ELSEVIER

Pattern Recognition Letters 17 (1996) 67-76

Pattern Recognition Letters

Face recognition using a hybrid supervised/unsupervised neural network

Nathan Intrator*, Daniel Reisfeld, Yehezkel Yeshurun

Department of Computer Science, Tel-Aviv University, Ramat Aviv 69978, Israel

Received 1 January 1995; revised 18 July 1995

Abstract

A system for automatic face recognition is presented. It consists of several steps. Automatic detection of the eyes and mouth is followed by a spatial normalization of the images. The classification of the normalized images is carried out by a hybrid (supervised and unsupervised) Neural Network. Two methods for reducing the overfitting – a common problem in high-dimensional classification schemes – are presented, and the superiority of their combination is demonstrated.

Keywords: Face recognition; Neural Networks; Interest points; Symmetry operator

1. Introduction

Automatic face recognition has gained much attention in recent years, due to the variety of potential applications, and the increase in computational power which enables effective implementation of algorithms.

Traditionally, face recognition was based on extracting certain features (e.g. spatial location of facial features and their geometrical relations) (Bledsoe, 1966; Kanade, 1973). These features are detected either manually, or by automatic algorithms (Yuille et al., 1992; Craw et al., 1992; Brunelli and Poggio, 1992; Reisfeld and Yeshurun, 1994). Another approach (Kirby and Sirovich, 1990; Turk and Pentland, 1991), is based on direct processing of the grey-level images. A review of face processing systems could be found in (Bruce and Burton, 1989; Samal and Iyengar, 1992).

The task of recognizing faces is inherently a classification problem in high dimensional feature space, and thus subject to the "curse of dimensionality" (Bellman, 1961) which essentially says that the number of training patterns needed for robust classification, should be restrictively high.

Regarding an image merely as a matrix and looking for algebraic invariants (Hong, 1991) reduces the dimensionality. However, such algebraic constraints can be designed to be invariant to practically any transformation but they are too general. For instance they are not affected by upside down inversion, while biological systems are (Moses et al., 1993). A recent approach to this problem (Bischel and Pentland, 1994) is based on the finding that facial images are projected to connected domains (an extension of clusters) and thus could be used to reduce dimensionality. An alternative approach for the reduction of the dimensionality is to use a limited set of biologically motivated receptive fields (Manjunath et al., 1992; Edelman et al., 1992). Yet another way to overcome the "curse"

* Corresponding author.

SSDI 0167-8655(95)00089-5

is to base the recognition on a small number of linear combinations (projections) of the high dimensional space. This approach is at the heart of projection pursuit methods (Huber, 1985) and neural network methods. Taking this approach, one is then confronted with the task of finding such an optimal projection. A commonly used approach is based on second order statistics of the data where one extracts the directions in which the variance is maximized - also called the principal components of the data (Kirby and Sirovich, 1990; Turk and Pentland, 1991).

In this paper we adopt a different approach to dimensionality reduction and classification, based on a combination of supervised and unsupervised learning (Intrator, 1991). We first automatically detect the eyes and the mouth in face image by using the *Generalized Symmetry Transform*, and use this information to normalize the images by affine transformation. We then proceed to classification. The supervised learning seeks projections that minimize mean squared error between the output of a feed-forward network and the class label of the image. The unsupervised learning seeks projections which demonstrate some interesting structure in the data essentially by measuring deviation from Gaussian distribution in the form of multimodality. We conclude by comparing our method with principal components based recognition, and by discussing the interpretability of our results.

2. Facial preprocessing and normalization

The main goal of the preprocessing step in our method is the reduction of the dimensionality problem by spatial normalization of the face image. We detect the eyes and mouth in face images using the Generalized Symmetry Transform, and then we warp the face image to a "standard" location using these points.

The Generalized Symmetry Transform is described in (Reisfeld et al., 1995; Reisfeld et al., 1990). The main idea behind it is the following: starting with an edge map, every pixel is assigned a magnitude M that estimates the probability that there is a symmetric spatial configuration of edges around it, and an orientation α, that points in the direction of the main axis of symmetry around the pixel. Thus, for example, the pixel (or pixels) in the center of a circular, elliptic or rectangular area surrounded by edges, will be assigned a high value of M. This results in a Symmetry Map, where every pixel has a value, and the highest peaks of symmetry could be detected. The main difference between our method and other symmetry estimation methods (or even straight forward detection of centers of gravity), stems from the fact that the symmetry map is computed prior to the segmentation stage, while most other methods are performed only when the contours of specific object are already available.

