A TWO-STAGE NEURAL NETWORK FOR TRANSLATION, ROTATION AND SIZE-INVARIANT VISUAL PATTERN RECOGNITION

A.Ravichandran and B.Yegnanarayana,
Department of Computer Science and Engineering,
Indian Institute of Technology,
MADRAS- 600 036, INDIA.

ABSTRACT In many practical situations, such as in sensor array imaging, there is need to recognise objects from noisy images in spite of spatial transformations like scaling, translation and rotation. We describe a two stage neural network for transformation-invariant visual pattern recognition. In the first stage, features are extracted after normalizing the image. We show how parameters of spatial transformation can be estimated even in the presence of noise by using knowledge about rigid objects. Circular arcs in the normalized image are used as generalised features to describe the input pattern. Each image pixel contributes to the features which it can constitute. Contributions from noisy pixels are distributed over the feature space, whereas meaningful parts contribute to clusters that correspond to features of the image. In the second stage, the image is classified on the basis of these features by a multilayer perceptron network trained using backpropagation algorithm.

I. INTRODUCTION

In many practical situations there is need for recognising an object from its image in spite of spatial transformations like scaling, translation and rotation. Besides, the image may be partial and noisy. For instance, images reconstructed from sparse and noisy data collected by an array of sensors are of poor quality and are noisy. Moreover, when the object moves, the images obtained are transformed versions of the image which we have in our knowledge base. When the number of expected targets is very large, it becomes difficult for a human observer to identify the object correctly from such an image. Hence there is a need for a knowledge-based system for identification of objects from noisy images in spite of spatial transformations and distortions.

It is reasonable to expect that in noisy environments, the performance of such a system will be sometimes superior to that of a human being. This is because our finite attention span makes it difficult to bring in the total knowledge and evaluate all possibilities. These restrictions do not exist for a machine. Thus given a noisy partial image as the input, the goal of the knowledge-based system is to correctly identify the pattern as one of those in its knowledge base that matches most. In classical approaches, pattern recognition is done by description matching. Here, the pattern is represented in terms of its characteristic features and subfeatures. Given the noisy input pattern, its features are extracted and the pattern is described in terms of these features. Classification is done by matching this description with those of the patterns in the knowledge base. The input pattern is classified as the pattern whose description matches most.

Description matching approach is superior to the direct template matching approach because it can handle partial patterns and also variations and distortions. However, it has two disadvantages: (i) if the input pattern is noisy, then it may be difficult to extract the features reliably due to the presence of spurious pixels and absence of relevant points. (ii) Choice of suitable knowledge representation (features and descriptions) has to be worked out.

Neural networks, which consist of highly interconnected multiple processing nodes, offer hope in solving these problems. A neural network using distributed representation and distributed processing has inherent fault tolerance, and hence can handle noise and partial input. It can adaptively learn to extract features in the presence of noise. In addition, it can evolve its own internal representation by learning from examples. Learning usually consists of modifying its structural parameters till the goal is achieved. The backpropagation algorithm is a very powerful technique for training such networks[1].

Let us consider the problem of transformation-invariant pattern recognition. Theoretically, we can build a fully connected feedforward multilayer perceptron network and train it by error backpropagation in such a way that it can perform successful object recognition. However, the size of the network required for this job may be very large. More importantly, such an unconstrained superfluous structure may result in the network blindly memorizing the input-output relationships without attempting to generalise. Such a network cannot extend its knowledge to handle new examples of the same input pattern. Generalisation also depends on the type of training examples. The question of how to constrain the structure to force the network to generalise and how to select and sequence the training samples so that the generalisation will be a valid one, are questions yet to be answered.

When we consider the tasks of feature extraction and classification, feature extraction is a data-oriented task and has the function of extracting relevant and meaningful information from the input image data, whereas classification tries to intrepret this information according to knowledge of the pattern set. For feature extraction to be useful over different sets of patterns, it must not employ task-specific knowledge. However, since classification can be performed only on the basis of knowledge, neurons with same structural and functional organisation cannot be used for feature extraction and classification tasks. In this paper, we describe a two-stage neural network for transformation-invariant pattern recognition. The first stage performs the feature

CH2977-7/91/0000-2393 $1.00 © 1991 IEEE

extraction after normalization of the image, and the second stage performs the classification.

When we consider images of an object that has undergone spatial transformations, normalizing the image becomes very important issue during feature extraction. This becomes more so in the presence of noise. We also show in this study that if normalization is done correctly, then the tasks of feature extraction and classification become simpler.

