How Should We Represent Faces for Automatic Recognition?

Ian Craw, *Member*, *IEEE Computer Society*, Nicholas Costen,
Takashi Kato, and Shigeru Akamatsu, *Member*, *IEEE Computer Society*

Abstract—We describe results obtained from a testbed used to investigate different codings for automatic face recognition. An eigenface coding of shape-free faces using manually located landmarks was more effective than the corresponding coding of correctly shaped faces. Configuration also proved an effective method of recognition, with rankings given to incorrect matches relatively uncorrelated with those from shape-free faces. Both sets of information combine to improve significantly the performance of either system. The addition of a system, which directly correlated the intensity values of shape-free images, also significantly increased recognition, suggesting extra information was still available. The recognition advantage for shape-free faces reflected and depended upon high-quality representation of the natural facial variation via a disjoint ensemble of shape-free faces; if the ensemble was comprised of nonfaces, a shape-free *disadvantage* was induced. Manipulation within the shape-free coding to emphasize distinctive features of the faces, by *caricaturing*, allowed further increases in performance; this effect was only noticeable when the independent shape-free and configuration coding was used. Taken together, these results strongly support the suggestion that faces should be considered as lying in a high-dimensional manifold, which is locally linearly approximated by these shapes and textures, possibly with a separate system for local features. Principal Components Analysis is then seen as a convenient tool in this local approximation.

Index Terms—Automatic face recognition, eigenfaces, face shape, shape-free faces, caricaturing, face manifold.

1 Aims

In machine-based face recognition, a *gallery* of faces is first enrolled in the system and coded for subsequent searching. A *probe* face is then obtained and compared with each coded face in the gallery; recognition is noted when a suitable match occurs. The challenge of such a system is to perform recognition of the face despite transformations: changes in angle of presentation and lighting, common problems of machine vision; and changes also of expression and age, which are more special to faces. The need is, thus, to find appropriate codings for a face which can be derived from (one or more) images of it and to determine in what way, and how well, two such codings shall match before the faces are declared the same.

A number of face recognition systems have become available in the laboratory recently which propose solutions to these problems and a natural concern has been the overall performance of the system [32], [13], [18], [5], [25], [19]. Accordingly, test sets have been constructed and a recognition accuracy computed. In practice, published recognition results are very good, but notoriously difficult to compare [30]. Although the choice of coding and matching strategies differ significantly between systems, the greatest source of variability is probably the least relevant: the selection of the particular collection of faces on which to carry out tests and, in particular, the choice of transformation between target and probe over which the system is supposed to perform recognition. The FERET database may eventually provide a standard, but is only just becoming widely available and does not claim to test recognition over a comprehensive set of transformations.

We seek to avoid some of these difficulties by fixing a matching strategy and a testing regime, and concentrating on the first of the problems just discussed to find effective codes for recognition. Our concern is then no longer how well we can recognize; indeed, for our purposes, a testing regime with a low recognition rate is of most interest: Our interest instead is in *comparing* different coding strategies, both overall and with regard to specific psychological manipulations.

2 Principal Component Analysis

Our main concern is to contrast simple image-based codings with subspace codings, particularly *eigenface* codings, derived from Principal Component Analysis (PCA) and to investigate the effects of content-based preprocessing on these methods. Eigenface codings were used to demonstrate pattern completion in a net based context [17 p. 124], for representing faces economically [16], and explicitly for recognition [32]. Much subsequent work has been based on eigenfaces, either directly or after preprocessing [11], [31], [25], [19].

While undoubtedly successful in some circumstances, the theoretical foundation for the use of eigenfaces is less clear. Formally, PCA assumes that face images, usually

- *I. Craw is with the Department of Mathematical Sciences, University of Aberdeen AB24 3UE, UK . E-mail: i.craw@maths.abdn.ac.uk.*
- *N. Costen is with the Department of Medical BioPhysics, University of Manchester, Oxford Road, Manchester M13 9PT, UK. E-mail: n.costen@.man.ac.uk.*
- *T. Kato and S. Akamatsu are with ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0, Japan. E-mail: {akamatsu, kato}@hip.atr.co.jp*

Manuscript received 17 Oct. 1998; revised 23 Mar. 1999.
Recommended for acceptance by K. Ikeuchi.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 108083.

Fig. 1. Conditions 1, 4, 5, 8, and 10.

normalized in some way, such as co-locating eyes to make them comparable, are usefully considered as (raster) vectors. A given set of such faces (called an *ensemble* here), is then analyzed to find the "best" ordered basis for their span. Some psychological theories of face recognition have such a norm-based coding as their starting point [34]; we now consider such theories, and the relationship between an eigenface coding and human face recognition in more detail.

Aligned, but not distorted, faces have been used [23] to show that, since the accuracy of an eigenface code will depend upon the similarity between the face and the system's knowledge, it is possible to reproduce the well-known "other-race effect" (that members of a different race are notoriously difficult to tell apart) by varying the proportions of Japanese and Caucasian faces used to make up the ensemble. Different eigenfaces correlate differentially with the orthogonal human recognition dimensions "general familiarity" and "memorability" [36] so that these are, respectively, coded by early, global, features and late, local features [24]. The perceived gender of similar images may be manipulated by altering their "weight" on the second and third eigenfaces derived from an ensemble of male and female faces [12], which were previously identified as proving the best gender discrimination [22].

It is thus natural to seek a more principled justification for the use of eigenface or, at least, subspace codings. Considerations of the transformations over which recognition must be performed suggest that an appropriate model may be a "face manifold" [10], and the usual normalization is then seen as a local linear approximation, or chart, for this manifold. Since a chart is a local diffeomorphism, and has its range in a linear space, the average of two sufficiently close normalized faces should also be a face. Clearly, then, existing normalization techniques approximate this interpolation property; it can be argued that, since PCA itself is a linear theory, this is precisely why they are useful.

Other more elaborate normalization techniques can be identified which better approximate this interpolation property and, so, are natural candidates for improved charts. One such has recently become prominent as the way in which a "morph" between two faces is performed [33]. Landmarks are located on each face to provide a description of the face shape or configural information; there is a natural way to average landmark positions and then to map an average face texture onto the resulting shape. More details of such a normalization before PCA are available [11]; we describe it as a decomposition into a shape vector or configuration and a shape-free, or texture, vector. The main aim of our paper is to show that this more elaborate coding can produce significantly better recognition results, as well as advantages in explaining psychological phenomena, as is seen in [15] which found that ,while memorability was determined by the texture, general familiarity was described by the face shape.

