Journal of Experimental Psychology: Applied
2002, Vol. 8, No. 3, 180–193

1076-898X/02/$5.00 DOI: 10.1037//1076-898X.8.3.180

A Direct Measure of Facial Similarity and Its Relation to Human Similarity Perceptions

Colin Tredoux
University of Cape Town

Research is reported on a measure of facial similarity in which the similarity of 2 faces is defined as the *Euclidean distance* between them in a principal-component space. Five studies were conducted in which participants rated sets of facial images, and in which the measure was applied to 2 problems in the eyewitness literature. Comparisons of ratings with distances derived from the principal-component analysis suggest that the measure corresponds reasonably well to perceptions of facial similarity. In addition, the measure correlates strongly with empirical measures of lineup fairness and is related to eyewitness identification performance. Further potential applications include a software tool for constructing arrays of faces of varying similarity, and a software tool for reconstructing facial images from memory.

The recent explosion of research on face perception and recognition has meant significant advances in both theoretical and applied cognitive psychology. One of the under-researched issues in this field is facial similarity. Little is known (or postulated) in theoretical models about how perceptual and cognitive structures deal with the high degree of similarity that faces exhibit, or how the perceptual system manages to retain identity information when faces are transformed (e.g., in orientation or pose; see Bruce, 1994). This problem also weakens applied research. Without a suitable conceptualization and measure of facial similarity, theoretical research on face perception and recognition is palpably impoverished. Many of the key findings may turn out quite differently once researchers are able to treat facial similarity as an independent variable, and much the same may hold true for applied face recognition research.

The present research reports an attempt to conceptualize and implement a measure of facial similarity that is based on the physical properties of faces. The task is to find a measure that corresponds in reasonable degree to the perceptions and judgments of facial similarity made by human participants.

Cognitive psychologists have attempted in the past to measure facial similarity, but these attempts have usually been one-shot solutions to emergent problems in the research design. They are worth mentioning here, as they are a source for some of the validity tests used in the present research.

A Priori Methods

In these techniques, researchers use a criterion that is presumed to distinguish faces on the basis of their similarity. Thus, Patterson and Baddeley (1977) created groups that ostensibly differed in the facial similarity of their members, using photographs of people from very different social categories; to wit, actors (low similarity, because there is no defining attribute of this group to ensure similarity) and soldiers (high similarity, due to common characteristics determined by shared age, haircuts, etc.). Malpass and Devine (1983) created lineups of varying similarity by selecting individuals according to their height, weight, hair color, hair length, and eye color. Laughery, Fessler, Lenorovitz, and Yoblick (1974) operationalized similarity in terms of a set of trials in which faces were paired, and where participants were required to discriminate old from new faces. The proportion of mistaken ("old") responses was used to define similarity.

The weaknesses inherent in these types of techniques include (a) the untested nature of the assumptions used to determine similarity and (b) their impreciseness. Although it may be reasonable to assume that groups of actors and soldiers will show different variability in facial similarity, this is a gross division and of little use in most situations where similarity needs to be measured or manipulated.

Rating Techniques

Most psychological studies that attempt to measure facial similarity do so by obtaining ratings of faces from independent participant judges. Bruce (1979) required participants to rate stimulus faces in relation to target faces on a 4-point scale; in Milord's (1978) study, participants rated pairs of faces on a 7-point scale for similarity–difference. Harmon (1973) based an early, computer-driven face recognition system on ratings of face descriptors. Usually, ratings of similarity are made globally; that is, participants are asked to rate faces on a single scale ranging from, for example, *not at all similar* to *very similar*. Alternative conceptualizations and operationalizations are relatively unexplored. Researchers have not investigated whether "highly similar" and "easily mistakable" are coterminous or correlated, nor have they systematically examined the dimensions governing similarity judgments. The psychometric properties of these similarity ratings

I acknowledge the financial assistance of the Centre for Science Development (Human Sciences Research Council, South Africa) for this article. Opinions expressed or conclusions arrived at in the work are mine and should not be attributed to the Centre for Science Development.

Correspondence concerning this article should be addressed to Colin Tredoux, Department of Psychology, University of Cape Town, Private Bag, Rondebosch 7701, South Africa. E-mail: plato@humanities.uct.ac.za

are also very rarely reported, and there are some indications that this is an important failure. Lindsay (1994), for example, reported that facial similarity judgments show great interparticipant variability—an array of faces that appear highly similar to one observer may not appear at all similar to another observer.

This type of approach has the advantage of retaining a hold on the cognitive aspect of facial similarity: What is important, after all, for most face recognition research is *perceived* similarity. Rating studies obtain a "direct" measure of this perceived similarity (notwithstanding the unexplored psychometric problems). The chief drawbacks of this technique are the dependence on participant ratings and the statistical ramifications of this dependence.

Scaling Techniques

Although the use of participant ratings ensures the connection of similarity measures to cognitive process, such ratings are typically only useful for a small set of comparisons. Several authors have recognized the need to formulate similarity measures for larger stimulus samples and utilized forms of scaling technique to this end. Hirschberg, Jones, and Haggerty (1978) obtained similarity ratings of all pairs of faces in a large sample and entered these into a multidimensional scaling (MDS) analysis, incorporating individual differences into the analytic model. Because MDS generates a dimensional basis, spatial distance measures can be used as a measure of similarity. Other cognate approaches include Rhodes's (1988) study, which used a "tree sorting" algorithm. Young and Yamane (1992) took extensive anthropometric measurements from each of a set of faces, found Euclidean distances for each face on these axes, and then submitted the distances in matrix form to MDS. Davies, Shepherd, and Ellis (1979) measured facial similarity as an interim step in an application of (hierarchical) cluster analysis.

These approaches, in principle, present the most satisfactory solution in the literature to the problem of measuring facial similarity. The recognition that a similarity metric must be based on a representational scheme capable of simultaneously representing all faces in a set is particularly important. The further recognition that similarity must be conceptualized as inherently multidimensional is also significant. Valentine has argued extensively for such a conceptualization of "face space" (Valentine, 1991a, 1991b; Valentine & Endo, 1992; Valentine & Ferrara, 1991).

However, the schemes discussed here do not go far enough: The dimensions of the representational space are implicit, and it is not clear that they can generate faces that are not in the set submitted to MDS in the first place. Nevertheless, the issue is not broached in these studies.

A Possible Solution: Principal-Component Analysis and Multidimensional Space

An alternate approach, which does not use MDS but retains the notion of a multidimensional representational scheme, is the principal-component analysis (PCA) exemplified in studies by Sirovich and Kirby (1987); O'Toole, Abdi, Deffenbacher, and Valentin (1993); and Craw and Cameron (1991). This approach appears capable of providing a set of generating dimensions that can accurately represent faces, which were not included in the initial PCA. This is a fruitful procedure to follow in the quest of a facial similarity metric.

The starting point of the PCA approach is to conceptualize a digitized face image as a two-dimensional image pixel width (M) × image pixel height (N) array of intensity values. An ensemble of images maps to a collection of points in this M × N space. Because face images would bear considerable resemblance to each other, this space would be relatively low dimensional. PCA finds the vectors that generate this subspace.

