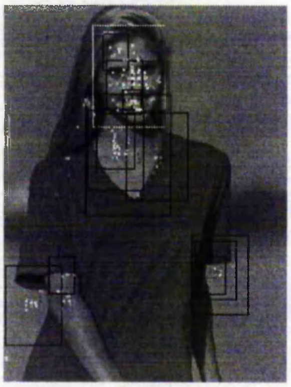
Test Image 37



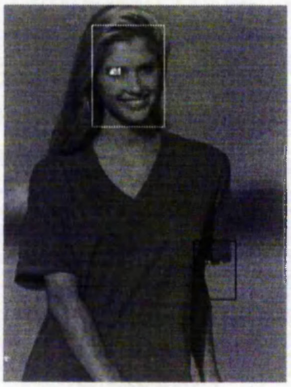
Test Image 38



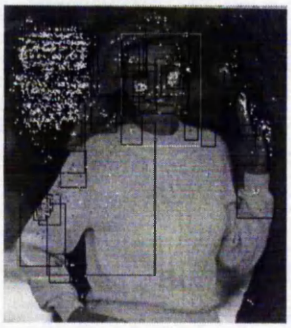
Test Image 39



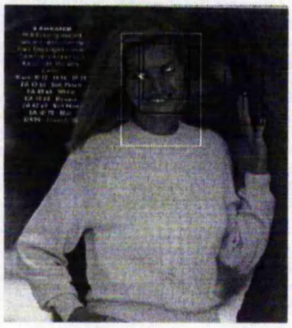
Test Image 40



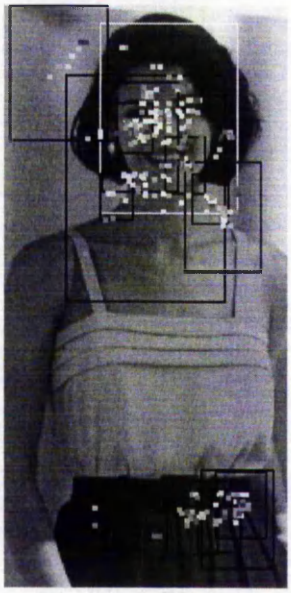
Test Image 41



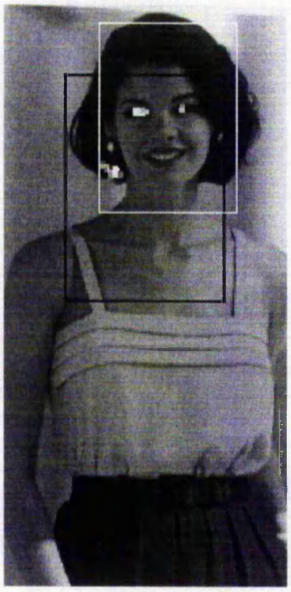
Test Image 42



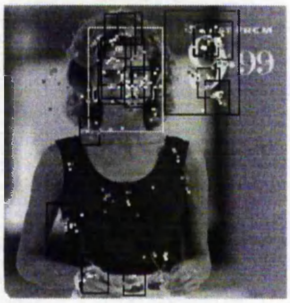
Test Image 43



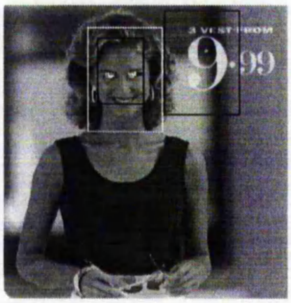
Test Image 44



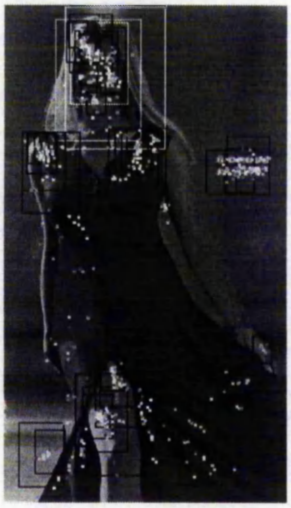
Test Image 45



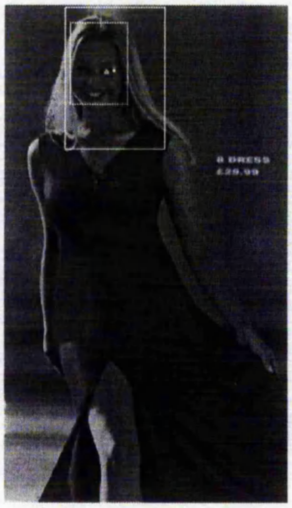
Test Image 46



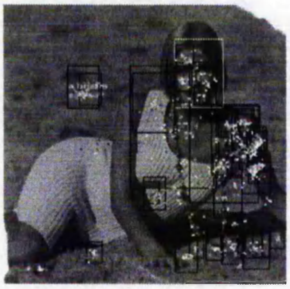
Test Image 47



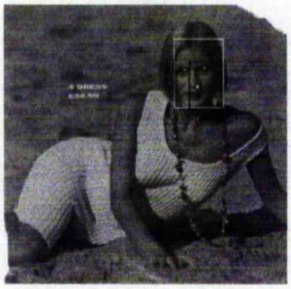
Test Image 48



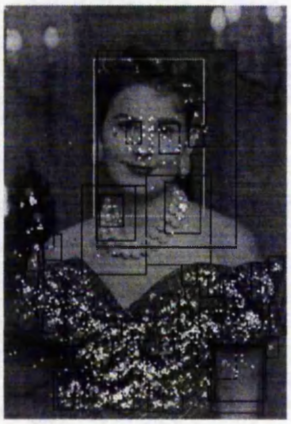
Test Image 49



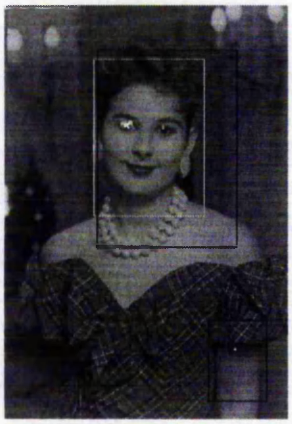
Test Image 50