A very similar decomposition, also to assist PCA coding, is given in [6], while [19] also presents results whose motivation is very like our own. Configuration and texture are available separately, coded using PCA, and it is shown that the combination was more effective than either alone. However, they choose different images of the same faces as ensemble and, as such, address neither the more general coding issue, essential for larger collections of images, nor the problem of recognition from a single example. The use of larger scale versions of certain face-features in addition to the whole face as an input to a "normal" eigenface-based system can also be considered as another way of approximating the shape-free transformation.

3 Methodology

Our methodology starts with face images on which a collection of landmarks have been located. Our tests are with manual location; automatic location of landmarks is possible, most notably using active shape models [7], [8], or optic flow techniques [35], but, in this study, we avoid the confusion which would arise if incorrectly located landmarks were used in subsequent coding. For a given probe, there is exactly one *target*—another image of the probe face—in the gallery and our interest is in when the probe matches the target better than it matches all other members of the gallery. The restriction to having only one target in the gallery means we are not concerned with learning, based on a set of training examples of the face, leading to a sophisticated matching strategy; here, we concentrate purely on coding issues.

3.1 Images

We work with images of size 128×128, writing N for the number of pixels in each image (so, initially, $N = 16,384$) and n for the number of images in the ensemble; in our case, $n = 50$ or $n = 100$. A total of 14 images of each of 27 people,

Fig. 2. Conditions 11, 12, 13, and 14.

provide our test material. Each image was acquired under fairly standardized conditions; we refer to them as Condition 1 through Condition 14. An initial set of 10 images of each person was acquired on a single occasion. Those in Condition 1 through Condition 4 were lit with good flat controlled lighting, with acquisition times a few seconds apart. Later conditions have increasingly severe lighting variations as well. In Fig. 1, we indicate something of the variability involved.

A subsequent set of four images of each of these 27 faces was acquired between one and eight weeks afterwards: The first such, Condition 11, in lighting conditions similar to those obtaining for Condition 1; subsequent ones in increasingly different conditions. Condition 14 is the only image to be lit with a significant amount of natural and, so, uncontrolled light. In Fig. 2, we show the same subject as in Fig. 1 in each of these remaining conditions.

The 27 images in Condition 1 provide our gallery, which remains fixed throughout. The decision to eliminate condition variation in the gallery was a deliberate simplification. The remaining 13 images of each subject provide our probes; this gives 27×13 or 351 potential probes, each with a corresponding target in the gallery. Using *each* of the 27 faces as probes avoids the possibility that faces in the gallery not used as targets may be hard to recognize: We do this except when calculating acceptance parameters, when a gallery with no target is required; in that case, we used a gallery of 26 faces. Rather than pool results over condition numbers, we keep the conditions distinct, expecting essentially perfect recognition from images in Conditions 2 and 3; those in Condition 14 provide a more varied test.

An additional 50 images of individuals not in the gallery were collected in Condition 1 alone and are used as ensemble images. We describe their span as the "face subspace" and use this when we consider subspace methods. We choose a preferred basis of eigenfaces for this subspace using PCA; this is described in more detail below. The ensemble and the gallery and probes are mutually exclusive, there is no training set per se, and recognition is based on a single target. This approach differs from that employed when eigenfaces are used for representation and reconstruction is done using only a few of the early eigenfaces. In our tests, we use *all* the eigenfaces, although, in Section 6, we discuss the effect of ignoring some. Examples of images in the ensemble are shown in Fig. 3.

Certain tests were performed using an ensemble of nonfacial images. The corresponding eigenimages are, essentially, psychophysically useful two-dimensional derivative-of-Gaussian filters, which can be used for face-recognition [1], [26]. A collection of 50 scenes (resembling holiday snaps) were selected at random. Although the images here are typically less detailed than the faces, [14] found that the nature of the principal components was not affected by the magnification of similar images, which lacked a consistent scale. In contrast, PCA performed upon images of text, which *did* have a consistent scale, produced components which were magnification-dependent. Thus, the particular selection of images is relatively arbitrary. A selection of these images, masked to exclude those regions not considered as part of the "face" are shown in Fig. 4.

3.2 Processing

Each image is processed in the same way before being used in the ensemble, gallery, or as a probe. A total of 34 landmarks, both true and deficient (e.g., the edge of the chin "half way" between two true landmarks) were found manually on each image, giving a triangulation, or face *model*, part of which can be seen in Fig. 5. A (uniformly) scaled Euclidean transformation of the image is derived to minimize the error between the actual positions and the corresponding points on a reference face, here, the average of the ensemble faces, retaining the aspect ratio of the face. Such images are called *normalized*; this removes the effect of image variation associated with different camera locations and orientations and is an alternative to positioning subjects carefully before the images are acquired. The background can then be identified and has no further role in the process; the remaining pixel values are adjusted so the resulting histogram is as flat as possible. Greater sensitivity is attained when our data has a zero mean; to give this, the average image is calculated and subtracted from each member of the ensemble; in practice, the resulting face subspace has dimension $n - 1$.

When a face image includes significant portions of the hair, the available featural information can often give good short term recognition results. However, the hair is not invariant over periods of months during which a practical

Fig. 3. Ensemble images before processing.

system must maintain useful recognition performance. To avoid this problem, we concentrate on a smaller part of the face whose appearance is more invariant; the available landmark data enables such an image, containing "inner features" only, to be extracted, as in Fig. 5. In the example given, the mask has been expanded slightly to display the facial locations and connecting triangulation better; in the results, the border coincides with the outer black line. Essentially, all the results we report are for such images in which the hair has been excluded. These images have $N = 2,557$ pixels; in contrast, the full face image (excluding the background and shoulders) has $N = 5,533$.

In order to ensure comparability, our nonface images were processed in the same way as the face-ensemble. There is no useful way that these images can be normalized; however, we randomly associated them with the 50 ensemble models. Since the point-locations are essentially arbitrary, the resulting distortions caused by normalizing these images do not change the relationship between the images, but ensure that they are processed by the same programs.

3.3 Coding and Matching

The resulting normalized ensemble is subjected to a PCA in which eigenvalues and unit eigenvectors (or *eigenfaces*) of the image cross-correlation matrix are obtained, thus generating a basis for face space. The orthonormality of the basis means it is simple to compute the component of any (normalized) face in the direction of each eigenface and, hence, obtain an $(n-1)$-tuple or code. A coded probe image is then compared with each gallery code to determine the best match. One way to do this uses nearest neighbor matching in $\mathbf{R}^{n-1}$, the span of the ensemble, and a natural choice of metric is the usual Euclidean distance. Since our basis of $\mathbf{R}^{n-1}$ is orthonormal in $\mathbf{R}^N$, this is just the usual Euclidean metric in $\mathbf{R}^N$ and such recognition is effective template matching. Another natural choice of metric on $\mathbf{R}^{n-1}$ which utilizes the fact that our basis is derived by PCA is the Mahalanobis distance, in which

$$d(\mathrm{x},\mathrm{y})^2 = \sum \lambda_i^{-1}(x_i - y_i)^2, \tag{1}$$

where $\{\lambda_i\}$ is the sequence of eigenvalues. This treats variations along all axes as equally significant by weighting components corresponding to smaller eigenvalues more heavily and is arguably appropriate since our aim is discrimination, rather than representation.

