Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International 179 (2008) 211–218
Forensic Science International
www.elsevier.com/locate/forsciint

Exact solutions to Bayesian and maximum likelihood problems in facial identification when population and error distributions are known

Rory J. Allen *

Department of Psychology, Goldsmiths College, New Cross, London SE14 6NW, UK

Received 9 November 2007; received in revised form 21 May 2008; accepted 28 May 2008
Available online 17 July 2008

Abstract

The reliability of traditional photogrammetric identification techniques using a small number of facial landmarks has recently come in for criticism. However, the transformation of parameters into a new face space in which the error distributions are orthogonal, yields a maximum likelihood solution to the problem of identifying a photographed face from a small, known, population which, in a simulated example, raises the success rate from 20% to 93%. A full transformation yielding simultaneously independent population and error distributions can be derived from raw population and error data using a straightforward computer procedure. Such a transformation facilitates computations for the situation where a single suspect is held in custody and the likelihood ratio of his being identical with a photograph is desired. It seems premature to condemn photogrammetry until the more efficient data-analysis approach outlined in this paper has been applied and tested.

Keywords: Forensic Science; Forensic photogrammetry; Facial identification; Principal component analysis; Bayesian analysis

1. Introduction

There have recently been some major advances in the application of advanced statistical methods in forensic science, such as the use of Bayesian Networks in quantitative and qualitative situations [1,2] and the application of Bayesian techniques to the analysis of manipulated evidence [3], to the inference of identity in speaker recognition [4], and to the analysis of fingerprint, face and signature evidence [5]. A novel and complex score-normalization technique, KL-Tnorm, was developed as an aid to automatic speaker recognition [6]. In addition, a group at the University of Edinburgh led by Professor Colin Aitken has carried out important and pioneering work in the application of multivariate analysis to the development of significance tests and likelihood ratios (LRs) for the assessment of trace evidence, such as glass fragments found at a crime scene and on a suspect [7].

This paper addresses one of this group of problems, namely that of identifying faces from photographs, such as stills taken from CCTV video footage. The traditional approach to this, with historical roots in the work of the French pioneer of anthropometry, Alphonse Bertillon, involves identifying a number of well-defined points on the image, such as the left and right ectocanthions, the stomion, and the nasion, and measuring the distances between them. These measurements may be standardized by dividing, for example, by the interpupillary distance, to produce a number of proportion indices, or as they are termed in Ref. [8], PIs; the angles between lines joining pairs of landmarks can also be measured [8,9]. A more complex technique involves principal component analysis of image pixels, such as in Ref. [5], where a combination of eigenfaces and fisherfaces involving 180 dimensions was used.

Both methods require an estimate of within-source variability, i.e. the extent to which images from a particular individual would tend to vary if the image were taken repeatedly; without such an estimate, there can be no certainty as to the range within which the true values of the parameters for that individual may lie, and so the degree of confidence to be placed in any identification is impossible to evaluate. Good results were obtained in Ref. [5] from computing a minimum variance estimate from the mean of all within-source variabilities in the database.

Kleinberg et al. [8] adopted the converse strategy. They conducted a series of tests in which they attempted to identify a

* Tel.: +44 207 9197225.
E-mail address: r.allen@gold.ac.uk.

0379-0738/$ – see front matter © 2008 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2008.05.010

video still of a subject by taking an individual PI measure from that photo, and comparing it with the PI measurements for a set of high-quality photos of 10 people including the subject, using a closest fit criterion. They demonstrated that even the best performing PI only identified the individual correctly in 25% of cases, and concluded that the accuracy of the method was too poor for use in the identification of criminals.

This paper examines the effectiveness of a multivariate approach to the use of PIs in identification, in a situation analogous to that in Ref. [8], and compares the results. It also looks at the use of Bayesian methods for the case where there is a single suspect and a photograph of a known perpetrator. This approach differs from that in Ref. [7] in that we focus less on the theoretical approach and more on providing practical details of how to convert a set of raw photogrammetric data into simultaneously orthogonalized population and error distributions with no assumptions of multivariate normality, using a commonly available statistical program (SPSS). These methods can be applied without too much effort to situations involving many variables (we have trialled them with 16 variables) whereas the methods examined in Ref. [7] were applied to only three variables of interest. We do, moreover, make a significant additional invariance assumption, spelt out in Section 2.2 below: the assumption seems plausible, but has not yet been rigorously tested.

The Bayesian approach is sometimes taken to be synonymous with the provision of likelihood ratios. In the words of Ref. [5]: ''the Bayesian approach provides ... results in the form of likelihood ratios (LR) from the forensic laboratory to court''. We conform with this approach to the extent that we provide means of calculating LRs for face identification. The approach has the advantage of being philosophically uncontroversial and mathematically precise. We also point out, however, that it is at least theoretically possible to imagine circumstances in which the bald provision of an LR without qualification could be misleading. Our reason for confining discussion to LRs and not attempting to apply Bayes' theorem to the calculation of posterior probabilities, is that this requires an estimate of prior probability. This can be a highly controversial area, and one in which agreement, especially in an adversarial forensic context, is highly unlikely. Nevertheless, it seems worthwhile to point out that in the event that a suspect is apprehended in the neighbourhood of a crime and subsequently found to closely resemble a perpetrator, conclusions as to his guilt are likely to be much better founded if his apprehension was independent of the identification than if, for example, his arrest followed a trawl of a digitized photographic database to find the individual with best fit, even though the LRs in these two hypothetical instances might be identical.

