Automatic Face Recognition: What Representation?

Nicholas Costen[1], Ian Craw[2] *, Graham Robertson[2] and Shigeru Akamatsu[1]

[1] ATR Human Information Processing Research Laboratories,
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02, Japan.
[2] Department of Mathematical Sciences, University of Aberdeen,
Aberdeen AB9 2TY, Scotland.

Abstract. A testbed for automatic face recognition shows an eigenface coding of shape-free texture, with manually coded landmarks, was more effective than correctly shaped faces, being dependent upon high-quality representation of the facial variation by a shape-free ensemble. Configuration also allowed recognition, these measures combine to improve performance and allowed automatic measurement of the face-shape. Caricaturing further increased performance. Correlation of contours of shape-free images also increased recognition, suggesting extra information was available. A natural model considers faces as in a manifold, linearly approximated by the two factors, with a separate system for local features.

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 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, such as changes in angle of presentation and lighting, common problems of machine vision, and changes also of expression and age which are more special. 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 match, before the faces are declared the same.

A number of face recognition systems have become available recently which propose solutions to these problems, and a natural concern has been the systems's overall performance [13, 7, 10, 3, 12, 11]. Although the choice of coding and matching strategies differ significantly, the greatest source of variability is probably the selection of the faces to test, and the choice of transformation between target and probe over which the system performs recognition. The FERRET database, a potential standard, is currently only available within the USA.

In this paper we seek to avoid some of these difficulties by fixing a matching strategy and testing regime, and concentrating on the first of these problems; to find effective codes for recognition. Our concern is then no longer how well we

* This work was in part supported by EPSRC (GR/H75923 and GR/J04951 to IC).

can recognise; indeed for our purposes, a testing regime with a low recognition rate is of most interest: our interest is in *comparing* different coding strategies.

2 Coding via Principal Component Analysis

We contrast simple image-based codings with *eigenface* codings, derived from Principal Component Analysis. Eigenface codings were used to demonstrate pattern completion in a net based context [9, Page 124], to represent faces economically [8], and explicitly for recognition [13]. Much subsequent work has been based on eigenfaces, either directly, or after preprocessing [5, 12, 11].

While undoubtedly successful in some circumstances, the theoretical foundation for the use of eigenfaces is less clear. Formally, Principal Component Analysis assumes that face images, usually normalised in some way, such as co-locating eyes, are usefully considered as (raster) vectors. A given set of such faces (an *ensemble*), is then analysed to find the "best" ordered basis for their span. Some psychological theories of face recognition start from such a norm-based coding; an appropriate model may be a "face manifold" [5], and the usual normalisation 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 normalised faces should also be a face.

Clearly existing normalisation techniques approximate this property, but a more elaborate one [14] has recently become prominent as the way to perform a "morph" between two faces. Landmarks give a description of each face's shape; there is a natural way to average landmark positions, and then to map the face texture onto the resulting shape. We describe this as a decomposition into shape *configuration* and shape-free *texture* vectors. The main aim of our paper is to show that this coding produces significantly better recognition results.

Our methodology starts with face images on which a collection of landmarks have been located. Our first tests use manual location; automatic location, and the corresponding results are discussed in Section 4. Eigenfaces are computed from an ensemble of faces which have no further rôle; the gallery and probe faces are coded in these terms. Each probe face has one other image in the gallery, the *target*; our interest is in when the target best matches the probe.

Fourteen images of each of 27 people were acquired under fairly standardised conditions, referred to as Conditions 1 to 14. An initial set of 10 images was acquired on a single occasion: those in Conditions 1 to 4 were lit with good flat controlled lighting; later conditions have increasingly severe lighting and pose variations. Four images were acquired between one and eight weeks later: the first, Condition 11, in lighting conditions similar to Condition 1; subsequent ones with increasing differences. Condition 14 is the only image lit with a significant amount of natural, uncontrolled, light. Fig. 1 shows some of the variability.

The 27 images in Condition 1 provide our fixed gallery. The remaining 13 conditions provide 351 probes. Using all faces as probes avoids that differences in ease of recognition. An additional 50 faces were collected only in Condition 1 and are used as ensemble images, from which the eigenfaces are generated.