In section II, we discuss normalisation of a noisy input image with respect to spatial transformations. Section III describes the feature extraction process. Classification is discussed in Section IV. Results of experimental studies are discussed in section V.

II. NORMALIZATION WITH RESPECT TO SPATIAL TRANSFORMATIONS

Before classification can be done, it is necessary to retransform the image to a fixed reference point (translation), correct its size (scaling) and its orientation (rotation). As the parameters of these spatial transformations will not be available for us to use for normalisation, methods have to be found out to estimate these parameters directly from the image data itself.

When an object undergoes spatial transformation, the spatial positions vary. However, the spatial interrelationships between the points on the object are maintained. For example, after translation or rotation, the positions of the individual pixels change, but their relative distances remain unaltered. Even after scaling, the ratios of relative distances remain same. Hence, normalization can be done by defining relationships that are invariant to spatial transformations[2]. These are functions that map the set of image pixels to an unique invariant point and such functions can be mathematically defined. Definition of the centre of gravity of an image is an example of an invariant function defined on the image pixels. Besides, we choose two centres of moments that are invariant to spatial transformations.

It can be proved that the following points are invariant points with respect to translation and rotation:

(i) The centre of gravity (CG) of an object is defined as

$$x = \Sigma\ x_i / N,\ y = \Sigma\ y_i / N$$

(ii) The first centre of moments (P_x, P_y) about the CG is defined as

$$P_x = \Sigma\ (F_i x_i) / \Sigma\ F_i$$

and

$$P_y = \Sigma\ (F_i y_i) / \Sigma\ F_i$$

where

$$F_i = \sqrt{}\,(\,(x_i - x\,)^2 + (y_i - y\,)^2\,).$$

(iii)The second centre of moments (Q_x, Q_y) about the CG is defined as above with

$$F_i = 1 / \sqrt{}\,((x_i - x\,)^2 + (y_i - y)^2\,).$$

Viewed in another way, the three invariant points together define a set of reference axes with respect to which the image pixels can be described in an invariant manner. Hence, normalization consists of computation of image pixel coordinates with respect to the new reference axes.

The normalization algorithm has the following steps:

Step 1 : Calculate the centre of gravity,
$x = (\Sigma\ x_i) / N,\ y = (\Sigma\ y_i) / N$

Step 2 : Calculate the weights as
$w_i = \sqrt{}\,((x_i - x)^2 + (y_i - y)^2)$

Step 3 : Calculate centres of moments as

$$X_{w1} = (\Sigma\ (w_i x_i)) / \Sigma\,(w_i))$$
$$Y_{w1} = (\Sigma\ (w_i y_i)) / \Sigma\ (w_i))$$
$$X_{w2} = (\Sigma\ (x_i / w_i)) / (\Sigma\ (1 / w_i))$$
$$Y_{w2} = (\Sigma\ (y_i / w_i)) / (\Sigma(1 / w_i))$$

In all these cases, points with $w_i = 0$ are excluded.

Step 4 : Compute the distance d,
$d = \sqrt{}\,(\,(X_{w1} - X_{w2})^2 + (Y_{w1} - Y_{w2})^2\,)$

Step5 : For all the points P_i (i = 1,2,..n), redefine the coordinates by the following transformation rules:

$$X_i = (X_i - X_{w1})\,(X_{w1} - X_{w2})/d + (Y_i - Y_{w1})\,(Y_{w1} - Y_{w2})/d$$

$$Y_i = (X_{w1} - X_i)\,(Y_{w1} - Y_{w2})/d + (Y_i - Y_{w1})\,(X_{w1} - X_{w2})/d$$

Step 6 : To obtain scale invariance we further divide each coordinate by d.

A noisy image may contain pixels that do not belong to the object and it may leave out pixels that actually belong to the object. Estimation of the transformation parameters will fail if all the wrong pixels are allowed to contribute to the decision. Thus to ensure the correct estimation of the parameters, only the valid pixels must be allowed to contribute to the measurements. To decide whether a pixel represents a valid point on the object, we apply our knowledge about rigid objects. We exploit the fact that rigid objects are composed of points that are held close together without gaps in between. In other words, more neighbours a pixel has, more it is a valid part of a rigid object. An isolated noisy pixel cannot be a valid part.

To allow for bigger chunks of the object getting lost, and to avoid bigger chunks of noise from being accepted, this idea is extended to pyramidal hierarchies, i.e, a pixel's validity is decided by looking at its 3x3 neigbourhood, and whether this block itself is a valid part is decided by looking at the 3x3 block neighbourhood.