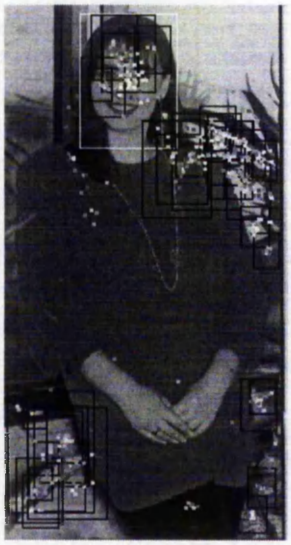
Test Image 51



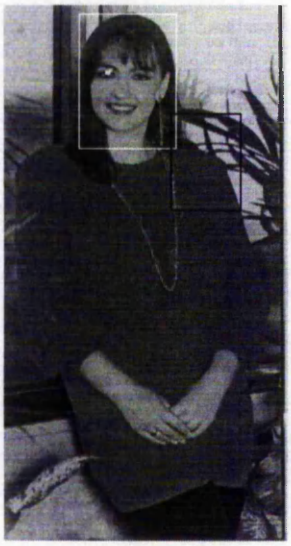
Test Image 52



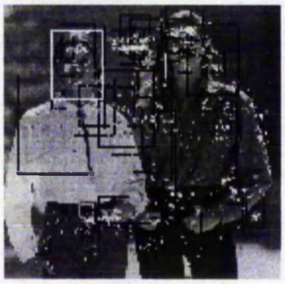
Test Image 53



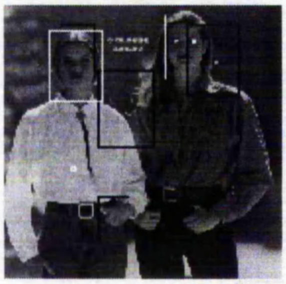
Test Image 54



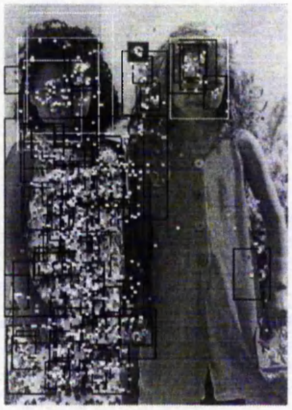
Test Image 55



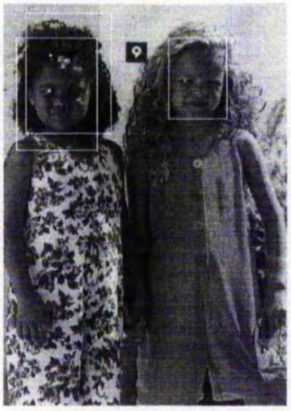
Test Image 56



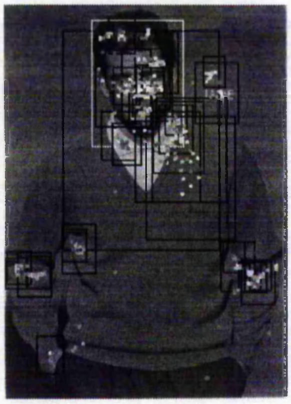
Test Image 57



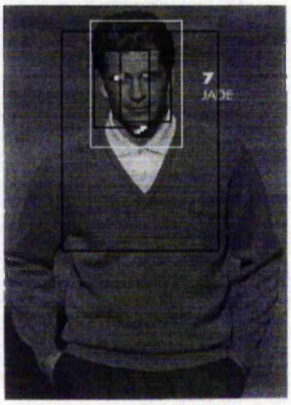
Test Image 58



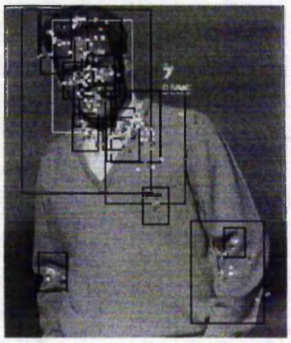
Test Image 59



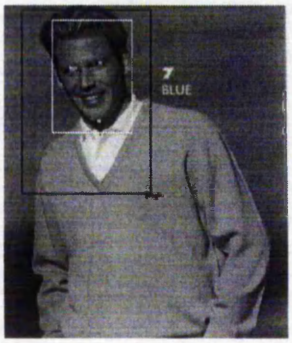
Test Image 60



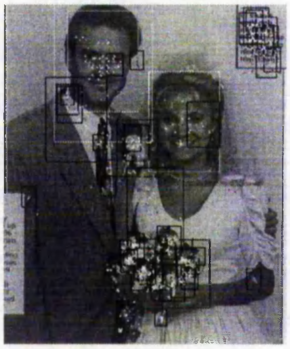
Test Image 61



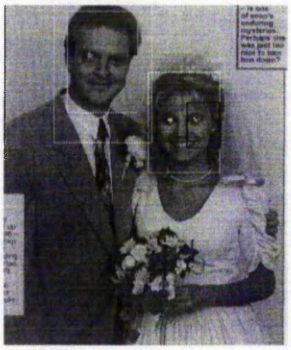
Test Image 62



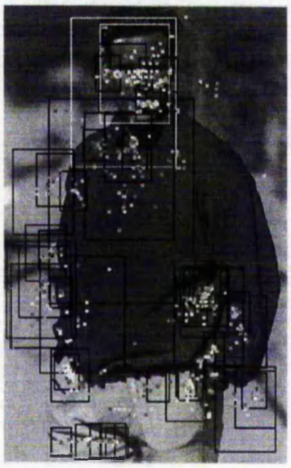
Test Image 63



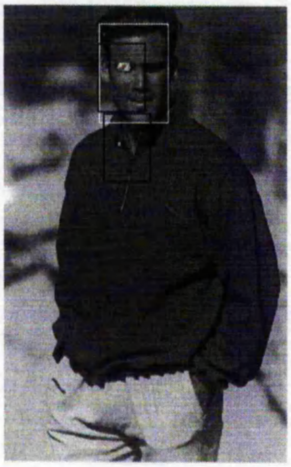