A more robust scheme balances false acceptances with false rejections and allows the possibility of no match being acceptable. One such, has a match score c_j between each image in the gallery and the probe image [18]. The best match corresponds to the lowest score and interest centers on the sequence $\{c_j\}$, together, with the lowest value c_0 and the next lowest value c_1. The mean μ and standard deviation σ of the sequence obtained by removing the target image from the gallery are calculated and used to define two inequalities, $c_0 < c_1 - t_1\sigma$ and $c_0 < \mu - t_2\sigma$, for fixed thresholds t_1 and t_2, which must both be met to accept a match.

We adopt this, reporting a correct match as a *clear* hit if the target passes this acceptance or *separation* criterion, and *just* a hit otherwise, with a similar terminology for misses. To set thresholds, the distances between the probes in Condition 2 and a reduced gallery, from which the target had been deleted, were found. The two parameters t_1 and t_2 were then calculated for each probe and the largest values independently chosen as the fixed t_1 and t_2 to be used in subsequent tests. This procedure ensured that, in the best base condition, there was not a false recognition; although particularly conservative, there are cases (e.g., the "Shape-only" results in Table 1), where "clear misses" do occur.

4 Recognition

We group Conditions 2, 3, and 4 together, and describe this as "Immediate" recognition. Conditions 5, 6, and 7 form a very similar set with a small change in lighting and position and these are described as the "Variant" group. More fundamental lighting changes distinguish Conditions 8, 9, and 10, and these are combined as the "Lighting" group. Finally, the four conditions in which the images were acquired after a delay are grouped together as the "Later" set. To give a feel for the overall performance, a weighted average is given as the "Overall" value. Although more images are available in the sets with low condition numbers, greater interest attaches to the more difficult conditions. Accordingly, while the "Lighting" group has a

Fig. 4. Masked nonface ensemble images before processing.

weight twice that of the "Immediate" and "Variant" groups, "Later" has four times the weight.

Our main interest is in the comparison between (scaled) Euclidean normalization and the more intrusive shape-free form and the contrast between these two and a pure correlation approach. However, we first discuss other choices which make up our testing regime.

4.1 Subspace Coding?

The use of subspace coding and subsequent PCA means that only the information in the relevant images, which is preserved when projected onto the "face subspace" is used during matching. In this projection, idiosyncratic information, which could aid recognition, may be lost. An appropriate baseline to asses the effects of such manipulations should use the whole of the relevant image information. To investigate this, a template-based recognition procedure was implemented using the whole of the (masked) face image, excluding the hair. Matching was done on the basis of the best correlation between the probe and gallery, both normalized by a (scaled) Euclidean transformation derived from the landmarks. Pixel value normalization, as throughout, is by histogram equalization. This yields recognition performance given in the "Correlation" section of Table 1. Clearly, the probe images are sufficiently different from the gallery that recognition occurs with a notably low frequency.

4.2 Ensemble Size

Initial testing was done using the ensemble of 50 faces described above. It is not clear that an ensemble of 50 faces is adequate and we borrowed an idea [16], making use of vertical symmetry, or rather the lack of it, in individual faces, by creating 50 "mirror" faces whose images and landmarks were created by reflection about the vertical facial midline. The resulting improvement in recognition shown in Table 2 was sufficiently noticeable to suggest that this enlargement of the ensemble was worthwhile and all subsequent tests are reported with this "doubled" ensemble.

4.3 Mahalanobis or Euclidean Distance?

Our "baseline" recognition in the Euclidean part of Table 1 gives results against which subsequent performance is to be compared. Each normalized image is projected onto the subspace spanned by this enlarged ensemble and matching simply uses the Euclidean metric in this subspace, and so is effectively template matching. Pixel values were processed by histogram equalization to give a crude correction for lighting. Other pixel value normalization methods were tested, including just setting the length of the image vector to be constant (more useful when the average was *not* subtracted, which can, in effect, negate this preprocessing, but still with a total of 15 fewer hits than histogram equalization), using an edge image, an H-transform [37], and restricting to psychologically important spatial frequencies [9]; these last three gave significantly worse results.

Our first comparison is between the "Correlation" section of Table 1 and the "Euclidean" section in which recognition uses only the projection of the faces onto the face subspace. It is clear that a significant amount of information has been lost by this process. In contrast, our second comparison uses the same set of tests in which the match is based on Mahalanobis distance (1), given in the "Mahalanobis" section of Table 1. Scaling the metric in accordance with the variance within the ensemble ensures that matching is in terms of expected variations rather than absolute values.

The use of the Mahalanobis distance is clearly more effective than either straight matching, or projection using Euclidean distance. This confirms that the eigenface

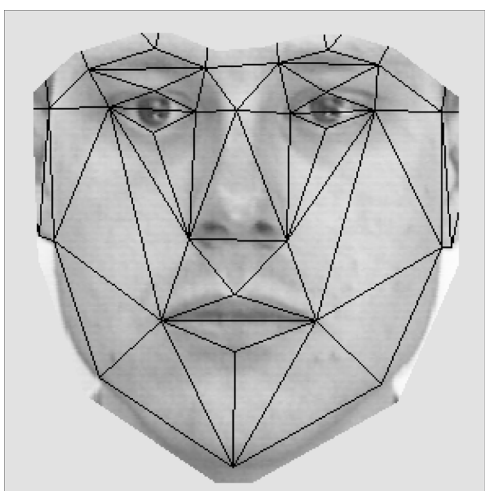

Fig. 5. Inner face showing the facial locations.