Each face in the set of face images can then be represented as a set of coordinates on these eigenfaces, or axes: the face would be perfectly reconstructed as a sum of the coordinate-weighted eigenfaces, provided that all the eigenfaces are used. If only a subset of the eigenfaces is used, the face would be imperfectly reconstructed, although this reconstruction may still be very accurate. By way of example, Figure 1 presents an ensemble of eight eigenfaces, generated from the set of 278 frontal views of faces referred to in Studies 2a and 2b of the present article. Figure 2 shows 10 images and their reconstruction from increasingly large sets of eigenfaces. It is clear that the approximation to the original images becomes better as more eigenfaces are used. The benefits of this approach to the task at hand are considerable, four of which are worth discussing.

The eigenfaces generated by the PCA allow the representation of the images in terms of a common set of reference axes (the eigenfaces). Individual facial images are linear combinations of these eigenfaces. This is a direct implementation of the type of multidimensional model discussed earlier in this article. It also provides a solution to the problem of identifying the dimensions for such a model: the dimensions are just the eigenfaces identified by the PCA.

The multidimensional space generated by the eigenfaces has the associated advantage that well-developed measures of spatial distance can immediately be used to determine nearest neighbors, relative density around particular points, and a variety of other useful indices. Of particular interest in the present article is the use of Euclidean distance as an index of facial similarity. Although there are a number of other measures of image similarity, it is not my intention to investigate these here.

If we admit some degree of error into the representation of faces in the space, we can represent the set of faces with considerably fewer eigenfaces than there are faces in the set (here, *error* refers to the variance explained by the eigenfaces not included in the

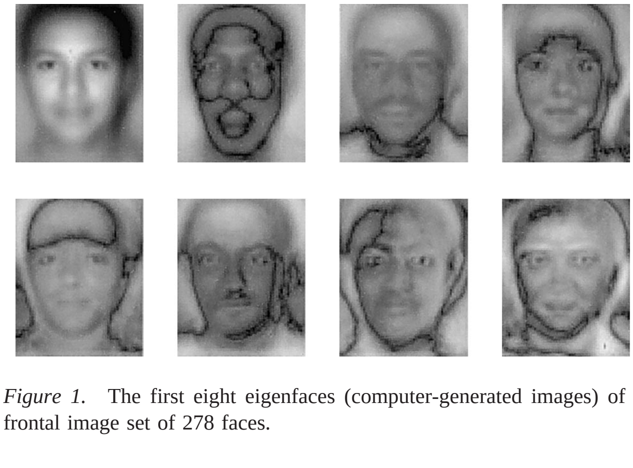

Figure 1. The first eight eigenfaces (computer-generated images) of a frontal image set of 278 faces.

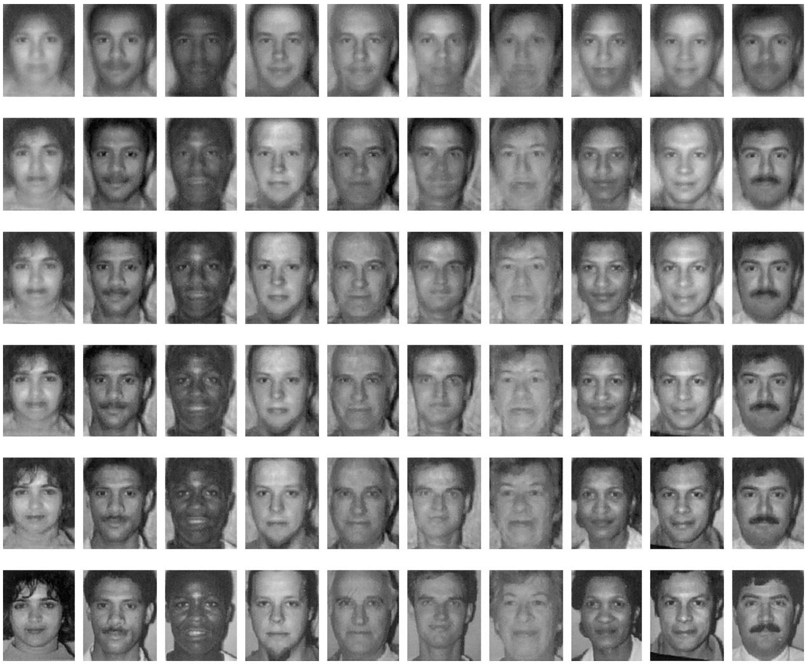

Figure 2. Ten images and their reconstruction by increasing numbers of eigenfaces.

model). Thus, Craw and Cameron (1991) argued that about 40 eigenfaces is adequate to represent 100 faces, with about 5% error, and Kirby and Sirovich (1990) argued likewise. However, there is no agreement in the literature about the number of faces needed to adequately represent a population of faces (but see Penev & Sirovich, 2000), which may be required if we want a PCA implementation of Valentine's (1991a, 1991b) multidimensional theory of face perception.

Attractive as this approach appears, there are several problems. First, the approach unreasonably assumes the structural equivalence of images. Prior to analysis, images must be standardized so that eyes in one image correspond in location to the eyes of other images—there would be no point in averaging ears and eyes, for example. This objection is made forcefully by Craw and Cameron (1991), who pointed out that PCA assumes linearity, and this assumption will not be met unless images are structurally equivalent. The standardization of images is not only difficult to accomplish with respect to location in the raw image space, but also with respect to aspects such as facial expression, ambient lighting, lighting of the face, and so forth. However, the variation across faces on these latter variables would not be nearly as great as the variation between the standardized images themselves, and Sirovich and Kirby (1987) have shown that several of these variables do not severely affect matters.

In practice, images are aligned so that the pupils match (i.e., the left and right pupils occupy the same spatial locations on each image). This ensures a close match of most faces. Craw and Cameron (1991) outlined an alternative approach to the problem, which uses elements of the caricaturing method developed by Benson and Perrett (1991a, 1991b). That is, fiducial points are defined for a "standard" image, and triangular tesselations are created by joining certain areas of these points. Each face is then mapped onto this image, using bilinear interpolation, before the PCA. Hancock, Burton, and Bruce (1996) used both methods (i.e., PCA on "shape-free" images, and PCA on "shaped" images) and found a moderate advantage for a combination of the shape-free and shaped methods in predicting context-free familiarity.

From a practical point of view, standardizing images is laborious (much more so in transforming images to shape-free form), albeit necessary. An algorithm to automate the standardization would be a useful addition: moderately successful attempts are reported by Bowns and Morgan (1993); Li, Qiao, and Psaltis (1993); and Cootes and Taylor (2001).

PCA and Similarity

We may be able to express the similarity between two faces as a function of the Euclidean distance between them. This is predicated on the notion that the perceived similarity of faces would be a function of the physical properties they share, which is what makes up the principal-component space of facial images. Figure 3 shows three sets of highly similar faces, as identified in the PCA of 278 frontal images collected for the present research.