Fig. 1. Conditions 1,4,8,12 and 14.

All the images are processed in the same way. Thirty-four landmarks are found manually, giving a triangulation, or face *model*, part of which can be seen in Fig. 2. A (uniformly) scaled Euclidean transformation is applied, minimising the error between these and corresponding points on a reference face, giving *normalised* images. The background is removed and the histogram of the remaining pixel values flattened. When the face image includes the hair, featural information in the hair can allow good short term recognition. To avoid this, the face model is used to extract an image, containing "inner features" only, as in Fig. 2. Essentially all the results reported are for such images of 2557 pixels.

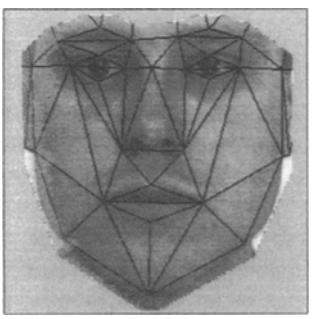

Fig. 2. Inner face showing the facial locations. The mask is enlarged to show the points.

A Principal Component Analysis is performed on the resulting ensemble, obtaining eigenvalues and unit eigenvectors (or *eigenfaces*) of the image cross-correlation matrix. The orthonormality of the eigenfaces allows the computation of the weight of any (normalised) face on each eigenface, giving an n-tuple or code. A coded probe image and the gallery codes are then compared. One method uses nearest neighbour matching in the ensemble span, and a natural metric is the Euclidean distance, leading to template matching within the span. Another natural choice is the Mahalanobis distance, where $d(\mathbf{x},\mathbf{y})^2 = \sum \lambda_i^{-1}(x_i - y_i)^2$, where $\{\lambda_i\}$ is the sequence of eigenvalues. This treats variations on each axis as

equally significant, arguably better for discrimination.

A more robust scheme uses match strength to reduce false acceptances. One such has a sequence of match scores $\{c_j\}$ between the gallery images and the probe[10]. The best match gives the lowest score, c_0; the next c_1. The mean μ and standard deviation σ of the sequence excluding c_0 are calculated and define two inequalities, $c_0 < c_1 - t_1\sigma$ and $c_0 < \mu - t_1\sigma$, for fixed thresholds t_1 and t_2, which must both be met to accept a match. A correct match is reported as a *clear* hit if this criterion is met, and *just* a hit otherwise, similarly for misses. The distances between the Condition 2 probes and the gallery (minus the target) set t_1 and t_2 which were calculated for each probe and the largest values independently chosen. This ensured that in the best, base, condition there were no "clear misses"; although conservative, there are cases where these do occur.

3 Results

We group Conditions 2, 3 and 4, describing this as "Immediate" recognition. Conditions 5, 6 and 7 form a similar set, called "Variant", with small changes in lighting and position. More fundamental lighting changes distinguish the "Lighting" group, Conditions 8, 9 and 10. Finally the four conditions with delayed image acquisition, are called "Later". A weighted average gives the "Overall" value; since the latter conditions are more important, the "Lighting" group has twice the weight of "Immediate" or "Variant", and "Later" four times the weight.

Our main interest is the comparison between scaled Euclidean normalisation, and the more intrusive shape-free form; and the contrast between these and a pure correlation approach. Initial testing used the ensemble of 50 faces described above. However, using the approximate vertical symmetry in individual faces by creating 50 "mirror" faces, reflected about the vertical facial mid line [8] gave a noticeable improvement in recognition, and all results use this "doubled" ensemble. Table 1, Method 'Mah' gives results against which subsequent performance is compared, obtained using all 99 eigenfaces from this ensemble.