The Generalized Symmetry Transform is context-free, in the sense that it operates directly on pixels and not on known objects. However, it is possible to incorporate application specific information to enhance its performance. In this paper we have used the following operations on face images in order to detect the location of the eyes and mouth:

- Computation of the symmetry magnitude and orientation. This is the standard Symmetry operation described before.
- Computation of the Radial Symmetry (*RS*) (Reisfeld et al., 1995). While the regular symmetry definition does not depend on the specific spatial organization of the edge that contribute to the symmetry measure, this measure assigns high value to pixels that are surrounded by circular contours.
- Detection of the highest peaks of the regular and radial symmetry in the image.
- Detection of the midline of the face image by finding the peak in the autocorrelation function of the edge image.
- Detection of the eyes and mouth by including geometric considerations. This is carried out by finding the location of the highest peaks of the symmetry values, with the assumption that the eyes should be on both sides of the midline, and the mouth should intersect it (Fig. 1).

Once we have detected the location of the eyes and mouth, we warp the image by using affine transformation based on 3 anchor points: the centers of the eyes and the mouth. The images are warped such that the eyes and center of mouth are translocated to predefined locations, thus forming a normalized grey-level image of the face (Fig. 2). We further try to reduce the variability by considering the gradient of these images, to compensate for variable lighting conditions.

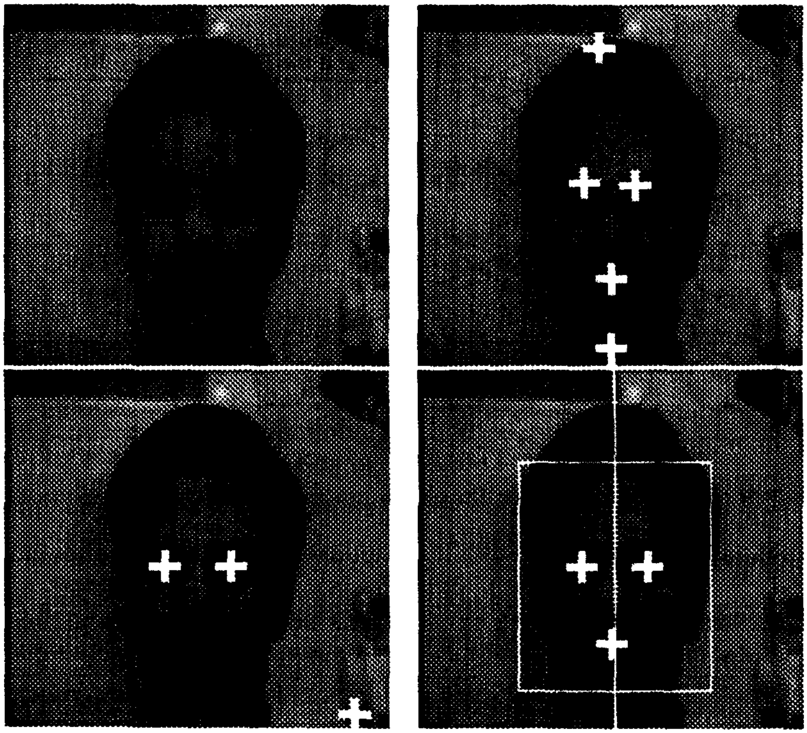

Fig. 1. Locating facial features. Top left to bottom right: An image, the highest peaks of the symmetry magnitude, the highest peaks of *RS*, the midline and the features after applying geometric constraint.

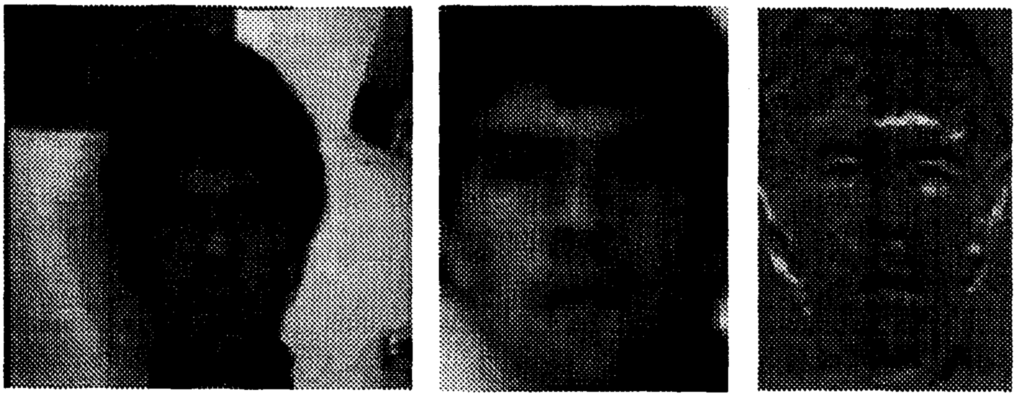

Fig. 2. Normalization transformations. Left to right: Original image; its warping; gradient of the warped image.