Although the PC implementation of the multidimensional conceptualization of face perception is mathematically appealing, researchers need to show that it is a suitable analogue to human perception of facial similarity. There has been some research in this respect. Hancock, Bruce, and Burton (1998) reported a significant but low correlation of approximately .2 between the Euclidean distance between PCA vector representations of faces and their rated similarity by human participants, in a sorting task. Their face images were drawn from a relatively homogenous group (young adult male Caucasians), though, and this may have placed a range restriction on both PCA and human ratings of similarity,

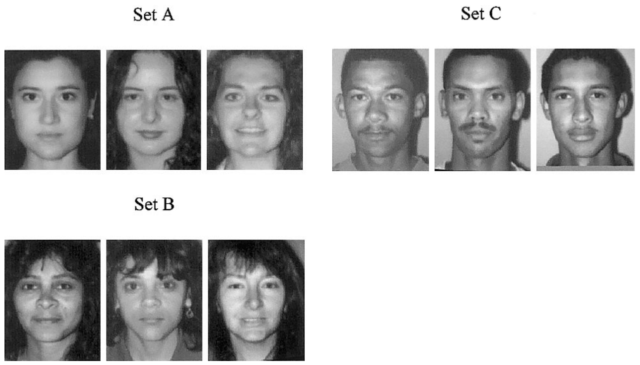

Figure 3. Examples of highly similar face images, using Euclidean distance as a metric.

reducing the size of the correlation coefficient relating them. In addition, they used a total of 50 faces for their PCA, and it is unlikely that their PC solution generalizes to large collections of faces from the same population. An unpublished study by Kalocsai, Zhao, and Elagin (1998) reported much higher correlations between a PCA-based measure and a similarity index ($r \approx .45$), but it is not clear from that study what PCA measure was used. In addition, the human similarity index was a same–different judgment in a psychophysical paradigm task rather than a judgment of perceived similarity, and the face images used as stimulus material had their hair removed.

In the present set of studies, research is reported that investigates the relationship between similarity measures derived from a principal-component model and human estimations of face similarity. A variety of tasks are used to obtain the estimations of similarity, and the principal-component models are based (in some of the studies, at least) on relatively large and heterogenous sets of faces, which are not edited to remove hair. In addition, a practical site of application for the measure is tentatively explored in the measurement of lineup fairness (which ostensibly depends on the facial similarity of suspects and foils) and the measurement of eyewitness identification performance. In particular, it is possible that there may be a trade-off between the fairness of a lineup (higher for greater facial similarity) and eyewitness identification ability (lower for greater facial similarity). The ability to predict unfair lineups can also be seen as a practical test of validity of the proposed measure.

Study 1: Correspondences Involving Sorting Judgments

Method

Stimuli

Sixty-two volunteer students and staff at the University of Cape Town (UCT) were photographed against a uniform, dark background. Participant characteristics are reported in Table 1.

Ambient lighting was standardized, and a flash unit was used to provide a direct source of light and was coupled to an automatic 35-mm Leica SLR camera, which was used to capture photographs. Participants were asked to adopt a neutral expression and to look straight ahead at the camera. Photographs were developed by a commercial photographic service and digitally scanned at 300 dots per inch (dpi) on a Hewlett Packard IIx scanner to 256 gray-level images. The images were edited digitally to remove jewelry and other extraneous items. The 62 images in the stimulus set were standardized with respect to the position of the left and right pupils (i.e., images were cropped, enlarged, or reduced so that the pupils occupied the same coordinate positions in a common pixel space). Image size was equated by cropping to a uniform size of 154 × 205 pixels. (This image size, and the use of gray-level images, is typical for studies that have conducted PCA on face images, because PCA puts great demands on processing resources.) The image set was then submitted to PCA, using SPSS software. Face images were submitted as variables, each constituted by 31,570 "observations." Principal components and their coefficients were derived from this analysis, and these were used to generate a matrix of Euclidean distances between faces in the image set.

For the sorting task in question, 9 arrays of 20 facial images were constructed at random by selecting images from the original set. Arrays were printed on a Hewlett Packard Laserjet printer at a resolution of 600 dpi, which produced face images of acceptable quality. (Our criterion was whether the image was clearly identifiable as the person in the original photograph.)

Participants

One hundred eleven undergraduate students of psychology at UCT participated in the sorting task.

Procedure

Each of the study participants was asked to create similarity pairings of the images in one of the arrays by choosing (a) the most similar pair of faces; (b) the next most similar pair of faces, and so on, until all 10 possible (exclusive) pairings had been effected. In all of these tasks, participants were given a booklet containing the arrays, with instructions, and were asked to complete the tasks during a 20-min period at the beginning of a lecture.

Results

For each pairing made by participants, a corresponding Euclidean distance was calculated: This was the distance in the principal-component space, between the pair of faces selected by the participant. These distances were averaged over participants so that mean distances were obtained for the 10 exhaustive possible pairings in the task (i.e., each participant produced 10 pairings; for each of these pairings, a PC-based distance was calculated, and

Table 1
Participant Characteristics of the 62 Faces Submitted to Principal-Component Analysis in Study 1

Variable	Group	*n*
Age	18–29	45
	30–39	6
	40–49	7
	>50	4
Sex	Male	21
	Female	41
Race	Black	8
	Coloured	13
	White	41

Note. Race groups reported here are based on those defined in the (now defunct) South African Population Registration Act. They should not be taken to indicate distinct genetic or physiognomic populations, although the groups do differ considerably in physical appearance.

then mean distances were found for each pairing, averaging over participants). Because the task required participants to pair the most similar faces in the array, sequentially, and because each array had a determined sequence of closest pairings in terms of the principal-component coefficients, it was possible to calculate expected Euclidean distances. The expected distances were thus just the distances of face pairings in the PC space, arranged in increasing order, and the observed distances were the average distances corresponding to pairings made by participants. Table 2 reports the average observed distances and expected distances.

There was a very strong correlation between the expected and obtained distances ($r = .94$, $df = 8$, $p < .01$). The size of this correlation is misleading, though, because the obtained distances were averaged over 111 participants, and there was considerable variability between participants (the average individual correlation was .29). A better indication of the strength of the relation may be the effect size calculated from an appropriate analysis of variance (ANOVA).

Accordingly, a repeated measures single-factor ANOVA was conducted across pairings, taking observed Euclidean distance of each face pairing made by participants as the dependent variable, and ordinal sequence (which had 10 levels) as the independent variable. The omnibus test in this analysis was not of much interest, because the PCA measure of facial similarity provided fairly precise predictions of differences between levels of the independent variable. Table 2 shows the predictions, and it is clear from visual inspection that the relation between ordinal sequence and predicted distance is nearly linear. (It is important to note that the predicted distances are averages, because multiple arrays were used in the pairing task, and each array has a unique pairing sequence.) The ANOVA was therefore conducted by partitioning sums of squares with a set of orthogonal polynomials, which allows one to estimate linear, quadratic, cubic, and other higher order terms. A summary is reported in Table 3.

It is clear from the trend analysis that the linear effect provides a good fit to the observed pairings data: The effect size was substantial ($\omega^2 = 0.39$) and also statistically significant, $F(1, 94) = 74.13$, $p < .01$. However, the degree of fit was clearly not perfect, and there is also a discernible nonlinear component to the relationship. Because the expected performance in terms of the PCA similarity measure was reasonably close to constituting a linear relation, the observed pairings data correspond quite well to that predicted by the PCA measure.

Table 2
Average Obtained and Expected Euclidean Distances for the Pairings Task

Pair no.	Expected	Obtained (and *SD*)
1	0.55	0.84 (0.15)
2	0.60	0.85 (0.16)
3	0.65	0.83 (0.14)
4	0.70	0.86 (0.19)
5	0.73	0.90 (0.18)
6	0.76	0.91 (0.19)
7	0.78	0.92 (0.20)
8	0.83	0.96 (0.21)
9	0.92	0.93 (0.21)
10	1.08	1.01 (0.22)

Discussion

The results of Study 1 suggest that the PC-based measure of facial similarity corresponds reasonably well to human similarity judgments. These results should be regarded as preliminary because there are several aspects of Study 1 that require further investigation.