Test Image 64



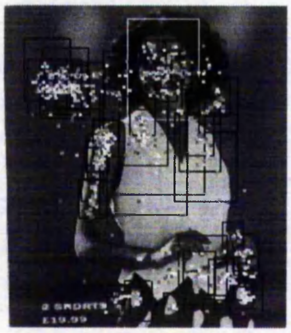
Test Image 65



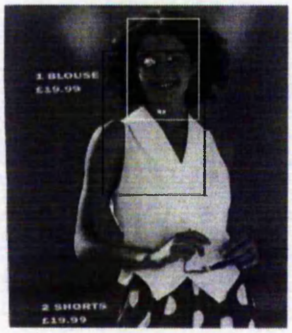
Test Image 66



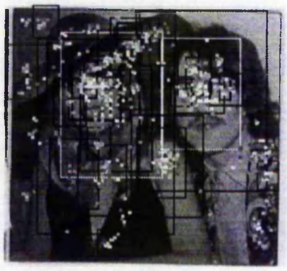
Test Image 67



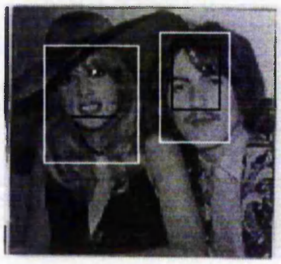
Test Image 68



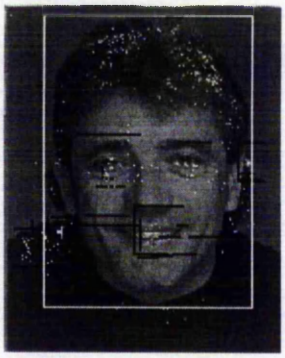
Test Image 69



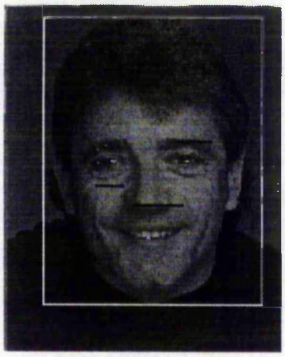
Test Image 70



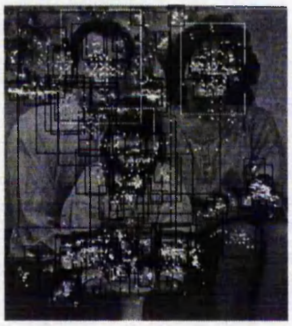
Test Image 71



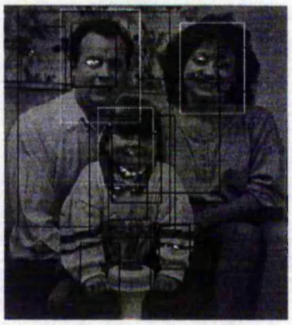
Test Image 72



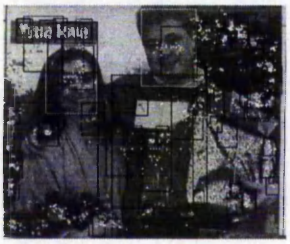
Test Image 73



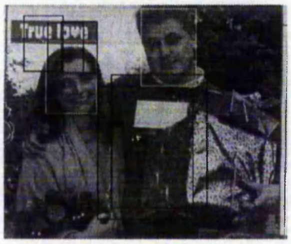
Test Image 74



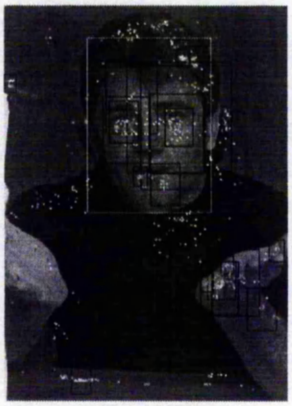
Test Image 75



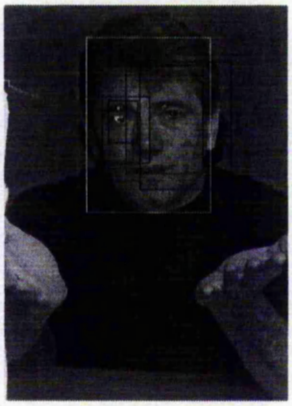
Test Image 76



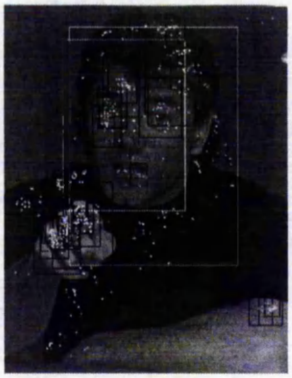
Test Image 77



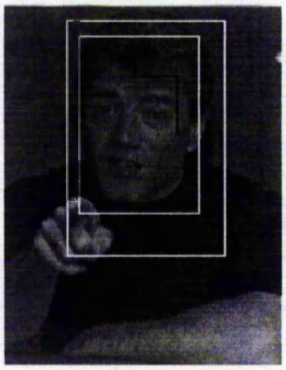