2. Methodology

2.1. Assumptions about the basic system

The situation envisaged here has three basic components. Firstly, there is an agreed system of PIs based on facial landmarks, which can be measured by an operator from a still photograph. (The term ''PI'' is used for convenience to include any facial measurement, if necessary both distances and angles involving landmark points, which is the sense in which it is used in Ref. [8].) Secondly, the assumption is made that there is a source of background data, consisting of high-quality photographs of a large number of members of a particular population, which can be measured accurately and their PIs determined (as might be the case for example if there is a national, digitized database of passport photographs taken according to standard criteria). Thirdly, there will be a system, such as a security camera, in a particular location, which is the source of forensic data.

The first difficulty is to estimate the within-source variability of the camera system. It is envisaged employing a purely empirical method for estimating these errors, involving the calibration of a camera/operator system. This would be complicated though not impracticable; it could involve for example testing the system by repeatedly photographing volunteers having known PIs, using the photogrammetric measurement system in question to obtain a scatterplot of points within the multidimensional space represented by the PI measurements as axes, and calculating the corresponding scatterplot of error vectors by subtracting the known accurate PI parameters for the particular volunteer in question, from those parameters estimated from the stills, to obtain an error reading for that calibration point. The variance due to operator error could be estimated (and reduced) if a number of operators were asked to measure the same photo repeatedly. The errors will in practice depend on the distance from the camera to the face, on the angle between the camera and the face and on lighting levels. However, in many practical cases, where a camera is covering a particular position (say the entrance to a building) from a particular angle and where security lighting is installed, it is likely that these factors will have little effect in comparison with other errors of measurement. The transformation outlined below, which is the first step in our method, does not incidentally depend on the calculation of parameters from this raw data; it can use the data in their raw form.

It is assumed in what follows that the system is unbiassed, and that the error distribution is the same for all faces in the population: photographing face ''A'' 100 times, say, will yield the same distribution or scatterplot of PI measurements around the mean values for face ''A'' as photographing face ''B'' round the mean for face ''B''. A second, more immediately plausible assumption is that the quality of PI data available in the population database is error-free in comparison with the error introduced by the camera/operator system, and can be ignored. This was not the assumption in Ref. [8], where all photos were of high quality, and where the variance arose from the target photograph being taken on a different occasion, and by a different camera, to the comparison photographs. However, if target photographs are taken on different occasions and compared with a fixed database of suspects, all the variance will be due to differences in the target images, and the analysis in this paper can still be applied.

2.2. Preliminary analysis of the basic system

Even before dealing with the analysis of any 'live' data, considerable work can be done on the population and error databases to enable the subsequent analysis to proceed more easily. There are two sets of random variables (RVs) to be considered. The first set is that of the values of the PIs for the faces in the general population. Each face gives rise to a set of n parameters, or a vector in the n-dimensional space with the PIs as the axes, and this vector can be considered as itself being an RV, giving rise to a scatterplot in the PI-space. Call this the population distribution in PI-space.

The second set of RVs consists of the errors in the PIs. This is also represented by a series of vectors in the n-dimensional PI-space, and in this case the scatterplot is, by assumption of zero bias, centred on the origin. Call this the error distribution in PI-space. This is available from the process of error calibration.

The assumption introduced above is that the error distribution is independent of which face is being measured. Suppose $\underline{\mathbf{v}}$ is a vector in the population distribution corresponding to a particular face. Imagine that this face is now measured repeatedly by the camera system in question. The distribution of face vectors obtained will be the set of vectors $\{\underline{\mathbf{v}} + \underline{\mathbf{e}}_i\}$, where $\underline{\mathbf{e}}$ is a random variable vector representing the error distribution. If $\underline{\mathbf{w}}$ is another face vector, the set of measurements that would be obtained from observing that face would of course be $\{\underline{\mathbf{w}} + \underline{\mathbf{e}}_i\}$. The key point is that even if the face vector changes, the distribution of the set of points $\{\underline{\mathbf{e}}_i\}$ remains the same. In what follows, the

vectors in the face space, both error vectors and face vectors, will be subjected to a linear transformation; by virtue of linearity, it will leave invariant the property just mentioned: if f is linear, it takes $\{\underline{\mathbf{v}} + \underline{\mathbf{e_i}}\}$ to $\{f(\underline{\mathbf{v}}) + f(\underline{\mathbf{e_i}})\}$, and likewise $\{\underline{\mathbf{w}} + \underline{\mathbf{e_i}}\}$ to $\{f(\underline{\mathbf{w}}) + f(\underline{\mathbf{e_i}})\}$, so that the property of the invariance of the error term is preserved in the transformed space.

In general, there will be non-zero product–moment correlations between pairs of PI variables, for both the population and error distributions. The use of Bayes' theorem requires the calculation of the probability density functions (pdfs) for both distributions. These are much simpler to calculate if the PI-space has first been transformed, by choosing new axes such that for both distributions, variation along these new axes is independent. This can be achieved by appealing to the well-known theorem of linear algebra which states that for a pair of real, symmetric matrices, one of which is positive definite, it is possible to choose a coordinate system in which both are simultaneously diagonalized, with the positive definite matrix assuming the form of the identity (see for example Ref. [10], p. 58). In the present instance, the covariance matrices of both the distributions are positive definite, so either could be chosen as the identity, but there are advantages in choosing the error distribution for this.