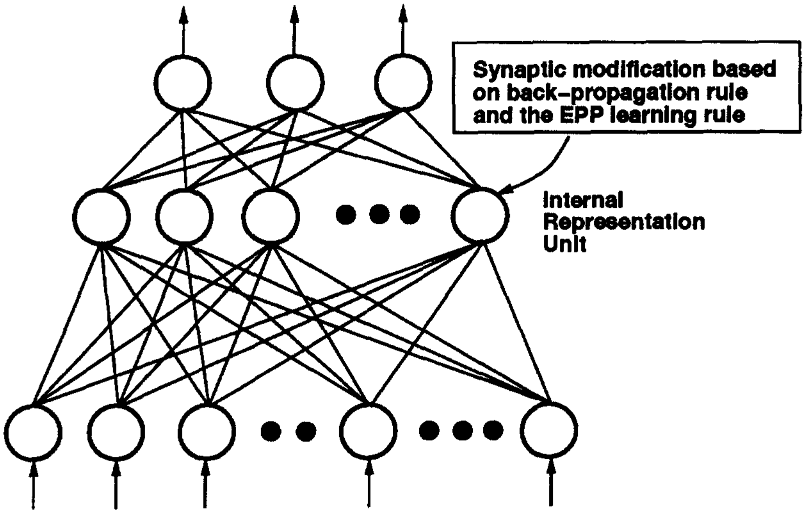
Synaptic modification based on back-propagation rule and the EPP learning rule
Internal Representation Unit

Fig. 3. A hybrid EPP/PPR neural network (EPPNN). The unsupervised learning rule is applied to the hidden layer units only.

3. The feature extraction/classification

We have employed several variations of the frequently used feed-forward artificial neural network for classification. We have chosen to use feed-forward artificial neural networks due to their ability to cope with very high-dimensional data, thus making them excellent candidates to perform recognition from pixel values. The class of functions that can be approximated by a back-propagation type network is very large. This architecture (with an unlimited number of projections) can uniformly approximate arbitrary continuous functions on compact sets, as well as their derivatives (Hornik et al, 1990). The ability to approximate a function and its derivatives will be used below for model interpretability.

The error is propagated backwards to the previous (hidden) layer for modification of its synaptic weights (projections). The single hidden layer architecture is of the form

$$f(x) = \sum_j \beta_j \sigma\Big(\sum_{k=1}^{d} \omega_{jk} x_k + w_{j0}\Big),$$

where σ is an arbitrary (fixed) bounded monotone function. The form

$$f(x) = \sigma\Big(\sum_j \beta_j \sigma\Big(\sum_{k=1}^{d} \omega_{jk} x_k + w_{j0}\Big)\Big),$$

is more suitable for classification tasks. Since this method can approximate any continuous function, great care should be taken so that the variance of the estimated weights is small, and the model does not "overfit" the training data (see (Geman et al., 1991) for discussion). This is often done using some form of complexity regularization such as weight decay (see (Weigend et al., 1991) for review).

The performance of a single back-propagation network can be easily enhanced by training several different networks and averaging their result (Lincoln and Skrzypek, 1990). On this network ensemble, we have used a hybrid training method (Intrator, 1991). This method is based on a formulation that combines unsupervised (exploratory) methods for finding structure (extracting features) and supervised methods for reducing classification error. The unsupervised training portion is aimed at finding features such as clusters. The supervised portion is aimed at finding features that minimize classification error on the training set. The combination of both methods may give better generalization performance (under "good" a priori assumptions about the structure of the data). The application of the hybrid training in a feed-forward neural network is done by modifying the learning rule of the hidden units to reflect the additional constraints (Fig. 3).

The unsupervised feature extraction which we used, is based on the biologically motivated BCM neuron (Bienenstock, 1982). This method essentially seeks clusters in the data distribution by seeking multimodality in the projected distribution via a robust measure that is based on the third and second order statistics of the data.

3.1. The unsupervised constraint

The network implementation described below can find several projections in parallel while retaining its computational efficiency. It was found to be applicable for extracting features from very high dimensional vector spaces (Intrator and Gold, 1991).