TABLE 1
Match Percentages from 351 Trials

		Hit		Miss	
		Clear	Just	Just	Clear
Correlation	Immediate:	31.3	7.4	1.2	0.0
	Variant:	55.6	35.8	8.6	0.0
	Lighting:	11.1	42.0	46.9	0.0
	Later:	23.1	40.7	36.1	0.0
	Overall:	31.3	36.9	31.7	0.0
Euclidean	Immediate:	82.7	14.8	2.5	0.0
	Variant:	34.6	45.7	19.5	0.0
	Lighting:	3.7	29.6	66.7	0.0
	Later:	16.7	28.7	54.9	0.0
	Overall:	22.9	29.2	47.9	0.0
Mahalanobis	Immediate:	90.1	9.9	0.0	0.0
	Variant:	67.9	22.2	9.9	0.0
	Lighting:	17.3	48.1	34.6	0.0
	Later:	34.3	32.4	33.3	0.0
	Overall:	40.2	32.3	27.5	0.0
Shape-free	Immediate	95.1	4.9	0.0	0.0
	Variant	64.2	29.6	6.2	0.0
	Lighting	18.0	51.9	29.6	0.0
	Later	28.7	46.5	25.0	0.0
	Overall	37.4	41.3	21.3	0.0
Laplacian	Immediate	85.2	14.8	0.0	0.0
	Variant	67.9	30.9	1.2	0.0
	Lighting	38.3	50.6	11.1	0.0
	Later	28.7	62.0	9.3	0.0
	Overall	41.0	51.2	7.8	0.0
Shape-only	Immediate:	39.5	46.9	13.6	0.0
	Variant:	23.5	58.0	17.3	1.2
	Lighting:	27.2	54.3	18.5	0.0
	Later:	19.4	59.3	21.3	0.0
	Overall:	23.7	56.7	19.4	0.1

Scaled Euclidean normalized, matching by correlation of the images, or with Euclidean or Mahalanobis distance in the face subspace. Shape-free normalized, matching with Mahalanobis distance, or by full correlation of Laplacian images. Shape only, matching with Mahalanobis distance on the 20 eigenvectors with the largest variance. Hair has been excluded from the matches.

formulation, with its variance properties, is worthwhile here even though we are not truncating the representation, and that we are not simply using the orthogonality properties of the basis. Note also that the advantage is least evident in the "Immediate" group, where naive template matching is expected to perform well; but that, even in this case, the effect on the separation of weighting the later components is noticeable. The comparison is very similar when the hair is included in the image area.

4.4 Shape-Free Normalization

The theoretical considerations in Section 2 suggest that the decomposition of a face into a shape-free or texture vector and the configuration or shape vector of landmark locations may provide more effective coding for recognition. We discuss first the case in which only the shape-free face is used, deliberately ignoring configuration. Thus, our normalization, rather than using a (scaled) Euclidean transformation, involves texture mapping each face to a standard shape; in this case, the average shape of the set of ensemble images. We used linear interpolation based on the model in Fig. 5; although simpler than Bookstein's thin plate spline warps [19], we found the procedure more effective. As usual, the pixel values are then normalized by histogram equalization. The results given in "Shape-free" (part of Table 1) are directly comparable to the "Mahalanobis" section; only the method of the image normalization is different.

TABLE 2
Hit percentages from 351 Trials

	50	$(50 \times 2)/2$	50×2
Immediate	95.1	96.3	98.8
Variant	86.4	82.8	87.7
Lighting	56.8	60.5	71.6
Later	57.4	64.8	71.3
Overall	64.4	69.2	76.1

Scaled Euclidean normalized, matching with Mahalanobis distance. Hair has been excluded from the match. Comparison between an ensemble with 50 faces, the first 50 eigenfaces from the "doubled" ensemble of 100 images (50 faces and their mirrors), and the full "doubled" ensemble.

The comparison suggests that shape-free normalization appears to be slightly *better* than the (scaled) Euclidean version, despite the fact that the shape information has been deliberately ignored. It may be that we have implemented the (scaled) Euclidean normalization inappropriately, but other procedures, including one in which an appropriate ellipse was generated to mask the exterior features of the face and subsequently scaled to a fixed size, were tested and proved less effective.

4.5 Shape-Free instance-Based Codes?

These tests show that there is a real advantage in representing faces in terms of shape-free texture variation. We are thus led to ask if it is also an advantage for instance-based recognition, in which matches are performed on the images themselves, avoiding the need for an ensemble. We thus seek the analog for shape-free faces of "Correlation" section of Table 1. Because our normalization methods use a relatively small number of points, the quality of the match between images may be underestimated; to compensate, the correlation between a probe and each gallery image was optimized separately by choosing the (scaled) Euclidean transform (assumed already very close to the identity), which maximized the image correlation. We found a significant effect of the type of image-processing used; there was a very noticeable advantage for preprocessing using a Laplacian transformation, a 3×3 matrix with a positive center and zero corners, often thought of as a sharpening operator.

The complete recognition rates for the shape-free Laplacian images are given in the "Laplacian" section of Table 1, showing very good and constant recognition. It should, however, be noted that this is very slow, even with the relatively small gallery used here; optimizing the match meant that each image was compared with each gallery

TABLE 3
Recognition by Correlation of Laplacian-Transformed and Histogram-Equalized Images.

			Gallery & Probe Normalization	
			Shape-free	Euclidean
Correlation matching and grey-level transformation.		Match Type		
	Laplacian transform	Without Movement	85.8	74.7
		With Movement	92.2	76.2
	Histogram-equalization	Without Movement	57.5	68.1
		With Movement	57.3	68.3
Ensemble type and normalization agreement.		Ensemble Normalization		
	Face ensemble	Euclidean	70.6	72.5
		Shape-free	78.7	77.0
	Non-face ensemble	Euclidean	56.9	63.7
		Shape-free	57.0	61.2

Hit rates for Euclidean-normalized and shape-free images, with and without correlation-optimization. Recognition with different standardization methods for ensemble and gallery and probes: hit rates for Euclidean-normalized and shape-free images. Using an ensemble of faces or a nonface ensemble.

member 50 times. In contrast, the Laplacian preprocessing is not useful when first projecting onto the face subspace; the shape-free transformation yields a correct hit-rate overall of 26.7 percent.

These results again show the advantage of using a shape-free representation; in this case, it ensures that all sections of the Laplacian-processed images can be aligned at once. In contrast, when the images retain their natural shape, different sections of the probe face compete to match corresponding sections of the gallery images and the effect on the ratio between the distances to the target and distractors is reduced. However, this "shape-free advantage" is dependent on appropriate preprocessing to allow the matching; indeed, the effect is reversed when histogram equalization, our "standard" method is used instead. Table 3 shows these differential effects of optimizing the correlations when subjected to Laplacian and histogram equalization transformations, both for shape-free images and those retaining their natural shape.

We suggest below that the shape-free advantage for PCA appears to reflect better representation, rather than superior matching; thus, we would not expect it to remain in a matching task. Certainly, the shape-free manipulation removes information; the emergence of a shape-free matching advantage following a Laplacian preprocessing has a different basis and serves to emphasize again that notably different operations are occurring in PCA and in correlation.