Appendix A provides details of how the transformation to simultaneous principal axes can be carried out for real data, using a widely available statistical package, SPSS. The consequent matrix multiplications are simple enough to be executed using the basic functions and formula-dragging facilities in a program like Excel. An example of this method applied to a sample dataset is given in Appendix B. It is assumed in what follows that this transformation has been carried out, so that the PE covariance matrix is the identity, and the population distribution matrix is diagonalized.

I make the final assumption that the distribution of all individual variables is Gaussian multivariate normal. Failure of normality would not be fatal to the method: as demonstrated for example in Ref. [7], it is possible in the event of failure of normality to use a kernel density estimate to approximate the actual distribution found. In the analysis of an initial data set involving 16 photogrammetric variables within a limited experimental population of 100 individuals and a single camera, we found that once the population and error distributions had been simultaneously diagonalized, all 32 variables were normal using the univariate Kolmogorov–Smirnov statistic ($p \geq .05$, Bonferroni correction). Univariate tests are of course sufficient to determine multivariate normality at this stage, all variables being orthogonal. We might also expect on general principles that the linear transformations involved in diagonalization would tend to produce new variables likely to approximate normality more closely than the original ones, if only because the central limit theorem would suggest that linear combinations of many random variables are likely to approach normality. However, failure of normality would not condemn the method. We make the assumption of normality here if only because it enables mathematically exact solutions to be described more easily, and shortens the discussion. However, the simultaneous orthogonalization procedure works in SPSS without any assumptions of normality, and this procedure is used here, to our knowledge, for the first time in this application.

2.3. Nature of the problem: challenges in identification

As remarked in Ref. [8], it is important to be clear about the objectives when a surveillance system is established. It is assumed that its main purpose is that, in the event that a crime is committed, the person responsible (the 'perpetrator') is caught on camera. Suppose that independently, a 'suspect' is arrested in the vicinity by police, perhaps investigating an alarm or report of crime in the area. The arrest might be on grounds of suspicious behaviour and possession of items or equipment suggesting criminal intent. The suspect is then taken to the police station and photographed to establish their PIs accurately. The forensic expert is asked to provide evidence on whether or not the suspect is identical with the perpetrator. Alternatively, the expert may be provided with the crime scene photograph and a database of criminal suspects, and asked to rule on which if any of them might be identical with the perpetrator caught on camera.

We wish by examination of the photographs of the perpetrator and the suspect to determine whether the corresponding individuals (i.e. the suspect and the perpetrator) are identical. In this first case, where a suspect has already been arrested, there are two mutually exclusive and exhaustive hypotheses to be considered: H, that the suspect is the perpetrator, and A, that he is not. If A is true, it is assumed that the suspect is a random sample from the population on which the population distribution is based. The strict Bayesian approach requires the estimation of a prior probability for H, before the video evidence is taken into account, in accordance with the formula:

$$Pr(\mathrm{H}|\underline{\mathbf{v}}) = \frac{Pr(\mathrm{H}) \times Pr(\underline{\mathbf{v}}|\mathrm{H})}{\{Pr(\mathrm{H}) \times Pr(\underline{\mathbf{v}}|\mathrm{H}) + Pr(\mathrm{A}) \times Pr(\underline{\mathbf{v}}|\mathrm{A})\}}.$$

Here the symbol 'Pr' refers either to finite probabilities or to pdfs; the context makes it obvious which is intended.

In some treatments (e.g. Ref. [5]) the formula

$$Pr(\mathrm{H}|\underline{\mathbf{v}}) = \frac{Pr(\mathrm{H}) \times Pr(\underline{\mathbf{v}}|\mathrm{H})}{\{Pr(\mathrm{H}) \times Pr(\underline{\mathbf{v}}|\mathrm{H}) + Pr(\mathrm{A}) \times Pr(\underline{\mathbf{v}}|\mathrm{A})\}}$$

is rewritten in terms of the odds and the likelihood ratio, defined as $\Pr(\underline{\mathbf{v}}|\mathrm{H})/\Pr(\underline{\mathbf{v}}|\mathrm{A})$, as follows:

$$\mathrm{O}(\mathrm{H}|\underline{\mathbf{v}}) = \mathrm{O}(\mathrm{H}) \times \mathrm{LR}, \text{ where the odds of an event E are } \mathrm{O}(\mathrm{E}) = \frac{Pr(\mathrm{E})}{(1 - Pr(\mathrm{E}))}$$

$\Pr(\underline{\mathbf{v}}|\mathrm{H})$ is simply the pdf of $\underline{\mathbf{v}}$ on condition that the suspect is the perpetrator, i.e. it is given by the error distribution centred on the suspect's parameters, which we will call vector $\underline{\mathbf{w}}$ say. It is the probability of measuring $\underline{\mathbf{v}}$ in the photo, conditional on the true value of the PI vector for the suspect being $\underline{\mathbf{w}}$. It can be calculated from the Euclidean distance between $\underline{\mathbf{v}}$ and $\underline{\mathbf{w}}$ in the transformed PI-space, because the Mahalanobis distance giving the multinomial normal error distribution is equal to the Euclidean distance: all error variances are unity, and independent. $\Pr(\underline{\mathbf{v}}|\mathrm{A})$ is the distribution of $\underline{\mathbf{v}}$ on the assumption that it is from an unknown member of the population, and not the suspect. But this is practically equivalent to stating that $\underline{\mathbf{v}}$ is measured from a random member of the population, so $\Pr(\underline{\mathbf{v}}|\mathrm{A}) = \Pr(\underline{\mathbf{v}})$ in the absence of any other information.