It is clear from the similarity rating results that there was considerable interparticipant variability (or disagreement) in similarity judgments. This is an interesting finding in its own right. The phenomenon is unreported in previous research, apart from a casual observation made by Lindsay (1994), which is surprising. Possible explanations should be considered. First, it may be that people will agree only when faces show substantial similarity or dissimilarity, but not when they show levels of similarity between the extremes. This is explicable if judgments of similarity are made on multiple bases; that is, if they are multidimensional and if these differ across participants. Highly similar faces would resemble each other on many dimensions, and judgments based on a subset of these dimensions are more likely to be congruent. Highly dissimilar faces would resemble each other on very few dimensions, and would be likely to attract ratings of dissimilarity. Other faces are problematic: Because they would resemble each other on some dimensions but not on others, there is considerable room for disagreement when participants use only a subset of available dimensions. Second, and more problematically, it may be that judgments of similarity are inherently unstable; that is, judgments of the same stimuli made by an individual at different points of time would not concur. This possibility is explored in Study 3.

The judgment task used in Study 1 required participants to pair similar faces, but this is clearly only one operationalization of similarity perception. It should be shown that the PC-based measure correlates with other kinds of similarity judgments. This is attempted in other studies reported in this article.

A significant problem in much face recognition research is the use of single-viewing perspectives—typically photographs taken from a frontal perspective. Studies that use single-viewing perspectives cannot distinguish face perception and memory from picture perception and memory, and probably yield inflated estimates of recognition ability, in particular. What needs to be shown is a relation between similarity scores derived from analysis of a set of frontal images, and scores derived from analysis of a set of other views (say three-fourths profile). These scores should be strongly correlated, and the absence of a strong correlation would be evidence against the robustness of the PC-based similarity measure. Study 2b gathered evidence of this kind.

Study 2: Correspondences Involving Ratings of Similarity

Method

Stimuli

Study 1 used a set of images that proved limited in several respects. In particular, the set was relatively small (62 images) and disproportionately constituted by photographs of young White women. A much larger, and more representative, corpus of images is needed to study perceptions and ratings of facial similarity. The first task in Study 2 was therefore to collect such a corpus.

Table 3
Polynomial Trend Analysis for Results of the Face-Pairing Task

Effect	*SS* effect	*SS* error	*MS* effect	*MS* error	*F*	$p <$	ω^2
Linear	2.54	3.22	2.54	0.03	74.13	.01	0.39
Quadratic	0.04	3.82	0.04	0.04	1.04	.311	0.0004
All other	0.31	—	—	—	—	—	0.098

Note. Dashes indicate that data were not obtained or reported. *SS* = sum of squares; *MS* = mean square.

Photographs were collected by setting up photographic stalls in supermarket malls and offering passersby a free photograph in exchange for permission to use this photograph for research purposes. A total of 278 people agreed to pose for photographs. Their characteristics are summarized in Table 4.

Participants were positioned in front of a gray matte screen, and a flash unit provided a direct source of light. Two 35-mm format cameras were used to take photographs (a Canon EOS 500, and Canon EOS 100) at a focal length of approximately 80 mm. Exposure was controlled by the automatic through-the-lens metering system of each camera. Participants were asked to adopt a neutral expression (as in a passport photograph) and to look straight ahead at the camera. A second photograph was then taken, with participants adopting a three-fourths profile position.

Photographic film was later developed, contact printed on Ilford Pearl photographic paper, and digitally scanned at 300 dpi on a Hewlett Packard IIx flatbed gray-image scanner to 256 level gray-level images.

The 278 images in the stimulus set were standardized with respect to the position of the left and right pupils, and image size was equated by cropping to a uniform size of 120 × 150 pixels. The image set was then submitted to PCA, using SPSS software. Face images were submitted as variables, each constituted by 18,000 observations. Principal components and their coefficients were derived from this analysis, and the first 100 of these were used to generate a matrix of Euclidean distances between faces in the image set.

Participants

Participants in Study 2a were 76 Psychology 1 students who participated in the rating task during a lecture, and participants in Study 2b were 90 Psychology 1 students who also participated during a lecture.

Table 4
Participant Characteristics of the 278 Facial Images Collected in Supermarket Malls

Variable	Group	*n*
Age	16–19	22
	20–29	122
	30–39	74
	40–49	40
	>50	20
Gender	Male	148
	Female	130
Race	Black	14
	Coloured	139
	White	121

Note. Race groups reported here are based on those defined in the (now defunct) South African Population Registration Act. They should not be taken to indicate distinct genetic or physiognomic populations, although the groups do differ considerably in physical appearance.

Procedure

Study 2a: Rating similarity in arrays of frontal views. In the rating tasks described here, participants were given 3 arrays of 10 face images from the original set, printed at 600 dpi on plain paper. One face in each array was designated as "target." The other nine faces were chosen to systematically vary in spatial distance from the target face (regardless of sex, race, age, or any other gross attribute). Participants were asked to rate the similarity of each of the remaining nine faces to the target face, using a 10-point scale. Each participant thus rated nine faces in terms of their similarity to the target in each array, making a total of 27 ratings per participant (3 sets of 9 ratings). Twelve separate arrays were created, although individual participants received only three to rate.

Three arrays were included with a cover page of instructions in an experimental booklet. These instructions for the similarity-rating task allowed participants to use any facial quality they thought relevant but suggested that they take hair (length, darkness, and texture), hairline (relative position on the forehead), face shape, and skin texture (and color) into account, as well as differences between specific facial features (noses, mouths, and chins).

Participants were given a booklet containing the arrays, with instructions, and were asked to complete the tasks during a 20-min period at the beginning of a lecture they were attending.

Study 2b: Rating similarity in arrays of frontal, profile, and combined frontal-profile views. Viewing perspective is an important consideration for the type of measure of facial similarity proposed in this article—just as face recognition studies which test participants for their memory of face images run the risk of mistaking picture memory for face memory, so a similarity measure based on just one view of a face runs the risk of mistaking view similarity for face similarity. Accordingly, the rating tasks conducted on frontal views were repeated for profile views of the same stimulus images. There was also an additional condition in which participants were shown both frontal and profile views in the rating task. All other details were the same as those described for the frontal-rating task. Thus, participants received booklets containing 3 arrays of 10 faces and were asked to rate the similarity of each of nine members of each array to a designated target in the array. These arrays contained profile-view photographs of the same people shown in frontal pose to participants in Study 2a, or (in one condition) profile and frontal views of the same people shown in Study 2a only in frontal pose. Again, 12 arrays were constructed, but individual participants viewed and rated only three of these. The participants in Study 2b completed the task at the beginning of a course lecture.

Results

Study 2a

To assess the correspondence between participant ratings of similarity and the spatial distance measure, average similarity ratings were computed across participants and correlated with spatial distances for each of the 12 arrays. Table 5 shows correlation coefficients and Kendall concordance coefficients.