Test Image 78



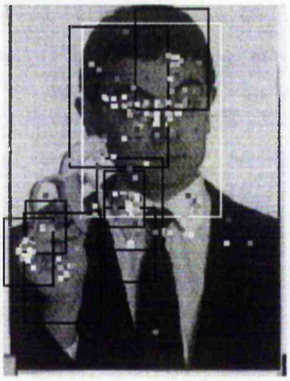
Test Image 79



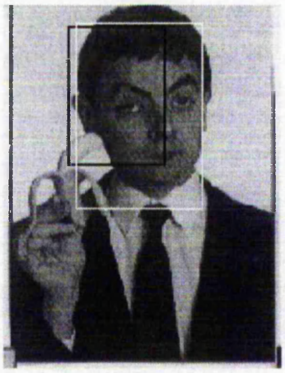
Test Image 80



Test Image 81



Test Image 82



Appendix F – Miscellaneous

Derivation of Equation 6 - Determining the Maximum Linear Steps

To establish the maximum reduction necessary to reduce the image size down to the input frame size (because this is the minimum size at which an object can be identified) the following is applied:

1. The reduction ratio determines the scaling factor to reduce the image to the size of the input frame.

   $$reduction\ ratio = \frac{window\,size}{picture\,size}$$

2. The maximum number of reductions can also be expressed as requiring *n* reductions.

   $$maximum\ possible\, n \equiv reduction^{n} = reduction\, ratio$$

3. To deal with the *n* power term requires both expressions being converted to logs.

   $$\log_{10}(reduction^{n}) = \log_{10}(reduction\, ratio)$$

4. Using the mathematical rule for logs the power term can be re-arranged.

   $$n\log_{10}(reduction) = \log_{10}(reduction\ ratio)$$

5. This finally allows the number of reductions to be determined.

   $$n = \frac{\log_{10}(reduction\, ratio)}{\log_{10}(reduction)}$$

6. Replacing the original notation provides the final equation given in Equation 6.

$$\textit{maximum possible } n = \frac{\log_{10}\left(\dfrac{\textit{window size}}{\textit{picture size}}\right)}{\log_{10}(\textit{reduction})}$$

window size refers to the maximum dimension of the input frame in either direction. *picture size* refers to the maximum dimension of the input image in either direction.