4.6 Shape Data

Since the shape-free normalization discards information on the shape of the face, recognition may be enhanced by independent consideration of the shape, performing a PCA on the landmark locations. This was done as already described, first applying a (scaled) Euclidean transformation to remove accidental position effects and then, if necessary, removing the points relating to the hair. The shapes of the ensemble images then provided suitable principal components (or *eigenshapes*, we reserve "eigenface" for texture components) as descriptors.

There are a maximum of 46 degrees of freedom in the shape data derived from the 34 landmarks, as some are at fixed angles to an axis defined by the eyes. However, it is to be expected that these data are highly correlated. This was born out during the PCA, when the eigenvalues became small after the first 15 or 20 eigenshapes had been derived. The number of principal components used to code the shape vector was thus varied and the associated hit rates are shown in Fig. 6 for the doubled ensemble with and without hair. These show that, in both cases, recognition peaks when 20 components are included in the analysis; the with-hair images are better than the no-hair images with more than 20 components. The decline in recognition performance when more components are used reflects the undue influence of low-variance noise components on the Mahalanobis distance.

The summary statistics for the doubled ensemble with 20 components used to code the shape are shown in the "Shape only" section of Table 1. Some of the recognition may be a result of variation in camera optics, which were unfortunately not controlled between subjects, but the relatively good "Later" performance argues against this.

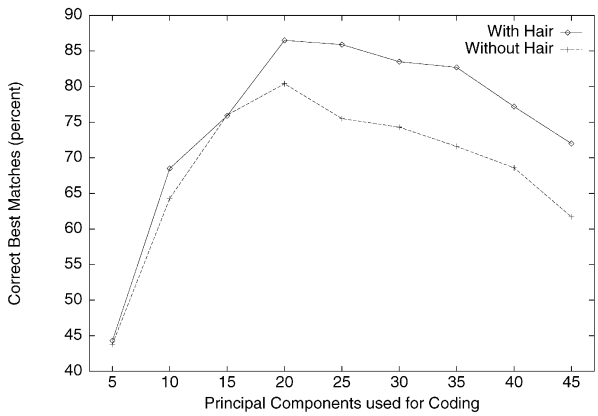

Fig. 6. Variation in correct classification from shape with available principal components.

TABLE 4
Match Percentages from 351 Trials

		Hit		Miss	
		Clear	Just	Just	Clear
Shape and Texture	Immediate:	92.6	7.4	0.0	0.0
	Variant:	70.4	24.7	4.9	0.0
	Lighting:	43.2	49.4	7.4	0.0
	Later:	50.9	36.1	13.0	0.0
	Overall:	55.4	34.7	9.5	0.0
S & T 156% caricature	Immediate:	95.1	4.9	0.0	0.0
	Variant:	76.5	23.5	0.0	0.0
	Lighting:	53.1	40.7	6.2	0.0
	Later:	61.1	30.6	8.3	0.0
	Overall:	64.7	29.2	6.1	0.0
S & T with Correlation	Immediate:	97.5	2.5	0.0	0.0
	Variant:	88.9	11.1	0.0	0.0
	Lighting:	67.9	29.6	2.5	0.0
	Later:	66.7	32.4	0.9	0.0
	Overall:	72.6	26.3	1.1	0.0

Combined shape and texture measures, matching with Mahalanobis distance. Shape and texture alone, with the images caricatured to 156 percent of their original value and veridical images combined with Laplacian preprocessed optimized correlation. Hair has been excluded from the match.

4.7 Shape and Texture

Coding using shape-free faces and coding using just the face shape each give reasonable recognition. If these two measures are relatively independent, an appropriate combination may be more effective than either. This was investigated by applying PCA separately to the shape and shape-free images, using the 20 most variable shape components but all the texture eigenfaces derived from histogram-equalized images. Independence was assessed by measuring correlations of the ranks of the distances between each probe and the *other* images in the gallery (this helped to avoid outliers effects). The distances were pooled across the different probes, to give a single set of distances for each lighting condition. The average Spearman rank correlations for the four groups of conditions ranged from 0.10320 (for "Lighting") to 0.2049 (for "Immediate"). These differ little from the Pearson correlations, which include the magnitude of differences as well as the order. This suggests that shape and texture describe dissimilar properties; the positive correlation may reflect a tendency for faces to be extreme in both measures. Such a tendency may be an artifact; slightly inaccurate location of some landmarks, would produce distinctive, badly coded, faces on both shape (the shape would be unusual) and texture (from bad normalization).

The shape and texture distances for each probe were combined using a root mean square, having first rescaled the individual distances so the sum of each set was unity. The results shown in Table 4, Shape-and-Texture are thus comparable with our (Euclidean) baseline in Table 1, but combine locally linearized shape with (shape-free) texture information.

Obviously, the combination of incompatible types of information is arbitrary. Three other methods were used; combining at the level of eigenface spectra so that a single distance was derived on the basis of a vector of 119 values, multiplying the pairs of distances, and excluding the worse half of the shape matches before matching on texture; each gave slightly worse performance. The number and range of texture-eigenfaces included in the coding was also varied, but this showed no clear pattern; variations in hit-rates were 1-2 percent.

4.8 Shape, Texture and Shape-Free Correlation

We conclude the section with a final result in which all three matching methods, shape, texture, and shape-free correlation, are combined. Thus, we recognize by combining (again using a root mean square) the evidence which led to the "Shape-free," "Shape-only," and "Laplacian" sections of Table 1. We work with images with the hair portion removed and combine PCA-based representations derived separately from shape and shape-free texture, each matched using Mahalanobis with an optimized correlation between shape-free images preprocessed with a Laplacian filter. The results in Table 4 suggest there remains relevant information which we have been unable to code using PCA techniques; but we again emphasize that the optimized correlation takes impractically long and, unlike PCA-based methods, does not scale well for larger gallery sizes.

5 Representational Issues

5.1 Matching or Representing?

We now present comparative recognition tests in which the *shape-free advantage*, seen in Table 1, is probed a little more deeply. For example, our argument for coding faces in terms of naturally occurring facial variation by using an ensemble of *faces* was based on an implied psychological model suggesting that coding occurs *after* the faces have been recognized as such. Thus, we would hope only to get an advantage for shape-free faces when they are coded in terms of faces. This was addressed by coding also in terms of the nonface ensemble described in Section 3. Such a vocabulary has been previously considered [1], showing that the Principal Components of natural images approximate tuning curves in early vision, while showing that such tuning curves can be used quite effectively for face recognition [26], even without correction for facial shape. The nonface ensemble was processed in exactly the same way as the face ensemble; thus, any differences in coding effectiveness should arise because of the content of the ensemble, rather than for more trivial image-based reasons.

