Internal Use Only
非公開

TR - H - 124 0020

Automatic Face Recognition: Combining Configuration and Texture

Ian Craw *Nicholas Costen*
Takashi Kato *Graham Robertson*
(University of Aberdeen)

1995. 1. 31

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2 ☎ 0774-95-1011

ATR Human Information Processing Research Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan
Telephone: +81-774-95-1011
Facsimile: +81-774-95-1008

Automatic Face Recognition: Combining Configuration and Texture

Ian Craw Nicholas Costen Takashi Kato
Graham Robertson*
ATR Human Information Processing Lab†

Abstract

We describe in detail a baseline set of procedures for face recognition using Principal Component Analysis. Results are obtained for cues faces in 13 different conditions, with varying lighting and interval between gallery and cue acquisition. We suggest potential improvements to the coding and show how they may be investigated using the baseline setup. Tests show that paying detailed attention to the configuration of faces, and separating the shape and texture into separate vectors significantly improves recognition and allows the modeling of human face recognition phenomena. A theoretical account in terms of a manifold model of facial variation is proposed.

1 Aims

In machine based face recognition, a *gallery* of faces is first enroled in the system and coded for subsequent searching. A *cue* 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, 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 shall match, before the faces are declared to be 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 (Turk and Pentland 1991, Edelman, Reisfield and Yeshurun 1992, Lades, Vorbrüggen, Buchmann, Lange, v.d. Malsburg, Würtz and Konen 1993, Brunelli and Poggio 1993, Pentland, Moghaddam and Starner 1994, Lanitis, Taylor and Cootes 1994). Accordingly, test sets have been constructed and a recognition accuracy computed. In practice, published recognition results are usually very good, but are notoriously difficult to compare (Robertson and Craw 1994). Although the choice of coding and matching strategies differ

*University of Aberdeen

†Much of this work was done between July 1994 and January 1995 when Ian Craw was an Invited Researcher at, and funded by, ATR (Advanced Telecommunications Research International), Kyoto, and on Research Leave from the University of Aberdeen, Scotland. Parts were also funded by the UK Science and Engineering Research Council (now EPSRC or BBSRC) grant GR/H75932 to Ian Craw supporting Graham Robertson, and grant GR/JO4951 to Mike Burton, Vicki Bruce and Ian Craw. Nicholas Costen is normally emplyed on this latter grant, but was funded by ATR from October 1994 to January 1995.

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 cue over which the system is supposed to perform recognition.

In this paper 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 recognise; 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.

1.1 This Technical Report

In this report we try to give full details of testing done in the period October 1994 to January 1995 at ATR. The content in part supersedes that of (Craw, Kato, Costen and Robertson 1994b), although that report does give more details of potential coding improvements which are not discussed here. It also extends the work reported in (Craw, Costen, Kato, Robertson and Akamatsu 1994a), and contains more reliable and recent testing than does that submission.

2 Existing Face Recognition Results

A number of early studies on face recognition culminated in the book of Kanade (1977) which described a sequential system in which eye, nose and mouth locations were sought on more than 500 images. Although primarily designed as a system to locate features, the resulting locations were used as a crude recognition system. For its time it was remarkably thorough, and seems to have been the only work of substance done during that decade. Another early attempt it recognition, or at least face matching, was the FRAME database (Shepherd 1986) in which witnesses endeavored to retrieve mugshots interactively using subjective descriptions of the face. The mugshots themselves were coded, in part, on shape measurements. It was necessary to investigate obtaining these automatically (Craw, Ellis and Lishman 1987), and following Kelly (1971) this used a multiresolution approach, but still had a sequential control structure.

Feature-based approaches to recognition have continued to attract interest, and recent work includes an extension of the earlier FRAME-based work (Craw, Tock and Bennett 1992), and more hybrid systems which use features during the encoding stage (Edelman et al. 1992, Brunelli and Poggio 1993).

There were also early net-based approaches to recognition, including Stonham (1986), and Kohonen, Oja and Lehtiö (1981). Although these were demonstrated effectively, there appears to have been no extended study of their properties on large populations of faces, while issues associated with the number of images needed to train the system may also limit their utility.

The use of more general holistic features originated with Sirovich and Kirby (1987) and Kirby and Sirovich (1990) who were interested in representing faces economically, and Turk and Pentland (1991) who used an "eigenface" coding for recognition. Much subsequent work has been based on eigenfaces, either directly, or after preprocessing (Craw and Cameron 1992, Shackleton and Welsh 1991, Pentland et al. 1994). Lanitis et al. (1994) have a slightly different approach, in that they are modifying a purely shape based system to use grey-level information as well. They deal separately with shape and texture, but both are coded using a form of Principal Component Analysis. Finally a successful recognition system which matches on distortion and grey level is Lades et al. (1993). Each object is coded as a graph derived from its image. A grid is placed roughly on the image; the vertices of the graph are then filtered (using Gabor wavelets) versions of the image at each grid point, and the edge vectors describe the link to (some) adjacent vertices. Matching is elastic graph matching between cue and each gallery member, and involves optimization of a (simple) matching costs function. Performance is assessed matching (naturally) distorted face images of 87 people.

3 PCA - based recognition.

In this paper we consider only *eigenface* codings, derived from Principal Component Analysis. Such codings were used to demonstrate pattern completion in a net based context (Kohonen et al. 1981, Page 124), representing faces economically (Sirovich and Kirby 1987), and explicitly for recognition (Turk and Pentland 1991). Much subsequent work has been based on eigenfaces, either directly, or after preprocessing (Craw and Cameron 1992, Shackleton and Welsh 1991, Pentland et al. 1994, Lanitis et al. 1994).

An early use of eigenfaces was for data compression (Sirovich and Kirby 1987). All the faces to be subsequently processed were known initially and were taken as the ensemble (in our language). A complete set of eigenfaces was computed, thus providing an alternative basis for the face subspace, the linear span in $\mathbf{R}^N$ of the ensemble images. Thus each face in the ensemble can be perfectly represented, to within computational error, as a linear combination of eigenfaces. However a gain occurs when eigenfaces corresponding to small eigenvalues are ignored; the resulting lower dimensional approximation is then the best possible of that dimension; that is precisely how Principal Components are defined; and clearly this procedure makes essential use of this property of eigenfaces.

A simple recognition system can be built from this, taking the gallery and ensemble to be the same set of faces, and coding a cue in terms of the eigenfaces being used. However there seems little evidence that there is much gain in representing the gallery faces less than perfectly, and the resulting recognition procedure is then simply template matching unless some more sophisticated distance, such as the Mahalanobis distance is used. A fundamental difficulty here is the need to use the full gallery as ensemble. While this is feasible when $n = 100$ or $n = 200$, significantly larger gallerys require unreasonable computational resource to obtain the eigenvalues and eigenfaces, and a different approach is needed.

One such is that of Pentland et al. (1994), who use an ensemble of 128 faces, and code using only the top 20 such eigenfaces. The ensemble was drawn from their gallery of approximately 3000 people, and was chosen to be representative of that gallery. This approach then aims to obtain a "generic" basis for the "face-subspace", although of course there is now no guarantee that it will be an actual basis. We argue in section 11 that this is an appropriate procedure in that it is in accord with a natural model of such an object, and we shall follow it throughout, except that, to avoid confusion, our eigenfaces are computed from an ensemble of faces which have no further rôle in the process; the gallery and cue faces are then coded in terms of these.

An alternative approach uses an approximation procedure for calculating the required set of eigenfaces. Given a strategy in which the only the first few components are used for coding, this can be computationally effective, and Jungman, Levi, Aperman and Edelman (1994) describe such a system, motivated by a mugshot retrieval task They use only the first

100 components, and generate these from a collection of 1500 faces. However in practice the resulting eigenfaces are used to describe faces not in the training set; as such there is perhaps less difference than first appears between this approach an that of Pentland et al. (1994). We discuss in Section 6.1 whether such a large training set is necessary, and whether the choice of the first tranche of eigenfaces as coding vocabulary is necessarily the most appropriate; while eigenfaces are ordered precisely by their importance in terms of reconstruction of the image, truncation may be less justifiable in terms of recognition; indeed the use of the spectral band corresponding to eigenfaces 20–55 has been found to be better at capturing identity specific features of individual faces, even in the case where the gallery and ensemble are distinct (O'Toole, Deffenbacher, Valentin and Hervé 1994).

In Lanitis et al. (1994) results are presented whose motivation is very like our own. Configuration and texture are available separately, coded using Principal Component Analysis, 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.

4 Methodology

Our methodology starts with face images on which a collection of landmarks have been located manually, although automatic location is being implemented. Eigenfaces are computed from an ensemble of faces which have no further rôle in the process; the gallery and cue faces are then coded in terms of these. For a given cue, there is exactly one *target* — another image of the cue face — in the gallery, and our interest is in when the target best matches the cue.

4.1 Choice of images

We work with images of size 128 × 128, writing N for the number of pixels in each image (so initially $N = 16384$) 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 provide our test material. The pictures were taken with a colour CCD camera connected to a framegrabber which digitised 24 bit colour at a resolution of 576 × 768 pixels. In order to have control of the lighting, images were acquired in a blacked out room. A white board, illuminated by an overhead projector was behind the camera, and served to reflect an even light on the subject. This lighting was supplemented by two strip lights on the ceiling about six feet to the left and right of the subject. A desk lamp, placed on the floor, provided additional lighting to the chin area. None of the subjects had moustaches or beards, and those who wore glasses were asked to remove them.

Each of the images used either as cue, or in the gallery, was obtained as part of a series of 14 images of the subject. In addition a number of images of distinct individuals were grabbed, all in Condition 1. These provided the images which were used in the ensemble. No attempt was made to keep the images homogeneous in any other way; images were simply acquired of whoever could be persuaded to give their time. Nor was there an attempt to control for clothing during the second session; it is in general different from that worn in the first session.

The conditions are as follows.

Condition 1. The image was grabbed with all the lighting on.

Condition 2. A second image was grabbed a few seconds after image 1.

Condition 3. A third image was grabbed a few seconds after image 2.

Although these three images were grabbed in essentially the same conditions, we use them in different ways. The first set of images are used to provide images for the gallery, while the second is used to set rejection thresholds. This condition, and *all* of the remaining images are available as cues during testing.

Condition 4. The subject was asked to move around and then reseat. The image was then grabbed, with all lighting on, as in the previous three conditions.

Condition 5. The floor light was switched off and the image grabbed with all the remaining lights on.

Condition 6. A further image was grabbed a few seconds after the previous one, in the same lighting conditions.

Condition 7. The subject was again asked to move around and reseat. The image was then grabbed, with the same lighting as in the previous condition. Thus this condition bears the same relationship to Condition 6 that Condition 4 has to Conditions 1 to 3.

Condition 8. The overhead projector which was the main source of even illumination, was switched off, leaving only the two overhead strip lights on, and the image was grabbed.

Condition 9. A further image was grabbed a few seconds after the previous one, in the same lighting conditions.

Condition 10. The door of the office was opened to let in ambient light. This was done by the subject, who reseated. The final image of the first session was then grabbed.

In Fig. 1 we give examples of these conditions to indicate something of the variability involved.

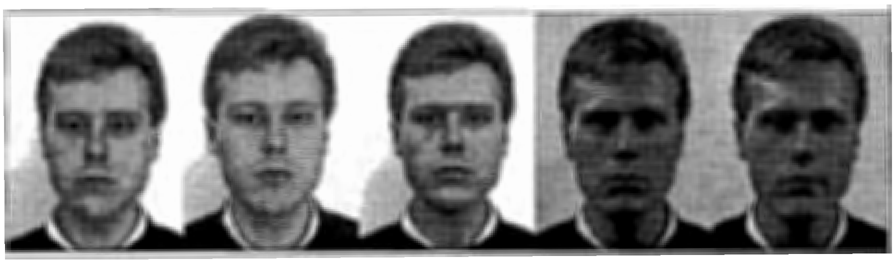

Figure 1: *One subject in each of conditions 1,4,5, 8 and 10.*

The next four images were acquired at least one week, and up to eight weeks after the previous ones. Lighting conditions were similar on both occasions, but no attempt was made to make them "identical".

Condition 11. An image was grabbed in conditions comparable with Conditions 1 to 4.

Condition 12. The subject was asked to move around and then reseat. The floor light was switched off and the image grabbed with all the remaining lights on, thus providing comparability with Condition 5.

Condition 13. The overhead projector was switched off and the image grabbed with only the two overhead strip lights remaining on, thus providing comparability with Condition 8.

Condition 14. The ceiling lights were left on, the black-out material was removed from the window to let in daylight, and a further image was grabbed. This image is not directly comparable with any of the previous conditions, and provide the largest amount of distraction.

In Fig. 2 we show the same subject as in Fig. 1 in each of these remaining conditions.