Below is a brief description of the unsupervised portion of the network (see (Intrator and Cooper, 1991) for details.) The activity of neuron k in the network is

$$c_k = \sum_i x_i w_{ik} + w_{0k}.$$

The *inhibited* activity and threshold of the kth neuron are given by

$$\tilde{c}_k = \sigma\Big(c_k - \eta \sum_{j \neq k} c_j\Big), \quad \tilde{\Theta}_M^k = E[\tilde{c}_k^2].$$

The threshold $\tilde{\Theta}_M^k$ is the point at which the modification function ϕ changes sign. The function ϕ is given by

$$\phi(c, \tilde{\Theta}_M) = c(c - \tilde{\Theta}_M).$$

The risk (projection index) for a single neuron is given by

$$R(w_k) = -\{\tfrac{1}{3}E[\tilde{c}_k^3] - \tfrac{1}{4}E^2[\tilde{c}_k^2]\}.$$

The total risk is the sum of each local risk. The negative gradient of the risk that leads to the synaptic modification equations is given by

$$\frac{\partial w_{ij}}{\partial t} = E\Big[\phi(\tilde{c}_j, \tilde{\Theta}_M^j)\,\sigma'(\tilde{c}_j)x_i - \eta \sum_{k \neq j} \phi(\tilde{c}_k, \tilde{\Theta}_m^k)\,\sigma'(\tilde{c}_k)x_i\Big].$$

This last equation is an additional penalty to the energy minimization of the supervised network. Note that there is an interaction between adjacent neurons in the hidden layer. In practice, the stochastic version of the differential equation can be used as the learning rule. In the results reported here, a feed-forward architecture with a single hidden layer of 12 units was used in all the experiments. Training was done using the back-propagation algorithm (Rumelhart et al., 1986) for the supervised part and using the projection pursuit learning (Intrator and Cooper, 1991) for the unsupervised part.

For comparison, we also report classification results based on other classification techniques.

The calculation of significance of the object features for recognition was done via a newly introduced method for interpreting neural networks which is described elsewhere (Intrator and Intrator, 1993). This method extends the interpretability associated with linear or logistic regression to feed-forward neural networks.

4. Experimental methodology

We have used a subset of the MIT Media Lab database of face images, courtesy of Turk and Pentland (1991). Previous results using the same preprocessing and dimensionality reduction using receptive fields and radial basis function networks have been described in (Edelman et al., 1992).

The database we used contained 27 instances of each of 16 different persons. The images were taken under varying illumination and camera location. Of the 27 images, 17 were randomly chosen for each person to be used in training, while the remaining 10 were used for testing. The images were preprocessed as described in Section 2 namely, the eyes and mouth were transformed to standard locations via the symmetry transformation. The size of the warped image was 40 × 60 pixels, the eyes locations at (13, 27) and (27, 27), and the mouth location at (20, 44). The processed images are shown in Fig. 4; on the left, a representation of each of the 16 faces is shown, and on the right 16 instances of a single face are presented to demonstrate the variability between instances of a single image.

5. Results and discussion

We have performed many experiments aiming at analyzing different components of the face recognition scheme. First, we describe the main building blocks