There is one final twist in the calculation of $\Pr(\underline{\mathbf{v}})$. The distribution of $\underline{\mathbf{v}}$ in PI-space is not exactly that of the set of transformed faces, i.e. the population distribution in this space. This is because $\underline{\mathbf{v}}$ represents the PIs taken from a random member of the population using the noisy camera system. The distribution of $\underline{\mathbf{v}}$ therefore comprises two elements: the variance of the population, and the variance introduced by the video system itself. It is therefore the sum of two RVs, one from the population distribution, and one from the error terms. This means that if the standard deviations (S.D.s) of the population distribution along the transformed axes are $\{\sigma_i\}$ for $i = 1, \ldots, n$, the distribution of $\underline{\mathbf{v}}$ is similar but with S.D.s of $s_i = \{(\sigma_i^2 + 1)^{1/2}\}$ for $i = 1, \ldots, n$, along the axes. This gives all the information now required to compute the posterior likelihood that the suspect is the perpetrator.

The LR has been recommended as a means of presenting forensic evidence in an understandable manner [5,11–13]. Indeed, most treatments of forensic applications of the Bayesian approach prefer to avoid the controversial area of prior probabilities, i.e. the estimation of O(H) in order to give $\mathrm{O}(\mathrm{H}|\underline{\mathbf{v}})$ by multiplication. It seems to be assumed that the calculation of LRs is sufficient to encapsulate the effective message of forensic analysis, while at the same time steering clear of any awkward controversy.

Unfortunately, however, there may be circumstances in which at least a very rough estimate of O(H) may be unavoidable. Suppose for example the suspect has been identified not on the basis of independent evidence but because the investigating authorities have trawled a database of digitized photographs and found him to be the best fit, and found, say, an LR of 10,000 on this basis. Mere common sense would suggest that such an identification is much less secure than if an individual had been arrested at the scene of a crime that had occurred shortly before, and at a place and time when there were few if any other likely suspects, and only subsequently been found to fit the photograph of the perpetrator, with an LR of 10,000. Common sense, often a misleading guide, can here be given a theoretical justification. We should observe that when the suspect has been identified only after examination of the PI database, and therefore there was no prior information against him, O(H) should be equated with $1/(N - 1)$ where N is the size of the digitized database. Whereas when the suspect was arrested on account of other suspicious circumstances, the prior probability might be estimated (admittedly very roughly) at not less than 0.01: imagine a deserted industrial estate at 3 a.m. where the suspect has been picked up acting suspiciously. This suggests that even a very crude estimation of priors may give important information on how to interpret LRs which we ignore at peril of miscarriages of justice.

It may be urged as an objection to the first scenario that such a trawl is at present beyond the resources of any law enforcement agency. One might respond that passport photographs are already digitized in some countries, and technical developments may well make it possible to measure PIs automatically within, say, the next 15 years if not sooner. At any event, if LRs are employed exclusively, it is worth pointing out the need to use them with care and discretion.

In what follows we focus on the LR, for which exact estimates can be made that evade the troublesome philosophical controversies, with the caveat that evasion of this problem does not necessarily equate to its avoidance.

Assuming as above that $\Pr(\underline{\mathbf{y}}|\mathrm{A}) = \Pr(\underline{\mathbf{y}})$, the expression for the LR can be expanded as follows, using well-known expressions for the pdfs of independent multivariate normal distributions:

$$\mathrm{LR} = \frac{\exp(-d^2/2) \times \Pi_i(s_i)}{\exp(-D^2/2)} = \exp\left\{\frac{(D^2 - d^2)}{2}\right\} \times \Pi_i(s_i)$$

where d is the Mahalanobis (or Euclidean) distance of the suspect face from the perpetrator in the error distribution, D is the Mahalanobis distance of the perpetrator from the centroid of the distribution in the face distribution (with S.D.s s_i as above) and the product term is to adjust the pdf for the non-unity S.D.s in the face distribution. Ignoring the term in D for the moment (and it will differ little from 1 if the target face is near the average, i.e. the centroid of the population distribution), the LR is seen to depend not only on the distance of the faces in error space, but also on the product of the S.D.s of the face/picture distribution. This product is a good measure of the sensitivity of the system, and could be used to compare different camera/operator systems.

3. Results

3.1. Identification of a suspect already detained

In the case of the system considered in Appendix B, the product term is 19.51989 and assuming for simplicity that $D = 0$, the maximum value of the LR is just under 20, when $d = 0$. This is unlikely to be of much use for securing a conviction, however close the perpetrator face is to the suspect's face. The system is simply not sensitive enough to give proof of identity beyond reasonable doubt.

However, the minimum value of the LR is of course bounded below only by zero, so in this case clear evidence of non-identity is possible. Acting for the defendant, we may consider an LR of 1/1000 as sufficient to cast grave doubt on the suspect's guilt: even with a prior likelihood of guilt of 99.9%, an LR of this value will reduce the posterior probability of guilt to just 50%.

What is the minimum value of d for which LR $< .001$? This gives (assuming again $D = 0$) $\exp(-d^2/2) < .001/19.52$, and $d^2 = 2\ln(19520) = 19.76$, and finally $d = 4.45$. So even if proof of guilt is unlikely to be obtained by using this system, in many cases proof of innocence could be demonstrated.

