Learning to Recognize Faces from Examples

Shimon Edelman,[1] *Daniel Reisfeld,*[2] *Yechezkel Yeshurun*[2]

[1] Dept. of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel (edelman@wisdom.weizmann.ac.il)
[2] Dept. of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel (reisfeld@math.tau.ac.il)

Abstract. We describe an implemented system that learns to recognize human faces under varying pose and illumination conditions. The system relies on symmetry operations to detect the eyes and the mouth in a face image, uses the locations of these features to normalize the appearance of the face, performs simple but effective dimensionality reduction by a convolution with a set of Gaussian receptive fields, and subjects the vector of activities of the receptive fields to a Radial Basis Function interpolating classifier. The performance of the system compares favorably with the state of the art in machine recognition of faces.

1 Learning from Examples as Function Interpolation

Classifying the image of a face as a picture of a given individual is probably the most difficult recognition task that humans carry out on a routine basis with nearly perfect success rate. It is not too surprising, therefore, that advances in face recognition by computer fail to match recent progress in the recognition of general 3D objects. The major problem in face recognition appears to be the design of a representation that, on one hand, would be sufficiently informative to allow discrimination among inputs that are all basically similar to each other, and, on the other hand, would be efficiently computable. One way around this problem is to *learn* the required representations, e.g., by examining and remembering several instances of the input.

How can such a simple scheme generalize recognition to novel instances? In a standard formulation of pattern recognition, a characteristic function is defined over a multidimensional space, so that its value is close to 1 over the region corresponding to instances of the pattern to be recognized, and is close to 0 elsewhere [2]. If the characteristic function is smooth, recognition may be generalized to novel patterns of the same class by interpolating the characteristic function, e.g., using splines. An efficient scheme for interpolating (or approximating) smooth functions was proposed recently under the name of HyperBF networks [9,6]. Within the HyperBF scheme, a multivariate function is expanded in terms of basis functions, with parameter values that are learned from the data. For a scalar-valued function, the expansion has the form $f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|^2)$, where the parameters $\mathbf{t}_\alpha$ that correspond to the centers of the basis functions and the coefficients c_α are unknown, and are in general much fewer than the data points ($n \leq N$). The parameters $\mathbf{c}, \mathbf{t}$ are searched for during learning by minimizing the error functional defined as $H[f] = H_{\mathbf{c},\mathbf{t}} = \sum_{i=1}^{N}(\Delta_i)^2$, where $\Delta_i \equiv y_i - f(\mathbf{x}) = y_i - \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2)$. If the centers $\mathbf{t}_\alpha$ are fixed (e.g., are a subset of the training examples), the coefficients c_α can be found by pseudo-inverting a matrix composed of center responses to the training vectors [9] (other, iterative, methods such as gradient descent or stochastic search can be used for the minimization of H). HyperBF interpolation has been previously applied with success to 3D object recognition [7,3,1].

2 Learning Face Recognition

2.1 Preprocessing

Three-dimensional objects change their appearance when viewed from different directions and when the illumination conditions vary. We used alignment [13] to remove the variability in the input images due to changing viewpoint. Our program starts with the identification of *anchor points:* image features that are both relatively viewpoint-invariant and well-localized. Good candidates for such features in face images are the eyes and the mouth. The input image is then subjected to a 2D affine transformation that normalizes its shape and size, so that the two eyes and the mouth are situated at fixed locations. The parameters of the transformation are computed from the desired and the actual locations of the anchor points in the image. We remark that the central assumption behind the choice of 2D affine transform as the normalizing operation is that faces are, to a first approximation, two-dimensional.

Our method of detecting the eyes and the mouth in face images is based on the observation that the prominent facial features are highly symmetrical, compared to the rest of the face [10]. We proposed in [11] a low-level operator that captures the intuitive notion of such symmetries and produces a "symmetry map" of the image. This map is then subjected to clustering. Geometrical relationships among the clusters, together with the location of the midline (as defined by a cross-correlation between two halves of that portion of the image that presumably contains a face), allow us to infer the position of the face, and of the eyes and the mouth in it. These positions are then used as anchor points for affine normalization.

After normalization, the input is a standard-size array of (8-bit) pixels, in which the value of each pixel is determined both by the geometry of the face and by the direction of the illumination. We next reduce the influence of illumination, by the usual method of taking a directional derivative of the intensity distribution at each pixel. The input is then subjected to dimensionality reduction, to increase both the efficiency and the effectiveness of the HyperBF classifier.

A well-known statistical method for dimensionality reduction, principal component analysis, has been applied recently to face recognition with some success [5,12]. In the present work we chose to explore a considerably simpler method, based on the neurobiological notion of receptive field (RF), defined as that portion of the retinal visual field whose stimulation affects the response of the neuron. Assuming that the neuron performs spatial integration over its RF, its output is a (possibly nonlinear) function of $\iint_{RF} K(x,y)I(x,y)dxdy$, where $I(x,y)$ is the input, and $K(x,y)$ is a weighting kernel that we took to be Gaussian (cf. [8]). As noted in [4], pattern classification requires that dimensionality reduction facilitate discrimination between classes, rather than faithful representation of the data. Indeed, the vector of RF activities proved to be adequate for representing face images for recognition, although it would be impossible to recover from it the original structure of the image.