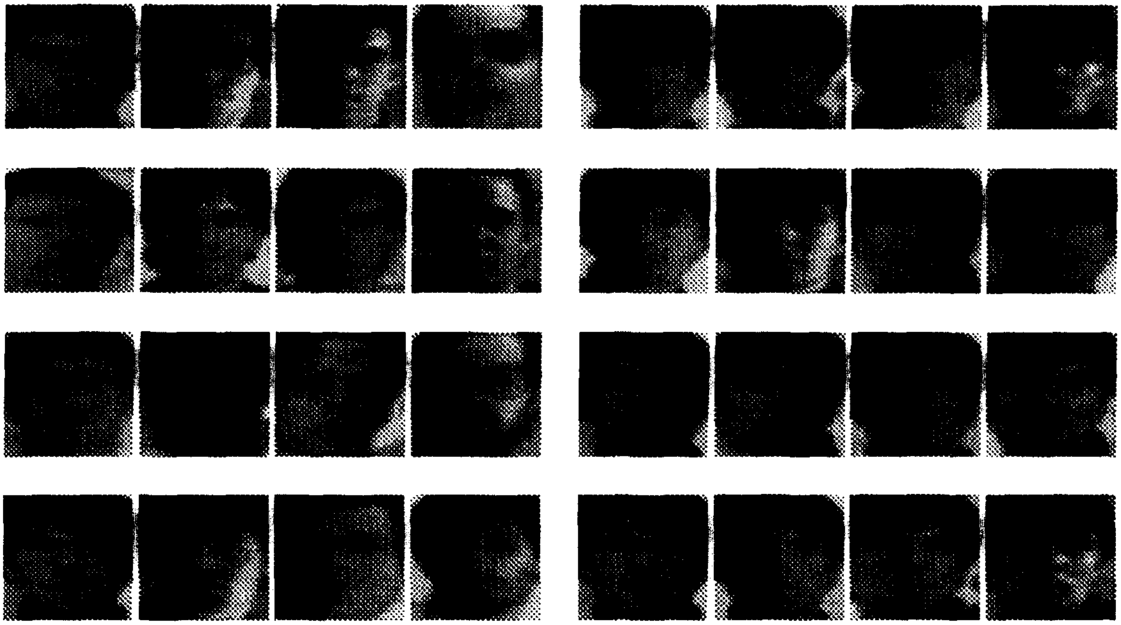

Fig. 4. One normalized image of each class (left) and the variability within normalized subjects for a single face (left)

of our recognition scheme, and then we proceed with the more refined and unique methods which further improved the results. We start with a discussion on the image background of the Turk and Pentland (1991) database.

5.1. Preprocessing

Contribution of the facial background

The faces in the Turk and Pentland data-set are part of a larger picture that contains some background scenery usually a laboratory room of some sort. Generally, the facial background should have a negative effect on recognition performance due to its high variability. However, in this particular data-set, it appears that different subjects were photographed at different places, thus the background such as a corner of a blackboard, picture on the wall etc., was common to the same subject but not to all of them. The positive effect of the image background on recognition results is best demonstrated in Table 1 which includes classification results using one nearest neighbor classifier. Performance on the full images without any preprocessing is already very good. However, this performance drops sharply when the image background is removed, and only slightly when the faces are removed.

Table 1
Effect of the background on nearest neighbor classification. Success ratio is given using the number of images correctly classified and their percentage

Preprocessing type	Success ratio	
Full image	156/160	97.50%
No background	115/160	71.88%
No face	154/160	96.25%

Contribution of the image warp

Once the image background is removed we are left with the facial image only of a fixed size. The image warp, i.e., the affine transformation which takes the eyes and tip of mouth to a fixed location, has now a strong effect on recognition performance. To account for varying light source locations, we also normalized the images so that the mean *gradient* of pixel intensities is zero in all directions. The effect of image-warp transformation and gradient normalization on nearest neighbor classification results is exhibited in Table 2. Classification results were very sensitive to the classification scheme. For example, the gradient normalization was helpful in recognition via nearest neighbor classification, but not needed for the neural network classification schemes. It was surprising to find that three nearest neighbors performed much worse that one nearest neighbor (Tables 3, 5), suggesting that even after the removal of background and image nor-

Table 2
Nearest neighbor performance on the warped faces (with no background)

Warp	Gradient	Success ratio	
No	No	115/160	71.88%
No	Yes	118/160	73.75%
Yes	No	131/160	81.88%
Yes	Yes	147/160	91.88%

Table 3
The effect of classification method on the removed-background images with warping and gradient normalization

Classifier	Success ratio	
1 NN	147/160	91.88%
3 NN	83/160	51.88%
RBF	147/160	91.88%

malization, the variability between images of the same class was still large enough, so that under simple Euclidean metric, there were closer neighbors from other classes. The implication of this finding is that another transformation was needed to reduce the dimensionality to a more invariant image representation. Earlier work with Radial Basis Function (RBF) classification (Edelman et al., 1992) produced similar results to one nearest neighbor scheme on the warped images (Tables 3, 5).

5.2. *Principal component extraction*

Due to the success of principal components for face recognition (Kirby and Sirovich, 1990; Turk and Pentland, 1991), we have studied the classification performance based on projections onto a varying number of principal components extracted from the data (PC features). In all cases, the principal components were the eigenvectors of the covariance matrix of the pixel correlations with highest eigenvalues. Nearest neighbor performance under the various preprocessing described above, is given in Table 4 for a varying number of PC features. The extraction of features via principal components did not eliminate the strong dependency on the number of nearest neighbors (Table 5) however improved the results compared with the original images (Table 3), suggesting that the image data representation is redundant, and that dimensionality reduction, such as the one done by projecting onto the principal components, can reduce the error to a half, while reducing the representation from 40 × 60 dimensions to 44. In the next section we study the classification performance achieved by a much stronger dimensionality reduction based on neural network approaches.

Table 4
One nearest neighbor classification using varying number of principle components extracted from 17 images for each person, without background, with warping and gradient

Number of eigenvectors	Success ratio	
1	29/160	18.13%
5	115/160	71.88%
20	142/160	88.75%
40	150/160	93.75%
50	152/160	95.00%
62	151/160	94.38%

Table 5
Various classification techniques using 17 learned images for a person, without background, with warping and gradient for 44 eigenvectors

Classifier	Success ratio	
1 NN	152/160	95.00%
3 NN	99/160	61.88%
RBF	152/160	95.00%

Table 6
Classification error on a test set from the Turk/Pentland database. Average is done on 5 networks. Figure of merit is calculated as 100 − rejections − 10 × substitutions.

