[27] S. Ullman and R. Basri. Recognition by linear combinations of models. A.I. Memo No. 1152, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[28] T. Valentine. Representation and process in face recognition. In R. Watt, editor, *Vision and visual dysfunction*, volume 14, chapter 9, pages 107–124. Macmillan, London, 1991.

[29] A. L. Yuille. Feature extraction from faces using deformable templates. In *Proc. CVPR-89*, San Diego, CA, 1989.

[14] D. G. Lowe. *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston, MA, 1986.

[15] J. L. Mundy and A. J. Heller. The evolution and testing of a model-based object recognition system. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 268–282. IEEE, Washington, DC, Osaka, 1990.

[16] T. Poggio. A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910, 1990.

[17] T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

[18] T. Poggio, M. Fahle, and S. Edelman. Synthesis of visual modules from examples: learning hyperacuity. A.I. Memo No. 1271, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[19] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[20] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[21] 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.

[22] D. Reisfeld and Y. Yeshurun. Facial feature detection based on symmetry pronciples: a first step toward recognition. 1991. in preparation.

[23] G. Rhodes, S. Brennan, and S. Carey. Identification and rating of caricatures: implications for mental representations of faces. *Cognitive Psychology*, 19:473–497, 1987.

[24] A. Saha and J. D. Keeler. Algorithms for better representation and faster learning in Radial Basis Function networks. In D. Touretzky, editor, *Neural Information Processing Systems*, volume 2, pages 482–489. Morgan Kaufmann, San Mateo, CA, 1990.

[25] M. Turk and A. Pentland. Eigenfaces for recognition. *J. of Cognitive Neuroscience*, 3:71–86, 1991.

[26] S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989.

[2] R. Brunelli and T. Poggio. HyperBF networks for real object recognition. In *Proceedings IJCAI*, 1991.

[3] P. Burt. Smart sensing within a Pyramid Vision Machine. *Proc. IEEE*, 76:139–153, 1988.

[4] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.

[5] S. Edelman and T. Poggio. Bringing the Grandmother back into the picture: a memory-based view of object recognition. A.I. Memo No. 1181, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990. to appear in Int. J. Pattern Recog. Artif. Intell.

[6] D. H. Foster and P. A. Ward. Asymmetries in oriented-line detection indicate two orthogonal filters in early vision. *Proceedings of the Royal Society of London B*, 243:75–81, 1991.

[7] B. K. P. Horn. Determining lightness from an image. *Computer Vision, Graphics, and Image Processing*, 3:277–299, 1974.

[8] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. *J. Physiol.*, 148:574–591, 1959.

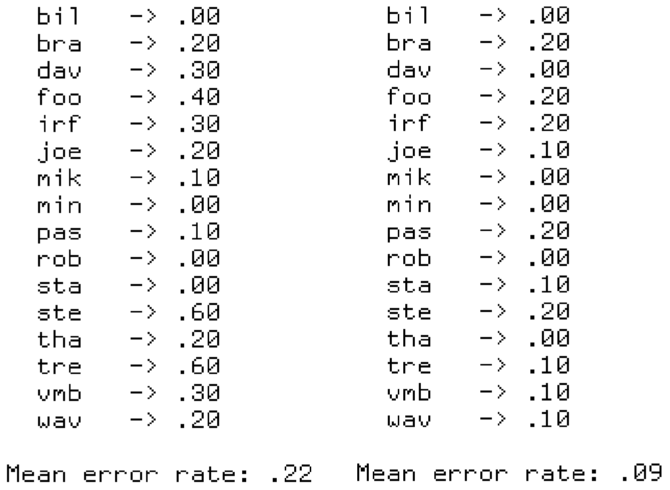
[9] P. J. Huber. Projection pursuit (with discussion). *The Annals of Statistics*, 13:435–475, 1985.

[10] N. Intrator. A neural network for feature extraction. In D. Touretzky, editor, *Neural Information Processing Systems*, volume 2, pages 719–726. Morgan Kaufmann, San Mateo, CA, 1990.

