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

Abstract

Face recognition schemes that are applied directly to gray level pixel images are presented. Two methods for reducing the overfitting – a common problem in high dimensional classification schemes – are presented and the superiority of their combination is demonstrated. The classification scheme is preceded by preprocessing devoted to reducing the viewpoint and scale variability in the data.

1 Introduction

The need for robust face recognition methods has increased in recent years due to increase in possible applications. One can envision having plastic money identification, electronic keys of various sorts, and security identification, all based or supported by face recognition. For these purposes it is crucial that performance will be as close as possible to optimal with minimal number of substitution errors since those may be very costly.

A successful method has to be insensitive to the visual environment in which faces are presented, thus illumination conditions, background, and location variability should not degrade performance. Orientation variability to some extent should be treated as well so that the system could automatically produce feedback to control orientation, or else, reject the image. Another issue that we might expect from such a system is indication to those regions in the pixel image which are most important for recognition. This can be most important for selective compression of facial images for the sol purpose of recognition.

Interpretability of network results is not clear, and in general does not shed much light on the way recognition is performed. In this paper we present a recognition scheme based on artificial neural network that addresses the issues raised above. In addition we apply a recently-introduced method for neural network interpretation to study the effect on recognition performance of different regions in the pixel images.

1.1 Recognition from pixel images

Face recognition form grey level images is highly desired on one hand but very difficult on the other. It is desired since images contain the richest 2D representation of faces. It is also one representation from which good human performance is obtained. Thus, it is reasonable to expect that recognition based on the original grey level images may find rich structure that will lead to better recognition performance. Such recognition is difficult since grey level images are vectors in a very high dimensional space and are thus subject to the "curse of dimensionality" [1] which essentially says that the number of training patterns needed for robust classification, should be ridiculously high.

One way to overcome this is to normalize the facial images over changes of viewpoint and then to base the recognition on a small number of linear combinations (projections) of the high dimensional space. The search for projections is at the heart of projection pursuit methods [2] and artificial neural networks (ANN). Taking this approach, one is then confronted with the task of finding optimal projections. A commonly used method is based on second order statistics of the data where one extracts the directions maximizing the variance – the principal components of the data [3, 4].

In this paper we adapt a different approach for dimensionality reduction and classification, based on a combination of supervised and unsupervised learning [5]. 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

1051-4651/94 $04.00 © 1994 IEEE

data, essentially by measuring deviation from normal distribution in the form of multi-modality. In section 2 we describe the preprocessing done on the images for improved invariant recognition. We then describe the architecture and methodology used in our experiments. Results and discussion including comparison with several other approaches is followed by a section describing the interpretability of our results from the point of view of features for face recognition.

2 Facial Normalization

Biological and machine vision systems have to cope with enormous amounts of information. The mechanisms of *attention* and *fixation* enable primates to reduce the amount of information and processing. Most of the photo-receptors of the retina are located at the *fovea* – the part of the eye with the highest resolution and the eyes rapidly move from one *fixation point* to another [6]. Moreover, resources are not allocated uniformly over the field of view: When a primate focuses his *attention* on a location, events occurring at that location are responded to more rapidly, give rise to enhanced electrical activity, and can be reported at a lower threshold [7].

We have introduced an interest operator, inspired by the intuitive notion of symmetry, as a computer vision analogue to attention and fixation [8, 9]. Our interest operator – the *generalized symmetry transform* [9] assigns a *symmetry magnitude* and a *symmetry orientation* to every pixel. The input to the transform is an edge map – the gradients of intensity at each pixel, and its output is a *symmetry map*, which is a new kind of an edge map, where the magnitude and orientation of an edge depends on the symmetry associated with the pixel. Strong symmetry edges are natural interest points, while linked lines are *symmetry axes*.

Although the symmetry transform is a general purpose low level transform, it effectively locates interest points in images without using *a priori* knowledge of the world. However, when supplied with such knowledge, it can be turned into an efficient feature detector. In particular, a face recognition system can take advantage of the fact that it is usually confronted with face images and the symmetry map can be further processed to locate faces and facial features, such as the eyes and mouth. We have turned the generalized symmetry transform into an effective facial features detector using various transformations on the symmetry map along with some basic geometrical relations [10, 11]. The transformation of the symmetry map includes various operations that can be applied also to edge maps such as projecting the symmetry values on the horizon, edge linking, and using local maxima for locating anchor points. The geometrical relations include trivial knowledge on faces such as the fact that the eyes are above the mouth.

Equipped with facial feature detector, the image preprocessing involves a normalization procedure based on an affine transformations which is determined by the locations of the eyes and mouth. A demonstration and further details are given in [9]. The usage of the eyes and mouth anchor points is supported by the fact that humans fixate mainly on these features [6].