Table 5 also shows that there is a reasonably good correspondence between participant ratings of facial similarity and the spatial distance measure of facial similarity. In each of the 11 array tasks, the relationship is in the expected direction, and the (absolute) correlation is always greater than .40 in size. The median absolute correlation is .70, which is strong. The consistency of both size and direction are convincing demonstrations that the spatial distance measure corresponds in reasonable degree to human judgments of similarity. Nevertheless, it should be remembered that average ratings tend to inflate correlations, and that participants were far from consistent in their ratings, even with the modifications to the task instructions used in this study. (The inconsistency is shown in the size of the Kendall coefficients: These show that agreement was better than chance expectation, but not nearly perfect.)

Study 2b

Profile and frontal similarity scores were obtained for each face image for which both profile and frontal views were available; that is, the (PC-based) similarity of each face to every other face in the image set was determined for both frontal and profile image sets. Each face was present in a frontal view, and in a profile view, in separate databases, and PCA was conducted on the separate databases after standardizing face images. This allowed the determination of the similarity of any two faces in the frontal database, and the similarity of the same faces in the profile database. Because the aim was to determine whether the image sets generated equivalent similarity relations, a correlation was calculated between the set of frontal and profile similarity scores for each face image. The distribution of this correlation is shown in Figure 4.

The similarity relations are clearly not equivalent across frontal and profile views, but are certainly not insubstantial. The important question here is how strong a relation is acceptable? A perfect relation is improbable, as faces would show differences when viewed from different angles, and this can be expected to attenuate the strength of the relation. However, it is unlikely that similarity relations would change dramatically with a change of viewing perspective. To assess the strength of the relation shown in Figure 4, the relation was determined between similarity ratings made by participants when shown frontal views and ratings made when the same faces were shown in three-fourths profile view. At the same time, the relation between rated similarity and the PC-based measure of similarity was investigated.

Table 5
Relations Between Spatial Distance and Human Ratings of Facial Similarity in Study 2a

Array	*r*	*W*
1	−.66	0.41*
2	−.81	0.5*
3	−.66	0.27*
4	−.61	0.53*
5	−.70	0.43*
6	−.84	0.38*
7	−.69	0.51*
8	−.72	0.38*
9	−.42	0.2*
10	−.86	0.4*
11	−.83	0.22*

Note. Kendall's coefficient (*W*) is calculated for participant ratings only and reflects the degree of agreement among participants. Negative correlations are expected because the rating and spatial distance scales take reverse directions.
* $p < .01$.

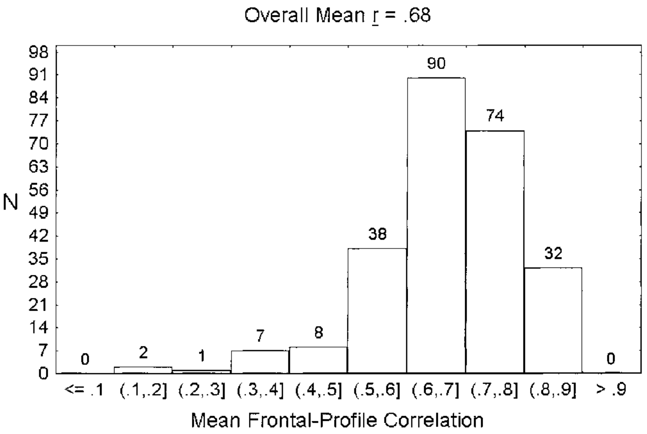

Figure 4. Distribution of correlations between frontal and profile similarity scores. To calculate the overall mean correlation, individual correlations were Fisher-transformed before computation, and the mean of these transformed correlations was inverse-transformed. Individual correlations represent the relation between frontal and profile similarity scores, calculated for each face image.

Similarity ratings of frontal, profile, and frontal + profile views were strongly related, and the correspondence between these ratings and spatial distances was again fairly high. These results again point to the usefulness of the PC-based similarity measure.

Table 6 shows relations between frontal, profile, and frontal + profile ratings, and relations between spatial distances calculated for each viewing perspective from the PCA. In each case where the results for participant ratings are reported, reported correlations involve mean rather than individual participant ratings. Note that arrays were structured in accordance with a facial distinctiveness manipulation (which was unsuccessfully derived from the PC solution, and is not discussed here), and that each array was presented in two sequences (i.e., faces occupied different positions in the arrays, across sequences).

Several things are clear from the table. First, it is clear that intercorrelations of spatial distances across frontal, profile, and combined views are at least as strong as those between participant ratings across the same views. Second, correlations between participant ratings of similarity and spatial distance estimates of similarity are again high (albeit not uniformly) and in the expected direction.[1]

Discussion

In both Studies 2a and 2b, there was a strong relation between similarity ratings of faces made by humans and spatial distances

[1] I am grateful to Peter Hancock for pointing out that because similarity ratings were not made by the same participants in the profile and frontal conditions, differences between individuals are likely to have reduced the strength of the correlation between ratings made in profile and frontal conditions.

Table 6
Intercorrelations of Similarity Ratings and Spatial Distances

	Low-distinctive target						High-distinctive target					
	F		P		F + P		F		P		F + P	
Rating	Seq 1	Seq 2	Seq 1	Seq 2	Seq 1	Seq 2	Seq 1	Seq 2	Seq 1	Seq 2	Seq 1	Seq 2
Intercorrelations of participant ratings of similarity												
Frontal (F)												
Seq 1	1.00						1.00					
Seq 2	0.92	1.00					0.57	1.00				
Profile (P)												
Seq 1	0.77	0.83	1.00				0.38	0.57	1.00			
Seq 2	0.58	0.78	0.85	1.00			0.34	0.81	0.83	1.00		
F + P												
Seq 1	0.85	0.94	0.94	0.90	1.00							
Seq 2	0.56	0.58	0.84	0.78	0.74	1.00	0.53	0.71	0.45	0.41		
Intercorrelations of spatial distance similarity scores												
F												
Seq 1	1.00						1.00					
Seq 2	0.92	1.00					1.00	1.00				
P												
Seq 1	0.86	0.86	1.00				0.82	1.00	1.00			
Seq 2	0.86	0.86	1.00	1.00			0.82	1.00	1.00	1.00		
F + P												
Seq 1	0.95	0.95	0.97	0.97	1.00		0.95	0.95	0.95	0.95	1.00	
Seq 2	0.95	0.95	0.97	0.97	1.00	1.00	0.95	0.95	0.95	0.95	1.00	1.00
Correlations of similarity ratings and spatial distance similarity scores												
F												
Seq 1	−.70						−.20					
Seq 2	−.70	−.79					−.20	−.25				
P												
Seq 1	−.70	−.82	−.91				−.28	−.56	−.88			
Seq 2	−.70	−.82	−.91	−.78			−.28	−.56	−.88	−.63		
F + P												
Seq 1	−.71	−.83	−.92	−.86	−.91		−.23	−.41	−.91	−.78	−.03	
Seq 2	−.71	−.83	−.92	−.86	−.91	−.77	−.23	−.41	−.91	−.78	−.03	−.09

Note. The wrong member in Sequence (Seq) 1 of the F + P condition, high-distinctiveness array, was inadvertently labeled as the "target," and intercorrelations are therefore not reported for this condition.

between face images, as derived from a PCA. In addition, Study 2b showed that there is a reasonable correspondence between spatial distances derived from different viewing perspectives of the same faces—at least as good as that shown by human participants. Taken together, these studies suggest that the spatial distance between two facial images provides real information about the similarity of the faces.