[11] 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.

[12] 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.

[13] S. Lin and B. W. Kernighan. An effective heuristic algorithm for the traveling salesman problem. *Operations Research*, 21:498–516, 1973.



```
 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
```

Figure 6: Left: performance record of the one-stage recognition scheme (see section 3). Right: performance record of the two-stage scheme that uses ensemble knowledge (see section 4). Taking into account ensemble knowledge reduces the mean error rate more than by a factor of 2.

[17, 5, 18]. 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 9% generalization error rate under changes of orientation, size and lighting, compares favorably with the state of the art in face recognition [25]. We are currently working on extending the system to use the full HyperBF scheme [20], and to support continuous learning by data-driven acquisition and modification of the HyperBF centers (cf. [18]). In addition, algorithmic improvements should soon enable the operation of the system at a rate of several frames per second. The outcome of this effort has the potential of contributing to the evaluation of a recently proposed theory of brain function [16], and of making practical impact in machine vision.

References

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

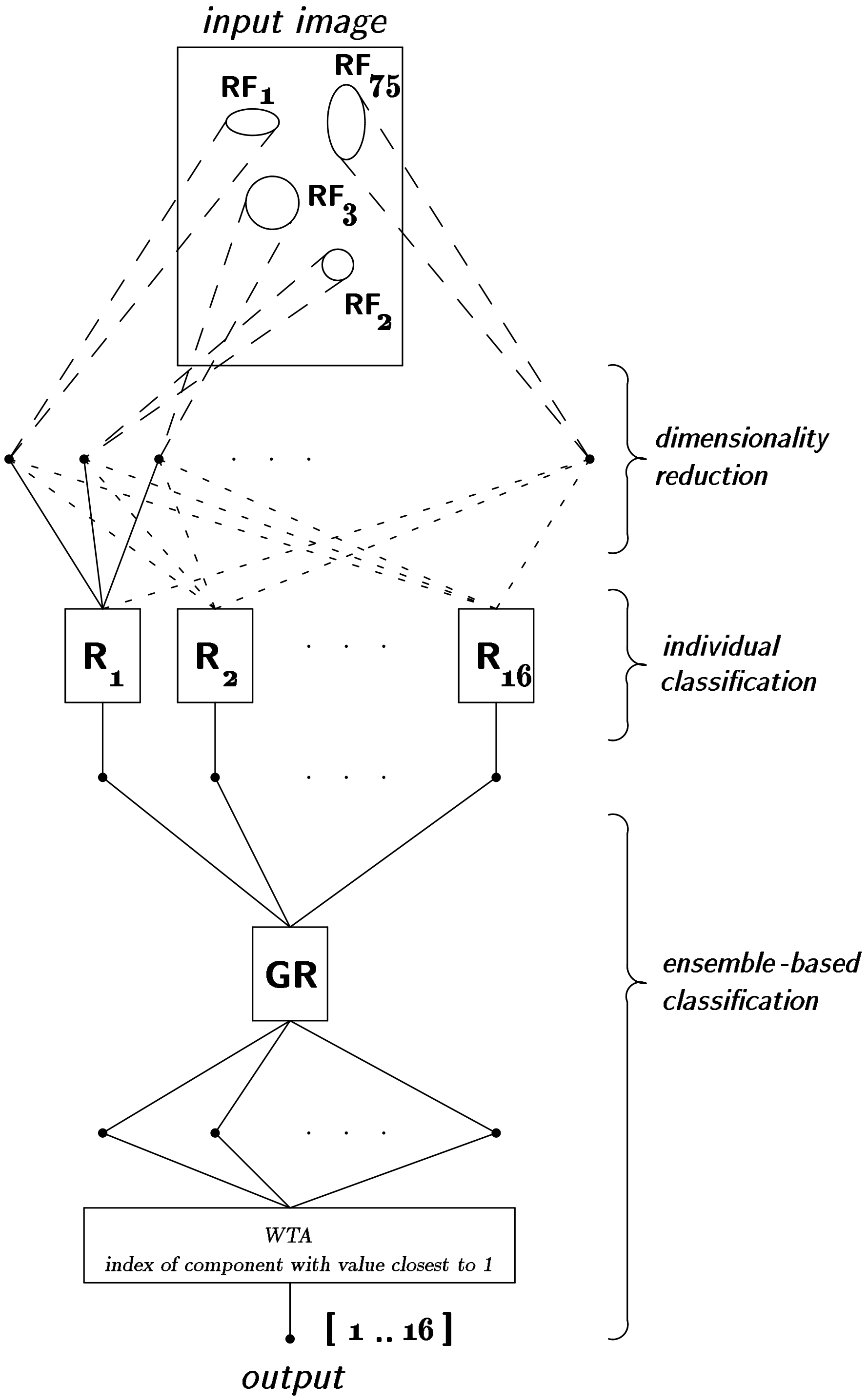

input image
RF1
RF75
RF3
RF2
dimensionality reduction
R1
R2
R16
individual classification
GR
ensemble-based classification
WTA
index of component with value closest to 1
[1 .. 16]
output

Figure 5: The entire two-stage recognition scheme (see section 4 for explanation).

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. The overall error rate, averaged over all the entries of the table im Figure 4, was 7.9%.

3.2 Recognition of specific faces

The method used to build the confusion table described in the previous section does not reflect faithfully the performance of a system based on one recognizer per person, because for any input the activity of more than one recognizer can conceivably exceed its threshold, resulting in a tie for recognition. To break such ties, we carried out another experiment, in which no recognition thresholds were used. 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 6 (left).