3 Hybrid Feature Extraction and Classification

We have employed several variations of the frequently used feed-forward artificial neural network for classification. In addition to plain vanilla feed-forward net trained with back-propagation of error, we have trained several networks to get ensemble network results. This performed significantly better than each of the networks separately. We also used a hybrid training method [5]. 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. Their combination is likely to 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.

The unsupervised feature extraction which we used, is based on the biologically motived BCM neuron [12, 13]. This method essentially seeks clusters in the data by seeking multimodality in the projected distribution via a robust measure that is based on the third and second order statistics of the data. The combined method has already been successfully used for feature extraction and classification in speech recognition [14], using a detailed high dimensional cochlear model speech representation.

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 [15] for the supervised part and using the projection pursuit learning [16] 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 [17]. This method extends the interpretability associated with linear or logistic regression to feed-forward neural networks.

4 Experimental Methodology

We used a subset of the MIT Media Lab database of face images (see Figure 1) courtesy of Turk and Pentland [4]. Previous results using the same preprocessing and dimensionality reduction using receptive fields and radial basis function networks have been described in [18]. 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 center of eyes and tip of the mouth were fixed at a predefined location using the affine mapping determined by the location of the eyes and mouth. Then, a portion of 60×40 pixel image was extracted. Figure 1 shows the processed images. 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

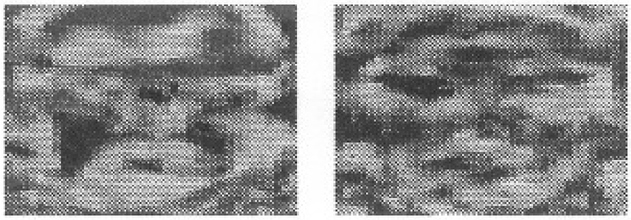

Figure 2: The significance of the features extracted by using a Back-Propagation Network (left) and by using a Hybrid BCM/Back-Propagation Network (right)

Classification results are summarized in Table 1. 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 method corresponds to using a uniform prior on the weight space under Bayesian setup, but can simply be considered as reducing the variance of the network outputs (considered as random variables) by summing over an ensemble of networks [19].

Two points are worth mentioning in the results. First, as is often found, network ensemble reduces classification error. However, the surprising result is that although the mean performance of networks trained with additional (bias) constraints, which are supposed to seek structure in the form of multi-modality, is 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 tradeoff [20, for review]. The effort to control the bias via bias constraints, increases the variance in single networks, however, the network averaging which does not affect the bias, reduces the variance so that the ensemble result is better. An indication of the increased variance can be seen 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 [21].

Interpretability of the networks

Although a total of 12 features were extracted (using 12 hidden units), only 7 of the projections appear to be different. This gives the surprising result that an efficient dimensionality reduction can give good classification performance of 16 different faces using only 7 features.

Figure 2 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 mostly 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 non-uniqueness property of ANN solutions. Full details of the method are described in [17]. The extremum parts of the images (both negative – dark, and positive – bright) indicate the important features. Notice that the head

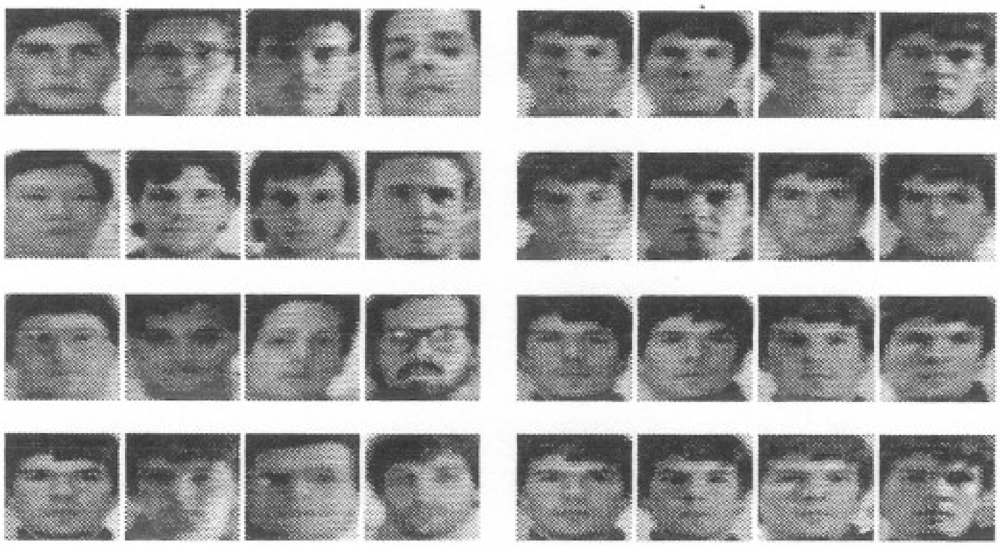

Figure 1: One normalized image of each class (left) and the variability within normalized subjects for a single face (left)

Method	% Error	% Figure of Merit
Back-Propagation	3.28± .31	73.36± 13.43
Hybrid BCM/BP	3.96± .96	71.14± 17.33
Averaged Back-Propagation	1.25	96.3
Averaged Hybrid BCM/BP	0.62	98.1

Table 1: 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.