Three qualifications need to be made. First, the tasks in Studies 2a and 2b were structured to allow considerable variation in both spatial distance and human similarity ratings. Face arrays were composed of a mixture of images of male and female and of Black and White faces. This was done in order to maximize variation and to allow the relation between spatial distance and similarity ratings to show itself. There may be a much weaker relation between spatial distance and similarity ratings within more homogenous sets of face images, which would have implications for certain potential sites of application for the spatial measure. Second, despite an attempt to constrain undesirable sources of variation in human ratings of facial similarity, there was still considerable interrater disparity in ratings of facial similarity. Third, correlations between similarity ratings and PC-based distance measures were once again based on average similarity ratings, inflating the size of the coefficient over that which would be obtained from individual participants. However, this may be unavoidable in the presence of considerable participant variability: If participants do not agree on the similarity of particular faces, one cannot expect consistent correlations between distance measures and similarity ratings, and the sensible way to calculate the correlation may be by removing or reducing the individual variability.

Study 3: Stability of Similarity Ratings Over Time

In Studies 1 and 2, ratings of similarity showed considerable interparticipant variance. The most bothersome explanation for this is that people are inherently inconsistent in perceptions of similarity. If that is the case, perceptions of similarity at Time 1 would differ substantially from those at Time 2. In the present study, the reliability of similarity ratings over time is investigated by using a standard test–retest procedure.

Method

Participants

Participants were 21 Psychology 3 students. They participated in a face-rating task during the final tutorial of the term and a follow-up rating task some 3 weeks later.

Materials

Three face-rating tasks were prepared for use in the study. Each rating task followed the format used in Study 2 (i.e., an array of 10 faces was printed on a sheet of paper, one of which was designated as the "target" face). Participants were then asked to rate each face in terms of similarity to the target face. The first of the three arrays was used in the initial rating completed by participants, and the remaining two arrays were created for administration at the follow-up stage. Each of the additional arrays was created by removing a number of the faces presented in the initial array, replacing them with different faces and changing the order of presentation of the five original faces. Five faces were removed in one of the arrays, and the other four faces (constituting the original array of nine nontarget faces) in the second of the arrays. As in the previous studies, faces were chosen to systematically vary in spatial distance from the target face (regardless of sex, race, age, or any other gross attribute).

Procedure

Participants were approached during the last tutorial session of the term and asked to participate in a face perception and recognition study. Two groups of participants, attending different tutorial sessions, participated in the study. Each participant was given a rating task, which had the necessary instructions appended as a cover page. Participants were asked to provide names and student numbers, but were not told for what purpose they would be used. After a period of 2 weeks, participants were contacted by mail and asked to complete a second rating task (postal details for 2 participants were missing from university records, and these participants were not asked to complete the follow-up task). Rating tasks were attached to letters requesting their participation: there were two versions of the follow-up rating task, and these were randomly distributed among participants. Of the 19 participants, 12 were asked to participate in the follow-up stage of Study 5, and they submitted completed rating tasks.

Results

Test–retest reliability of similarity ratings proved to be fairly good, although this varied substantially across participants. The median correlation was .7, and the correlation between mean ratings (i.e., over participants) of initial and follow-up arrays was .94. The correlations are acceptably high, especially when computed over participants. However, it is prudent to bear in mind the methodological problems usually associated with test–retest designs. In particular, participants may show demand effects and attempt to recreate ratings from memory, rather than from a fresh scrutiny of the task. On one hand, the 3-week period between test and retest is some security against this threat, as is the insertion of five (or four) new faces and the rearrangement of remaining faces within arrays. On the other hand, the use of only five (or four) faces across tasks renders the estimate of the test–retest correlation coefficient a little unreliable.

Study 4: Applying the Facial Similarity Measure to the Task of Measuring Lineup Fairness and Lineup Bias

The results of Studies 1 through 3 suggest that a spatial distance measure of facial similarity may be a reasonable analogue to ratings of facial similarity made by human participants. There is much work required before researchers could use the measure with great confidence, but it is worth considering some practical applications of the measure at this point. To this end, Studies 4 and 5 attempt to apply the measure to an important site of practice where facial similarity is known to be important—the police lineup. In Study 4, the measure is applied to the problem of measuring lineup fairness, and in Study 5, the measure is used to explore the relationship between facial similarity and identification accuracy among simulated eyewitnesses. This work is a first attempt to apply the proposed measure to a practical problem and should be considered a preliminary investigation. If the measure predicts either mock witness performance or witness identifications, then further research will be needed to show how the measure could be systematically used to build lineups, or to evaluate them.

Some of the most influential and significant psycholegal research concerns a method of evaluating lineup fairness. This is the method of the mock witness. In essence, a number of research participants who have no direct knowledge of either the perpetrator(s) or the suspect(s) are given a verbal description and asked to identify the suspect from a lineup. To the extent that they are able to do this at rates better than chance expectation, the lineup is taken to be biased, or to have too few plausible foils. (Interested readers are referred to the special issue of *Applied Cognitive Psychology,* 1999, on measures of lineup fairness and lineup bias.) Several lineup measures are associated with this method, in particular "proportion choosing the suspect" (Doob & Kirshenbaum, 1973), "functional size" (Wells, Leippe, & Ostrom, 1979), and "effective size" (Malpass, 1981).

Facial similarity is directly relevant to the task of lineup construction. Police instructions in many countries emphasize the importance of ensuring that foils are matched to suspects on general aspects of physical appearance as well as facial characteristics (Shepherd, Ellis, & Davies, 1982). However, very little research has investigated the relationship of facial similarity to the performance of eyewitnesses or mock witnesses. In the case of mock witnesses, there appears to be only one relevant study, Malpass and Devine (1984), which reports a strong relationship between physical similarity and indices of mock witness performance. However, this study operationalized similarity in rather general physical terms and used no facial criteria except hair length, style, and color. Accordingly, in Study 4, the relationship of a more specific facial measure of similarity to measures of lineup fairness is investigated.

Method

Participants

Participants were 169 first-year medical students at UCT. They completed mock witness tasks at the beginning of a year-end course evaluation.

Materials

Three face images were selected from the frontal set of 278 to serve as suspects in mock witness tasks. (The images were selected so as to vary on a distinctiveness measure, as for Study 4, but because the question of distinctiveness is outside the scope of the present article, there is no further discussion of it here, except where unavoidable.) These images were then given to three raters, in different orders, to get verbal descriptions for the mock witness tasks. Raters were shown each image for 20 s, and after presentation of the image were instructed to describe the person they had just seen. Raters were asked to make this description highly accurate—other people should be able to identify the person on the basis of the description alone. Written descriptions were then carefully examined and combined to form a description of each suspect.

Six 8-person photospread lineups were then created for each of the suspects. These lineups were constructed so as to structure the similarity of lineup members in relation to the target. This was achieved by selecting images that were in the first 6th, second 6th . . . , sixth 6th of the distribution of similarity scores, calculated in relation to the suspect. In this way, three sets of six lineups, of differing target–member similarity, were constructed, making 18 total lineups. Three lineups—one selected from each of the three sets, at random—were combined, along with a cover page of instructions, into an experimental booklet. Each participant was therefore exposed to each of the targets only once. Six of these booklets were created in total. Instructions required participants to identify suspects on the basis of the descriptions provided to them.