4 Second stage: incorporating ensemble knowledge

An examination of the table in Figure 4 reveals that some of the individuals tended to be confused with almost any other person in the database. In fact, this observation, which pertains to the database as a whole rather than to any particular individual, constitutes a kind of "ensemble knowledge", and can be used to improve the performance of the winner-take-all scheme described in the previous section. To take advantage of this knowledge, we have augmented the basic recognition scheme by another processing stage, in which an RBF interpolation module was trained 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 5). The training set for the second-stage RBF 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 (see Figure 6, right).

5 Summary

The approach to face recognition described in this paper was made possible by recent advances in model-based object recognition [26], in automatic detection of spatial features [21, 22], and in applications of learning and of function approximation to recognition and other visual functions

train\test	bil	bra	dav	foo	irf	joe	mik	min	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			
mik		0.1					**0.1**				0.1					
min	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**

Figure 4: A confusion table representation of the performance of our scheme. Entries along the diagonal correspond to "miss" error rates; off-diagonal entries signify "false-alarm" error rates. Entries equal to 0 are omitted for clarity.

the database (ideally, the output for those images should be 0). The success of the approach depended on this property holding true also for the test images, to which the recognizers were never exposed during training. No attempt was made to incorporate negative knowledge by forcing each recognizer to output 0 for images not belonging to the person for which it was being trained. Instead, negative knowledge was used in the second stage of the recognition process, where the vector composed of the output activities of all 16 recognizers was further processed by a winner-take-all module (see section 4). This decision was motivated by our reluctance either to train each recognizer on an unbalanced mixture of 17 positive and $15 \times 17 = 255$ negative examples, or to use an arbitrary subset of these 255 negative examples.