2.2 First Stage: Recognizing Individual Faces

We tested our recognition program on the subset of the MIT Media Lab database of face images made available by Turk and Pentland [12], which contained 27 face images of each of 16 different persons. The images were taken under varying illumination and camera location. Of the 27 images available for each person, 17 randomly chosen ones served for training the HyperBF recognizer, and the remaining 10 were used for testing.

A different recognizer was created for each person, and was trained to output 1 for the images in the training set.

The performance of the individual recognizers was assessed by computing a 16 × 16 confusion table, in which the entries along the diagonal signified mean miss rates and the off-diagonal entries — mean false alarm rates. The table (see Figure 1, bottom) was computed row by row, as follows. First, recognizer for the person whose name appears at the head of the row was trained. Second, the recognition threshold was set to the mean output of the recognizer over the training set less two standard deviations. Third, the performance of the recognizer on the test images of the same person was computed and the miss rate entered on the diagonal of the table. The above choice of threshold resulted in a mean miss rate of about 10%. Finally, the false alarm rates for the recognizer on the images of the other 15 persons were computed and entered under the appropriate columns of the table.

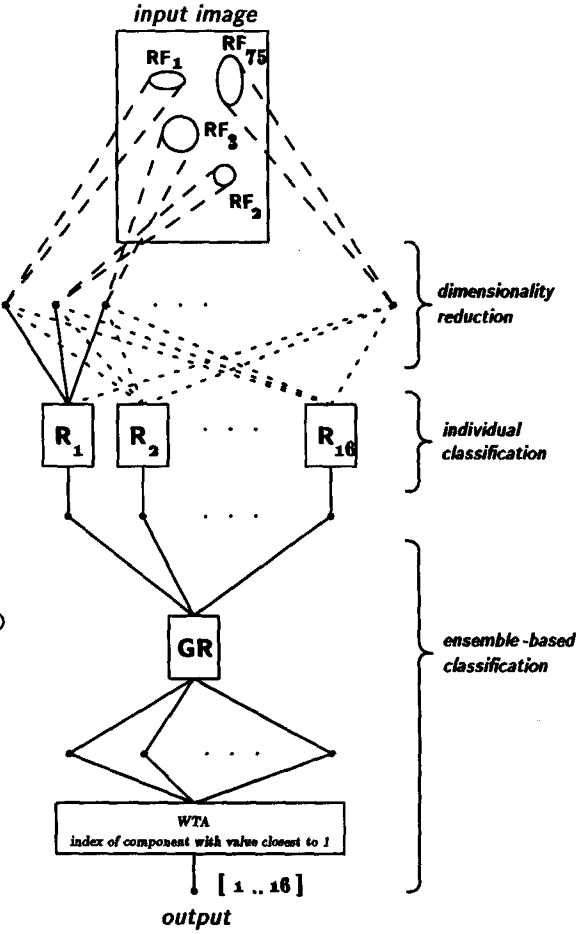
Our second experiment used no thresholds. Instead, recognition was declared for that person whose recognizer was the most active among the sixteen. The performance of this winner-take-all scheme is shown in Figure 2 (left).

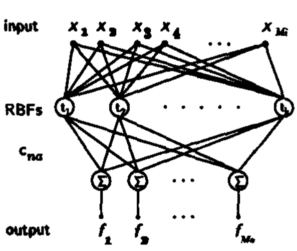
2.3 Second Stage: Incorporating Ensemble Knowledge

An examination of the confusion table reveals that some of the individuals tended to be confused with almost any other person in the database. To take advantage of this "ensemble phenomenon", we trained another HyperBF module to accept vectors of individual recognizer activities and to produce vectors of the same length in which the value corresponding to the activity of the correct recognizer was 1, and all other values were 0 (see Figure 1, right top). The training set for the second-stage HyperBF module was obtained by pooling the training sets of all 16 first-stage recognizers. The outcome of the recognition of a test image was determined by finding the coordinate in the output vector whose value was the closest to 1. The performance of the two-stage scheme was considerably better than that of the individual recognizer stage alone (9% error rate, compared to 22%), demonstrating the importance of ensemble knowledge for recognition (Figure 2, right).