If the case where D is large is now considered, a heuristic argument can be used to show that even an insensitive system may give a high LR when d is sufficiently small and D sufficiently large. To take an extreme example, suppose a perpetrator has so untypical a set of PIs, and lies so far from the centroid of the population distribution, and in such a sparsely populated portion of it, that a sphere of radius $d = 4.45$ around the target contains no other individual from the entire population. Suppose for simplicity also that the suspect happens to have identical measurements to the perpetrator's image. Then it seems clear that the suspect must be the perpetrator. For any other individual than the suspect must lie outside the critical region of radius 4.45, and therefore be ruled innocent by the argument given above.

The critical value of D to ensure a high LR can be calculated for the simulated system. If we demand an LR of at least 10,000, we require that

$$\exp\left\{\frac{(D^2 - d^2)}{2}\right\} \times \Pi_i(s_i) > 10,000,$$

and so (assuming the most favourable case, $d = 0$)

$\exp(D^2/2) > 10{,}000/19.52$, and finally $D > 3.54$. Examination of the chi-square distribution with 3 d.f. with chi-square = (D^2) shows that this will happen in approximately 0.5% of cases (one-tailed $p = .005$); the method will thus be sufficiently sensitive in this, admittedly very small, proportion of cases.

3.2. Identification from face data alone

A situation that may occur increasingly often is that where there is no suspect in custody, and it is necessary to attempt to identify an individual from a photograph alone. If the transformed parameters for the faces whose data are given in Appendix B are jittered by the transformed values of the errors, and then compared with the originals and the 'best fit' found in terms of the Euclidean distance in image space, the correct identification is made in all but four cases. The total number of trials was 60, with each face given 6 error 'jitters' and then tested against the 10 exact values. Face 1 was misidentified as face 2, face 4 as face 6, face 6 as face 4 and face 10 as face 8, in each case for just one value of the error jitter. That meant that 56 out of 60 trials, or 93%, produced correct identifications.

When the method in Ref. [8] was used, each untransformed PI was tested in turn. The method was given the benefit of the doubt when it identified two faces, one of which was correct, as equally close to the probe. The most successful PI was PI2, which gave a success rate of 12 out of 60 trials or 20%. PI1 and PI3 had successes on just 6 trials each, or 10%. These figures are comparable with those in Ref. [8].

It might be thought that if individual PIs are unsatisfactory, then taking all three PIs simultaneously and judging closest fit by Euclidean distance in the untransformed face space, might give all the benefits of using the transformed space, without the tedious matrix manipulations. This method was in fact tested on the simulated faces. Out of 60 trials (each face presented six times, i.e. jittered by the six error terms) 40 successfully identified the correct face, or 42 if dead heats were given the benefit of the doubt. This represents a success rate of just 70%. Admittedly it did give much better results than for using individual PIs to judge identity, but the outcome was still much inferior to working in the transformed space.

By comparison, taking a single PI formed by adding the three original parameters and taking as the identification criterion Euclidean distance with respect to this single dimension, 46 faces were correctly identified out of 60 trials, giving a 77% success rate. (For an explanation of why this

parameter outperformed both the individual PIs and the use of a Euclidean distance criterion in the untransformed PI-space, see Section 4.)

4. Discussion

4.1. Sensitivity of simulated example

The PCA-based analysis of the simulated example seems almost unreasonably powerful in view of the fact that only three PIs were used. The reason for its success in this case can be seen by looking at the transformation matrix,

10.07838	5.941495	0.778506
16.249	−2.14996	−3.36617
11.17439	−3.153	2.844475

The striking thing about it is that the first column is evidently larger than the other two. Very approximately, it is a multiple of the matrix with 1's down the first column and zeros elsewhere. The S.D.s of the new variables are:

10.86191 1.571878 1.143279

and these confirm that the first component seems to be the most effective in producing variance that is large compared with the error variance (which is unity along each axis). In fact, little would be lost by taking this first component alone and discarding the rest. But this component is approximately a multiple of (PI1 + PI2 + PI3). The reason for the effectiveness of doing this is seen if this simple transformation is applied to the PIs and the error terms (see Appendix B). The variance between the faces for the three original PIs and this new one are:

0.302765 0.31693 0.313404 0.856673

whereas for the error terms, the variances are:

0.130384 0.148324 0.186548 0.107703

where the first three terms refer to the old PIs and the final one to their sum. The variance has increased for the face population, but decreased for the corresponding error term. The ratios between the two, which determine the sensitivity of the system, have increased from

2.322102 2.13674 1.680023
to
7.954011

and this shows that even with this crude approximation, it is possible to find a new PI which is much more effective than any of the original ones taken individually, with a success rate of 77% compared with 70% with the use of all three PIs in untransformed space, and 10–20% for the individual PIs.