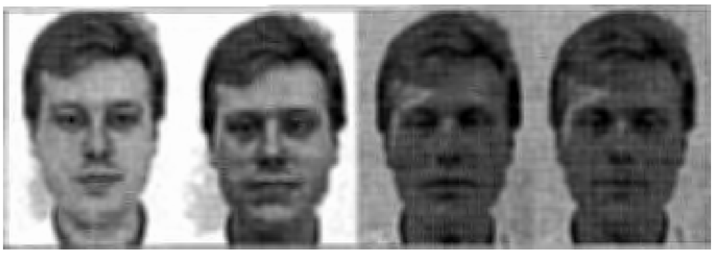

Figure 2: *The same subject in Conditions 11,12,13 and 14.*

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, avoiding the possibility that recognition is following condition rather than identity; an example where this can happen is described by Robertson and Craw (1994). Of course there are circumstances in which it is important to have recognition independent of the condition of the face in the gallery; nevertheless this *can* increase the apparent recognition rate, as potential matches are eliminated on the basis of their condition. Our choice here then is to simplify by eliminating condition variation in the gallery. The full gallery of 27 images consisted of:

amellanby1 andrew1 barry1 catherine1 david1 dhands1 dlow1 dpearson1 ian1
jenni1 jim1 kay1 kirsty1 lisa1 louise1 marie1 martin1 michael1 mmanson1 nick1
pat1 paul1 peter1 simon1 stephen1 tock1 trevor1.

The remaining 13 images of each subject then provide our cues; this gives 27 × 13 or 351 potential cues, each of which has a corresponding target in the gallery. Using *each* of the 27 faces as cue avoids the possibility that faces in the gallery which are not used as targets, may be hard to recognise; we do this except when calculating acceptance parameters, when it is necessary to use a gallery with no target, and hence we used a gallery of 26 faces. Rather than pool results over condition number, 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 72 images were available only in Condition 1, and as such provided images which could be used in the ensemble, but were not available for recognition tests. From these 50 images were selected informally with the aim of making the ensemble fairly homogeneous. Non-Caucasian faces were removed as were those which were obviously distinctive. Examples from the remaining images are shown in Fig. 3.

As discussed in Section 3, none of the faces used in the ensemble is used in any other part of this process; insisting that the ensemble and gallery are disjoint is a simplification, since

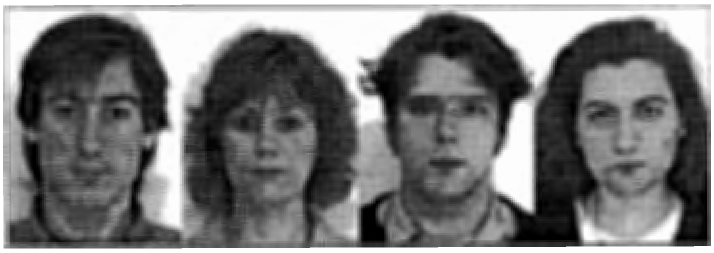

Figure 3: *Four images from the ensemble before processing.*

for a gallery of realistic size precludes using all the faces in the ensemble. It is not enough to ensure that the images in the gallery and ensemble are distinct. If an image $\mathbf{x}_2$ in Condition 2 is coded using an ensemble containing an image $\mathbf{x}_1$ of the corresponding face in Condition 1, the spectrum giving the decomposition of $\mathbf{x}_2$ in terms of eigenfaces appears normal. However when this spectrum is rewritten to show the corresponding decomposition in terms of the original images in the ensemble, the loading on $\mathbf{x}_1$ is overwhelming, and suggest again that doing recognition in such a situation is effectively template matching.

The ensemble of 50 images consisted of:

> adrian1 alan_rose1 alison1 alister1 annanena1 anon_one1 arlene1 bfegan1 chris1 chris_harbron1 chris_pin1 dave_faquhar1 david_imray1 derek1 dougal_grant1 dsmith1 fiona1 fiona_hogarth1 george1 gfindley1 gillian1 graham_brown1 grant_cumming1 heather1 joanna1 johannes1 john_mccal1 johnny_page1 kieran1 kim1 liz1 lynn1 lynn_james1 mark1 martin_smith1 meggan1 merilyn1 mnicholson1 neil1 pao1 peter_macgeorge1 pkyle1 richard_hardwick1 robert_deegan1 ruth1 scott1 stewart1 stuart_brown1 terry_johnstone1 tracy1.

4.2 Finding Landmarks

A fully automatic face recognition system based on the type of coding discussed here is a two-stage process; landmarks are first located and then used to generate appropriate codings. Automatic landmark location ensures that landmarks are chosen "honestly", and has the advantage that they are repeatable on the same image. Indeed the aim is to achieve "repeatability" over different images of the same face, and in many cases this will be done with greater reliability than had the landmarks had been picked manually. However there is no guarantee that the same landmarks, as obtained automatically, on different images of the same face are true homologues, even if the location program is working "correctly'; and in addition it may make more gross errors.

Because of the clear distinction between landmark location, and subsequent coding, there is thus interest in studying the processes separately, and the results given here are all obtained using manually located landmarks. Location was done using a self-contained "picking" program, which displayed the wire frame model described in A and provided an updated position when each landmark was moved. Locations were stored in individual files which remains fixed. For the testing we report here, each of the 34 landmarks on the 27 images comprising each of our cue and target sets were located manually; with 14 image sets, this involves 12852 individual locations; including the ensemble increases this to 14552.

Care has to be exercised to ensure there is no contamination between landmark locations in different images of the same face. A natural simplification is to choose landmarks on one instance of each face, and use the resulting files as a starting configurations from which landmarks on other images of the same face are obtained, simply by adjusting those landmarks seen to be in the "wrong" position in the new image. Since there is a facility to move models rigidly, this provides rapid locations; but there is the possibility that some landmark configurations may remain the same on different images of he same face. This has the effect that any form of "shape-based" recognition appears to be overly effective. We believe an earlier draft of our results suffered from this problem; as such landmarks for all the faces in the (fixed) gallery were measured independently by two people, neither of whom measured the bulk of the cue images, using a fixed neutral initial model. An attempt has been made to define landmark locations unambiguously, but in practice, locating on images of size 512 × 512 lead to locations differing by 4 or more rows or columns. We thus averaged the resulting positions and used them in subsequent testing. Possible contamination in the remaining measurements is irrelevant, since only a single set is used at any given time.

4.3 Including hair

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 system must maintain a useful recognition performance.

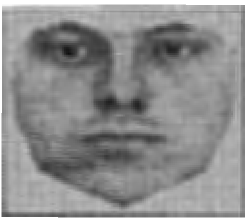

Figure 4: Normalised Face with hair excluded.

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. 4. Another advantage of concentrating on such inner features is that landmarks can be located with greater confidence; locations such as the "top of the head" are natural, and appear in most sets of landmarks; however they are clearly very dependent on hairstyle, and can be hard to choose repeatably if the hair is fine, or a parting is present. Essentially all the results we report are for such images in which the hair has been excluded; or main interest in "with hair" results is as confirmation of more general results obtained first when ignoring the hair. These images have $N = 3211$ pixels; in contrast, the full face image of Fig. 5 has $N = 6630$.

4.4 Processing

Each image is processed in the same way before being used in the ensemble, gallery or as a cue. A total of 34 landmarks, both true and deficient (eg the edge of the chin "half way" between two true landmarks), are found manually on each image, and an affine transformation of the image derived to minimise the error between the actual positions and those of the corresponding points on a reference face, here the average of the ensemble faces. The aspect ratio of the face is preserved, and no rotation in the plane was used. We shall call such images *normalised*; such normalisation is sometimes done by carefully positioning subjects before the images are acquired. The background can then be identified and has no further rôle 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 have zero mean; to give this, the average image is calculated, and subtracted from each member of the ensemble; in practice

$(n-1)$ eigenfaces are then available. The complete normalisation processing is illustrated in Fig. 5, although for display purposes, the average of the corresponding ensemble has *not* been subtracted.

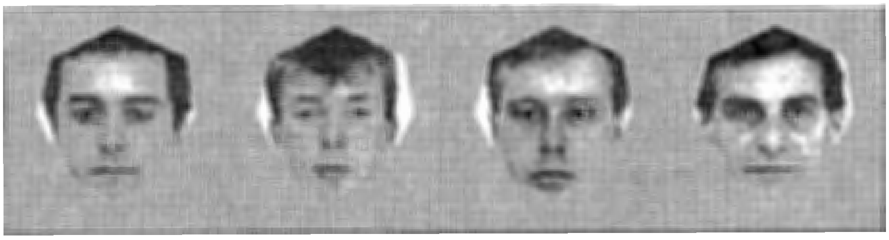

Figure 5: *Processed gallery faces.*

4.5 Eigenfaces and Coding

The resulting preprocessed ensemble is then subjected to a Principal Component Analysis, and the corresponding eigenvalues and eigenvectors obtained. Again, only the data within the face are considered; that outside the mask is ignored in this, and all subsequent processing. We describe the underlying mathematics of Principal Component Analysis in more detail in appendix B; in a practical case the input images will be linearly independent, and the outcome will be a set $\{\mathbf{e}_i \mid i = 1, \ldots, n-1\}$ of eigenvectors each of unit length, with corresponding (distinct non-zero) eigenvalues. This is the full rank case, since we have deliberately first subtracted the mean, and we assume this in the description that follows. This then fixes the eigenvectors uniquely to within a choice of sign. Because of our application, we refer to these eigenvectors as *eigenfaces*, and their span, or equivalently the span of the ensemble images, as the *face subspace.* In Fig. 6 we show examples of such eigenfaces.

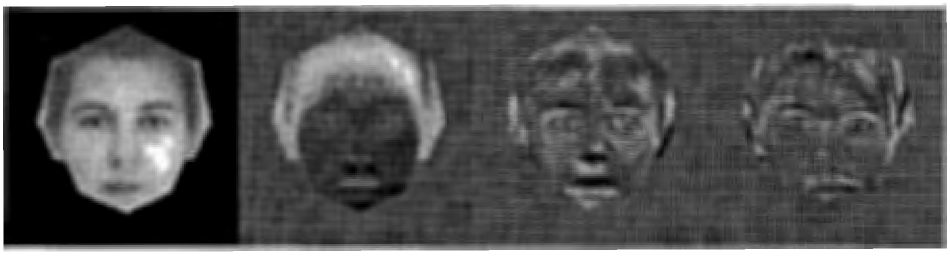

Figure 6: *The average face from an ensemble of size 50 on the left, together with the first, eleventh and 21st eigenface from the ensemble. The grey levels of the images have been scaled individually for display purposes.*

By construction, these eigenvectors are orthonormal; we can thus calculate the orthogonal projection of a new (processed) image onto the subspace spanned by these eigenvectors; indeed the projection of a (normalised) face $\mathbf{x}$ onto the face subspace is given by

$$P(\mathbf{x}) = \hat{\mathbf{x}} = \sum_{i=1}^{n-1} \langle \mathbf{x}, \mathbf{e}_i \rangle \, \mathbf{e}_i,$$

where $\langle ., . \rangle$ is the usual inner product in $\mathbf{R}^N$. For display purposes, we will then add the average face to $\hat{\mathbf{x}}$ in order to obtain the "true" reconstruction of $\mathbf{x}$.

This projection is used to code each new image for recognition purposes. Rather than storing the projection itself, we code each new image, preprocessed as described above, with the component vector

$$\{x_i = \langle \mathbf{x}, \mathbf{e}_i \rangle \mid i = 1, \ldots (n-1)\}$$

describing the image in terms of the eigenface basis. By analogy with the Fourier spectrum, we shall refer to this as the *eigenface* spectrum, and note that such a spectrum is only defined relative to some fixed ensemble. Our aim then in this section is to characterise the utility of this spectrum as a (relatively naive) code for recognition. In Fig. 7 we show the spectrum of the first image from Fig. 1; the arbitrary sign of each eigenvector has been chosen so that all the coefficients in this case are positive. In general the coefficients can be positive or negative.

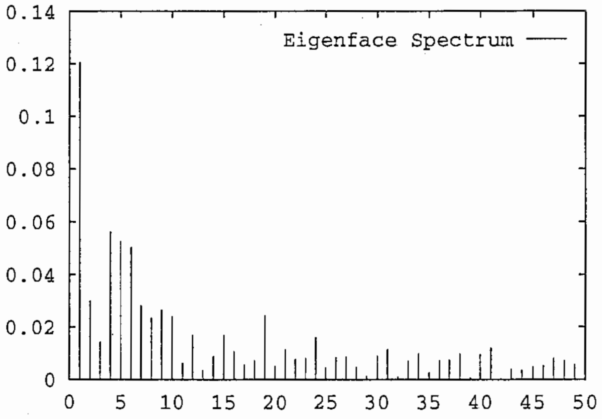

Figure 7: *The spectrum of a face image, with 49 components.*

4.6 Matching

Given a gallery of faces coded as above, the code from the cue image is compared with each gallery code to determine the best match. One way to do this is to put a metric on $\mathbf{R}^{n-1}$ and use nearest neighbour matching, and a natural choice of metric is the usual Euclidean metric on $\mathbf{R}^{n-1}$; since our basis of $\mathbf{R}^{n-1}$ is orthonormal in $\mathbf{R}^N$, the metric is just the usual Euclidean metric in $\mathbf{R}^N$; in particular given faces $\mathbf{X}$ and $\mathbf{y}$, both normalised so that $||\mathbf{x}|| = ||\mathbf{y}|| = 1$, we have

$$d(\mathbf{x},\mathbf{y})^2 =< \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{y} >= ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2 < \mathbf{x},\mathbf{y} >= 2(1- < \mathbf{x},\mathbf{y} >).$$

We thus can describe the distance as the departure from perfect correlation between $\mathbf{x}$ and $\mathbf{y}$, and so recognise nearest neighbour recognition as effectively template matching, a method which is know to be effective in simple cases, although necessarily slow if implemented directly in $\mathbf{R}^N$. However, this choice of metric means that the use of eigenfaces *per se* is irrelevant. The only property necessary is orthonormality, and any other such basis, obtained for example from Gram - Schmidt orthogonalisation of the ensemble image, would give the same results.