	Hit						Miss					
	Clear			Just			Just			Clear		
Method:	Mah	Euc	Cor	Mah	Euc	Cor	Mah	Euc	Cor	Mah	Euc	Cor
Immediate:	90.1	82.7	31.3	9.9	14.8	7.4	0.0	2.5	1.2	0.0	0.0	0.0
Variant:	67.9	34.6	55.6	22.2	45.7	35.8	9.9	19.5	8.6	0.0	0.0	0.0
Lighting:	17.3	3.7	11.1	48.1	29.6	42.0	34.6	66.7	46.9	0.0	0.0	0.0
Later:	34.3	16.7	23.1	32.4	28.7	40.7	33.3	54.9	36.1	0.0	0.0	0.0
Overall:	40.2	22.9	31.3	32.3	29.2	36.9	27.5	47.9	31.7	0.0	0.0	0.0

Table 1. Match percentages from 351 trials per method. Scaled Euclidean normalised in all cases . Method 'Mah': matching with Mahalanobis distance. Method 'Euc': matching with Euclidean distance. Method 'Cor': matching by correlation of the images. Hair has been *excluded* from the match.

Our first comparisons are between the Mahalanobis distance and identical tests using Euclidean distance or thirdly a correlation of the whole of the (masked) face images, all shown in Table 1. The Mahalanobis distance is clearly most effective, confirming that the eigenface formulation, with its variance properties, is worthwhile here. The advantage is smallest in the "Immediate" group, where simple template matching is expected to perform well; but even here, the effect on the "Clear Hits" is noticeable. The alternative baseline uses the whole of the relevant image information, including that lost when the images are projected onto the span of the ensemble. It is clear that this projection looses significant information. Matching using the Mahalanobis distance more than makes up for this loss and we thus adopt this as our baseline.

The theoretical considerations in Section 2 suggest that the distortion of a face into a shape-free or texture vector may provide more effective coding. The normalisation texture-maps each face to a standard shape, here the average of the ensemble images. We used linear interpolation based on the model in Fig. 2; although simpler than Bookstein's thin plate spline warps [11], the procedure was more effective. The results given in Table 2, Method 'T', are directly comparable to Table 1 and suggest that shape-free normalisation is slightly *better* than the scaled Euclidean version, despite deliberately ignoring the shape information.

	Hit						Miss					
	Clear			Just			Just			Clear		
Method:	T	S	S–T	T	S	S–T	T	S	S–T	T	S	S–T
Immediate:	95.1	39.5	90.1	4.9	46.9	9.9	0.0	13.6	0.0	0.0	0.0	0.0
Variant:	64.2	23.5	71.6	29.6	58.0	25.9	6.2	17.3	2.5	0.0	1.2	0.0
Lighting:	18.0	27.2	40.7	51.9	54.3	50.6	29.6	18.5	8.6	0.0	0.0	0.0
Later:	28.7	19.4	42.6	46.5	59.3	46.3	25.0	21.3	11.1	0.0	0.0	0.0
Overall:	28.7	23.7	50.4	41.3	56.7	41.1	21.3	19.4	8.5	0.0	0.1	0.0

Table 2. Match percentages from 351 trials per method. Matching with Mahalanobis distance in all cases . Method 'T': using shape-free texture. Method 'S': using shape or configuration (20 most variable eigenshapes). Method 'S–T': combining shape and texture. Hair has been *excluded* from the match.

The shape-free advantage may reflect superior matching of the distorted images, rather than superior coding. A shape-free normalisation was used on the ensemble and a scaled Euclidean normalisation on the gallery and probes, and *vice versa*. The results in Fig. 3 show that the determining factor is the ensemble standardisation method, suggesting the advantage for shape-free-faces reflects superior representation of the faces, not just better matching.

The data discarded by shape-free normalisation can also be used for recognition, performing Principal Component Analysis on the landmark locations. This was done as already described, applying a scaled Euclidean transformation to remove position effects, and then, if necessary, removing the points relating to the hair. The shapes of the ensemble images then provided suitable principal

components (*eigenshapes*). The data are highly correlated; after the first 15 or 20 eigenshapes the eigenvalues become small. The number of principal components used to code the shape was varied and the hit rates are shown in Fig. 4. Both with and without hair, recognition peaks when 20 components are included; the peak results for the configuration without hair are given in Table 2, Method 'S'.