A related question is whether the shape-free advantage simply reflects superior matching of distorted images, rather than superior coding using an apposite vocabulary. These are confounded since, in Section 4, coding and normalization were performed together. To investigate further, coding and normalization procedures were contrasted by separating the processing of the ensemble from that of the probe and gallery images. Thus, in addition to tests comparing shape-free normalization with (scaled) Euclidean normalization of all the images involved, we also combine a shape-free normalization of the ensemble with a (scaled) Euclidean normalization of the gallery and probes and vice-versa. The results for this set of four tests

Fig. 7. Effects of caricaturing an image at 41, 64, 100, 156, and 244 percent.

are shown in Table 3, where they are compared with the results from the same set of four tests based on a nonface ensemble.

It is clear that, for the ensemble of faces, the determining factor is the ensemble standardization method, with improved recognition following the use of a shape-free ensemble for coding. In contrast, as would be expected, the nonface ensemble only shows an effect of the probe standardization method with a notable advantage for (scaled) Euclidean normalization. In both cases, there is evidence of an interaction between the two factors, reflecting slight differences in the pixel gray-level interpolation algorithms for transforming the images.

The dependence on the method used to standardize the ensemble in these results, suggests that the advantage for shape-free-faces reflects superior representation of the faces. The very different pattern seen when nonfaces are used as a vocabulary suggest that these latter recognition results should be thought of as picture-based, rather than face-based.

5.2 Caricaturing

It thus appears clear that there is a significant advantage in terms of overall recognition rates when the faces are represented in term of other faces. It is also worth considering whether it will respond to manipulations which alter human performance. One such uses the technique of *caricaturing*. We code face shape as a set of position vectors, each of which is the displacement of a landmark from the location of the corresponding landmark in the average face. Scaling the set of displacements uniformly by an amount k gives a caricatured shape, with $k = 100$ percent representing the veridical; a caricatured face is then created by texture mapping the face image to this shape. When shown to humans, familiar faces are recognized better with modest caricatures for line-drawings [27], [29], [28] and, also, warped gray-scale images [3], [2]. In both cases, recognition is better for caricatures of about 150 percent if a naming or identification paradigm is used. Such a paradigm is plausibly similar to the one here. Image texture can be caricatured similarly by displacing the gray levels in a shape-free face away from the mean gray-level for that pixel and yields better human recognition for 140 percent, caricatures [20]. An example in which both the shape and texture of the image have been caricatured is shown in Fig 7.

Certain techniques automatically extract caricatures; modest caricatures are extracted in a Radial Basis Function network operating on feature-distances, [4]. Typically the representations were about 110 percent caricatures, as RBFs extract the most distinctive set of features. However, it is not clear that this is the most effective caricature for recognition; we now use the control given by our explicit veridical representation to investigate this further.

Our first tests explore the existence of a caricature effect within our machine-based paradigm. Each probe was preprocessed as before to remove lighting effects; the shape was then caricatured away from the shape of the ensemble mean and the texture caricatured away from the texture mean. The results, shown in Table 4, give a strong caricature effect peaking at around 150 percent; the proportion of confident hits is also increased.

Following [21], we conjecture that this effect may simply be a consequence of our naive nearest-neighbor matching strategy and that caricaturing, in moving the representation of the probe away from the position of the mean, is more likely to move into, rather than out of, the region within which it will be correctly recognized. This effect is shown geometrically in Fig. 8 and holds under the reasonable assumption that the distribution of faces in the gallery is peaked about the mean.

Our second tests on caricaturing seek to mimic more closely the conditions under which the human caricature effect is observed. Thus, the probe itself, rather than a preprocessed version, is caricatured; and the caricature center, from which the image is displaced, is taken to be the mean of an ensemble disjoint from that used for subsequent coding and is similarly not preprocessed. This distinction is made since, for humans, we reject the assumption of a unique absolute "average" face and, thus, use different versions for the processes of caricaturing and coding. And, avoiding the preprocessing stage ensures that a 100 percent caricature is precisely the veridical, as is the case with human studies. We repeat the tests that led to Table 1, based on a (scaled) Euclidean normalization, and to Table 4, in which shape and shape-free texture are combined, using these caricatured probes. The comparison between this and the combined shape and shape-free texture case is shown in Fig. 9 for confident recognition; the overall hit-rate shows a similar, but rather less extreme, pattern. Only the shape and texture case shows a significant caricature effect, as expected if this codes facial, rather than image, variation.

Note that in passing to a more realistic comparison with the human results on the caricature effect, we have reduced the absolute recognition rate, although the effect remained.

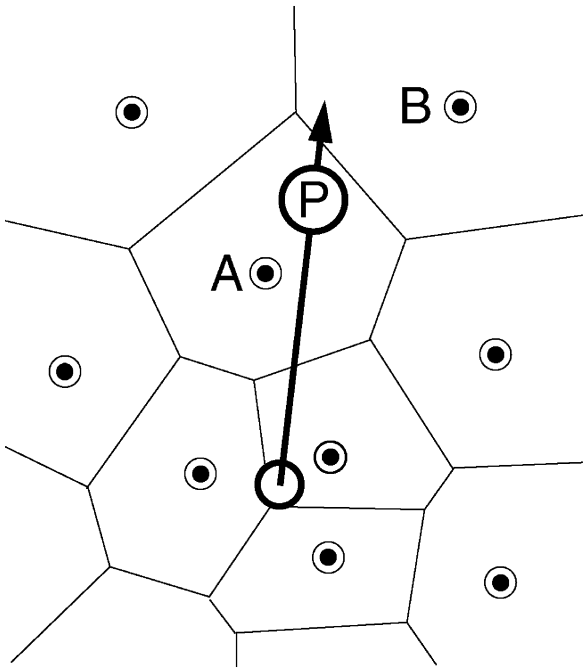

Fig. 8. Nearest neighbor matching is determined by the Voronoi tessellation shown; caricaturing the probe **P** changes the match from **A** to **B**.

6 Conclusions

This paper has explored the utility of a form of subspace coding which involves significant content-based preprocessing, dependent on the identification of the objects to be represented as belonging to a particular class. Thus, the subspace is not spanned by real images, but rather by distorted versions (shape-free faces) chosen for their expressive power in terms of intraclass discrimination.

This coding was derived from abstract mathematical principles and, in particular, a consideration of the appropriate structures to express the invariances needed for recognition purposes. This led to a manifold-based model of "face space" in which configural and textural information are separated. Although of interest theoretically, we have discussed it here purely in terms of effectiveness, comparing the performance on a fixed set of recognition tests with other codings.