Procedure

Participants were addressed at the beginning of a year-end course evaluation meeting and asked to participate in a study of face recognition and perception. Experimental booklets, which were prearranged in random order with respect to manipulations constituting the design of the study, were then distributed to participants. Participants completed the tasks independently, at their own pace. After which, booklets were collected.

Results

Several dependent measures were formed because several measures of lineup fairness are currently used in psycholegal research, and one of the aims of work reported in this thesis is to compare the measures. At the simplest level, a binary dependent variable was created according to whether participants had chosen the suspect. A log-linear analysis on a three-way table embodying the design of the study (Similarity × Distinctiveness × Correctness) showed that a model incorporating all main effects, and the interaction effects Similarity × Correctness, Distinctiveness × Correctness, produced a satisfactory fit to observed frequencies. (L.R.), $\chi^2(20, N = 169) = 25.4, p > .19$. In other words, it was not necessary to include the three-way interaction in the model, nor was it necessary to include the remaining two-way interaction (Distinctiveness × Similarity): The effects of similarity and distinctiveness were independent of each other. Figure 5 shows the proportion of accurate identifications per experimental condition.

The similarity effect is evident in Figure 5. It is clear that increasing similarity of lineup members to the suspect reduces the likelihood that the identity of the suspect can be guessed by witnesses armed with only a brief verbal description.

Alternate measures of lineup fairness were then computed. These included the measures known as "effective size" (Malpass, 1981), and "E" (Tredoux, 1998), which are intended to give an estimate of the number of plausible foils in a lineup. A lineup consisting of a Black male suspect, two Black male foils, and five White policemen may not be biased against the suspect (e.g., the Black male foils draw equivalent numbers of choices), but there is little doubt that it has very few plausible foils.

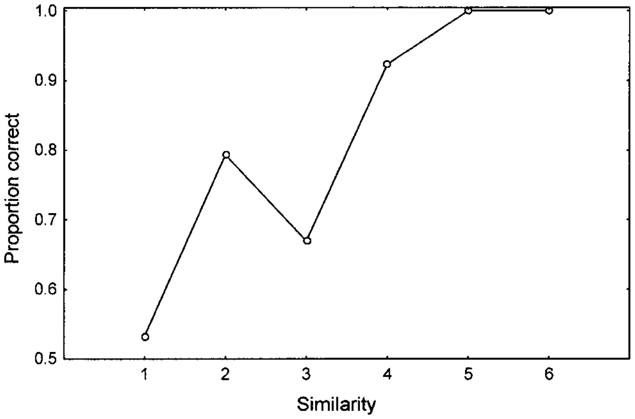

Figure 5. Similarity effects on mock witness accuracy.

Correlations between measures of lineup fairness and a pseudo variable, representing facial similarity, were computed over the 18 conditions of Study 4 and are reported in Table 7. Correlations between alternate measures of lineup fairness are very strong, and correlations of all measures with lineup similarity are strong and in the expected direction.

Study 5: Facial Similarity and Identification Accuracy

Facial similarity has received very little attention in either the face recognition or witness identification literatures. On occasions where it has been investigated, facial similarity has proved to be a variable of substantial import. We know that it is strongly associated with indices of lineup fairness (Malpass & Devine, 1984; Study 4 in the present set) and it appears to affect identification accuracy in simulated identification scenarios. Studies that have investigated the impact of similarity in identification scenarios have typically used indirect measures of similarity (e.g., ratings made by independent judges), or a priori similarity classifications. Of particular significance may be the common practice in eyewitness studies of forming perpetrator-absent lineups by designating an "innocent suspect." The innocent suspect is typically chosen to be (subjectively) similar to the perpetrator, and this may have disguised effects that varying degrees of similarity have on identification rates. In particular, it is likely to have increased false positive identifications.

The present study aims to provide some information on the effect that systematic variation of facial similarity has on identification ability of eyewitnesses, using a PC-based spatial distance measure of facial similarity.

A few important methodological considerations are worth outlining before the report of Study 5. Work reported in the identification literature has shown that simulated identification scenarios that use lineups as recognition tests need to bear in mind the distinction between target-present and target-absent lineups. Both are necessary if one wishes to correctly evaluate identification accuracy, and especially if one wishes to obtain "diagnosticity" estimates (Wells & Lindsay, 1980). Second, one of the most significant contributions of witness identification research has been the development of "sequential lineups" (Lindsay & Wells, 1985). Any assessment of the effect of facial similarity on identi-

Table 7
Correlations Between Measures of Lineup Fairness and Facial Similarity

Variable	1	2	3	4
1. Similarity	—			
2. Esize (1)	−.72	—		
3. Esize (2)	−.78	.97	—	
4. E	−.77	.96	.94	—
5. Functional size	−.73	.89	.86	.97

Note. Esize (1), Esize (2), E, and Functional are measures of lineup fairness.

fication ability needs to use both forms of lineup (i.e., simultaneous and sequential). Study 5 incorporates both of these considerations.

Method

Participants

Participants were 22 Psychology 3 students and 46 Psychology 2 students at UCT. Psychology 3 students participated in the experiment during a tutorial, and Psychology 2 students participated during a lecture.

Materials

Two sets of 3 face images were chosen from the frontal set of 278 images, along with corresponding profile views of these images. A set of face images was presented to each participant at the first stage of the experiment, in document form. The frontal images were also embedded in lineup arrays, which were presented to participants at the final stage of the experiment. The lineup arrays varied in terms of target-foil similarity, so as to constitute three levels of similarity. This was achieved by selecting foils who fell within 0–10, 45–55, or 90–100 percentile points of the target on the similarity measure.

Two documents were prepared to present to participants at the first stage of the experiment. Each document contained one of the sets of images, along with fictitious descriptions of each of the people represented by the images. Participants were required to read these descriptions, and having read them, to write three additional facts they believed to be probably true of each of the 3 people (this was intended as a filler task). These documents presented information that was later tested with lineup arrays.

Simultaneous and sequential lineup arrays were then created for each of the three images: Each of these arrays either contained or omitted the relevant face image, thus constituting the target-present, target-absent manipulation. Arrays were combined into booklets three at a time: each target was represented in one of these arrays (either in target-present or target-absent form). In the case of simultaneous parades, these were simply stapled together, along with an instruction page. In the case of sequential parades, a minibooklet was created for each of the three arrays. One image was printed on each page of the minibooklet. Three minibooklets and a page of instructions were inserted into an envelope. Instructions attempted to ensure that sequential lineup arrays were completed as sequential tasks.

Design

Study 5 was a 2 × 2 × 2 × 3 factorial experiment, with one dependent variable. Factors were (a) Lineup Structure (simultaneous, sequential); (b) Target Presence (present, absent); (c) Image Set (a, b); and (d) Target-Foil Similarity (high, moderate, low). Lineup Structure, Target Presence, and Image Set were between-participants factors, while Target-Foil Similarity was a repeated-measures factor. Assignment of Image Set, Target Presence, and Lineup Structure conditions was random.