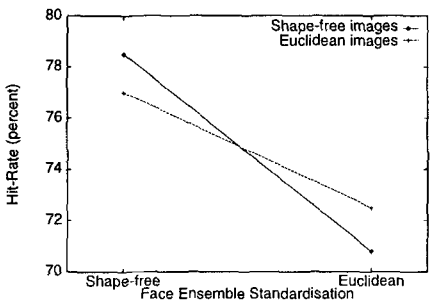

Fig. 3. Recognition using different normalisations for ensemble and test images: hit rates for Euclidean and shape-free.

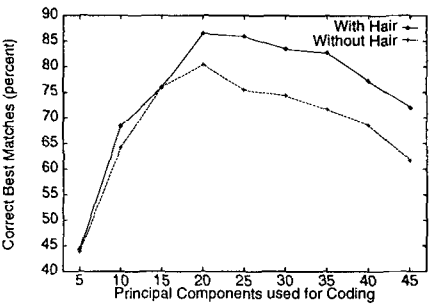

Fig. 4. Recognition using shape or configuration: hit rates for variable numbers of initial principal components.

These tests show a real advantage in representing faces by shape-free texture. This may extend to matches between the images themselves. Because the normalisation uses a relatively small number of points, the match may be underestimated; to compensate, the correlation between a probe and each gallery image was optimized by varying a scaled Euclidean transform of the normalised probe. There was a very noticeable advantage for preprocessing using a laplacian transformation, a 3×3 matrix often thought of as a sharpening operator. The results for the shape-free laplacian images are given in Table 3, showing very good and constant recognition. However, this is very slow even with the small gallery here; optimizing the match required comparing each image with each gallery member 50 times. These results again show the advantages of a shape-free representation; it ensures that all sections of the laplacian-processed images can be aligned at once. In contrast, when shape is still in the images, different sections of the probe face compete to match sections of the gallery images.

Coding using either texture or face shape gives reasonable recognition. If these measures are relatively independent, a combination may be effective. Principal Component Analysis was performed separately on the shape and shape-free images. Independence was assessed by rank correlations of distances between each probe and the *other* gallery images (reducing outlier effects). The average Spearman rank correlations are positive but modest with a maximum value (for the "Immediate" images) of 0.267. This suggests that shape and texture describe dis-similar properties; the positive correlation may reflect landmark location errors. The shape and texture distances for each probe were combined using a root

	Hit						Miss					
	Clear			Just			Just			Clear		
Method:	Cor	Car	Com	Cor	Car	Com	Cor	Car	Com	Cor	Car	Com
Immediate:	85.1	95.1	97.5	14.8	4.9	2.5	0.0	0.0	0.0	0.0	0.0	0.0
Variant:	67.9	80.2	88.9	30.9	18.5	11.1	1.2	1.2	0.0	0.0	0.0	0.0
Lighting:	38.3	48.1	67.9	50.6	46.9	29.6	11.1	4.9	2.5	0.0	0.0	0.0
Later:	28.7	58.3	66.7	62.0	34.3	32.4	9.3	7.4	0.9	0.0	0.0	0.0
Overall:	41.0	62.4	72.6	51.2	32.1	26.3	7.8	5.4	1.1	0.0	0.0	0.0

Table 3. Match percentages from 351 trials per method. Method 'Cor': shape-free normalised, matching by full correlation of laplacian images. Method 'Car': shape and texture, matching with Mahalanobis distance with the images 156 % caricatures. Method 'Com': shape and texture, matching with Mahalanobis distance combined with full correlation of laplacian images. Hair has been *excluded* from the match.

mean square, after rescaling the individual distances so the sum of each set was unity. The results in Table 2, Method 'S–T', are thus comparable with Table 1, but combine locally linearised shape with (shape-free) texture information.

The distinct shape and texture components of the face allow it to be *caricatured*. Face shape is coded as a set of position vectors, each the displacement of a landmark from its position in the average face. Scaling the displacements by an amount k gives a caricatured shape, with $k = 100\%$ representing the veridical; the face image is then texture-mapped to this shape. In humans, familiar faces are recognised better with modest caricatures (about 110 %) [1]. Image texture can also be caricatured by displacing the grey levels in a shape-free face away from the mean for each pixel; an example is shown in Fig. 5. Similar modest caricatures are extracted by a Radial Basis Function network using feature-distances [2], as RBFs extract distinctive sets of features. However, a Principal Components Analysis technique, with veridical coding, has greater freedom as it allows investigation of the coding giving the most effective caricatures.