It is possible that with a small number of variables and data, a suitable combined PI could be found by inspection. But the merit of the PCA approach adopted here is that it automatically discovers the combinations of PIs which yield the best results and gives them full credit in the analysis. There are other benefits. To quote from Ref. [8]: ''one important factor that may limit the reliability of anthropometric proportions is changes in facial expression''. However, if the system is error-calibrated using volunteers who are instructed to assume a variety of expressions, the method will automatically take this into account and find linear combinations of face measurements which separate out the sources of error into independent components. It is likely that PIs will exhibit a degree of correlation when different expressions are assumed, particularly if there are, say, 10 or more PIs; most expressions will be accounted for by changes in a small number of these. The effects will be limited by the fact that facial muscles tend to reflect the basic emotions of happiness, sadness, anger, fear, surprise, disgust [14]. It has been argued by Schlosberg and his successors that there are perhaps only two underlying dimensions of variation [15,16], in which case an analysis involving 10 PIs may 'use up' two of them in accounting for spurious variance due to changes in expression, still leaving ample scope in the remaining variables to track genuine individual differences in facial characteristics.

4.2. Application to identification of an unknown suspect

On the question of identifying a perpetrator from a database of possible suspects, the comparison between using Euclidean distance in transformed space and the comparison of raw PI data as in Ref. [8], shows that the latter method is indeed as faulty as the authors suggest, but it also shows that even in this very simple example with just three parameter variables, use of the transform raises the success rate from a dismal 10–20% to a respectable 93%. In practice, anthropometry would almost certainly use more than three parameters, and one could expect the superiority of this method over that based on individual PIs to increase monotonically, if not proportionally, with the number of variables considered. Moreover, as mentioned in Ref. [8], other data such as comparison of eye and eyebrow shapes or mouth and nose sizes may distinguish individuals with near or identical PI measurements. With a method that is no more than 20% accurate this would be of little value, but the simulated data gave complete accuracy in 93% of cases, and even where the closest face was incorrect, the correct face was second closest. A procedure which involved checking both first and second choices for secondary characteristics such as eyebrow shape or mouth size would plausibly improve the success rate for the present system to something close to 100%.

Note that identifying a face on the basis of closest fit in the transformed space can be seen as an approximation to a Bayesian decision procedure, at least in the multivariate normal case. To show this, assume that the perpetrator must be one of a set of N suspect faces, each suspect having equal prior probability of being the perpetrator. If we write H_i for the hypothesis ''suspect i is the perpetrator'' and A_i for the logically contrary alternative hypothesis ''suspect j is the perpetrator for some $j \neq i$'', then Bayes' theorem states that

$$O(H_i|\underline{\mathbf{v}}) = O(H_i) \times LR_i,$$

where $LR_i = Pr(\underline{\mathbf{v}}|H_i)/Pr(\underline{\mathbf{v}}|A_i)$.

We now make the plausible approximation that all the denominators in the expressions for LR_i are equal as i varies. That is, we assume that if we remove one individual from the list of suspects and calculate the value of the probability function (assuming the known error distribution) at the actual obtained perpetrator face **v**, this value will not vary significantly whichever face we decide to remove. This assumption can of course be tested if necessary in particular cases.

Making this assumption, we conclude that

$$O(H_i|\underline{\mathbf{v}}) = \lambda \times Pr(\underline{\mathbf{v}}|H_i),$$

where λ is a constant that is independent of i.

If we now adopt the Bayesian decision procedure ''identify that individual i as the perpetrator for which the posterior probability of i's guilt is the maximum out of the list of suspects'', this procedure amounts to maximising $Pr(\underline{\mathbf{v}}|H_i)$.

But in our face space, in which the variables have been chosen so that the errors are independent standard normal distributions, $Pr(\underline{\mathbf{v}}|H_i)$ has a distribution whose value is a monotonically decreasing function of the Mahalanobis distance between **v** and $\mathbf{v}_i$, the face vector for suspect i. Therefore, finally, we have shown that our Bayesian decision procedure amounts to choosing that suspect whose face is closest, using the error-based Mahalanobis distance, to the perpetrator's image.

Therefore there are not two distinct methods, but really only one. The maximum likelihood method is an approximation to the full method, and will yield a useful result only in cases where it is known that the perpetrator comes from a given pool of suspects, and where there is no reason to favour any suspect over any other. Even then, it is a somewhat crude instrument, in that like any ''first past the post'' system, it may not lead to a fair result. It has the advantage of simplicity, however, in that because it does not require the explicit calculation of $Pr(\underline{\mathbf{v}}|A_i)$, or indeed of $Pr(\underline{\mathbf{v}})$, it does not necessitate the double PCA that is needed if the actual value of the LR is needed for any particular suspect. In other words, the simplification arises because only relative, not absolute values of LR suffice for this particular procedure.

We did not therefore really test the full method when, above, we showed that the transformed variables gave greater ability to identify faces than either the use of univariate PIs or a simple Euclidean distance based on untransformed variables. This cut down version of the method could in fact have been carried out using a simple one-stage transformation to independent standardized error variables. The full two-stage method would only show its value in cases where it is necessary to calculate the LR for a suspect, rather than simply to compare his LR with that of other suspects. Where the full LR is needed, our method provides a valid method of calculating it, which is why we have demonstrated all the steps needed to do this, in Appendices A and B.

5. Conclusions

The use of a double PCA transformation enables an exact solution to be found to statistical questions involving face data with a limited number of parameters. In particular, the method allows the calculation of LRs when comparing a photograph with a known suspect, and automatically gives the maximum likelihood solution, equivalent to a comparison of LRs, when an identification is to be made from a pool of suspects. It appears significantly more effective than either identification using an individual parameter, or the use of pooled parameters in an untransformed face space. The relative simplicity of the method, compared with sophisticated modern techniques such as eigenfaces, may recommend it when explaining the outcome of a photogrammetric analysis in a courtroom setting.