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 [22, 23, 24]. 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.

Summary

We have presented a system for face recognition that addresses several of the important issues needed for 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 ensemble of networks improves recognition performance and reduces substitution errors.
- The use of BCM feature extraction during training, 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

[1] R. E. Bellman, *Adaptive Control Processes.* Princeton, NJ: Princeton University Press, 1961.

[2] P. J. Huber, "Projection pursuit. (with discussion)," *The Annals of Statistics*, vol. 13, pp. 435–475, 1985.

[3] M. Kirby and L. Sirovich, "Application of the karhunen-loève procedure for characterization of human faces," *IEEE Transactions Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.

[4] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[5] N. Intrator, "Combining exploratory projection pursuit and projection pursuit regression with application to neural networks," *Neural Computation*, vol. 5, no. 3, pp. 443–455, 1993.

[6] A. Yarbus, *Eye Movements and Vision*. New York: Plenum Press, 1967.

[7] M. Posner and S. Peterson, "The attention system of the human brain," *Annual Review of Neuroscience*, vol. 13, pp. 25–42, 1990.

[8] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Detection of interest points using symmetry," in *Proceedings of the 3rd International Conference on Computer Vision*, (Osaka, Japan), pp. 62–65, December 1990.

[9] D. Reisfeld, H. Wolfson, and Y. Yeshurun, "Context free attentional operators: the generalized symmetry transform," *International Journal of Computer Vision*, 1994. special issue on qualitative vision.

[10] D. Reisfeld and Y. Yeshurun, "Robust detection of facial features by generalized symmetry," in *Proceedings of the 11th IAPR International Conference on Pattern Recognition*, (The Hague, The Netherlands), pp. A117–120, August 1992.

[11] D. Reisfeld, *Generalized symmetry transforms: attentional mechanisms and face recognition*. PhD thesis, Computer Science Department, Tel-Aviv University, January 1994.

[12] E. L. Bienenstock, L. N. Cooper, and P. W. Munro, "Theory for the development of neuron selectivity: orientation specificity and binocular interaction in visual cortex," *Journal Neuroscience*, vol. 2, pp. 32–48, 1982.

[13] N. Intrator, "A neural network for feature extraction," in *Advances in Neural Information Processing Systems* (D. S. Touretzky and R. P. Lippmann, eds.), vol. 2, pp. 719–726, San Mateo, CA: Morgan Kaufmann, 1990.

[14] N. Intrator and G. Tajchman, "Supervised and unsupervised feature extraction from a cochlear model for speech recognition," in *Neural Networks for Signal Processing – Proceedings of the 1991 IEEE Workshop* (B. H. Juang, S. Y. Kung, and C. A. Kamm, eds.), pp. 460–469, New York, NY: IEEE Press, 1991.

[15] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing* (D. E. Rumelhart and J. L. McClelland, eds.), vol. 1, pp. 318–362, Cambridge, MA: MIT Press, 1986.

[16] N. Intrator and L. N. Cooper, "Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions," *Neural Networks*, vol. 5, pp. 3–17, 1992.

[17] O. Intrator and N. Intrator, "Using neural networks for interpretation of nonlinear models," in *American Statistical Society: 1993 Proceedings of the Statistical Computing Section*, pp. 244–249, American Statistical Association, August 1993.

[18] S. Edelman, D. Reisfeld, and Y. Yeshurun, "Learning to recognize faces from examples," in *Proceedings of the 2nd European Conference on Computer Vision*, (Santa Margherita Ligure, Italy), pp. 787–791, May 1992.

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

[20] S. Geman, E. Bienenstock, and R. Doursat, "Neural networks and the bias-variance dilemma," *Neural Computation*, vol. 4, pp. 1–58, 1992.

[21] O. Comay and N. Intrator, "Ensemble training: Some recent experiments with postal zip data," in *Proceedings of the 10th Israeli Conference on AICV* (R. Basri, U. J. Schild, and Y. Stein, eds.), pp. 201–206, Elsevier, 1993.

[22] G. Davis, H. Ellis, and J. Shepherd, "Face recognition accuracy as a function of mode of representation," *Journal of Applied Psychology*, vol. 63, pp. 180–187, 1978.

[23] N. Haig, "Investigating face recognition with an image processor computer," in *Aspects of Face Processing* (H. Ellis, M. Jeeves, F. Newcombe, and A. W. Young, eds.), Dordrecht: Martinus Nijhoff, 1986.

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