There other metrics that can be placed on $\mathbf{R}^{n-1}$ to do nearest neighbour matching. One choice can be understood by considering a face image which has uniformly average deviations from the mean, and so has co-ordinates

$$(1) \qquad (\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda}_{n-1}),$$

where the $\lambda's$ are the variances obtained when the eigenface basis is constructed. The construction forces these co-ordinates to be decreasing, yet it can be argued that variations along all axes are equally significant, since our aim in this coding is discrimination rather than representation. This can be achieved by differentially rescaling the axes in such a way that our "typical" variant face (1) has co-ordinates $(1, 1, \ldots, 1)$, and we are led to a weighted norm, in which the distance between eigenface spectra $\{x_i\}$ and $\{y_i\}$ is computed as

$$(2) \qquad d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{n-1} \lambda_i^{-1} (x_i - y_i)^2 \right)^{1/2},$$

where $\{\lambda_i\}$ is the sequence of eigenvalues. The effect of this rescaling not usually large since the non-zero eigenvalues in practice differ by a factor whose square root (the relevant measure) is at most about 6; however this metric does make essential use of the Principal Component methodology. There is significant debate about scaling Principal Components (Jolliffe 1986, Page 225), and published work on face recognition is not always explicit about what scaling, if any, has been used. We give matching results with both the usual Euclidean metric, and for the above weighted norm.

4.7 Matching and rejection

The resulting normalised ensemble is subjected to a Principal Component Analysis, and the corresponding eigenvalues and unit eigenvectors (or *eigenfaces*) obtained. The orthonormality of the eigenfaces means it is simple to compute the component of any (normalised) face in the direction of each eigenface, and hence obtain an $(n-1)$-tuple or code. A coded cue image is then compared with each gallery code to determine the best match. One way to do this uses nearest neighbour 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 effectively template matching.

Another natural choice of metric on $\mathbf{R}^{n-1}$ is the Mahalanobis distance, in which where $\{\lambda_i\}$ is the sequence of eigenvalues. This treats variations along all axes as equally significant, 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 (Lades et al. 1993) has a match score c_j between each image in the gallery, and the cue image. The best match corresponds to the lowest score, and interest centres 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 statistics

$$r_1 = \frac{c_1 - c_0}{\sigma} \qquad \text{and} \qquad r_2 = \frac{\mu - c_0}{\sigma}$$

with associated thresholds t_1 and t_2; a match is accepted if $r_1 > t_1$ and $r_2 > t_2$ and otherwise rejected.

We adopt this, reporting a correct match as a *clear* hit if the target passes this acceptance criterion, and *just* a hit otherwise, with a similar terminology for misses. To set thresholds, the distances between the cues in Condition 2 and the gallery images with the target deleted were found. The two statistics, r_1 and r_2 were then calculated for each cue-image and the largest values independently chosen as t_1 and t_2. This procedure ensured that in the best, base condition, there was no false recognition. We have found cases in which "clear misses" occur, necessarily in conditions other than Condition 2. In fact none occur in the results we report here; this is an indication that our criterion is particularly conservative.

5 Baseline Recognition

5.1 Presentation

Because of the volume of data we group similar recognition results and present summaries. Conditions 2, 3 and 4 are grouped together, and described 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 overall performance, the four groups have been combined in an "Overall" group. Although more images are available in the sets with low condition numbers, greater interest attaches to the performance of the "Later" group, and accordingly we weight this latter group more heavily. The weights used are given in Table 1, together with the contribution that a single trial makes to the overall results.

	Trials	Weight	Individual
Immediate	81	1	0.11%
Variant	81	1	0.11%
Lighting	81	2	0.21%
Later	108	4	0.53%

Table 1: Weightings used for overall performance.

Our main interest is in the comparison between affine normalisation, and the more intrusive shape free form; however we first discuss other choices which make up our testing regime.

5.2 Results

Our "baseline" recognition in Table 2 gives results against which subsequent performance is to be compared, and was obtained using all 99 eigenfaces from this enlarged ensemble. Pixel value normalisation, as throughout, is by histogram equalisation.

6 Improved coding

A number of choices were made in fixing the testing regime of Table 2. We discuss those choices below, show in each case that the choice was made to obtain the most effective recognition, and draw conclusion from the observed behaviour. That more intrusive processing described subsequently improves recognition noticeably is thus of greater interest.

	Hit		Miss	
	Clear	Just	Just	Clear
Immediate	74.1	23.5	2.5	0.0
Variant	28.4	58.0	13.6	0.0
Lighting	12.3	51.9	35.8	0.0
Later	11.1	55.6	33.3	0.0
Overall	20.0	51.6	28.4	0.0

Table 2: Affine normalised, matching with Mahalanobis distance. Hair has been *excluded* from the match. Match percentages from 351 trials.

6.1 Choice of ensemble

Initial testing was done using the ensemble of 50 faces described above. It is not clear that an ensemble of 50 faces is adequate, but gathering more images and subsequent landmark location was inconvenient. We thus borrowed an idea from Kirby and Sirovich (1990) who made use of vertical symmetry, or rather the lack of it, in individual faces. Because we have landmark data, the vertical facial mid line is available on each image, and so it is possible to create, corresponding to each face in the ensemble, a "mirror" face, whose image and landmarks are obtained by reflection about the vertical axis of symmetry. The decision to include such faces can even be considered as a piece of expert knowledge; despite the lack of symmetry in an individual face, we believe there is no overall bias in the set of allowable faces. Such a reflection, although simple geometrically, is a non-linear operation, and so genuinely enlarges the span of the ensemble. In contrast creating new faces simply by averaging pairs of existing ones, would not do so; the effect would simply be to create more zero eigenvalues in the cross-correlation matrix.

	50	$(50 \times 2)/2$	50×2
Immediate	96.3	95.1	97.5
Variant	79.0	81.5	86.4
Lighting	49.4	61.7	64.2
Later	54.6	65.7	66.7
Overall	60.6	69.7	71.6

Table 3: Hit percentages from 351 trials. Affine normalised, 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 resulting improvement in recognition shown in Table 3 was sufficiently noticeable to suggest that the original ensemble was indeed too small, and all subsequent tests are reported with this "doubled" ensemble. The central column in Table 3 suggest that it was perhaps worth while using more than 50 of the available eigenfaces, but the difference is too small to attempt reliable conclusions.

6.2 Matching Method

Our first comparison given in Table 4 is between the baseline results of Table 2 and the same set of tests in which the match is based on Euclidean distance.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Euclid	Mahal	Euclid	Mahal	Euclid	Mahal	Euclid	Mahal
Immediate	59.3	74.1	35.8	23.5	4.9	2.5	0.0	0.0
Variant	11.1	28.4	70.4	58.0	18.5	13.6	0.0	0.0
Lighting	3.7	12.3	42.0	51.9	54.3	35.8	0.0	0.0
Later	3.7	11.1	42.6	55.6	53.7	33.3	0.0	0.0
Overall	10.4	20.0	44.7	51.6	44.8	28.4	0.0	0.0

Table 4: Comparing Euclidean and Mahalanobis distances for matching. Match percentages from 351 trials. Affine normalised; hair has been *excluded* from the match area.

The use of the Mahalanobis distance as given in equation (2) is clearly more effective. This is an important observation, confirming that we are making essential use of the variance properties of eigenfaces, rather than simply that they provide an orthonormal basis for the span of the ensemble images. Note also that the advantage is least evident in the "Immediate" group, where simple 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 relative advantage in doing so increases as one moves away from direct image matching ideas, either by not using the hair, with its rich source of local pattern, but invariant only over short time scales, or by moving to shape free standardisation. A prediction would be that the effect is less marked on those cue sets with a low condition number, where straight image matching is most successful. However the identification task here is sufficiently easy that it is hard to detect any significant difference.

The comparison between the two metrics is very similar when the hair is included in the image area.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Euclid	Mahal	Euclid	Mahal	Euclid	Mahal	Euclid	Mahal
Immediate	64.2	86.4	30.9	7.4	4.9	6.2	0.0	0.0
Variant	46.9	76.5	43.2	19.8	9.9	3.7	0.0	0.0
Lighting	33.3	55.6	54.3	38.3	12.3	6.2	0.0	0.0
Later	17.6	38.0	43.5	33.3	38.9	27.8	0.0	0.9
Overall	29.1	51.1	44.4	30.2	26.5	18.3	0.0	0.5

Table 5: Comparing Euclidean and Mahalanobis distances for matching. Match percentages from 351 trials. Affine normalised; hair has been *included* in the match area.

6.3 Preprocessing methods

Because of the dependency of any form of image matching on light levels in the original image, it is important to reduce the effects of such irrelevant variation as much a possible.

A simple way is to set the Euclidean norm of the vector of pixel values, to a constant value, typically 1. If V is the value of a given pixel of the original image, and R is the set of pixels under consideration after parts of the image have perhaps been masked, the corresponding transformed pixel value V' is give by

$$V' = V \Big/ k \left(\sum_{R} V^2 \right)^{1/2}$$

This will ensure that the total brightness or strength of the image is consistent across treatments, but it does not ensure that consistent across the image; an image with a consistent grey-level will be treated in the same way as one with a gradient across it, should the values of $\sum_R V^2$ be equal in the two images.

An alternative algorithm, which would tend to control for this problem is to histogram-equalise the image. This will set the distribution of grey-levels as close to equal across the range of possible values as is possible; the range of possible grey-levels is divided up into a number of bin (this is typically the same as the number of possible grey-level, in an 8-bit byte situation this is 256) and the number of pixels per bin is counted. In the particular implementation here, a map from the old grey-levels to the new ones is constructed by adding the number of pixels in successive bins until the total number unassigned is greater than average number of pixels per bin, and the map is set to the average of the currently contributing. The 'unassigned pixel' counter is then re-set to the number over the pixels which could be assigned and the process resets, with the map set to it's initial value, plus the number of average bins assigned until process repeats. This will then spread the grey-levels out so that they are uniformly distributed across the range, but only joining bins, not splitting them. Such an algorithm will not preserve the image-energy or vector length, and thus the vector length has to be set both before and after the histogram equalisation procedure. The process will again tend to reduce mean and variance differences, but will retain some of a brightness gradient and may actually produce an abnormal brightness pattern in the image under some circumstances.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Histo	Len	Histo	Len	Histo	Len	Histo	Len
Immediate	74.1	80.2	23.5	12.3	2.5	7.4	0.0	0.0
Variant	28.4	61.7	58.0	18.5	13.6	18.5	0.0	1.2
Lighting	12.3	35.8	51.9	30.9	35.8	30.9	0.0	2.5
Later	11.1	32.4	55.6	26.9	33.3	39.8	0.0	0.9
Overall	20.0	41.4	51.6	25.3	28.4	32.1	0.0	1.2

Table 6: Comparing fixing the total length of each image (Len) with histogram equalisation (Histo) as intensity normalisation techniques Match percentages from 351 trials. Matching with Mahalanobis distance; hair has been *excluded* from the match area.

There appears to be some gain in using histogram equalisation, although this reduces the number of clear hits. In fact this pattern is repeated over a number of combinations — we give another example in table 13 — and while the advantage is slight we choose to use histogram equalisation as the "baseline" condition. In fact the comparison is even more difficult to perform usefully, because there is some interaction between the choice of intensity normalisation method, and whether or not the average image is first subtracted before subsequent processing, as discussed in Section 4.4, since subtracting the average can in effect negate this preprocessing.

6.3.1 Other Methods

A alternative form of pre-processing, which may have a better relation with the psychology of early face recognition, is to explicitly remove the largest variations, perhaps by using a local-average algorithm (Watt 1994). This works in a number of stages, first calculating the local average at each point of the image, by convolving with a suitable mask, in this case a Gaussian:

$$V_m = \sum_R VW \Big/ \sum_R W,$$

and then finding the local standard deviation

$$V_{s.d.}^2 = \sum_R (V - V_m)^2 W \Big/ \sum_R W.$$

There are then two methods of normalising the image; the first applies

$$V' = \frac{V - V_m}{V_{s.d.}}, \tag{3}$$

but this linear equation is unstable when the standard deviation is small; if the image is uniform it will amplify the variations, but reduce them when the standard deviations are larger. An alternative is to use;

$$V' = \begin{cases} \dfrac{V - V_m}{V - V_m + kV_{s.d.}} & \text{if } V > V_m \\ \dfrac{V - V_m}{V - V_m - kV_{s.d.}} & \text{if } V < V_m \\ 0 & \text{if } V = V_m \end{cases} \tag{4}$$

where k is a constant, in this case set to 1. This algorithm has the effect of effectively binarising the image; it much more non-linear when $V_{s.d.}$ is small than when it is large.