The contribution of the pixel is weighted by the pixel distribution in its neighbourhood. Use of knowledge makes the measurement robust to noise. Fig.1 shows two images used in our study. Figs.2 and 3 show several images of these objects when they have undergone spatial transformations. Figs.4 and 5 show the corresponding images after normalisation by the above technique. This shows how estimation can be made robust by incorporating the knowledge of rigid objects.

III. FEATURE EXTRACTION

Feature extraction is the process in which the image is represented by a set of numerical features. This is done for two reasons:

(i) To reduce the dimensionality of the input pattern. If pixels have totally uncorrelated and random distributions in all the input images, then features cannot be seen. However, in the image of any object, the constituent pixels are related to their neighbourhoods and this makes a large number of pixel distributions meaningless. This order and the resulting redundancy can be exploited. Pixel distributions that occur often can be named as features. With a suitable set of such features, the image can be described in terms of these features leading to reduced dimensionality. Features are equivalent to verbal concepts in that they allow symbolic processing. Unless we have such abstract structures at intermediate levels, complex structures cannot be visualised or understood.

(ii) To allow for variations and distortions : By abstracting away the pixel distributions from their exact physical locations, features provide for invariant description of objects.

Various types of feature sets have been tried out in pattern recognition experiments. These fall into two broad categories:

In the first category, often occuring subpatterns in a given knowledge base are chosen as features. For example, in handwritten character recognition, arcs and line segments have been used as features for describing a handwritten character. The difficulty with this approach is that the feature set thus selected may turn out to be very restrictive and specific to the fixed knowledge base. For the same reason, it may not work when the objects undergo spatial transformations.

In the second category, exhaustive measurements are made, so that all possible distributions can be represented. Use of geometric moments, Fourier descriptors, etc., fall into this category. These features do not lead to as much reduction in dimensionality as those in the first category. Hence they are computationally intensive and also may not exploit redundancy. Stated in a different way, the first approach is a knowledge-based one, and hence it may turn out to be too restrictive, whereas the second approach is too general that it may not extract any knowledge.

In this study, we have tried to use a set of features that are general as well as knowledge-based. As discussed earlier, normalization is an important step during feature extraction and is based on the transformation-invariant nature of the centre of gravity of an object. This can be intuitively explained by the fact that spatial transformations of an object like rotation and scaling occur on a circle with the centre of gravity as its centre. Hence circular arcs centred at the CG are meaningful features in transformation-invariant object recognition. Another reason for this choice is that estimation of rotation and scaling parameters is less perfect than the translation parameter(which follows directly from CG). A circular grid is more robust to small errors in these parameters than a rectangular grid.

Considering an object of size 64x64 pixels, with the image restricted to the circle of radius 30 centred at pixel (32,32), 15 digital circles of two pixel thickness can be drawn to cover the image. On each circle eight arcs are identified. An image can be described in terms of these arcs. It should be noted that nonuniformly placed circular arcs that have equal areas may be preferable. Similarly, to improve accuracy, more arcs may have to be identified. A simple feedforward network with 64x64 input nodes and 15x8 output nodes performs feature extraction. Each input node represents a pixel on the input image. Each output node represents a digital arc on the image and it sums up the contributions from the image pixels.

Noisy pixels have an uncorrelated distribution and hence their contributions are distributed over the feature space, whereas pixels on the rigid object form parts of digital arcs and hence their contributions cluster in the feature space, instantiating the feature.

IV. CLASSIFICATION

Classification consists of associating the feature vectors with the corresponding output symbols. This consists of a learning phase in which the system is taught the input-output relationships. Based on this knowledge, it classifies the input images later. A multilayered feedforward neural network can form arbitrary decision regions in multidimensional vector spaces. In this process they also generalise better than conventional techniques[3].

In this study, a multilayered neural network is used for classification. It is a feedforward network with 120 nodes in the input layer, 2 nodes in the output layer corresponding to the objects. It has one hidden layer in between the input and the output layers. For classification, the network is trained to the set of noise free normalised patterns. During training, features derived from each of the input patterns are presented at the input layer along with its label at the output layer. Training consists of finding a set of weights for all the connections such that the desired output is generated for each input.

Backpropagation algorithm is used for training the multilayer network, in which learning is performed by computing the error between the desired and actual outputs, and then feeding back this error signal level by level to the inputs, changing each weight in proportion to its responsibility for the output error. Thus the algorithm is an iterative gradient descent procedure in the weight space, which minimizes the total error between the desired and actual outputs of all the nodes in the system.