Method	Success (%)	Figure of merit (%)
Back-Propagation	96.72 ± 0.31	73.36 ± 13.43
Hybrid BCM/BP	96.04 ± 0.96	71.14 ± 17.33
Averaged Back-Propagation	98.75	96.3
Averaged Hybrid BCM/BP	99.38	98.1

5.3. *Neural network classification*

Training neural networks was described in Section 3. Classification results are summarized in Table 6. We first give results of the images obtained after eliminating the background and leaving an image of 60 × 40 pixels. The last two lines in the table correspond to results obtained by averaging over the outputs of 5 networks before producing the classification results. This network ensemble average method can simply be considered as reducing the variance of the network outputs (considered as random variables) by summing over an ensemble of networks (Lincoln and Skrzypek, 1990). This technique has been shown to be a good stabilizer for neural network results (Breiman, 1992).

Often the cost of making a mistake (substitution error) is larger than the cost of no decision (rejection). In digit recognition, a frequent classification measure suitable in such cases is the *figure of merit*, in which the cost of substitution is 10 times the cost of rejection. The figure of merit, s, is calculated as

$$s = 100 - \text{rejections} - 10 \times \text{substitutions}.$$

Two points are worth mentioning in the results. First, as is often found, network ensemble reduces classification error. The results of networks trained with additional bias constraints in the form of BCM (Intrator and Cooper, 1991; Bienenstock, 1982) are intriguing; While the mean performance of networks trained with additional (bias) constraints, which are supposed to seek structure in the form of multimodality, is slightly worse compared to networks that were not trained with such constraints, the ensemble performance of such networks yields better performance. These results are best explained by the bias/variance trade-off (see (Geman et al., 1991) for review); the effort to control the bias via bias constraints, increases the variance in single networks, however, the ensemble network averaging does not affect the bias, but reduces the variance leading to an overall improvement in classification results. An indication of the increased variance can be seen in Table 6 by the increased standard deviation of the results for the hybrid method. These results complement a different set of experiments which tried to study the effect of variance constraints on feed-forward neural networks. In that work, variance constraints in the form of weight decay (Weigend et al., 1991) were used in a real-world character recognition problem. While performance of single networks improved on average, the performance of the network ensemble was worse than the performance of an ensemble of networks that were not trained using variance constraints. This is because the variance control via weight decay introduced bias which could not be removed by the network ensemble.

5.4. Interpretability of the networks

Fig. 5 shows a hidden unit representations of a plain back-propagation network (left) and of a hybrid BCM/Back-propagation (right) each taken from one of the corresponding networks. The different networks had various initial conditions and various relative strength of the unsupervised contribution. Although a total of 12 features were extracted, only 7 of the projections are different, which gives the surprising result that an efficient dimensionality reduction can give good classification performance of 16 different faces using only 7 features.

Fig. 6 presents another way to interpret the results of either network. The mean derivative with respect to the inputs for each of the 16 persons is shown. This form of interpretation is very useful when considering the network architecture as a non-linear regression function approximation. In this case it indicates which parts of the image are most useful in improving the classification results, (the white areas) and which parts are mostly contributing to classification errors (the dark areas). There are various robustification issues related to the fact that the models which a network converges to are not unique. The full details of the method are described in (Intrator and Intrator, 1993). The images presented in Fig. 6 give the relative importance of parts of the images for the recognition of that specific prototype. The extreme parts of the image (both negative – dark, and positive – bright) indicate the important features. Notice that the head outline, eyes and mouth are more salient on the Hybrid BCM/BP method (right) than on the BP method (left). This is more consistent with psychophysical experiments (Davis et al., 1978; Fraser and Parker, 1986) that show that more attention is devoted to prominent facial features such as eyes and mouth. Such interpretability method may be useful for human psychophysics studies, and for possible comparison between human and machine recognition, and for the study of object features.