During testing, we have used many images of each individual, but each separate test involves a cue, a single target, and a gallery containing a single image of each other face. In this simple form, appropriate perhaps for mugshot retrieval, there is no scope for training nor for sophisticated matching strategies based on learning, which ensures we concentrate solely on the effects of our coding methods. There is also no overlap between the ensemble and any of the images used for testing. Such an overlap would result in a perfect (to within machine accuracy) description of the corresponding image. More generally, we have avoided using the same *faces* in the ensemble as in our testing, even though this may be more appropriate if interested in, say, verification of a cue as being from a small fixed population; again, our choice allows concentration on the coding effectiveness with an "abstract" face vocabulary.

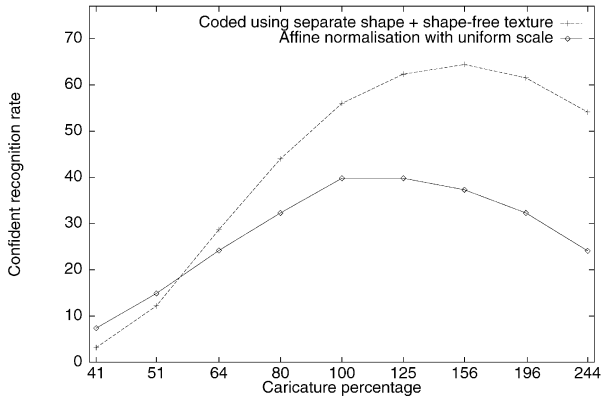

Fig. 9. Confident hit rates for shape-and-texture caricatured faces, recognized by a scaled Euclidean-normalised PCA, and by independent shape and texture PCA s Hair has been *excluded* from the match.

Our main result, that the shape-free transformation gives improved coding for recognition, is valid in this context, but a further improvement comes from adopting a preferred basis of shape-free eigenfaces for the face subspace, namely, that obtained by using PCA. The traditional advantage of PCA, allowing a maximally expressive truncated representation, only holds for images in the ensemble and this is of no direct concern here. However, the variances associated with eigenfaces allow the use of Mahalanobis distance in matching, which *does* give a worthwhile improvement in recognition rate.

Overall, the move from matching images normalized with a (scaled) Euclidean transformation to the combined configuration and texture images produces a three-fold reduction in misses without adding extra information. Rather, the configural information, which appears to overshadow the texture in the (scaled) Euclidean normalized images by requiring that facial features be approximated by a combination of different features from the different eigenfaces, has been treated separately so that it can have a positive effect. The corresponding number of clear hits has increased, but not to the same degree. Caricaturing the images can improve this by distorting them to emphasize their already atypical aspects. This may not change the ordering of matches, but it does increase the separation; thus, these two changes produce complementary improvements.

The clear advantage for Mahalanobis distance over Euclidean distance, consistent across conditions, provides evidence that PCA is a more appropriate method of coding faces than simply using raw images; and that something more sophisticated than simple template matching is occurring. Since the Mahalanobis distance aims to pay equal attention to all components, we expect no particular band of eigenfaces to best code the images; once variability is taken into account, the eigenfaces should all have the same importance. Within reasonable limits, this was found; for this reason, we have used all the eigenfaces in the tests described here.

Notably, this advantage for shape-free Principal Components remains even if the probe is not itself shape-free, reinforcing the conclusion that this is a representational

advance. Conversely, if the image has not been identified as a face (or potentially a member of some other class of objects sharing a configuration), such a manipulation is not useful. The processing and, thus, the representation are dependent upon the task at hand, changing from a general-purpose, spatial-frequency selective coding to a special-purpose, nonfrequency-selective coding as knowledge of facial shape increases. This observation is reinforced by the finding that, unlike the (scaled) Euclidean normalization, the shape and shape-free normalization allow equivalent transformations in a human face-space and that provided by the PCA.

Overall, we believe we have shown that coding using PCA, implemented under the influence of a manifold model of "face space," separating configural and textural information, has proven to be of value for recognition, and that this could be of relevance when constructing psychological models of face recognition. We are not advocating it as a universal code; the very high levels of recognition by shape-free contour matching and the increase in recognition when this is combined with the shape-and-texture output show that not all the facial information has been captured. This suggests that psychological implications of this work are late in the processing chain, when the face is being considered as a whole. One possible model has independent shape and texture derived from our local chart to select a small group of possible matches, with an ultimate recognition decision based on contour correlation.

ACKNOWLEDGMENTS

This work was in part supported by EPSRC Grants Numbers GR/H75923 and GR/J04951.

REFERENCES

[1] R.J. Baddeley and P.J.B. Hancock, "A Statistical Analysis of Natural Images Matches Psychophysically Derived Orientation Tuning Curves," *Proc. Royal Soc. B,* vol. 246, pp. 219–223, Dec. 1991.

[2] P.J. Benson, "Perception and Recognition of Computer-Enhanced Facial Attributes and Abstracted Prototypes," PhD thesis, Dept.-ment of Psychlogy, Univ. of St. Andrews, 1992.

[3] P.J. Benson and D.I. Perrett, "Visual Processing of Facial Distinctiveness," *Perception,* vol. 23, pp. 75–93, 1994.

[4] R. Brunelli and T. Poggio, "Caricatural Effects in Automated Face Perception," *Biological Cybernetics,* vol. 69, pp. 235–241, 1993.

[5] R. Brunelli and Tomaso Poggio, "Face Recognition: Features Versus Templates," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 15, no. 10, pp. 1,042–1,052, Oct. 1993.

[6] C.-Seok Choi, T. Okazaki, H. Harashima, and T. Takebe, "Basis Generation and Description of Facial Images Using Principal Component Analysis," Technical Report of IPSJ:Graphics & CAD, vol. 46, pp. 43–50, 1990. available in Japanese.

[7] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active Shape Models—Their Training and Application, *Computer Vision and Image Understanding,* vol. 61, pp. 38–59, Jan. 1995.

[8] N.P. Costen, I.G. Craw, G.J. Robertson, and S. Akamatsu, "Automatic Face Recognition: What Representation?" *Proc. of European Conf. Vision ECCV'96,* vol. l, B. Buxton and R. Cipolla, eds. pp. 504–513, 1996.

[9] P. Costen, M. Parker, and I. Craw, "Spatial Content and Spatial Quantisation Effects in Face Recognition," *Perception,* vol. 23, pp. 129–146, 1994.

[10] I.G. Craw, "A Manifold Model of Face and Object Recognition," *Cognitive and Computational Aspects of Face Recognition,* T.R. Valentine, ed. chapter 9, pp. 183–203, London: Routledge, 1995.

[11] I. Craw and P. Cameron, "Face Recognition by Computer," *British Machine Vision Conf.,* pp. 498–507, 1992.