Fig. 5. An image caricatured on shape and texture at 41, 64, 100, 156 and 244 percent.

The faces were caricatured on shape and texture before recognition using the inner face and deriving t_1 and t_2 from the veridical images. The tests show the effects of recognizing the images with a scaled Euclidean normalised Principal Components Analysis. This yields the notably small caricature effect shown by

Fig. 6, while independent shape and shape-free Principal Components Analysis, as shown in Fig. 7, give a strong effect with peak recognition at about 150 %. The peak recognition rates are shown in Table 3. This difference in the caricature effect is only seen if the images are caricatured against independent shape and texture averages. Caricaturing images against the average of the Principal Components Analysis, regardless of the type of normalisation used, gives approximately equal effects. The advantage of the shape-free manipulation is that it allows equivalent transformations in both image-space (as evidenced by the human data) and also in the Principal Components Analysis linearisation.

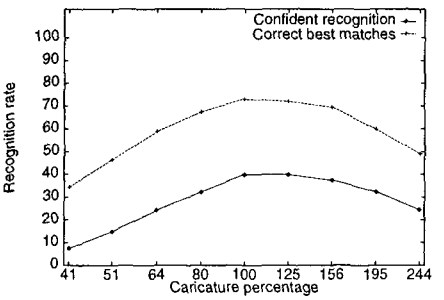

Fig. 6. Confident and total hit rates for shape-and-texture caricatured faces, recognised as Euclidean-normalised. Hair has been *excluded* from the match.

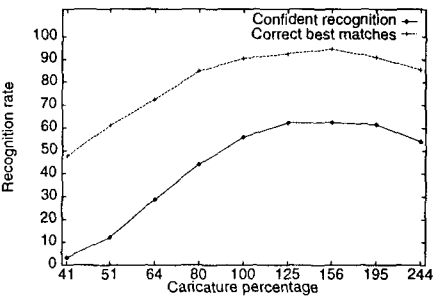

Fig. 7. Confident and total hit rates for shape-and-texture caricatured faces, recognised as separate shape and texture. Hair has been *excluded* from the match.

In a final result, all three matching methods; shape, texture and shape-free correlation, are combined using a root mean square. The results, in Table 3, Method 'Com', suggest there remains relevant information which has not been coded using caricature techniques; but we again emphasize that the optimized correlation takes impractically long, and does not scale well for larger galleries.

4 Facial shape-finding

If our coding process is to operate automatically the landmarks must be located. Given the location of enough landmarks to provide a scaled Euclidean normalisation of a new face, we sequentially generate refined shape estimates. A development of an earlier program, FindFace [6], provides the initial locations, and initialises a bootstrapping procedure to locate the remainder given the ensemble. Each set of landmark locations on a face defines a corresponding shape-free face; we choose those locations on our new face for which the shape-free version has the highest correlation with the average shape-free face.

To optimize efficiently needs the Principal Components Analysis orthogonal decomposition of face shape. Fitting these components successively gives an effective means of navigating in shape space. Starting with the average model,

new models were built by varying the shape on the first Principal Component over a range of up to two standard deviations, so applying an active shape model [4]. The resulting model was used to distort the probe to shape-free form; this was then correlated with the shape-free average texture to measure the appropriateness of the model. A simple hill-climbing algorithm sequentially derived the 20 most variable component. The fitting was performed upon the whole, masked, face, including the hair; this gave the most accurate and consistent point-definitions.

When the points so found were used as the input to the complete system, including a caricature of 156 %, it gave the values shown in Table 4. There was a significant caricature effect, suggesting location consistency on the same face; the recognition rate for veridical images was 65.7 %, with 34.7 % clear hits.