5.5. Summary

We have presented a system for face recognition that addresses several of the important issues in robust recognition:

- Location variability is addressed by the ability of the generalized symmetry transform to locate anchor points in the image and thus shift the image to a fixed location.
- The warping of the image using affine transformation such that the eyes and mouth are mapped to standard locations reduces variability between images, thus reducing the number of prototypes needed for training, and helps to overcome viewpoint variability.
- The use of neural network classification reduces

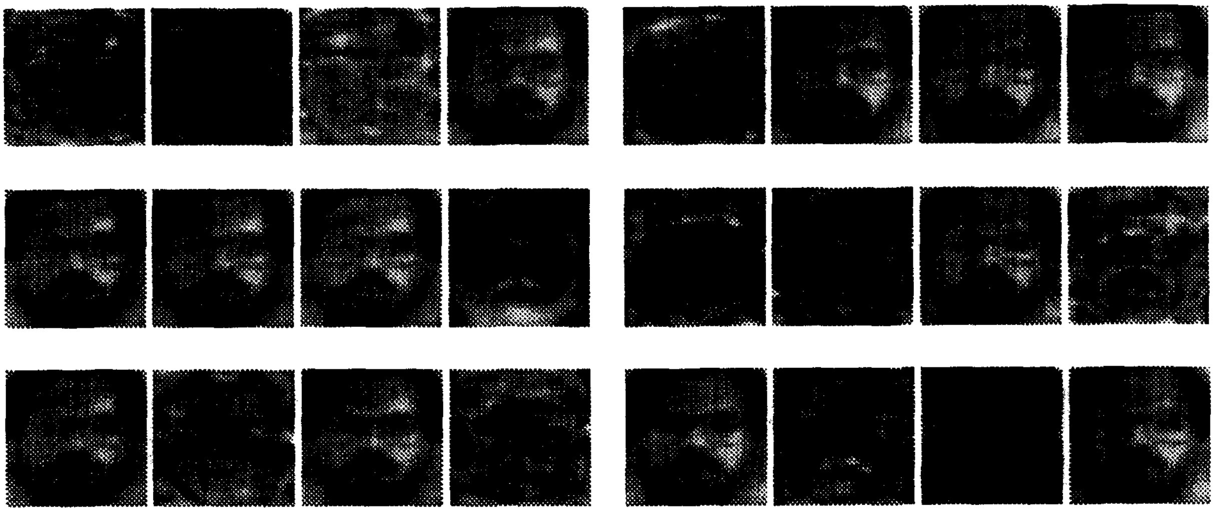

Fig. 5. Features extracted using Back-Propagation Network (left) and features extracted using a hybrid BCM/Back-Propagation Network (right).

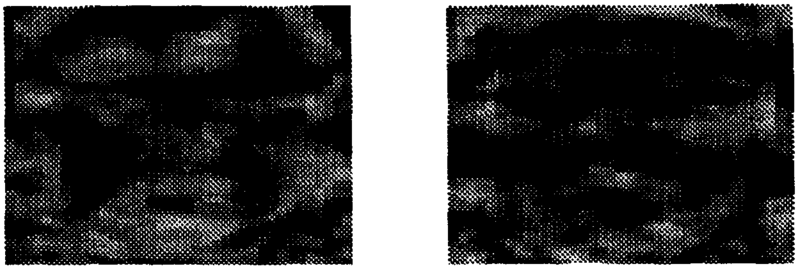

Fig. 6. The significance of the features extracted by using a Back-Propagation Network (left) and by using a Hybrid BCM/Back-Propagation Network (right).

dimensionality of image representation and improves recognition performance.

- The use of ensemble of networks improves recognition performance and reduces substitution errors.
- The use of BCM feature extraction, further improves recognition and reduces rejections for zero substitution errors.

Further work remains in studying the scaling properties of artificial neural networks to large data-sets of faces.

References

Bellman, R.E. (1961). *Adaptive Control Processes*. Princeton Univ. Press, Princeton, NJ.

Bienenstock, E.L., L.N. Cooper and P.W. Munro (1982). Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex. *J. Neuroscience* 2, 32–48.

Bischel, M. and A. Pentland (1994). Human face recognition and the face image set's topology. *Computer Vision, Graphics, and Image Processing* 59 (2), 254–261.

Bledsoe, W.W. (1966). Man-machine facial recognition. Technical Report Rep. PRI:22, Panoramic Research Inc., Palo Alto, CA, Aug. 1966.

Breiman, L. (1992). Stacked regression. Technical Report TR-367, Dept. of Statistics, Univ. of California, Berkeley, CA, Aug. 1992.

Bruce, V. and M. Burton (1989). Computer recognition of faces. In: A.W. Young and H.D. Ellis, Eds., *Handbook of Face Processing*. North-Holland, Amsterdam, 487–506.

Brunelli, R. and T. Poggio (1992). Face recognition through geometrical features. In: *Proc. 2nd European Conf. Computer Vision*, Santa Margherita Ligure, Italy, May 1992, 972–800.

Craw, I., D. Tock and A. Bennet (1992). Finding face features. In: *Proc. 2nd European Conf. Computer Vision*, Santa Margherita Ligure, Italy, May 1992, 92–96.