Both forms of this algorithm were applied to the images, without hair and using the Mahalanobis distance. This required a parameter to define W, the weighting function. This was a symmetric Gaussian and was thus determined by σ, the standard deviation. This was set at 10 pixels; since ce the width of the average tile face was 69 pixels, the standard deviation was 1/7th of the width. This was chosen intuitively, by inspecting images. Two other parameters were necessary; mask-size (the area over which the mask operated), this was set to four times the standard deviation, and the constant k for equation 4, again set by inspection. The results are shown in Table 7 and show that recognition is notably worse for these processing methods, compared with histo-equalisation with both lower hit-rates and smaller separations. This may well reflect the general removal of texture from the images, which tend to become line-drawings under these circumstances.

A second method of removing the large-scale differences between images, while retaining identity-specific information is to Fourier-transform the image and highpass and lowpass it

	Hit				Miss			
	Clear		Just		Just		Clear	
	Linear	Non-Lin	Linear	Non-Lin	Linear	Non-Lin	Linear	Non-Lin
Immediate:	40.7	29.6	45.7	65.4	13.6	4.9	0.0	0.0
Variant:	19.8	12.3	45.7	65.4	34.6	22.2	0.0	0.0
Lighting:	4.9	3.7	44.4	46.9	48.1	49.4	2.5	0.0
Later:	5.6	3.7	39.8	47.2	53.7	49.1	0.9	0.0
Overall:	10.7	7.4	42.1	51.1	46.2	41.5	1.1	0.0

Table 7: Comparing linear and non-linear versions of the lighting-normalisation functions. Match percentages from 351 trials. Matching with Mahalanobis distance; hair has been *excluded* from the match area.

(with appropriately chosen values). Studies of human face recognition show that almost all the performance of recognition can be attributed to a two-octave band of spatial frequencies, centred upon approximately 8 cycles per face (Costen 1994).

The Fourier spectrum was bandpassed with a radially symmetric Gaussian filter with half-powers of 3.17 and 20.06 cycles per face-width. The face-width was determined by the distance between points 24 and 33 in the model (the edges of the tile face, in line with the axis of the eyes). This ensured that the same size of variation was picked up in all the faces, and that the effects were not dependent upon the size of the faces within the image. The Gaussian filter was defined for each radii by,

$$f = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \frac{(r-\mu)^2}{2\sigma^2}. \tag{5}$$

Here σ was the standard deviation of the mask, μ the mean frequency, and r the frequency of the Fourier component. To ensure that the frequencies were given equal weight, this mask was applied to the logarithmic transform of the frequencies. Spatial frequency is a ratio scale, where successive doublings in frequency will add single bits of information to the location of a point. Thus a logarithmic scale will compensate from the increased number of frequencies possible, given a fixed frequency quantisation, at the higher frequencies. Thus in the case considered, the parameters were $\mu = 2.07$ log cycles per face (thus $\mu = 7.97$ cycles per face) and $\sigma = 0.67$ log cycles per face. This is not convertible, and is thus stated via the half-power points.

The results, shown in Table 8 again only display rather bad recognition. As well as suggesting that the results on the recognition of spatial-frequency filtered faces actually reflect difficulties in discovering the shape of faces, rather than in the coding of grey-level differences between the images, this poor recognition may well suggest that the Principal Component Analysis processing should be thought occurring relatively late within the processing stream, after shape-processing has occurred.

6.4 Standardisation Methods

The specific standard normalisation used here as a default to compare other normalisations involves using each of the located landmarks in the registration process. It performs a rigid, aspect-ratio retaining affine transformation to scale and move the face to minimise the total distance between all the corresponding facial landmarks. However, it is possible that this is

	Hit		Miss	
	Clear	Just	Just	Clear
Immediate:	34.6	43.2	22.2	0.0
Variant:	12.3	44.4	43.2	0.0
Lighting:	0.0	32.1	67.9	0.0
Later:	4.6	36.1	59.3	0.0
Overall:	7.7	36.9	55.4	0.0

Table 8: Match percentages from 351 trials for the Fourier filtering normalisation, matching with Mahalanobis distance. Hair has been *excluded* from the match.

an inappropriate method, especially as it gives equal emphasis to all sections of the face. This may reduce recognition, by ensuring that there are no portions in absolute alignment. Two other normalisation schemes which do not suffer from this problem were investigated. In the first of which which just the landmarks at the eyes were used, corresponding to the usual "normalise on the eyes" method and the affine transformation brought these into alignment. In the second, a smoother "oval" normalisation was used in which each images was presented with an oval outline and then scaled so that the outer points, and thus the contour, of the face were aligned and just inside the oval mask so the maximum degree of contour was visible.

The results of matching with these three sorts of normalisation are shown in table 9, with the hair included and Mahalanobis distance distance. It seems obvious here that the oval normalisation is notably worse than either of the affine transformations, with notably higher false alarm rates. The difference between the all-landmark and eye-landmarks is a little less clear, with slightly fewer false alarms for the eye-landmark normalisation than for the all-landmarks, although the position is reversed with regard to confident hits. It is notable that this ambiguity is entirely caused by the presence of the 'Later' images; this may reflect an enhanced ability to match on the rather variable facial contour in this case. It is notable that in general the results show better recognition for greater degrees of normalisation, suggesting that ensuring that equivalent features are combined leads to better representation and enhances recognition.

	Confident			Unreliable						Confident		
	Hit			Hit			Miss			Miss		
Method:-	A	B	C	A	B	C	A	B	C	A	B	C
Immediate	79.0	80.2	69.1	17.3	13.6	25.9	3.7	6.2	4.9	0.0	0.0	0.0
Variant	71.6	59.3	45.7	27.2	38.3	44.4	1.2	2.5	8.6	0.0	0.0	1.2
Lighting	51.9	40.7	34.6	44.4	49.4	43.2	3.7	9.9	22.2	0.0	0.0	0.0
Later	23.1	19.4	25.9	50.0	58.3	35.2	26.9	22.2	37.0	0.0	0.0	1.9
Overall	40.5	34.8	34.5	42.9	49.5	36.9	16.7	15.7	27.4	0.0	0.0	1.2

Table 9: Comparison between normalising on available landmarks (Method A) solely on eye landmarks (Method B) and normalising an appropriately shaped elliptical subimage (Method C). The hair has been *included* in the match, and the Mahalanobis distance has been used for comparison. Each figure is a percentage; the overall figure describes 351 trails.

7 Shape-Free Recognition

The discussion in Section 2 suggests 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, while in Section 11 such a decomposition is supported on theoretical grounds; in this section we explore the use of a configuration-free or texture vector alone for recognition.

The use of the word *texture* in this context comes from computer graphics, where objects are typically represented initially with a wire frame. The surface is then painted, with colour, with a regular texture, or more generally with (say) an image which is perhaps the appropriate reflection of the scene in the object. The word texture thus had its meaning extended to include more general image mappings, and we use it in this sense throughout.

7.1 General description

A very common way of distorting or warping an image is what we shall refer to as *linear* texture mapping. Distorting a triangle in an image to another triangle produces a natural distortion of the image, or texture, in the interior of the triangle. Formally this is done by describing each point in terms of barycentric co-ordinates with respect to the vertices of the triangle in which it lies. The output grey level at this point is then the grey level at the point in the original triangle with the same barycentric co-ordinates with respect to its vertices. Thus as soon as a portion of an image has been triangulated, specifying a distortion of each landmark or vertex of the triangulation then gives a linear texture mapping of the whole of the image lying within the set of triangles. In practice efficiency considerations can cause complications (Benson 1994), but the whole process, usually known as *morphing*, can be done relatively rapidly. However there are restrictions; in order that the relevant regions of both input and output images lie in triangulations, no triangle can "flip" or change its sense; this can sometimes limit the allowable distortions quite severely. An example of the output from linear texture mapping, on an exaggerated distortion, is given in Fig 10.

An alternative warping method is that of thin plate spline warping (Bookstein 1989). In this, rather than choosing a triangulation of the landmark sets, interpolation is by a pair of function of the form

$$F_x(x, y) = a_1 + a_x.x + a_y.y + \sum_{i=1}^{n} w_i.(r - r_i)^2 \log((r - r_i)^2),$$

where we write $r_i^2 = (x - x_i)^2 + (y - y_i)^2$. Here the constants a and w are chosen to ensure that landmarks map to landmarks and (x_i, y_i) gives the position of landmark i in the domain. This choice is made in order to minimise distortion energy; considered as the energy needed to distort a thin flat sheet of metal (a thin-plate spline) to assume the height $F_x(x, y)$ at the point (x, y); in particular, the energy of an affine deformation is zero.

The method is very natural in that it reduces to an affine map if the change in landmark positions can be described affinely, and it simply needs a set of landmarks specified in their original and distorted positions, rather than a triangulation as well. It can thus cope with a wider range of distortions than can linear texture mapping. However the resulting distortion takes significantly longer to compute and is much less localised. And unlike the map between triangulations, this distortion can only be inverted numerically; the new texture is thus usually computed by implementing the distortion from new to old positions, rather than what is perhaps the more expected way. An example of its use, on an exaggerated distortion, is given in Fig 10; the grey area in the background at the top of the image is an indication that no data from the original image was available at that point.

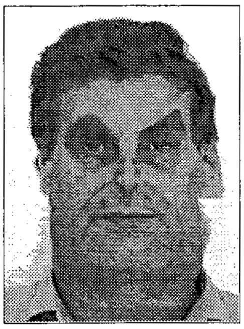

Figure 8: *Linearly warped image.*

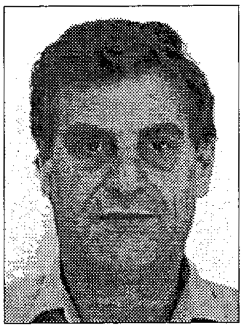

Figure 9: *Original Image.*

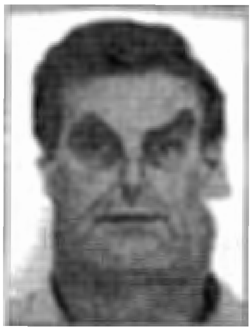

Figure 10: *Thin plate spline warped image.*

As described in Section 4.2, landmarks are available on all our images, and these warping methods thus suggest the possibility of a more intrusive shape normalisation than simply removing the effects of position, scale and orientation introduced by the imaging process. By warping each image to a fixed configuration of these landmarks we remove many of the gross effects of shape difference; to emphasis this we describe the resulting warped images as *shape-free faces.*

The passage to shape-free faces enables texture comparisons to be made on a finer scale; with an appropriate number of well chosen landmarks, gross features such as eyes or mouths will coincide on all shape-free faces and texture differences will arise from "second order" appearance differences between individuals' features. Such a strategy has been advocated as a possible mechanism by which humans perform within class discrimination (Rhodes, Brennan and Carey 1987), and as an effective object recognition strategy (Ullman 1989), and has been used for both coding (Choi, Okazaki, Harashima and Takebe 1990) and recognition of faces (Craw and Cameron 1992, Lanitis et al. 1994), and face features(Shackleton and Welsh 1991). It is thus natural to investigate it here and perform a more detailed comparison with more conventional methods of machine-based face recognition. We take our standard configuration to be the average landmark configuration (Bookstein 1991), although other configurations are just as valid, and each face image is warped to that configuration.

7.2 Linear distortions

Our first results use linear texture mapping of each face to the average shape of the set of ensemble images. As usual, the pixel values are then normalised by histogram equalisation. The results given in Table 10 compare this image normalisation with the baseline results, in which the image is simply affinely transformed to best match the average.

The comparison suggests that shape-free normalisation appears to be uniformly more effective than the corresponding affine version, even though the shape information has been deliberately ignored. There are both fewer misses and noticeably more confident hits in each of the condition categories. It may be that we have implemented the affine normalisation

	Hit				Miss			
	Clear		Just		Just		Clear	
	Affine	S-free	Affine	S-free	Affine	S-free	Affine	S-free
Immediate	74.1	95.1	23.5	4.9	2.5	0.0	0.0	0.0
Variant	28.4	65.4	58.0	28.4	13.6	6.2	0.0	0.0
Lighting	12.3	21.0	51.9	49.4	35.8	29.6	0.0	0.0
Later	11.1	29.6	55.6	46.3	33.3	24.1	0.0	0.0
Overall	20.0	38.6	51.6	40.6	28.4	20.8	0.0	0.0

Table 10: Comparing affine and linear texture mapping to the average face (S-free). Match percentages from 351 trials. Matching with Mahalanobis distance; hair has been *excluded* from the match area.

inappropriately, however we discussed variants of the procedure in Section 6.4, and our comparison is with the one that proved most effective.