[12] K.A. Deffenbacher, D.P. Huff, C. Hendrickson, A. O'Toole, and H. Abdi, "Manipulating Face Gender: Effects on Categorization and Recognition Judgements," presented at *Psychomics,* 1995.

[13] S. Edelman, D. Reisfield, and Y. Yeshurun, "Learning to Recognize Faces from Examples," *Proc. European Conf. Computer Vision ECCV-92,* pp. 787–791, 1992.

[14] P.J.B. Hancock, R.J. Baddeley, and L.S. Smith, "Principal Components of Natural Images," *Network,* vol. 3, pp. 61–70, Feb. 1992.

[15] P.J.B. Hancock, M.A. Burton, and V. Bruce, "Face Processing: Human Perception and Principal Components Analysis," *Memory and Cognition,* vol. 24, no. 1, pp. 26–40, Jan. 1996.

[16] M. Kirby and L. Sirovich, "Application of the Karhunen-Loève Procedure for the Characterization of Human Faces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 12, pp. 103–108, 1990.

[17] T. Kohonen, E. Oja, and P. Lehtiö, "Storage and Processing of Information in Distributed Associative Memory Systems," *Parallel Models of Associative Memory,* G. Hinton and J. Anderson, eds. chapter 4, Hillsdale, N.J.: Erlbaum, 1981.

[18] M. Lades, C. Vorbrüggen, J. Buchmann, J. Lange, C. v.d. Malsburg, R.P. Würtz, and Wolfgang Konen, . "Distortion Invariant Object Recognition in the Dynamic Link Architecture," *IEEE Trans. Computers,* vol. 42, pp. 300–311, Mar. 1993.

[19] A. Lanitis, C.J. Taylor, and T.F. Cootes, "An Automatic Face Identification System Using Flexible Appearance Models," *British Machine Vision Conf.,* pp. 65–74, 1994.

[20] K.J. Lee and D. Perrett, "Presentation Time Measures of the Effects of Manipulations in Color Space on Discrimination of Famous Faces," submitted to *Perception,* Dec. 1996.

[21] M.B. Lewis, "The Spitting Image of an Exemplar-Based Model of Face Recognition," paper presented to the *Experimental Psychology Soc.,* Apr. 1996.

[22] A.J. O'Toole, H. Abdi, K.A. Deffenbacher, and D. Valentin, "Low-Dimensional Representation of Faces in Higher Dimensions of the Face Space," *J. Optical Soc. Am.,* vol. 10, no. 3, pp. 405–411, Mar. 1993.

[23] A.J. O'Toole, K.A. Deffenbacher, H. Abdi, and J.C. Bartlett, "Simulating the "Other-Race Effect," as a Problem in Perceptual Learning," *Connection Science,* vol. 3, no. 2, pp. 163–178, 1991.

[24] A.J. O'Toole, K.A. Deffenbacher, Dominique V., and H. Abdi, "Structural Aspects of Face Recognition and the Other Race Effect," *Memory and Cognition,* vol. 22, no. 2, pp. 208–224, 1994.

[25] A. Pentland, B. Moghaddam, and T. Starner, "View-Based and Modular Eigenspace for Face Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 84–91, 1994.

[26] R.P.N. Rao and D.H. Ballard, "Natural Basis Functions and Topographic Memory for Face Recognition, " *Proc. of Int'l Joint Conf. Artificial Intelligence,* pp. 10–17, 1995.

[27] G. Rhodes, S.E. Brennan, and S. Carey, "Identification and Ratings of Caricatures: Implications for Mental Representations of Faces," *Cognitive Psychology,* vol. 19, 473–497, 1987.

[28] G. Rhodes and I.G. McLean, "Distinctiveness and Expertise Effects with Homogeneous Stimuli: Towards a Model of Configural Coding," *Perception,* vol. 19, pp. 773–794, 1990.

[29] G. Rhodes, S. Brake, K. Taylor, and S. Tan, "Expertise and Configural Coding in Face Recognition," *British J. Psychology,* vol. 80, no. 3, pp. 313–331, Aug. 1989.

[30] G. Robertson and I. Craw, "Testing Face Recognition Systems," *Image and Vision Computing,* vol. 12, pp. 609–614, 1994.

[31] M.A. Shackleton and W.J. Welsh, "Classification of Facial Features for Recognition," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 573–579, 1991.

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

[33] S. Ullman, "Aligning Pictorial Descriptions: An Approach to Object Recognition," *Cognition,* vol. 32, pp. 193–254, 1989.

[34] T. Valentine, "A Unified Account of the Effects of Distinctiveness, Inversion and Race in Face Recognition," *Quarterly J. Experimental Psychology,* vol. 43A, pp. 161–204, 1991.

[35] T. Vetter and T. Poggio, "Linear Object Classes and Image Synthesis from a Single Example Image," *Proc. of European Conf. Computer Vision ECCV-96,* B. Buxton and R. Cipolla, eds. vol. 1, pp. 652–659, 1996.

[36] J.R. Vokey and J.D. Read, "Familiarity, Memorability, and the Effect of Typicality on the Recognition of Faces," *Memory and Cognition,* vol. 20, no. 3, pp. 291–302, 1992.

[37] R. Watt, "A Computational Examination of Image Segmentation and the Initial Stages of Human Vision," *Perception,* vol. 23, pp. 383–398, 1994.

Ian Craw read Mathematics at Cambridge as an undergraduate and stayed on to take a PhD in functional analysis in 1969. He joined the University of Aberdeen, Scotland, where he still teaches mathematics. His interest moved to computer vision in the early 1980s and, shortly, thereafter, he started working on faces. He was an invited visiting researcher at ATR, Kyoto, for seven months in 1994-1995.

Nick Costen received a BA in experimental psychology from the University of Oxford before moving to Aberdeen, where he obtained a PhD in 1994. He spent the period 1994-1997 at ATR in Japan before moving to the Wolfson Image Analysis Unit at the University of Manchester.

Takashi Kato holds a PhD in cognitive psychology from the University of California, Los Angeles. He is currently a professor of cognitive science at Kansai University, Japan, and a visiting researcher at ATR. Previously, he managed cognitive engineering research at the IBM Tokyo Research Lab and taught at the University of Sydney. His current research interests include face processing, implicit learning, and humancomputer interaction.

Shigeru Akamatsu received the BE, ME, and PhD Eng degrees in mathematical engineering and instrumentation physics from the University of Tokyo, Japan. He is currently head of a department at ATR Human Information Processing Research Laboratories and a guest professor at Tokyo Institute of Technology. Before joining ATR in 1992, he was a senior research engineer and supervisor at NTT Human Interface Laboratories, where he conducted research on human image recognition and generation. His current research interests include computational and cognitive studies of high level vision, with a special focus on facial information processing by man and computer.