It would be premature to conclude, as some authors have done, that identification using facial landmarks is inefficient, until the most powerful methods of analysis have been tried and found wanting. This paper suggests theoretical approaches which maximize the value of photogrammetric information. A fair assessment of photogrammetry will only be possible once these methods have been applied in a practical context and the results evaluated; we should not lightly discard the methods introduced by Bertillon, and which were historically so important in law enforcement.

If evaluation of these exact analytical methods proves to be positive, then their applications should be of value not only in providing evidence for identity, but also in alerting law enforcement agencies to the dangers of unsafe identifications when they rely on camera/operator systems which are intrinsically unreliable.

International cooperation between law enforcement agencies and the exchange of intelligence on criminals and their activities is currently more necessary than ever [17,18]. Mathematically optimal methods of analysis in face identification might assist the standardization of methodology, thereby facilitating these highly desirable developments.

Appendix A

The first step is to standardize the error distribution. Suppose there are n PI measurements (variables $v_1, \ldots, v_n$), and therefore also n PI error terms, $e_1, \ldots, e_n$. Suppose the S.D.s of the error terms are $\varepsilon_1, \ldots, \varepsilon_n$, respectively. Multiply each error variable by the inverse of the appropriate S.D., so the new, rescaled error variables are $e_1/\varepsilon_1, \ldots, e_n/\varepsilon_n$; they will therefore be standardized (zero mean, unit variance). (This is needed because in SPSS, PCA using the correlation matrix automatically standardizes the variables before operating on them, and we need to ensure that this concealed transformation is made explicit so that it can be performed also on the population distribution variables.) Apply the same multiplication by ε_i^{-1} also to each of the population distribution variables v_i.

Now apply a principal component analysis to the transformed set of error terms. In SPSS, go to ANALYZE ≫ DATA REDUCTION ≫ FACTOR ANALYSIS.

Enter all rescaled error variables into the 'variables' box.

Ensure the following settings are used:

Rotation – Method – none.

Scores – check 'Display factor score coefficient matrix'.

Extraction – Method – Principal Components [not any other factor analysis method].
Analyze – correlation matrix.
Display – unrotated factor solution.
Extract – number of factors – enter total number n of variables.

The output gives the Component Score Coefficient Matrix. Extract this, for example to EXCEL or MATLAB, and apply it (by post-multiplication) to transform both the error distribution and the population distribution, to give a new error distribution and population distribution, in which the new variables are the components taken from the PCA. Because the error distribution variables were used in the PCA, the new error distribution will consist of independent components, each of unit variance. The new components of the population distribution will not, in general, be independent.

The final step is to transform to a second set of components, in which the population distribution axes will also be independent. To do this, go to

ANALYZE ≫ DATA REDUCTION ≫ FACTOR ANALYSIS.

Enter all population distribution components derived from the first PCA into the 'variables' box.
Ensure the following settings are used:

Rotation – Method – none.
Extraction – Method – Principal Components.
Analyze – covariance matrix [note, this is essential: using the correlation matrix, which is the default setting in SPSS, standardizes the variables prior to analysis and makes the error axes no longer of unit variance].
Extract – number of factors – enter total number n of variables.

The output will display the Component Matrix, subdivided into raw and rescaled versions. Copy the raw version into e.g. Excel. Before applying it to the variables, divide each column of the matrix by its 'length' (square root of inner product with itself). This ensures that the matrix is orthogonal. It is known that such a transformation will preserve the property of the error distribution variables, that they are statistically independent and of unit variance. At the same time, the matrix still, after, normalization, transforms the population distribution variables into independent components. We have therefore achieved the desired transformation.

Appendix B

The artificial set of data involves just three PIs, ten faces, and six error measurements. The situation is clearly unrealistic, but may serve as an indication of how much more powerful the Bayesian approach can be even for very sparse information sets than relying on individual PIs to discriminate faces.

The data are as follows. Face parameter measurements:

	PI1	PI2	PI3
Face 1	1.5	1.3	1.6
Face 2	1.3	1.4	1.4
Face 3	1.6	1.6	1.8
Face 4	1.7	2	1.9
Face 5	1.4	1.2	1.1
Face 6	1.8	2	2
Face 7	1.2	1.5	1.2
Face 8	1.9	1.8	1.7
Face 9	1.1	1.1	1.2
Face 10	2	1.7	1.5

Error terms:

PI1	PI2	PI3
0.1	−0.15	0.11
−0.15	0.1	0.11
−0.1	0.15	−0.24
0.15	−0.15	0.11
−0.1	−0.1	0.15
0.1	0.15	−0.24

Note that for each PI the error variables have, as required, a mean of zero.