3.1 Pairwise discrimination among faces

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 (a miss is an error in which a recognizer trained for a certain pattern rejects an instance of that pattern; a false alarm is an error in which an instance of a pattern other than the training one is accepted by the recognizer). The table (see Figure 4) 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

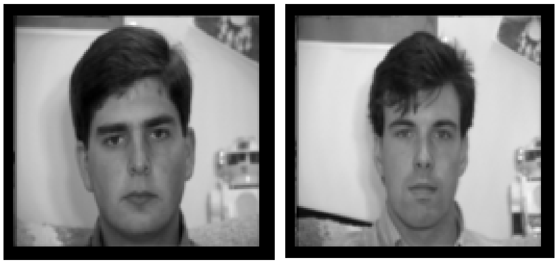


Figure 3: Two of the face images from the database we used in our experiments, courtesy of Turk and Pentland [25]. The original $120 \times 128 \times 8$ images were normalized by affine transformation and cropped to $40 \times 60 \times 8$, using the method described in section 2.2.

not seem to be suitable for distinguishing among highly similar patterns such as faces. First, their output is too precise and is prone to change with changing pose of the objects or lighting direction. Second, edge phase information is lost, unless explicitly considered (which would double the amount of image data to be processed). Indeed, using edge magnitude as the input to the classification stage proved in our case to lead to a severe deterioration in the performance.

Nevertheless, since the power of edge detection in discounting the effects of illumination comes from the derivative operation [7], we used a *directional* derivative, $\frac{\partial}{\partial y}$, aiming for a compromise between retaining the precise magnitude or the precise phase information. This proved to yield better performance than using either raw gray-level intensities, magnitude of intensity gradients, or Canny edges.

3 First stage: training recognizers for individual faces

We have tested our recognition program on the subset of the MIT Media Lab database of face images made available by Turk and Pentland ([25]; see Figure 3). The database we used contained 27 face images of each of 16 different persons. The images were taken under varying illumination and camera location (see [25] for details). Of the 27 images available for each person, 17 randomly chosen ones served for training the RBF 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 σ parameter of the Gaussian basis units, which governs the tradeoff between satisfactory performance on the training set and generalization to novel inputs, was set to the value of the mean distance among the members of the training set [24]. This choice of σ caused the output of a recognizer for a given person to be always greater for the (training set) images of the same person than for images of the other people in

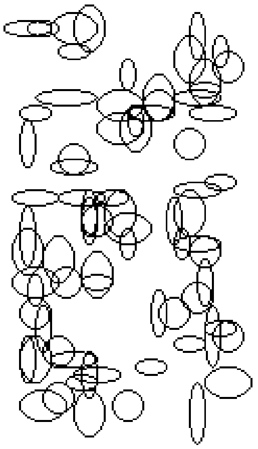


Figure 2: Coverage of the 40 × 60-pixel input image by a set of 75 Gaussian receptive fields (RFs), of different size and elongation. The size parameter σ of the RFs ranged from 2 to 4, and the ratio of x to y dimensions — approximately from 0.3 to 3.0. Using the vector of activities of the RFs instead of the original image for classification effectively reduced the dimensionality of the input from 2400 to 75. The x/y asymmetry was inspired by the shape of the receptive fields of the simple cells in the primary visual cortex of mammals (see, e.g., [8]), and made the units orientation selective. Only two orientations, horizontal and vertical, were explicitly encoded, because of the recent evidence that such encoding may suffice to model psychophysical data on orientation selectivity [6].

tasks, the human visual system exhibits spatial resolution that is far below photoreceptor spacing, presumably due to the integration of information over extended regions of the retina. A computational model based on the transduction of the input by a set of Gaussian receptive fields, followed by an RBF classification module, successfully replicated the major psychophysical findings on human hyperacuity performance. This gave us reason to believe that a similar approach would be useful in face recognition, a task in which the important information lies in the exact shape of the eyes, mouth, nose, etc., and in their precise locations, but not in the general layout and composition of these features (which are the same for all faces).