V. EXPERIMENTAL STUDY AND RESULTS

The test set consists of 64x64 binary images of two underwater objects shown in Fig.1. For each image, six different silhouettes are generated by arbitrarily varying scales, orientations and translations, thus giving a total of 12 input images. To these images, noise is added. The resulting images are shown in Figs.2(object 1) and 3(object 2). The noise free patterns are used as training input to the system. After successful training, each image in Figs.2 and 3 is normalised as described in section II, and the resulting images as shown in Figs.4 and 5 are presented to the network. For all the input image data, the network has correctly identified the target.

VI.CONCLUSIONS

Experimental results show that even for noisy images correct estimation of the transformation parameters can be done if knowledge about the rigid object shapes is incorporated in the computation procedure. We have stressed the need for logically separate stages for feature extraction and classification. Circular arcs have proved to be useful features to describe shapes. We have shown that if normalization can be done separately, feature extraction and classification become simpler. The use of neural networks for classification has emerged as an effective method especially for noisy images. Putting these ideas together, we have been able to develop a knowledge-based system for identification of objects from poor quality images despite translation, scale change and rotation.

ACKNOWLEGDEMENTS

We gratefully acknowledge Department of Electronics, Government of India for their support to this work.

REFERENCES:

1. D. E. Rumelhart and J. L. McClelland, 'Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol.1, Foundations,' Cambridge, MA: MIT Press, 1986.

2. S .Bhattacharya, 'Handprinted Character Recognition,' M.Tech Thesis, Indian Institute of Technology, 1987.

3. A. Khotanzad and J. Lu, "Classification of Invariant Image Representations Using a Neural Network," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. 38, No.6, June 1990, pp. 1028-1038.

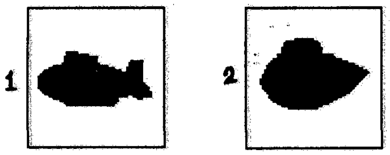

Fig.1. Two objects used in the study.

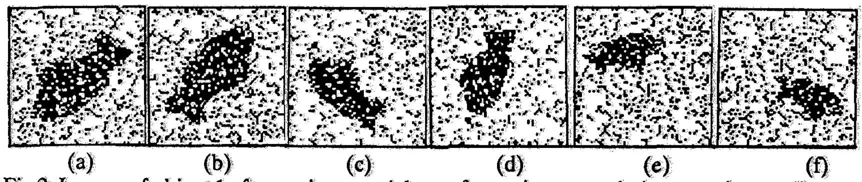

Fig.2. Images of object 1 after various spatial transformations- translation, rotation, scaling and noise.

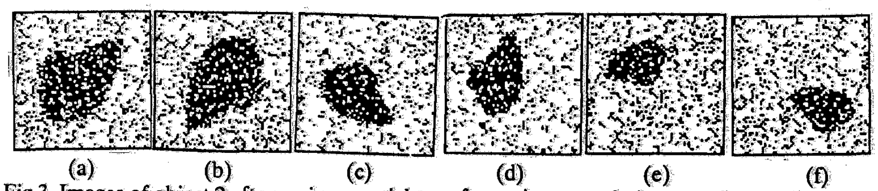

Fig.3. Images of object 2 after various spatial transformations-translation, rotation, scaling and noise.

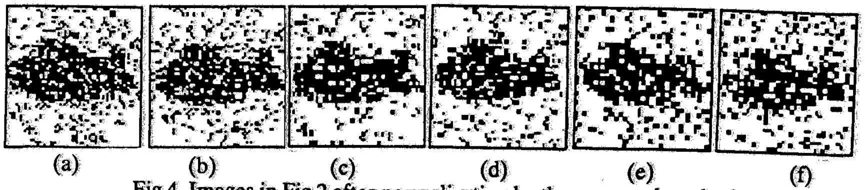

Fig.4. Images in Fig.2 after normalization by the proposed method.

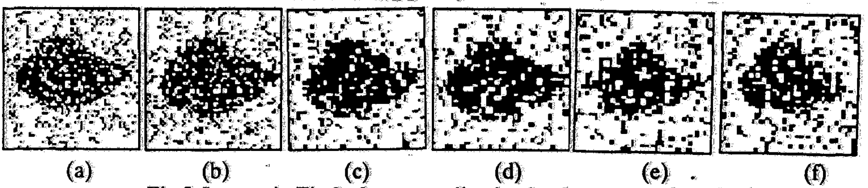

Fig.5. Images in Fig.3 after normalization by the proposed method.