Although we are less interested in results when the hair area is included in the match, Table 11 reinforces the suggestion that passing to a shape-free face is advantageous, although here, confident recognition is very much better in the affine case. As expected, overall rates rise, but the changing effect of the hair over time is also clearly seen; recognition in the "later" category matching over a significant period has declined substantially.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Affine	S-free	Affine	S-free	Affine	S-free	Affine	S-free
Immediate	86.4	67.9	7.4	32.1	6.2	0.0	0.0	0.0
Variant	76.5	34.6	19.8	65.4	3.7	0.0	0.0	0.0
Lighting	55.6	18.5	38.3	79.0	6.2	2.5	0.0	0.0
Later	38.0	11.1	33.3	63.0	27.8	25.9	0.9	0.0
Overall	51.1	21.3	30.2	63.4	18.3	15.3	0.5	0.0

Table 11: Comparing affine and linear texture mapping to the average face (S-free). Match percentages from 351 trials. Matching with Mahalanobis distance; hair has been *included* in the match area.

7.3 Thin Plate Spline distortions

We now give the corresponding results using thin plate spline distortions. The physical argument behind their creation is an attractive reason to believe the warp may be more "natural" and thus better fitted to our matching task. This warping method is that used by Lanitis et al. (1994), giving added interest to the comparison.

In Table 12, we present the comparison with our usual affine normalised images. Again we see an advantage in passing to the shape free form, and in general, the behaviour is very similar to that obtained with linear texture mapping. However the advantage over affine normalisation is not quite as marked as when using linear texture mapping, and there thus seems no good reason to accept the extra complication of thin plate spline warping, except perhaps for severe distortions when a linear warp fails.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Affine	TPS	Affine	TPS	Affine	TPS	Affine	TPS
Immediate	74.1	84.0	23.5	16.0	2.5	0.0	0.0	0.0
Variant	28.4	60.5	58.0	33.3	13.6	6.2	0.0	0.0
Lighting	12.3	24.7	51.9	44.4	35.8	29.6	0.0	1.2
Later	11.1	26.9	55.6	42.6	33.3	30.6	0.0	0.0
Overall	20.0	36.1	51.6	39.2	28.4	24.5	0.0	0.3

Table 12: Comparing affine and thin plate spline based texture mapping to the average face (TPS). Match percentages from 351 trials. Matching with Mahalanobis distance; hair has been *excluded* from the match area.

Finally we give one of a number of possible examples which serve to confirm earlier decisions about what the most useful form or preprocessing is; in particular our choice of histogram equalisation over simple length normalisation, as discussed in Section 6.3 is supported by Table 13, where we contrast the two methods in the context of shape-free faces.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Histo	Len	Histo	Len	Histo	Len	Histo	Len
Immediate	95.1	84.0	4.9	16.0	0.0	0.0	0.0	0.0
Variant	65.4	60.5	28.4	33.3	6.2	6.2	0.0	0.0
Lighting	21.0	24.7	49.4	44.4	29.6	29.6	0.0	1.2
Later	29.6	26.9	46.3	42.6	24.1	30.6	0.0	0.0
Overall	38.6	36.1	40.6	39.2	20.8	24.5	0.0	0.3

Table 13: Comparing fixing the total length of each image (Len) with histogram equalisation (Histo) as intensity normalisation techniques Match percentages from 351 trials. Matching with Mahalanobis distance; hair has been *excluded* from the match area.

8 Shape-based recognition

8.1 Description

Since the shape-free normalisation necessarily discards information on the shape of the face, recognition may be further enhanced by independent consideration of the shape. Obviously humans can recognize faces from line-drawings and the majority of early automatic face-recognition systems made use of the configural nature of faces (so all major points on any face can be assigned names or *local signs* and the ordering of these points will be consistent across faces) to derive lists of distances between feature points which were then matched. Assuming that the configural nature is retained, so there is no mis-ordering of points or changes of angle and scale, the face-space should be linear and by performing Principal Component Analysis on the landmark locations it should be possible to interpolate (and thus approximate) reasonable faces.

Processing was performed in exactly the same way as before; it was found possible to turn the X- and Y-coordinates of the landmarks into a 2×35 pixel image with the grey-levels set to the location-values. To do this, it was necessary to alter the base-line of the coordinates. The original model-files used an image-centred description, with landmarks described in terms of the number of pixels to the top left-hand corner. This would obscure differences as the images were not all the same number of pixels, nor were the faces of a uniform position within the images. In addition, these effects might interact, leading to differences in the apparent variablity of landmarks.

These problems were overcome by converting the models to face-based coordinates. The affine normalisation was applied to the model-files with all the landmarks locked into their average position, but retaining the face's aspect ratio. This removed scale and position effects, and also ensured that the (imaginary) images to which the models referred had a fixed number of pixels. The landmark locations were then converted to pixel values in 2×35 pixel images and the average image was subtracted; this ensured that each landmark (or now, pair of pixels) had a mean value of zero and any particular face was thus coded in terms of the distorsion from the ensemble mean. The hair could not be masked out in the usual way, but was excluded by uniformly setting landmark-coordinates to zero for the required points. The shapes of the ensemble images then provided suitable principal components (we reserve "eigenface" for texture components) as descriptors of the gallery and cue images.

8.2 Results

The model files contained 35 landmarks. However, of these, 14 were 'forced', so that one dimension was set by the operator and one derived from the positions of the other points. Thus for example, landmark 28 was defined as 'the point where a line at 135 degrees clockwise form the vertical, which passed through the right-hand corner of the mouth (landmark 19) intersects with the line of the chin'. In addition, one point had both dimensions defined, being necessary for the texture-map distorsion. Thus there are a maximum of 54 degrees of freedom in the whole-face data and a mere 45 when the hair is excluded. With 100 shapes in the ensemble, the whole of the space spanned by the shape data can be generated, and it became necessary to investigate whether this crude counting overestimated the dimensionality of the data. The number of principal components used to code the shape vector was thus varied, progressively removing the late-extraction, low variance components. The correct-classification rates are shown in Fig. 11 for the doubled ensemble for the with and without hair cases. These show that in both cases, recognition peaks when about 20 components are included in the analysis; the with-hair images are again consistently better than the hair-less ones even with the same number of components.

The results with the doubled ensemble, where the 75 eigenvectors which account for the least variance have been excluded are shown in Table 14, which compares Mahalanobis distance with Euclidean distance. This shows that, unexpectedly, the Euclidean distance is marginally more accurate on both clear and total recognition. This also shows that although reasonable recognition can be achieved with only the shape of the faces, it is an extremely impoverished representation, being determined by only a small number of factors. Although this is partly due to the relatively crude shape descriptors used here, there is also evidence that humans use relatively crude shape information for recognition, so large increases in performance may not be possible.

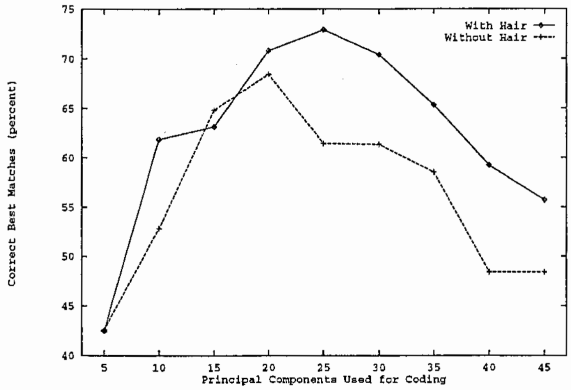

Figure 11: Variation in correct classification from shape with available principal components.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Mal	Euc	Mal	Euc	Mal	Euc	Mal	Euc
Immediate	29.6	33.3	43.2	42.0	27.2	24.7	0.0	0.0
Variant	17.3	22.2	60.5	34.6	22.2	42.0	0.0	1.2
Lighting	18.5	25.9	42.0	42.0	38.3	32.1	1.2	0.0
Later	15.7	12.0	40.7	47.2	42.6	40.7	0.9	0.0
Overall	18.0	18.4	43.4	44.2	37.8	37.3	0.8	0.1

Table 14: Comparing Mahalanobis and Euclidean spectrum scaling for matching on shape. Hair has been *excluded* from the match area.

8.2.1 With hair

The comparable with-hair test are shown in Fig. 15; this shows the usual behaviour with better recognition on both measures for the Mahalanobis distance. This also helps explain the reversal of the difference for the without-hair condition since there are fewer varying points and thus fewer degrees of freedom in that case. As both the Principal Component Analysis was performed with the same number of eigenvectors, there will be a greater number of irrelevant, redundant components in the without-hair condition, and the scaling of the spectrum will enhance these low-variance components.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Mal	Euc	Mal	Euc	Mal	Euc	Mal	Euc
Immediate	38.3	33.3	38.3	43.2	23.5	23.5	0.0	0.0
Variant	27.2	25.9	45.7	39.5	27.2	34.6	0.0	0.0
Lighting	23.5	28.4	51.9	40.7	24.7	28.4	0.0	2.5
Later	25.0	14.8	46.3	42.6	27.8	42.6	0.9	0.0
Overall	26.3	20.9	46.6	41.9	26.6	36.6	0.5	0.5

Table 15: Comparing Mahalanobis and Euclidean spectrum scaling for matching on shape. Hair has been *included* in the match area.

8.3 Thin Plate Spline Distances

As previously noted in Section 7, the thin plate spline manipulation will produce an energy value, which reflects the degree of distorsion required by the image to bring it into correspondence with the reference points. This has the advantage of weighting the effects of moving landmarks. Rather than a uniform weighting, as the principal component system used, the effect of moving a point will be determined by the ratio of the displacement to the distance from the adjacent points (Bookstein 1991). Thus the inner features of the face, where there are many points, each varying by a relatively small number of pixels will be given a greater weight than the outer features, which are relatively unconstrained. This may then give a better decription of face recognition, which in humans is known to depend upon the inner features to a greater extent when the faces are known to the subject and so can be recognized across a variety of condtions (Ellis, Shepherd and Davies 1979).

Since the thin plate spline distance only depends upon the particular landmark-locations, and not on the image being a face, there is no need to use an ensemble to code the faces. Each model file was treated in the same way as those used for the shape-based Principal Component Analysis; the rigid spline was applied to remove scale and position effects but retaining aspect ratio. The thin-plate spline distance was then calculated between each cue and each target. These distances were treated in the same way as the Principal Component Analysis distances to yield both best-matches and measures of confidence.

8.3.1 Results

The results of this process, compared with the 25-eigenvector Principal Component Analysis-shape match are shown in Table 16, for the without-hair case. Recognition here for the thin plate spline measures is quite noticeably worse, both for hit rate and confidence. This may well reflect the difference in the treatment of failures of configuration in the two cases; while the Principal Component Analysis should ignore them and treat the uncodable shape as noise, the thin pate spline will give the appropriate, very large, energy needed to produce the folded shape.

	Hit				Miss			
	Clear		Just		Just		Clear	
	PCA	TPS	PCA	TPS	PCA	TPS	PCA	TPS
Immediate:	29.6	17.3	43.2	65.4	27.2	17.3	0.0	0.0
Variant:	17.3	13.6	60.5	44.4	22.2	42.0	0.0	0.0
Lighting:	18.5	13.6	42.0	42.0	38.3	43.2	1.2	1.2
Later:	15.7	9.3	40.7	46.3	42.6	44.4	0.9	0.0
Overall:	18.0	11.5	43.4	47.2	37.8	41.0	0.8	0.3

Table 16: Comparing Principal Component Analysis and Thin Plate Spline distances for matching. Match percentages from 351 trials. Mahalanobis distance for the Principal Component Analysis, hair has been *excluded* from the match area.

8.3.2 With hair

The results for the same comparison performed with the hair left on the faces is shown in Table 17. It is notable that while the recognition for the Principal Component Analysis increases relative to the without-hair case, that for the thin plate spline decreases. This probably reflects the greater scope for failures of configuration due to the variability of the hair and ear points.

	Hit				Miss			
	Clear		Just		Just		Clear	
	PCA	TPS	PCA	TPS	PCA	TPS	PCA	TPS
Immediate:	38.3	8.6	38.3	70.4	23.5	21.0	0.0	0.0
Variant:	27.2	8.6	45.7	65.4	27.2	25.9	0.0	0.0
Lighting:	23.5	4.9	51.9	59.3	24.7	35.8	0.0	0.0
Later:	25.0	4.6	46.3	55.6	27.8	38.9	0.9	0.9
Overall:	26.3	5.6	46.6	59.0	26.6	34.9	0.5	0.5

Table 17: Comparing Principal Component Analysis and Thin Plate Spline distances for matching. Match percentages from 351 trials. Mahalanobis distance for the Principal Component Analysis, hair has been *included* in the match area.

9 Combining Shape and Texture

9.1 Description

The results on shape and shape-free images suggest that both give reasonable recognition. The combination of the two measures may increase recognition significantly, if they are notably independent. Principal Component Analysis was carried out separately on the shape and shape-free images, with the 25 most variable shape components, but all the texture eigenfaces for histogram equalised images, both measures using the inner face. Independence was assessed by measuring correlations for the ranks of the distances between each cue and the *other* images in the gallery (this helped to avoid outlier effects).