Davis, G.M., H.D. Ellis and J.W. Shepherd (1978). Face recognition accuracy as a function of mode of representation. *J. Appl. Psychology* 63, 180–187.

Edelman, S., D. Reisfeld and Y. Yeshurun (1992). Learning to recognize faces from examples. In: *Proc. 2nd European Conf. Computer Vision*, Santa Margherita Ligure, Italy, May 1992, 787–791.

Fraser, I. and D. Parker (1986). Reaction time measures of feature saliency in perceptual integration task. In: H. Ellis, M. Jeeves, F. Newcombe and A.W. Young, Eds., *Aspects of Face Processing*. Martinus Nijhoff, Dordrecht.

Geman, S., E. Bienenstock and R. Doursat (1991). Neural networks and the bias-variance dilemma. *Neural Computation* 4, 1–58.

Hong, Zi-Quan (1991). Algebraic feature extraction of image for recognition. *Pattern Recognition* 24 (3), 211–219.

Hornik, K., M. Stinchcombe and H. White (1990). Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks* 3, 551–560.

Huber, P.J. (1985). Projection pursuit (with discussion). *Ann. Statist.* 13, 435–475.

Intrator, N. (1991). Combining exploratory projection pursuit and projection pursuit regression with application to neural networks. *Neural Computation* 5 (3), 443–455.

Intrator, N. and L.N. Cooper (1991). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks* 5, 3–17.

Intrator, N. and J.I. Gold (1991). Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. *Neural Computation* 5, 61–74.

Intrator, O. and N. Intrator (1993). Using neural networks for interpretation of nonlinear models. In: *Amer. Statist. Soc.: 1993 Proc. Statistical Computing Section*. Amer. Statist. Assoc., Aug. 1993, 244–249.

Kanade, T. (1973). *Picture processing system by computer complex and recognition of human faces*. PhD thesis, Dept. of Information Science, Kyoto Univ., Nov. 1973.

Kirby, M. and L. Sirovich (1990). Application of the Karhunen–Loève procedure for characterization of human faces. *IEEE Trans. Pattern Anal. Mach. Intell.* 12 (1), 103–108.

Lincoln, W.P. and J. Skrzypek (1990). Synergy of clustering multiple back-propagation networks. In: D.S. Touretzky and R.P. Lippmann, Eds., *Advances in Neural Information Processing Systems*. Morgan Kaufmann, San Mateo, CA, Vol. 2, 650–657.

Manjunath, B.S., R. Chellappa and C. von der Malsburg (1992). A feature based approach to face recognition. Technical Report CS-TR-2834, Center for Automated Research. Univ. of Maryland.

Moses, Y., S. Ullman and S. Edelman (1993). Selective attention gates visual processing in the extrastriate cortex. Technical Report, The Weizmann Institute of Science.

Reisfeld, D., H. Wolfson and Y. Yeshurun (1990). Detection of interest points using symmetry. In: *Proc. 3rd Internat. Conf. Computer Vision*, Osaka, Japan, Dec. 1990, 62–65.

Reisfeld, D., H. Wolfson and Y. Yeshurun (1995). Context free attentional operators: the generalized symmetry transform. *Internat. J. Computer Vision* (Special Issue on Qualitative Vision) 14, 119–130.

Reisfeld, D. and Y. Yeshurun (1994). Facial normalization using few anchor points. In: *Proc. 12th IAPR Internat. Conf. Pattern Recognition*, Jerusalem, Israel.

Rumelhart, D.E., G.E. Hinton and R.J. Williams (1986). Learning internal representations by error propagation. In: D.E. Rumelhart and J.L. McClelland, Eds., *Parallel Distributed Processing*. MIT Press, Cambridge, MA, Vol. 1, 318-362.

Samal, A. and P.A. Iyengar (1992). Automatic recognition and analysis of human faces and facial expressions: A survey. *Pattern Recognition* 25 (1), 65–77.

Turk, M. and A. Pentland (1991). Eigenfaces for recognition. *J. Cognitive Neuroscience* 3 (1), 71–86.

Weigend, A.S., D.E. Rumelhart and B.A. Huberman (1991). Generalization by weight-elimination with application to forecasting. In: R.P. Lippmann, J.E. Moody and D.S. Touretzky, Eds., *Advances in Neural Information Processing Systems*. Morgan Kaufmann, San Mateo, CA, Vol. 3, 875–882.

Yuille, A.L., P.W. Halilinman and D.S. Cohen (1992). Feature extraction from faces using deformable templates. *Internat. J. Computer Vision* 8 (2), 99–111.