2.3.2 Discounting the illumination

One possible way to reduce the effects of changing illumination is to perform edge detection. Edge maps obtained with standard edge detection techniques such as the Canny operator do

2.3.1 Dimensionality reduction

From a mathematical viewpoint, classification of gray-level images is complicated by a phenomenon known as the curse of dimensionality. To enable the application of standard classification techniques [4] or of RBF interpolation, images must be represented by points in normed vector spaces. A straightforward way of obtaining such representation is to consider an image of size $n \times m$ pixels as a vector of length $n \times m$. In our application, the dimensionality of the representation space obtained in this fashion would be about 2400. Unfortunately, the number of patterns needed to train a classifier increases exponentially with the dimensionality [1, 11]. Thus, dimensionality reduction must take place before classification is attempted. An intuitive description of the goal of dimensionality reduction is the extraction of low-dimensional features from the original representation space.

The best-known statistical method for extracting features, principal component analysis, has been applied recently to face recognition with some success [25]. It has been argued, however, that principal component features do not necessarily retain the structure needed for classification [4, 9]. A more general and powerful method for feature extraction is projection pursuit (for a review, see [9]). The idea behind projection pursuit is to pick "interesting" low-dimensional projections of a high-dimensional point cloud, by maximizing an objective function such as the deviation of the projected distribution from normality [10]. In the present work we chose to explore a considerably simpler method of feature extraction, based on the notion of localized receptive field, borrowed from neurobiology of vision.

The receptive field (RF) of a neuron anywhere in the visual pathway is defined as that portion of the retinal visual field whose stimulation affects the response of the neuron. Computational models of perception usually assume that the neuron performs a spatial integration over its receptive field, and that its output activity is a (possibly nonlinear) function of $\int\int_{RF} K(x,y)I(x,y)dxdy$, where $I(x,y)$ is the input, and $K(x,y)$ is a weighting kernel that describes the relative contribution of different locations within the receptive field to the output. Here we have assumed that the shape of the kernel is Gaussian: $G(x,y) = \frac{1}{2\pi\sigma}e^{-\frac{(x-x_0)^2+(y-y_0)^2}{\sigma^2}}$.

Our program reduces the dimensionality of the input images by converting them into vectors of RF responses (see Figure 2). As noted by Intrator et al. in [11], for the purpose of pattern classification it is important to concentrate on dimensionality reduction methods that allow 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. Additional motivation for RF-based representation is provided by recent work on the modeling of human performance in a variety of visual tasks commonly referred to as hyperacuity [18]. In these be any point ($k = 1, \ldots, K$), and denote by $\nabla p_k = \left(\frac{\partial}{\partial_x} p_k, \frac{\partial}{\partial_y} p_k\right)$ the gradient of the intensity at point p_k. We assume that a vector $v_k = (r_k, \theta_k)$ is associated with each p_k such that $r_k = \log(1 + \|\nabla p_k\|)$ and $\theta_k = \arctan\left(\frac{\partial}{\partial_y} p_k / \frac{\partial}{\partial_x} p_k\right)$. For each two points p_i and p_j, we denote by l the line passing through them, and by α_{ij} the angle counterclockwise between l and the horizon.

We define the set $\Gamma(p, \psi)$, a distance weight function $D_\sigma(i,j)$, and a phase weight function $P(i,j)$ as

$$
\begin{aligned}
\Gamma(p, \psi) &= \left\{ (i,j) \left| \; \frac{p_i + p_j}{2} = p \; , \; \frac{\theta_i + \theta_j}{2} = \psi \right. \right\} \\
D_\sigma(i,j) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|p_i - p_j\|}{2\sigma}} \\
P(i,j) &= (1 - \cos(\theta_i + \theta_j - 2\alpha_{ij}))(1 - \cos(\theta_i - \theta_j))
\end{aligned}
$$

The directional symmetry measure $S_\sigma(p, \psi)$ of each point p in direction ψ is then defined as

$$S_\sigma(p, \psi) = \sum_{(i,j) \in \Gamma(p,\psi)} D_\sigma(i,j) P(i,j) r_i r_j$$

The rationale for this definition may be found in [21], where we also define a radial symmetry measure.

The map produced by this operator is then subjected to clustering. Geometrical relationships among these 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.