Unfortunately, there is no very obvious way of combining such arbitrary parameters as the shape and texture eigenfaces or distances. There is no reason why these should be given any particular weighting, or summed in any particular way. It was decided first that the shape and texture would be given an equal weighting, by normalising the cue-gallery distances so that both the two measures summed to one across the gallery. Three methods of distance-combination were then selected. The first was simply to multiply the two distances from the cue to each image in the gallery together. This had the effect of given equal weight to the two distances and if either was small, so was the resultant distance. A second method of combination was to add the values, taking the root square mean. This has the effect of emphasising whichever of the distances is the greater, so faces with large distances on one measure but small ones on the other were treated as being distant.

A third measure was chosen as being psychologically relevant; Vokey and Read (1992) suggest that a shape-based measure is used to make a familiarity decision, before identity is determined on the basis of texture. In this case, the shape-distances were ranked and the half with greater distances excluded before the image with the smallest texture distance was taken as being recognized. The logic was that if the shape and texture measures were uncorrelated, removing long-distance shape images would either remove an incorrect best texture match, or clse second-best match, thus allowing more confident recognition. The true match should be relatively good on both measures, and so should not be affected by the removal of the bad shape-matches.

9.2 Results

The average Spearman rank correlations (weighted as shown in Table 1) are provided in Table 18 and show that the correlations are positive but modest. This suggest that shape and texture describe dis-similar properties; the positive correlation may reflect a tendency for faces to be extreme in both measures.

Immediate	Variant	Lighting	Later
0.4349	0.4036	0.2898	0.2900

Table 18: Correlation on match rankings based on Mahalanobis distance for shape and texture from 351 trials. Hair has been *excluded* from the match.

A comparison of the effects of combining data by the methods of root mean square and multiplication is given in Table 19. It should be noted that these results are comparable with Table 2 but with the addition of linearised texture and also shape information. It is apparent here that the root mean square method is notably better than multiplication. The sequencial shape-sorting was dramatically worse with no correct clear hits.

	Hit				Miss			
	Clear		Just		Just		Clear	
	RMS	Mul	RMS	Mul	RMS	Mul	RMS	Mul
Immediate	86.4	63.0	12.3	35.8	1.2	1.2	0.0	0.0
Variant	56.8	38.3	39.5	58.0	3.7	3.7	0.0	0.0
Lighting	22.2	24.7	69.1	60.5	8.6	14.8	0.0	0.0
Later	31.5	21.3	57.4	68.5	11.1	10.2	0.0	0.0
Overall	38.1	28.3	53.2	62.2	8.7	9.5	0.0	0.0

Table 19: Comparing root mean square and multiplation methods of combining distances for Shape-and-Texture Principal Component Analysis for matching. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* from the match area.

This advantage for the root mean square combination reflects the need to combine relatively uncorrelated data. The consequences of combining values, which like the possible distances here, vary between 0 and 1 by root mean square and multiplication is shown in Fig. 12. This shows that the root mean square distance, which is generally greater than the multiplication, is most notably greater when the two scores diverge. If either is large the multiplication treats them as both small, while root mean square treats them as both big. This will have two consequences; firstly it makes it easyer to exclude faces, a unusually low value for one factor will not cause the other to be treated as spurious. Secondly, it will increase the range of values possible, relative to the multiplication method. This will a greater separation fo the data and thus more confident recognition.

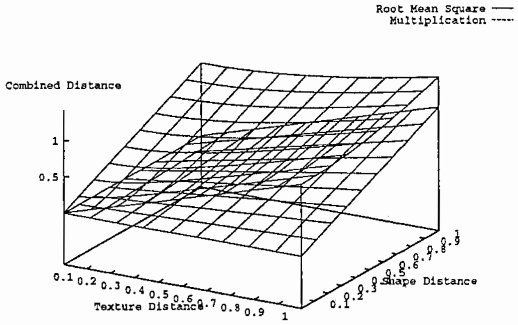

Figure 12: Variation in resultant distance for root mean square and multiplication combinations for a variety of shape and texture distances.

9.2.1 With hair

The results we have presented so far have concentrated on the face area alone as in Fig. 4; rather than those shown in Fig. 5. These latter correspond more to images often used to test for recognition; we give in Table 20 the rates comparable to those in Table 5. They are here primarily because it seems essential somewhere in such a paper to claim recognition rates of at least 95%!

	Hit		Miss	
	Clear	Just	Just	Clear
Immediate	98.8	1.2	0.0	0.0
Variant	93.8	2.5	3.7	0.0
Lighting	86.4	8.6	4.9	0.0
Later	62.0	32.4	5.6	0.0
Overall	74.6	20.8	4.6	0.0

Table 20: Match percentages from 351 trials. Shape and texture combined by root mean square, matching with Mahalanobis distance. Hair has been *included* in the match.

As usual, the pattern of results is the same whether or not the hair is included in the images. This leads to rather greater confidence in the results.

10 Caricaturing

10.1 Description

These methods of improving recognition are all concerned to produce a greater consistency in the data. This can be done by removing noise, as do the grey-level pre-processing methods and the shape-standardisation methods; these all attempt to remove variation between the images, so they can be compared more cleanly. Alternatively, additional information has been added in a form which would otherwise be lost, as with the shape and texture combination. However, all these cases are relatively neutral with regard to the specific Principal Component Analysis representation used, with the exception of the use of the Mahalanobis distance, which merely shows that it is an appropriate one.

An alternative method is to capitalise on the presence of an norm-based coding scheme. Recognition here is not dependent upon the absolute distance between the cue and target, but rather on the ratio of distances between these two and the other gallery images. Thus a method of altering position of the cue with the face-space will enhance recognition if it moves the cue away from the distractors, even if it also leaves the target. The obvious direction to move the cue is away from the origin without altering the angle of the vector; this is the one fixed point in the space. This can be done by manipulating the image in image-space; the pixel-values or feature-locations are moved away from the mean by a fixed (but variable) preportion of their pre-existing distance form the mean. This *caricaturing* technique has the effect of emphaising those aspects of the face which are already distinct from the average and is independent of the particular principal component system in which the faces are then coded.

10.2 Caricaturing and Human Recognition

As a technique, caricaturing derives from the study of the psychology of face recognition, where it has also been used to consider the presence of norm-based coding. Anti-caricatures, where the deviation from the norm is less than in the veridical image, are more difficult to recognize than the original image, while caricaturing leads to faster and more accurate recognition of faces, for both line-drawings (Rhodes et al. 1987), and also warped grey-scale images (Benson and Perrett 1991, 1994). Images manipulated in these ways also show effects of distinctiveness (so that those which are already unusual show larger caricature effects) and of expertise (Rhodes and McLean 1990). However, it should be noted that while there is typically a caricature effect (so that caricatured images are easier to recognize than equally-distorted anti-caricatures), caricature advantages, where recognition is better for the caricature than for the veridical image are typically only found in the case of line-drawings. Where grey-level caricatures do show caricature advantages, the degree of distorsion permissibly before recognition starts to decline is notably lower than for line drawings.

Once a shape-free manipulation has been performed, there is no reason why the grey-levels of the images should not also be caricatured, so that parts of the face lihter or darker than the average are made more extreme. However, this does not seem to have been performed, so there are no psychological data on the presence of a grey-level caricature effect. It should be noted that it is unlikely that there would be a very positive effect with texture caricatures on the outer part of the head and the hair, since there are likely to be large differences in the colour and the hair, and also because the variation in length will ensure that both hair and skin will be considered at the same location.

10.3 Shape Caricaturing

Since in theory, the application of a caricature distorsion should not alter the eigenfaces or eigenvectors extracted from an ensemble, the Principal Component Analysis was set up in the usual manner, with ensemble, gallery and the two response criteria derived from veridical images. An additional reason for adopting this stance was because one aim of the exercise was to reproduce the type of results found in humans, with the assumption that caricaturing is a relatively artificial manipulation, which taps aspects of facial recognition, but is not a process which occurs in people. This also avoids problems with the alteration of response criteria as the caricaturing percentage is changed. This will alter as the mean distance from face to face increases with increased caricaturing, and causes difficulties in interpreting results.

10.3.1 Results

The alteration in the two response criteria, t_1 and t_2 as the shape of the faces are caricatured can be seen in Fig. 13. Note that the abcissa is a logarithm-transfrom of the caricature percentage. This reflects the unequal range on the two sides of the two sides of the 100%, veridical image line as caricaturing is controled by a multiplicative scale. This allows comparison of equivalent distorsions on either sides of the veridical line and thus the determination of the presence of caricature effects as well as caricature advantages.

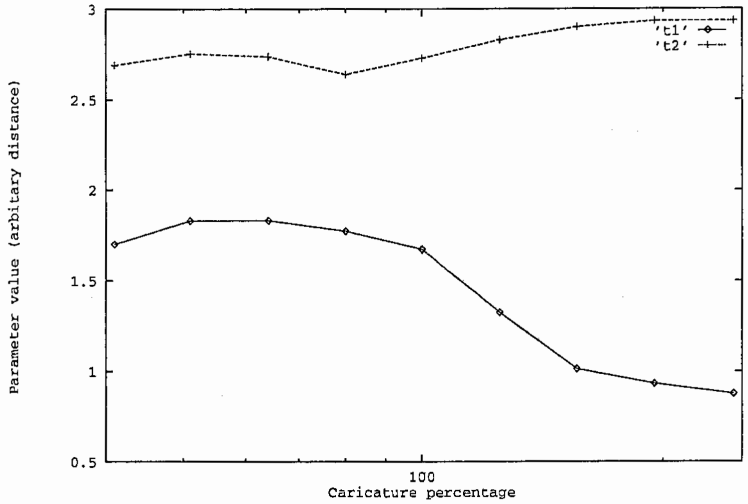

Figure 13: Variation in the response criteria as a function of caricaturing.

In particular, at least for positive caricatures, t_1 decreases, while t_2 stays constant or increases. This can be understood if the faces are considered as points on the surface of a sphere, the radius of which is increased as the faces are caricatured (given the number of dimensions, there is little difference between a sphere and a ball). Since the distances upon which the recognition judgements are made are on the surface, they will be affected by changes in area. Thus doubling the radius will increase the distances by 2^{n-1}, where n is the dimensionality of the current space. Thus already large distances will be increased at a far higher rate than small distances and so μ will increase faster than σ which will increase faster than $c_1 - c_0$.

This effect will interact with the changes in distance between the cues and targets on the recognition test, and our justification of this method of calculating the response parameters was that it reflected decision-making at the original encoding of the face into the gallery, which was not caricatured. To provide a closer fit between the model and the conception of the psychological situation, it sees best to use response criteria taken from the veridical images. This obviously will not affect the total recognition, but may well change the level of caricaturing which best supports confident recognition.

As a first step, the with-hair shape-images were caricatured; this is the traditional starting point for the use of caricatures with humans. The results are shown in Fig. 14, and show that peak confident recognition was obtained for a 156% caricature. This compares quite well with the value obtained with people.

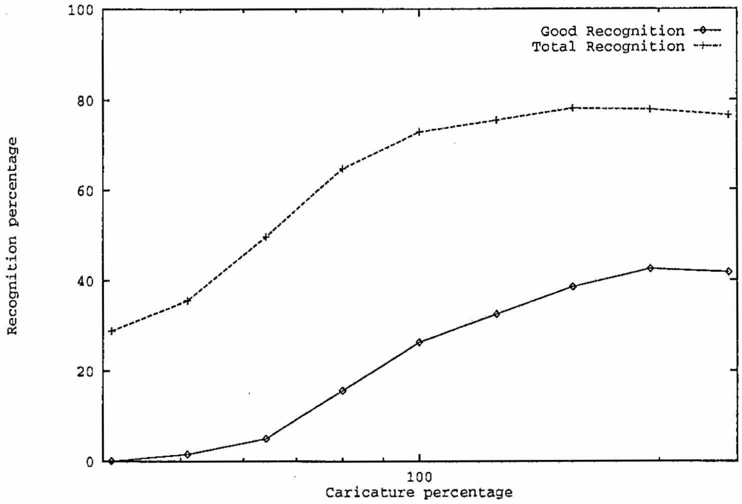

Figure 14: Variation in the recognition as a function of caricaturing, with response criteria taken from veridical images. Match percentages from 351 trials. Mahalanobis distance, hair has been *included* in the match area.

The full results for the 156% caricature are shown in Table 21. This shows that caricaturing consistently improves both the ordering and separation of the data.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Ver	Car	Ver	Car	Ver	Car	Ver	Car
Immediate	38.3	50.6	38.3	32.1	23.5	17.3	0.0	0.0
Variant	27.2	37.0	45.7	43.2	27.2	18.5	0.0	1.2
Lighting	23.5	39.5	51.9	38.3	24.7	22.2	0.0	0.0
Later	25.0	36.1	46.3	40.7	27.8	23.1	0.9	0.0
Overall	26.3	38.5	46.6	39.6	26.6	21.8	0.5	0.1

Table 21: Comparing veridical and 156% caricature Shape-Principal Component Analysis for matching. Match percentages from 351 trials. Mahalanobis distance, hair has been *included* in the match area.