Carrying out the three steps in the recipe given above results in the following successive operations, where all matrix multiplications are post-multiplications, and all vectors are row vectors (e.g. the vector representing face 1 is (1.5, 1.3, 1.6)):

1. Multiply by the diagonal matrix with entries

   7.66965 6.741999 5.360563

   representing the inverse of the error S.D.s.
2. Multiply by the component score coefficient matrix represented by the new error terms, which is found from SPSS to be:

	1	2	3
PI1 error	0.24	0.861	1.24
PI2 error	−0.518	−0.019	2.427
PI3 error	0.451	−0.48	2.131

3. Obtain the raw component matrix for the new face variables:

	1	2	3
FCP1	0.114	0.197	0.547
FCP2	0.961	1.192	−0.09
FCP3	10.772	−0.108	0.002

and divide the columns by their 'lengths' to obtain the following orthogonal matrix:

0.010541	0.162409021	0.986727
0.088855	0.982698238	−0.16235
0.995989	0.089036418	0.003608

which, post-multiplying both the error and the face parameters, gives, finally,

err1	−0.20033	0.569813	0.895668
err2	1.342326	−1.45305	−0.1405
err3	−1.25234	−0.15992	−1.26545
err4	0.30359	0.866888	0.934594
err5	−0.95658	−0.8521	0.685438
err6	0.763335	1.028377	−1.10975

and

face 1	54.1203	1.072489	1.342897
face 2	51.49464	0.299795	0.281684
face 3	62.23772	0.39105	0.979792
face 4	70.86259	−0.19009	−0.00438
face 5	45.90036	2.269837	0.179427
face 6	72.98787	0.088764	0.35792
face 7	49.87683	0.12125	−0.70168
face 8	67.39359	2.058807	0.255662
face 9	42.36939	0.387084	0.566939
face 10	64.54165	3.498553	0.101235

It can be verified that the error terms are independent and of unit variance and that the face terms are also independent, though not of course of unit variance. The manipulations are complete, except to calculate the S.D.s of the new face distribution, which are

10.81578 1.212766 0.554155

The S.D.s of the population of photos of faces now has the S.D.s

10.86191 1.571878 1.143279

taking into account the fact that the act of deriving a photo from a face adds an independent random variable of unit variance to it, i.e. the S.D.s $\{\sigma_i\}$ for $i = 1, \ldots, n$, must be corrected to $\{(\sigma_i^2 + 1)^{1/2}\}$ to represent the distribution of the photos, as required for Bayes' formula.

The three steps given above can of course be combined into a single post-multiplication by the product of the three matrices, namely

10.07838	5.941495	0.778506
16.249	−2.14996	−3.36617
11.17439	−3.153	2.844475

and it can be checked that applying this to the original set of data gives the final set directly. Moreover, using this matrix enables any target or probe face in the original set of variables to be transformed into the new variables and conclusions to be drawn using Bayes' theorem.

Acknowledgements

The author wishes to thank Dr. Josh Davis of Goldsmiths College for many helpful explanations of existing methods of biometric analysis as applied to face data processing, as well as Professor Tim Valentine of Goldsmiths College for first sparking the author's interest in statistical analysis of the identification process.

References

[1] P. Garbolino, F. Taroni, Evaluation of scientific evidence using Bayesian networks, Forensic Sci. Int. 125 (2002) 149–155.
[2] A. Biedermann, F. Taroni, Bayesian networks and probabilistic reasoning about scientific evidence when there is a lack of data, Forensic Sci. Int. 157 (2006) 163–167.
[3] G. Baio, F. Corradi, Handling manipulated evidence, Forensic Sci. Int. 169 (2007) 181–187.
[4] C. Champod, D. Meuwly, The inference of identity in forensic speaker recognition, Speech Commun. 31 (2000) 193–203.
[5] J. Gonzalez-Rodriguez, J. Fierrez-Aguilar, D. Ramos-Castro, et al., Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems, Forensic Sci. Int. 155 (2005) 126–140.
[6] D. Ramos-Castro, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, et al., Speaker verification using speaker- and test-dependent fast score normalization, Pattern Recognit. Lett. 28 (2007) 90–98.
[7] C. Aitken, D. Lucy, Evaluation of trace evidence in the form of multivariate data, J. R. Stat. Soc. Ser. C: Appl. Stat. 53 (2004) 109–122.
[8] K.F. Kleinberg, P. Vanezis, A.M. Burton, Failure of anthropometry as a facial identification technique using high-quality photographs, J. Forensic Sci. 52 (2007) 779–783.
[9] G. Porter, G. Doran, An anatomical and photographic technique for forensic facial identification, Forensic Sci. Int. 114 (2000) 97–105.
[10] R. Bellman, Introduction to Matrix Analysis, Society for Industrial and Applied Mathematics, Philadelphia, 1995.
[11] N.M. Egli, C. Champod, P. Margot, Evidence evaluation in fingerprint comparison and automated fingerprint identification systems – modelling within finger variability, Forensic Sci. Int. 167 (2007) 189–195.
[12] M. Horrocks, K.A.J. Walsh, Forensic palynology: assessing the value of the evidence, Rev. Palaeobot. Palynol. 103 (1998) 69–74.
[13] C.M. Triggs, J.S. Buckleton, Comment on: why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence, Law, Probability Risk 3 (2004) 73–82.
[14] P. Ekman, H. Oster, Facial expressions of emotion, Ann. Rev. Psychol. 30 (1979) 527–554.
[15] H. Schlosberg, The description of facial expressions in terms of 2 dimensions, J. Exp. Psychol. 44 (1952) 229–237.
[16] T. Takehara, N. Suzuki, Robustness of the two-dimensional structure of recognition of facial expression: evidence under different intensities of emotionality, Percept. Mot. Skills 93 (2001) 739–753.
[17] O. Ribaux, S.J. Walsh, P. Margot, The contribution of forensic science to crime analysis and investigation: forensic intelligence, Forensic Sci. Int. 156 (2006) 171–181.
[18] P. Esseiva, S. Ioset, F. Anglada, et al., Forensic drug intelligence: an important tool in law enforcement, Forensic Sci. Int. 167 (2007) 247–254.