	Hit		Miss	
	Clear	Just	Just	Clear
Immediate:	86.4	12.3	1.2	0.0
Variant:	67.9	24.7	7.4	0.0
Lighting:	37.0	34.6	24.2	1.2
Later:	30.6	37.0	32.4	0.0
Overall:	41.9	32.5	25.3	0.3

Table 4. Match percentages from 351 trials. Automatic shape and texture, matching with Mahalanobis distance on 156 % caricatures. Hair has been *excluded* from the match

5 Conclusions

We have attempted to show that a greater consideration of the nature of Principal Component Analysis yields advantages in recognition. Doing so, moving from scaled Euclidean normalised images to the combined configuration and texture images reduces misses three-fold without adding extra information. Caricaturing the images can improve this, by distorting them to emphasize their already atypical aspects. This may not change the ordering of matches, but does increase the separation. This advantage for shape-free Principal Components remains even if the probe is not itself shape-free, again suggesting that this is a representational advance. This decomposition of the face into configuration and texture, and then into Principal Components also allows the efficient location of facial features.

The clear advantage for Mahalanobis over Euclidean distance provides evidence that Principal Component Analysis is a more appropriate coding of faces than raw images; and that something more sophisticated than simple template matching is occurring. Since the Mahalanobis distance pays equal attention to all components, no particular band of eigenfaces should best code the images; once variability is accounted for, the eigenfaces should be equally important. Within limits, this was found; thus we used all the eigenfaces in the tests described here.

Overall we believe we have shown that Principal Component Analysis, implemented under the influence of a manifold model of "face space", separating configural and textural information, has proved of value in coding for recognition; this could be of relevance when constructing psychological models of face recognition. We do not advocate it as a universal code; the observations of very high levels of recognition with shape-free contour matching and 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 model selects a small group of possible matches with local chart-based shape and texture, and uses contour correlation for the final decision.

References

1. Benson, P. and Perrett, D.: 1994, Visual processing of facial distinctiveness, *Perception* **23**, 75–93.
2. Brunelli, R. and Poggio, T.: 1993a, Caricatural effects in automated face perception, *Biological Cybernetics* **69**, 235–241.
3. Brunelli, R. and Poggio, T.: 1993b, Face Recognition: Features versus Templates, *IEEE: Transactions on Pattern Analysis and Machine Intelligence* **15**, 1042–1052.
4. Cootes, T., Taylor, C., Cooper, D. and Graham, J.: 1995, Active shape models – their training and application, *Comp. Vis. and Image Understanding* **61**, 38–59.
5. Craw, I.: 1995, A manifold model of face and object recognition, *in* T. Valentine (ed.), *Cognitive and Computational Aspects of Face Recognition*, Routledge, London, chapter 9, pp. 183–203.
6. Craw, I., Tock, D. and Bennett, A.: 1992, Finding face features, *Proceedings of ECCV-92*, pp. 92–96.
7. Edelman, S., Reisfield, D. and Yeshurun, Y.: 1992, Learning to recognise faces from examples, *Proceedings of ECCV-92*, pp. 787–791.
8. Kirby, M. and Sirovich, L.: 1990, Application of the Karhunen-Loève procedure for the characterisation of human faces, *IEEE: Transactions on Pattern Analysis and Machine Intelligence* **12**, 103–108.
9. Kohonen, T., Oja, E. and Lehtiö, P.: 1981, Storage and processing of information in distributed associative memory systems, *in* G. Hinton and J. Anderson (eds), *Parallel models of associative memory*, Erlbaum, Hillsdale N.J., chapter 4.
10. Lades, M., Vorbrüggen, J., Buchmann, J., Lange, J., v. d. Malsburg, C., Würtz, R. and Konen, W.: 1993, Distortion invariant object recognition in the dynamic link architecture, *IEEE Transactions on Computers* **42**, 300–311.
11. Lanitis, A., Taylor, C. and Cootes, T.: 1994, An automatic face identification system using flexible appearance models, *BMVC 1994*, pp. 65–74.
12. Pentland, A., Moghaddam, B. and Starner, T.: 1994, View-based and modular eigenspace for face recognition, *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 84–91.
13. Turk, M. and Pentland, A.: 1991, Eigenfaces for recognition, *Journal of Cognitive Neuroscience* **3**, 71–86.
14. Ullman, S.: 1989, Aligning pictorial descriptions: An approach to object recognition, *Cognition* **32**, 193–254.