10.4 Texture Caricaturing

These tests were performed only upon measures of the *shape* of the face. However is equally possible to caricature the *texture* or pigmentation of the face. Here, all that need be done is the prototype, average image be calculated from the texture-mapped faces, and the deviation of the luminance of each pixel increased by the caricaturing percentage. The face can then recognized in the usual way.

One problem with this procedure is that it is based upon the assumption that all the corresponding pixels of the processed images will have the same meaning with regard to the faces. Although the texture-mapping will ensure that this is generally true, this will not be the case for the outer face, in particular the fore-head and ears. This stems from the imprecision of the definition of the landmarks in this region; the range of hair-lengths in the images ensure that both the forehead and ears of the average image are a blurred mixture of dark and light pixels. Including these regions in the caricatured image may well have a negative (or at least unrepresentative) effect upon recognition as it will emphase a aspects of the face which are not truly distinctive. Thus these texture caricatures were carried out upon the inner face, without the hair and ears.

10.4.1 Results

The same tests as for the shape images were performed, with a constant pair of criteria, set from the veridical images. The results are shown in Fig. 15; again the optimal recognition occurs at about 156%, but this only really applies to the good-recognition line, the total hit-rate declines from the 100% mark although there is still a large caricature effect.

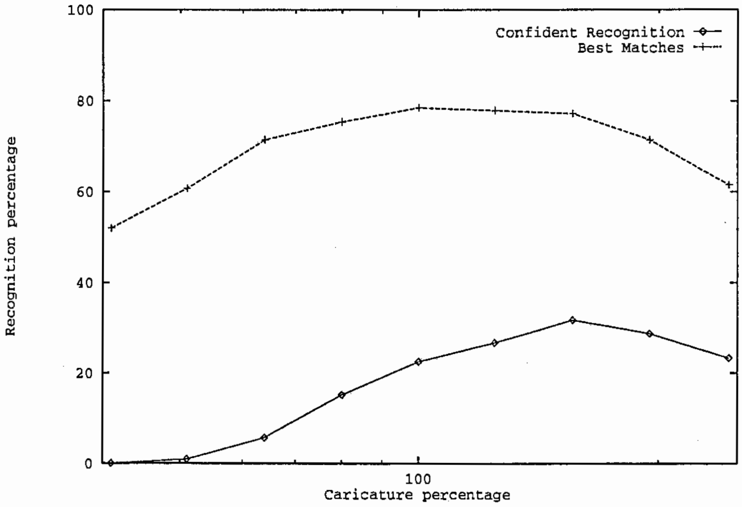

Figure 15: Hit-rates and good-hits rates as a function of caricaturing for texture-caricatures with constant response criteria. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* from the match area.

The full results for the 156% caricatures are given in Table 22 in comparison with the veridical results. It is notable here that although the separation has consistently been increased, the number of false alarms has increased sufficiently in the 'Lighting' condition to increase the total number of errors. This presumably reflects the use of the Condition 1, ensemble, average to caricature against. The large lighting differences between these images causes the caricaturing to change both the amplitude and angle of the vector. This may suggest that caricaturing of texture-images is not advisable, or alternatively suggest that the caricaturing should be done after lighting-normalisation.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Ver	Car	Ver	Car	Ver	Car	Ver	Car
Immediate	84.0	87.7	16.0	12.3	0.0	0.0	0.0	0.0
Variant	34.6	45.7	56.8	46.9	8.6	7.4	0.0	0.0
Lighting	8.6	17.3	59.3	42.0	32.1	40.7	0.0	0.0
Later	13.9	24.1	62.0	52.8	24.1	23.1	0.0	0.0
Overall	22.5	31.7	56.0	45.5	21.6	22.8	0.0	0.0

Table 22: Comparing veridical and 156% Texture-Principal Component Analysis for matching with shape-free images. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* from the match area.

10.5 Shape and Texture Caricaturing

Since both shape and texture give positive caricature effects, it is reasonable to caricature them together. However, this is a practice which seems to have been used with people, so there are no obvious comparisons with human data.

10.5.1 Results

As a consequence of the difficulties in caricaturing outer faces, the first comparison was the recognition of shape and texture caricatured inner-faces. The processing was performed in the same way as before, and the normalised distances were combined by the root mean square method, since this had already been shown to be the best. The results are shown in Fig. 16 and show that both the good and total recognition peak at about 156%, as would be expected from the individual results.

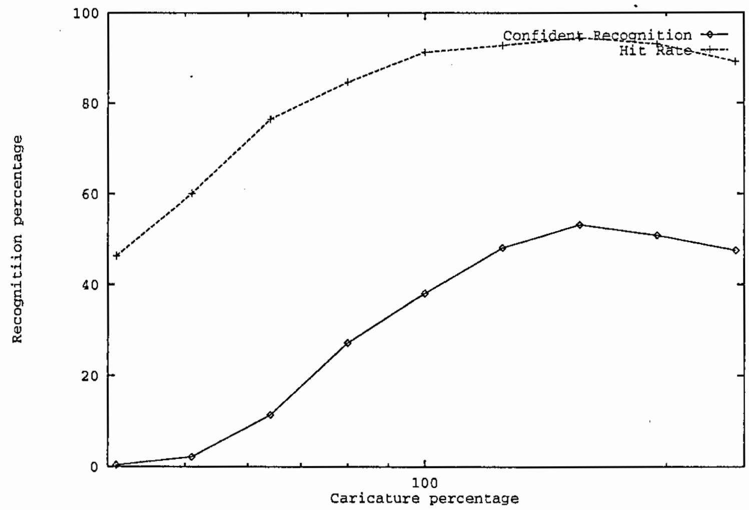

Figure 16: Hit-rates and good-hits rates as a function of caricaturing for combined shape and texture caricatures with constant response criteria. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* in the match area.

The full results, comparing the veridical and 156% caricature images are shown in Table 23. This shows a slight decrease in false alarm rate and a large increase, of 13.5% increase in the clear recognition rate. It is also notable that this effect is consistent across the image-classes. However, it is possible that the lighting effect on the texture has reduced total recognition for the caricatured images in the later image conditions.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Ver	Car	Ver	Car	Ver	Car	Ver	Car
Immediate	86.4	96.3	12.3	3.7	1.2	0.0	0.0	0.0
Variant	56.8	69.1	39.5	30.9	3.7	0.0	0.0	0.0
Lighting	22.2	37.0	69.1	56.8	8.6	6.2	0.0	0.0
Later	31.5	48.1	57.4	44.4	11.1	7.4	0.0	0.0
Overall	38.1	53.2	53.2	41.3	8.7	5.6	0.0	0.0

Table 23: Comparing veridical and 156% caricature Shape and Texture Principal Component Analysis for matching. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* from the match area.

An alternative comparison involves the use of veridical texture images combined with caricatured shape images. This procedure is essentially the same as that used by Benson and Perrett (1993), and may thus be expected to give slightly lower value for optimal recognition. Given the problems with supplying an appropriate average image for the texture-caricature images, this may also be a more appropriate combination method for enhancing recognition. The results for this combination with the hair excluded are shown in Fig. 17 and show that although the overall caricature effect is fairly minor, there is still a notable effect for the confident recognition and the peak is still at about 156%.

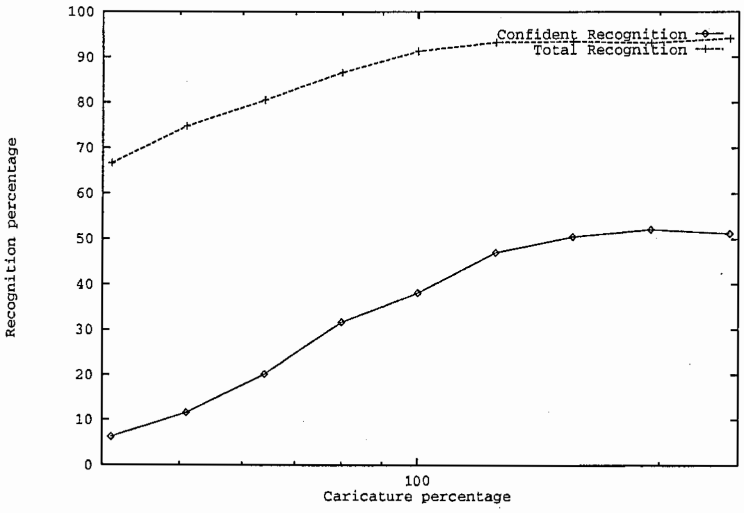

Figure 17: Hit-rates and good-hits rates as a function of caricaturing for shape with veridical texture and constant response criteria. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* from the match area.

The detailed results, comparing the two combination methods given in Table 17. The recognition for the veridical images for the shape-caricatured images is identical with that for the shape-and-texture caricature images as would be expected, while recognition is marginally worse for the shape-caricatured images compared with the shape-and-texture caricatures. It is is notable however, that this difference is dis-preportionate for the 'Miss' column, reflecting the somewhat larger effect of caricaturing upon the separation of the distances rather than the ordering. This may also reflect the the typical finding that caricature effects are predominately seen in reaction-time data when full grey-scale images are manipulated.

	Hit				Miss			
	Clear		Just		Just		Clear	
	S&T	S	S&T	S	S&T	S	S&T	S
Immediate	96.3	97.5	3.7	2.5	0.0	0.0	0.0	0.0
Variant	69.1	63.0	30.9	37.0	0.0	0.0	0.0	0.0
Lighting	37.0	37.0	56.8	55.6	6.2	7.4	0.0	0.0
Later	48.1	47.2	44.4	43.5	7.4	9.3	0.0	0.0
Overall	53.2	52.1	41.3	41.0	5.6	6.9	0.0	0.0

Table 24: Comparing 156% caricatures for Shape and Texture Principal Component Analysis and Shape with veridical Texture Principal Component Analysis for matching. Match percentages from 351 trials. Mahalanobis distance, hair has been *excluded* from the match area.

10.5.2 With Hair

The effects of shape and texture caricaturing on the whole face is given in Table 25, giving the veridical and 150% caricature results. Somewhat unexpectedly, this fails to show a very large effect of caricaturing. While the number of confident hits is increased by approximately 5.6% the total error rate actually increases. While it is possible that 150% actually overshoots the point of maximum caricature advantage, it is notable that a large part of the extra errors stem from the 'later' category, probably suggesting that texture caricaturing is even less sensible over long periods when the hair is included. This is understandable if the differences in local sign for the ear/hair and forehead/hair areas is considered.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Ver	Car	Ver	Car	Ver	Car	Ver	Car
Immediate:	98.8	100.0	1.2	0.0	0.0	0.0	0.0	0.0
Variant:	93.8	96.3	2.5	2.5	3.7	1.2	0.0	0.0
Lighting:	86.4	80.2	8.6	16.0	4.9	2.5	0.0	1.2
Later:	62.0	73.1	32.4	20.4	5.6	6.5	0.0	0.0
Overall:	74.6	80.0	20.8	15.3	4.6	4.4	0.0	0.3

Table 25: Comparing veridical and 150% caricature Shape and Texture Principal Component Analysis for matching. Match percentages from 351 trials. Mahalanobis distance, hair has been *included* in the match area.

The final comparison involves the use of veridical whole-face texture images combined with caricatured shape images. The results are shown in Fig. 18, and show that although there is no caricature advantage for the false alarm rate, there is a considerable one for the confident recognition rates. This reflects the common result that full grey-scale shape-caricature effects are generally only seen in response time measures. However, the optimal value still appears to be about 156%and the results are generly very like those for the same comparison without the hair, but with higher recognition rates.

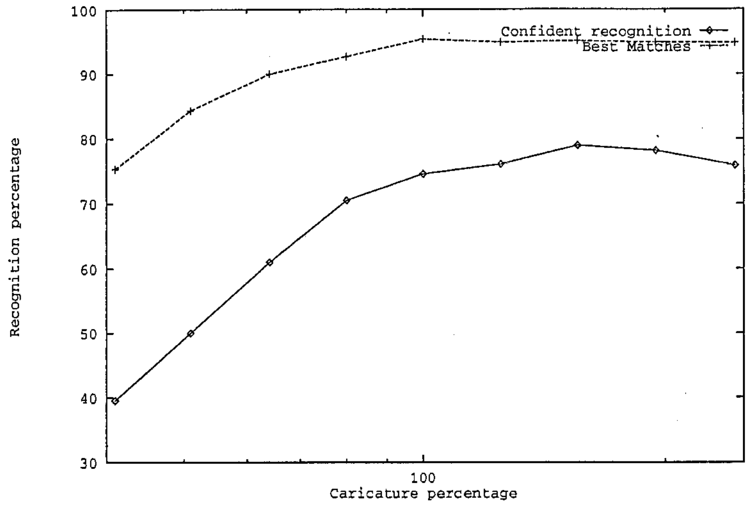

Figure 18: Hit-rates and good-hits rates as a function of caricaturing for shape with veridical texture and constant response criteria. Match percentages from 351 trials. Mahalanobis distance, hair has been *included* in the match area.