2.3 Feature extraction

Affine normalization removes much of the image variability due to changing viewpoint (pose), but does nothing to counter the effects of changing illumination. After normalization, each input image 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. At this stage, we had to address two different issues. First, the dimensionality of problem had to be reduced, to increase both the efficiency and the effectiveness of RBF interpolation. Second, the effects of illumination had to be discounted, so that the image to be compared with the stored prototypes would contain only the information regarding the identity of the face.

2D views, because the appropriate characteristic function was shown to be smooth [27, 5].[1] RBF-based recognition was subsequently tested on real 3D objects, with promising results [2]. The rest of the paper describes the details of our scheme for recognizing faces that uses RBF interpolation modules as the basic classification tool.

2 Preprocessing

2.1 General considerations

Three-dimensional objects change their appearance when viewed from different directions and when the illumination conditions vary. Images of faces, in addition to that, change with facial expression. In the experiments described in this paper we assumed that viewpoint and illumination direction are the only two factors that contribute to the variability among different images of the same face. Recall that the basic operation called for by the RBF interpolation approach to recognition is a comparison between the input image and a prototype. For this approach to be effective and meaningful, as much of the input variability as possible should be removed before the comparison is made.

To remove the variability due to changing viewpoint, we proceeded according to a generic two-stage scheme related to recently proposed recognition methods known as alignment [26] and viewpoint normalization [14] (cf. [28]). In the first stage of this scheme, our program identifies *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. In the second stage, the input image is 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.

2.2 Detection of anchor points using a symmetry operator

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 [21]. We proposed in [22] a low-level operator that captures the intuitive notion of such symmetries and produces a "symmetry map" of the image.

A symmetry measure for each point and direction is defined as follows. Let $p_k = (x_k, y_k)$

[1] Similar considerations of smoothness, which are outside the scope of this paper, apply to the case of face recognition under varying vantage point and lighting direction.

of Radial Basis Functions (RBF) and is related to multidimensional splines, can be found in [20, 16, 19]. Within the RBF scheme, a multivariate function (which can be vector-valued) 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

$$f(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|^2) \tag{1}$$

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 [20]. Alternatively, iterative methods such as gradient descent can be used for the minimization of H. An even simpler method that does not require calculation of derivatives looks for random changes (controlled in appropriate ways) in the parameter values that reduce the error (cf. [13]). Equation 1 can be implemented by the network of Figure 1, which can be interpreted as follows. The centers of the basis functions, which are points in the multidimensional input space, may be considered prototypical inputs for which the desired response is known. Each unit computes a (weighted) distance of the inputs from its center and applies to it the basis function. In the case of Gaussian bases, a unit will be the most active when the input exactly matches its center. The output of the network is a linear superposition of the activities of all the basis functions. In the limit case, when the bases are delta functions, the network becomes equivalent to a look-up table holding the examples.

1.2 Previous applications of RBF interpolation to recognition

RBF interpolation has been previously applied with success to the recognition of computer-generated 3D objects defined by collections of points in space from their 2D projections [17, 5]. In that case, it was possible to represent objects efficiently by a small number of prototypical