3 Summary

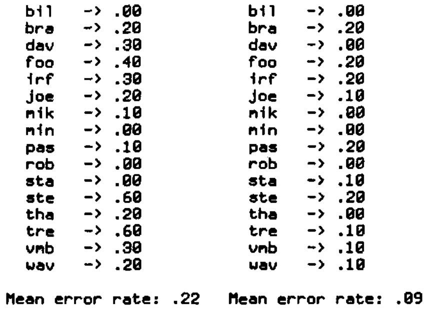
The approach to face recognition described in this paper was made possible by recent advances in model-based object recognition [13], in automatic detection of spatial features [10,11], and in applications of learning and of function approximation to recognition and other visual functions [7,3,8]. The architecture of our system (in particular, its reliance on receptive fields for dimensionality reduction and for classification) has been inspired by the realization that receptive fields are the basic computational mechanism in biological vision. The system's performance, which at present stands at about 5 – 9% generalization error rate under changes of orientation, size and lighting, compares favorably with the state of the art in face recognition [12]. These results have the potential of contributing to the evaluation of a recently proposed theory of brain function [6], and of making practical impact in machine vision.


input image
RF1
RF75
RF3
RF2
dimensionality reduction
R1
R2
R16
individual classification
GR
ensemble-based classification
WTA
index of component with value closest to 1
[1 .. 16]
output



train\test	bil	bra	dav	foo	irf	joe	nik	nin	pas	rob	sta	ste	tha	tre	vmb	wav
bil	0.1		0.2	0.1					0.1			0.2				
bra						0.4							0.4			
dav																
foo				0.1												
irf		0.3		0.1	0.1	0.4			0.1		0.5	0.3	0.2	0.5	0.2	0.1
joe		0.1											0.3			
nik		0.1					0.1				0.1					
nin	0.3		0.5	0.9		0.2		0.1	0.3	0.8		0.8		0.6		0.6
pas	1.0		1.0	0.2	0.1				0.1	0.5		0.8		0.4		0.5
rob				0.2						0.1		0.3		0.6		
sta		0.4				0.5					0.1	0.1	0.4		0.1	
ste												0.1		0.4		
tha						0.1										
tre				0.2				0.1		0.1		0.2		0.1		
vmb		0.3				0.2					0.1		0.4		0.1	
wav																0.1

Fig. 1. *Left top:* a face image from the database we used, courtesy of Turk and Pentland [12], before preprocessing. *Left middle:* a HyperBF network. Basis function centers $\mathbf{t}_i$ (points in the multidimensional input space) are prototypes for which the desired response is known. The output of the network is a linear superposition of the activities of all the basis function units. In the limit case, when the bases are delta functions, the network becomes equivalent to a look-up table holding the examples. *Right top:* The entire two-stage recognition scheme (see text for explanation). *Bottom:* A confusion table representation of the performance of the first stage. Entries along the diagonal correspond to "miss" error rates; off-diagonal entries signify "false-alarm" error rates (zeros omitted for clarity).



```
bil   -> .00         bil   -> .00
bra   -> .20         bra   -> .20
dav   -> .30         dav   -> .00
foo   -> .40         foo   -> .20
irf   -> .30         irf   -> .20
joe   -> .20         joe   -> .10
mik   -> .10         mik   -> .00
min   -> .00         min   -> .00
pas   -> .10         pas   -> .20
rob   -> .00         rob   -> .00
sta   -> .00         sta   -> .10
ste   -> .60         ste   -> .20
tha   -> .20         tha   -> .00
tre   -> .60         tre   -> .10
vmb   -> .30         vmb   -> .10
wav   -> .20         wav   -> .10

Mean error rate: .22  Mean error rate: .09
```

Fig. 2. *Left:* performance of the one-stage recognition scheme. *Right:* performance of the two-stage scheme that uses ensemble knowledge.

References

1. R. Brunelli and T. Poggio. HyperBF networks for real object recognition. In *Proceedings IJCAI*, pages 1278–1284, Sydney, Australia, 1991.
2. R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.
3. S. Edelman and T. Poggio. Bringing the Grandmother back into the picture: a memory-based view of object recognition. A.I. Memo No. 1181, AI Lab, MIT, 1990. to appear in Int. J. Pattern Recog. Artif. Intell.
4. N. Intrator, J. I. Gold, H. H. Bülthoff, and S. Edelman. Three-dimensional object recognition using an unsupervised neural network: understanding the distinguishing features. In D. Touretzky, editor, *Neural Information Processing Systems*, volume 4. Morgan Kaufmann, San Mateo, CA, 1992. to appear.
5. M. Kirby and L. Sirovich. Application of the Karhunen-Loève procedure for characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(1):103–108, 1990.
6. T. Poggio. A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910, 1990.
7. T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.
8. T. Poggio, M. Fahle, and S. Edelman. Synthesis of visual modules from examples: learning hyperacuity. A.I. Memo No. 1271, AI Lab, MIT, 1991. to appear in *CVGIP* **B**, 1992.
9. T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.
10. D. Reisfeld, H. Wolfson, and Y. Yeshurun. Detection of interest points using symmetry. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 62–65, Tokyo, 1990. IEEE, Washington, DC.
11. D. Reisfeld and Y. Yeshurun. Robust Detection of Facial Features by Generalized Symmetry 1991. in preparation.
12. M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3:71–86, 1991.
13. S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989.