The detailed results for the 156% caricature are shown in Table 26, showing that again the hit-rate has fallen as a consequence of the caricaturing, even if the confident hit-rate has risen. There are two probable causes for this lack of effect; the first is that recognition is simply at or about ceiling; typically the veridical miss-rate is under 5%. Under these conditions, there are very few manipulations which will have a positive effect upon the data. The second reason is that the large variance of the outer-head points and the ambiguity of their meaning ensures that the points will be moved a very great deal, and thus may loose their status as 'reasonable exagerations' rather quickly.

On a related point, it should be noted that the caricature effect here has a maximum at a rather larger degree of caricaturing than is typically found in human. Benson and Perrett (1991) claim a value of approximately 106% for the maximum point for shape-caricatured, full-texture images. However, these images have been caricatured against the 'true average'; the average image of the ensemble. Hence caricaturing will not alter the angle of the face-vector, only the magnitude. This would not be true if some other set of faces were used to derive a prototype, and it is notable that Rhodes and McLean (1990) found smaller effects of caricaturing (for line drawings of birds) when an inappropriate example was used as a prototype. Observations here also suggest a reduction in the level of caricaturing at the point of inflection if an image other than the true average is used as the prototype. It may well thus be the case the results here reflect the 'true' position, while those commonly reported are distorted by the atypicality of the faces used to derive the average.

	Hit				Miss			
	Clear		Just		Just		Clear	
	Ver	Car	Ver	Car	Ver	Car	Ver	Car
Immediate:	98.8	100.0	1.2	0.0	0.0	0.0	0.0	0.0
Variant:	93.8	96.3	2.5	2.5	3.7	1.2	0.0	0.0
Lighting:	86.4	87.7	8.6	9.9	4.9	2.5	0.0	0.0
Later:	62.0	68.5	32.4	24.1	5.6	7.4	0.0	0.0
Overall:	74.6	79.0	20.8	16.1	4.6	4.9	0.0	0.0

Table 26: Comparing veridical and 156% caricatured Shape and veridical Texture Principal Component Analysis for matching. Match percentages from 351 trials. Mahalanobis distance, hair has been *included* in the match area.

11 A Theoretical Model

We have already described the implementation of Principal Component Analysis based coding and shown that in practice, its performance is worth further investigation. In this section we try to place the use of Principal Component Analysis in a theoretical context, and give a model within which the potential advantages of the coding can be discussed.

At a fundamental level, the problem of machine based recognition can be considered as one of obtaining suitably invariant descriptions. It is thus natural to consider the process using language developed precisely for the study of such invariants. We are thus led to consider a set of face instances, together with transformations on this set which preserve identity. Allowing time to run backwards if necessary, this collection of transformations may be considered as a (not necessarily abelian) group, and the objects under study, faces, are the equivalence classes under this action. The nature of such a quotient object is essentially the same when face images are considered as primitives; the only change being the need to include the variation in imaging conditions in the set of allowable transformations. Such a structure is probably most usefully considered as some form of manifold modelled on a Hilbert space, while the identity preserving transformations should be considered as smooth. This then leads to a manifold structure for the quotient space; recognition is then the problem of deciding whether two points in this identity space coincide, something which of course can be done locally.

It is thus natural to look for appropriate localisations or charts, whose effect is then to give a local parameterisation. Although at this level of generality there is no formal way to construct these, a property of any such chart is that its range allows both local deformations and local averages within the class of faces: thus we seek a coding scheme which enables a pair of (similar) faces to be averaged to yield another face, and also allows small deformations of the parameters to yield another face. Of course one candidate consists of local sections back into the space of images, and one can thus consider the process of image pixelisation as a potential chart. However it is clear that this fails both the "local averages" and "small perturbations" test. Indeed these requirements occur in practice when attempting to morph faces for use in films or advertising; an appropriate coding is precisely one consisting of a shape vector together with the texture information. One is thus led to the hypothesis that this coding is natural for the problem, and should be considered as a chart on the "face manifold" (Craw and Cameron 1991).

From this viewpoint, the use of eigenfaces can be placed in its normal context. Such a chart maps into a (possibly infinite dimensional) Hilbert space, but in applications this must be approximated. Simply replacing an image by the corresponding discrete set of pixel values does much of this approximation, but there is still need for further dimensionality reduction. And now we are in a Euclidean space in which Principal Component Analysis is meaningful as a rotation and rescaling of the underlying axes; we thus recognise the application simply as one of approximation.

We are thus led to argue from the theoretical viewpoint that the dissociation of shape and texture is part of the expected structure of the problem, while the use of eigenfaces arises naturally as an approximation process. Existing eigenface based recognition techniques in which face images are first normalised in some way, such as co-locating eyes can then be seen as a first approximation to observing this dissociation.

12 Conclusions

We have attempted to show that a greater consideration of the nature of Principal Component Analysis yields advantages in recognition. Doing so, and moving from affine normalised images to the combined configuration and texture images produces a three-fold increase in performance. This advantage is particularly true of the most difficult conditions; a more than five-fold increase in the number of clear hits has been achieved without adding extra information. Rather, the configural information which appears to overshadow the texture in the affine normalised 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. Further increases in recognition can be gained by caricaturing the images, and the presence of this factorisation allows the ready explanantion of a range of psychological effects.

The shape vector is obviously impoverished relative to the texture. Adding points to the shape model may both enhance the shape-based recognition and also, by more accurately linearising the space, texture-recognition. Certainly commercially executed face-morphs use many more points to define shape. However, the very quality of the shape-free recognition suggest it is possible to get reasonable morphs with this model. The finding that flipping the ensemble has such a large positive effect upon recognition, defies obvious explanation. We are unable to see why, as Table 3 shows, 50 eigenfaces from 25 faces and their reflections allow better recognition than 50 eigenfaces from 50 faces. It may be that the presence of the reflected images enables faces which are asymmetrical or rotated to be coded more accurately, since the ensemble will have no preference for deviations in one direction. This awaits investigation, by varying the type and size of ensemble used. One particular form of asymmetry suggested by inspection of our images is a possible horrizontal lighting gradient; the ensemble images are slightly brighter on one side than the other. This may suggest that reflecting the ensemble is worthwhile as a general method of increasing lighting invarience by adding precisely equal, differing groups.

The clear advantage for Mahalanobis distance over Euclidean distance, consistent across conditions, provides evidence that Principal Component Analysis 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.

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 are not advocating it as a universal code; low level problems such as occlusion prove difficult for eigenfaces, while the extraction of the required configural information seems most easily done using distortion methods such as in (Lades et al. 1993), which themselves require biologically plausible operators to preprocess the image. This suggests that psychological implications of this work are late in the processing chain, when the face is being considered as a whole, and as such explain why we had most success with the simpler preprocessing methods.

References

Benson, P. J.: 1994, Morph tarnsformation of the facial image, *Image and Vision Computing* **12**, 691–696.

Benson, P. J. and Perrett, D. I.: 1991, Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images., *European Journal of Cognitive Psychology* **3**(1), 105–135.

Benson, P. J. and Perrett, D. I.: 1993, Extracting prototypical facial images from examplars, *Perception* **22**(3), 257–262.

Benson, P. J. and Perrett, D. I.: 1994, Visual processing of facial distinctiveness, *Perception* **23**, 75–93.

Bookstein, F. L.: 1989, Principal warps: Thin-plate splines and the decomposition of deformations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11**, 567–585.

Bookstein, F. L.: 1991, *Morphometric Tools for Landmark Data: Geometry and Biology*, Cambridge University Press, Cambridge UK and New York.

Brunelli, R. and Poggio, T.: 1993, Face Recognition: Features versus Templates, *IEEE: Transactions on Pattern Analysis and Machine Intelligence* **15**(10), 1042 – 1052.

Choi, C. S., Okazaki, T., Harashima, H. and Takebe, T.: 1990, Basis generation and description of facial images using principal component analysis, *Technical Report of IPSJ:Graphics & CAD* **46**(7), 43–50. in Japanese.

Costen, N. P.: 1994, *Spatial Frequencies and Face Recognition*, PhD thesis, University of Aberdeen, Aberdeen, Scotland.

Craw, I. and Cameron, P.: 1991, Parameterising images for recognition and reconstruction, *in* P. Mowforth (ed.), *British Machine Vision Conference 1991*, Springer Verlag, London, pp. 367–370.

Craw, I. and Cameron, P.: 1992, Face recognition by computer, *in* D. Hogg and R. Boyle (eds), *British Machine Vision Conference 1992*, Springer-Verlag, pp. 498–507.

Craw, I., Costen, N., Kato, T., Robertson, G. and Akamatsu, S.: 1994a, Automatic face recognition: Combining configuration and texture, Submitted to IWAFGR95, Zurich.

Craw, I., Ellis, H. D. and Lishman, J. R.: 1987, Automatic extraction of face-features, *Pattern Recognition Letters* **5**(2), 183–187.

Craw, I., Kato, T., Costen, N. and Robertson, G.: 1994b, Methods for improving principal component analysis coding of facse for recognition: A testbed, *Technical Report TR-H-104*, ATR — Advanced Telecommunications Research Institute International, 2-2, Hikaridi, Seika-cho, Soraku-gun, Kyoto 619-02, Japan.

Craw, I., Tock, D. and Bennett, A.: 1992, Finding face features, *in* G. Sandini (ed.), *Proceedings of ECCV-92*, number 588 in *Lecture Notes on Computing Science*, Springer-Verlag, pp. 92–96.

Edelman, S., Reisfield, D. and Yeshurun, Y.: 1992, Learning to recognise faces from examples, *in* G. Sandini (ed.), *Proceedings of ECCV-92*, number 588 in *Lecture Notes on Computing Science*, Springer-Verlag, pp. 787–791.

Ellis, H. D., Shepherd, J. W. and Davies, G. M.: 1979, Identification of familiar and unfamiliar faces from internal and external features: some implications for theories of face recognition, *Perception* **8**(4), 431–439.

Jolliffe, I. T.: 1986, *Principal Component Analysis*, Springer-Verlag, New York.

Jungman, N., Levi, A., Aperman, A. and Edelman, S.: 1994, Automatic classification of police mugshot album using principal component analysis, *SPIE Conference on Applications of Neural Networks, Orlando*, pp. 591–594. SPIE-2243.

Kanade, T.: 1977, *Computer Recognition of Human Faces*, Vol. 47 of *Interdisciplinary Systems Research*, Birkhäuser, Basel,Stuttgart.

Kelly, M. D.: 1971, Edge detection in pictures by computer using planning, *in* B. Meltzer and D. Michie (eds), *Handbook of research on face processing*, Edinburgh University Press, Edinburgh, pp. 397–409.

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**(1), 103–108.

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.

Lades, M., Vorbrüggen, J. C., Buchmann, J., Lange, J., v.d. Malsburg, C., Würtz, R. P. and Konen, W.: 1993, Distortion invariant object recognition in the dynamic link architecture, *IEEE Transactions on Computers* **42**(3), 300–311.

Lanitis, A., Taylor, C. J. and Cootes, T. F.: 1994, An automatic face identification system using flexible appearance models, *in* E. Hancock (ed.), *British Machine Vision Conference 1994*, BMVA Press, pp. 65–74.

O'Toole, A. J., Deffenbacher, K. A., Valentin, D. and Hervé, A.: 1994, Structural aspects of face recognition and the other race effect, *Memory and Cognition* **22**(2), 208–224.

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.

Rhodes, G. and McLean, I. G.: 1990, Distinctiveness and expertise effects with homogeneous stimuli: Towards a model of configural coding, *Perception* **19**, 773–794.

Rhodes, G., Brennan, S. E. and Carey, S.: 1987, Identification and ratings of caricatures: Implications for mental representations of faces, *Cognitive Psychology* **19**, 473–497.

Robertson, G. and Craw, I.: 1994, Testing face recognition systems, *Image and Vision Computing* **12**, 609–614.

Shackleton, M. A. and Welsh, W. J.: 1991, Classification of facial features for recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-91)*, pp. 573–579.

Shepherd, J. W.: 1986, An interactive computer system for retrieving faces, *in* H. D. Ellis, M. A. Jeeves, F. Newcombe and A. Young (eds), *Aspects of Face Processing*, Martinus Nijhoff, Dordrecht, chapter 10, pp. 398–409. NATO ASI Series D: Behavioural and Social Sciences - No. 28.

Sirovich, L. and Kirby, M.: 1987, Low-dimensional procedure for the characterization of human faces, *Journal of the Optical Society of America* **4**, 519–524.

Stonham, T. J.: 1986, Practical face recognition and verification with WISARD, *in* H. Ellis, M. Jeeves, F. Newcome and A. Young (eds), *Aspects of Face Processing*, Martinus Nijhoff, Dordrecht, pp. 426–441.

Turk, M. and Pentland, A.: 1991, Eigenfaces for recognition, *Journal of Cognitive Neuroscience* **3**(1), 71–86.

Ullman, S.: 1989, Aligning pictorial descriptions: An approach to object recognition, *Cognition* **32**, 193–254.

Vokey, J. R. and Read, J. D.: 1992, Familiarity, memorability, and the effect of typicality on the recognition of faces, *Memory and Cognition* **20**(3), 291–302.

Watt, R.: 1994, A computational examination of image segmentation and the initial stages of human vision, *Perception* **23**, 383–398.