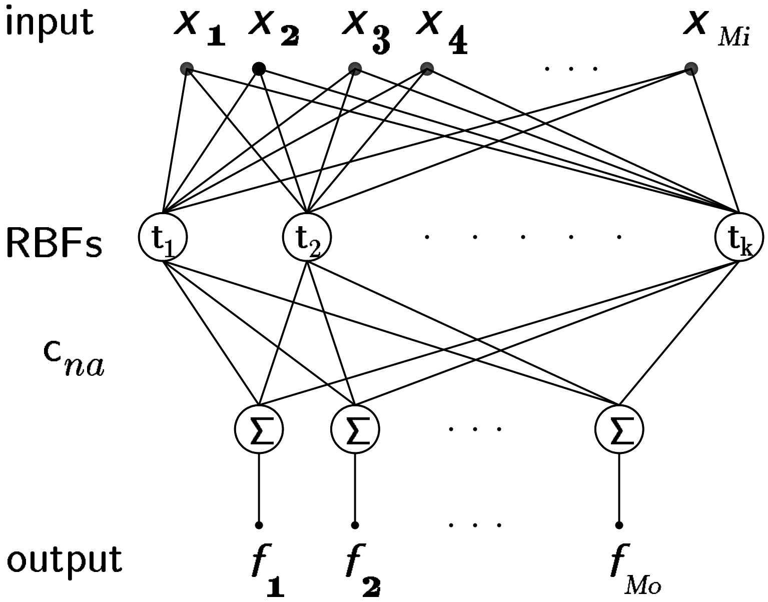

input
x_1 x_2 x_3 x_4 . . . x_{Mi}
RBFs
t_1 t_2 . . . . . . t_k
c_{na}
Σ Σ . . . Σ
output
f_1 f_2 . . . f_{Mo}

Figure 1: A network representation of approximation by Radial Basis Functions (RBFs). The centers of the basis functions, which are points in the multidimensional input space, may be considered as prototypical inputs for which the desired response is known. Each unit computes a (weighted) distance of the inputs from its center and applies to it the basis function. In the case of the Gaussian, a unit will be the most active when the input exactly matches its center. The output of the network is a linear superposition of the activities of all the basis functions. In the limit case when the bases are delta functions, the network becomes equivalent to a look-up table holding the examples.

mechanism known as Radial Basis Functions.

1.1 Function interpolation as a general approach to learning from examples

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 [4]. If the target region is well-behaved (i.e., the value of the characteristic function is known to depend smoothly on location), 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 [20]. Detailed descriptions of this scheme, which is a generalization of the well-known method

1 Introduction

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 (e.g., [14, 26, 15]). 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. Much of the search for a suitable representation method in the past was guided by the observation that humans can recognize faces reliably in cartoon-like drawings that contain very little 3D information, but preserve the shape and the location of the basic structural elements of a face such as eyes, nose and mouth. The possibility that the human visual system represents faces by parameters describing the shapes of the individual features and their spatial arrangement has been discussed recently in psychological literature [23]. Computational exploration of this approach, however, is still largely limited to the detection of individual features [29].

A radical alternative to the reduction of face images to cartoons before attempting recognition is to use all the available intensity information. Although recognition by matching raw images of faces has been successful under limited circumstances [3], it suffers from the usual shortcomings of straightforward correlation-based approaches, such as sensitivity to face orientation, size, and variable lighting conditions, as well as to noise. The reason for this vulnerability of direct matching methods lies in their attempt to carry out the required classification in a space of extremely high dimensionality: since in a direct comparison of two images all pixels make the same contribution to the result, the effective number of dimensions of the classification problem is equal to the number of pixels.

Some of the more recent work in this direction performs dimensionality reduction prior to matching, by computing the principal components of a set of faces and subsequently representing both the stored images and each new input as linear combinations of the estimated principal components (called eigenfaces [12, 25]). Our goal in the study reported here has been to evaluate a much simpler approach to dimensionality reduction, based on the notion of localized receptive fields, borrowed from biological vision. In our scheme, the image is represented by the vector of activities of a set of units that we call transducers. Each transducer can be regarded as a linear filter whose output is proportional to a convolution of the input image with a 2D Gaussian kernel of limited support. Transducer activity patterns for a small set of face images of a given individual are stored and regarded as prototypes. Classification of novel inputs is then done by interpolating among the stored prototypes, using a powerful general-purpose approximation

Learning to recognize faces from examples

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

October 15, 1991

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. Department of Applied Mathematics and Computer Science, The Weizmann Institute of Science, Rehovot 76100, Israel
e-mail: edelman@wisdom.weizmann.ac.il
Tel: +972 – 8 – 342856
Fax: +972 – 8 – 344122

2. Department of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel
e-mail: reisfeld@math.tau.ac.il