Procedure

The experimental procedure differed slightly across the groups of participants, and this is worth detailing. Psychology 3 students were recruited from groups in voluntary statistics tutorials at the end of the academic year. These groups varied in size, ranging from 2 to 7. At the beginning of the tutorial, they were given a document containing frontal and profile views of targets and were allowed 5 min to complete the task contained in the document. There were two documents designed for this stage of the experiment (as detailed in the *Materials* section), and participants were randomly assigned one of the documents. A statistics lesson followed and lasted 30 min. At the end of the lesson, participants were asked to complete the second part of the experiment (they did not know that there was a second part) and were randomly assigned an array booklet, which contained either simultaneous or sequential-lineup array tasks, corresponding to the image set they had received at the beginning of the experiment. Some of the arrays in these tasks contained the target, and others did not. The assignment of target-present and target-absent arrays had been effected randomly in the development of the experimental materials.

Psychology 2 students were addressed at the beginning of a year-end course lecture and asked to participate in a study of face recognition and perception. Each participant was then handed two envelopes. One envelope was marked “Open this envelope when you receive it. Complete the task inside it, place the completed task back in the envelope, and seal it.” This envelope contained the first stage of the experiment, namely the document discussed in the *Materials* section. The second envelope was marked “Do not open this envelope until instructed to do so,” and was sealed. Participants completed the first task, after which the lecture commenced. The lecture lasted 30 min, after which participants were instructed to open the second envelope. They then completed the lineup-array tasks, which were contained in the envelope. Lineup Structure, Target Presence, and Image Set conditions had been randomly distributed across envelopes, and the assignment of participants to conditions was also random.

Results

Similarity, Lineup Structure, and Target Presence all proved to have significant effects on participant performance in lineup tasks. Participants made better decisions with sequential lineups, but only when targets were absent; and low-similarity lineups improved accuracy of identifications in sequential and simultaneous lineups in target-absent and target-present conditions. Image set had no effect on participant performance, and because this manipulation was intended to check the generalization of findings across images, it is not discussed further.

Results from lineup tasks may be assessed in terms of correct identification decisions (i.e., to treat identifications of the perpetrator when he or she is present as equivalent to witness indications that the perpetrator is not present when he or she is indeed not present), but it is generally more useful to classify results in relation to target presence–absence.

Results are thus better considered separately for target-present and target-absent lineups. Chi-square tests of association between Lineup Structure and identification performance were calculated for the six combinations of target presence–absence (two levels) and similarity (three levels). Associations were not significant for any of the target-present conditions, but in two of the target-absent conditions there was a significant or nearly significant association between Lineup Structure (simultaneous vs. sequential) and identification decision (incorrect identification vs. correct rejection), that is for high similarity, $\chi^2(1, N = 31) = 7.04$, $p < .01$, and for low similarity, $\chi^2(1, N = 39) = 3.54$, $p < .06$.

The design of Study 5 demands an investigation of similarity effects across Target Presence and Lineup Structure manipulations. However, there are two problems in implementing such an analysis. The categories that the dependent variable assumes differ across target-present and target-absent lineups, making comparison difficult. This can be overcome by reclassifying identification decisions as "correct" or "incorrect" (which loses information, but achieves comparability), and in the analysis reported below, this variable is called *identification accuracy.* Second, the design of Study 5 incorporates a repeated-measures factor (similarity), which, like all other variables in the design, is categorical. Log-linear analysis is the analytic method of choice for exploring main and interaction effects of categorical independent variables on categorical dependent variables, but there is no general method for designs that use repeated measures. I proceeded to analyze the data for Study 5 with standard log-linear techniques (i.e., assuming all factors to be independent): this appeared to be the only option that would provide a full evaluation of the study, even though it does not take into account the dependency in the data.

Tests of all *k*-factor interactions suggested that two-, $\chi^2(9, N = 68) = 23.7$, $p < .01$, and four-factor interactions, $\chi^2(2, N = 68) = 4.65$, $p < .1$, should be included in the model, but tests of partial association suggested only a two-way interaction between similarity and identification accuracy, $\chi^2(9, N = 68) = 17.54$, $p < .01$. Specific models tested against each other revealed that the model (Similarity × Identification Accuracy) provided an adequate fit, and was simpler than any rival models.

The conclusion from the log-linear analysis is that the interaction between similarity and identification accuracy is a sufficient basis on which to understand the results of Study 5. This interaction is shown in terms of cell frequencies, in Table 8, and is explicable almost entirely in the deviation of the low-similarity condition from other conditions: In this condition, correct identifications were much more frequent than in high- and moderate-similarity conditions.

Discussion

Results of Study 4 showed that the PCA-based measure of facial similarity is strongly related to mock witness measures of lineup fairness, and may prove to be a useful proxy for these more expensive and less direct measures. Ideally, one should also be able to use the measure in foil selection, particularly for photo lineups. The measure would (theoretically) allow one to structure the similarity of a lineup in an objective manner.

Table 8
Frequencies for the Interaction Between Facial Similarity and Identification Accuracy

	Identification decision	
Similarity	Incorrect	Correct
High	41 (60%)	27 (40%)
Moderate	38 (56%)	30 (44%)
Low	17 (25%)	51 (75%)

Note. Row percentages appear in parentheses.

In Study 5, the PCA-based measure was used to investigate the effect of similarity on witness identifications, using a simulated identification scenario. Results from the study showed that similarity is strongly related to witness accuracy. Low-similarity lineups led to greater accuracy, in terms of hits and correct rejections, than moderate- or high-similarity lineups. This finding bears out Wells's contention (Wells, Seelau, Rydell, & Luus, 1994) that high-similarity lineups may not provide witnesses with "propitious heterogeneity." The finding also lies quite uneasily next to the finding from Study 4 that high-similarity lineups are associated with greater lineup fairness. It may be that the conventional way in which lineup fairness is understood and investigated in the psycholegal literature is mistaken. Fairness is usually assessed with mock witness tasks and is measured as bias toward (or against) the suspect, or in terms of the number of plausible foils in the lineup. The mock witness task assumes that a lineup is unfair if a witness is able to identify the suspect with only a brief description of the suspect. What constitutes "brief" has not been investigated, and this may be overdue: One expects a witness who has a detailed description to successfully identify a suspect, so we need to know more about the relation of the description to mock identification accuracy.

Study 5 also corroborated findings from previous studies in the field regarding the utility of sequential lineups. Sequential lineups were associated in this study with fewer false alarms than simultaneous lineups, while securing the same number of hits.

In sum, the key findings from Studies 4 and 5 are that the PC-based measure of facial similarity may be able to stand-in as a proxy for standard measures of lineup fairness, but that these standard measures may themselves need reexamination in light of the relation between lineup similarity and identification accuracy uncovered in Study 5.

General Discussion

On the basis of this empirical research, it appears that measures of facial similarity derived from principal-component representational bases are sufficiently closely related enough to participant ratings of similarity to use them as approximations of perceived similarity. In addition, the measure correlated positively with standard indices of lineup fairness, and with identification accuracy in simulated lineups. The measure shows potential as a research aid in face recognition studies, and may prove to be a useful direct measure of facial similarity in practical tasks such as lineup construction.

Some limitations of the test of the spatial distance similarity measure are worth reemphasizing here. First, the face images used in test arrays were allowed to vary maximally—male and female faces, Black and White faces, young and old faces appeared alongside each other in test arrays, and while the spatial distance measure showed a strong relation with human similarity judgments in these arrays, it is not certain that this would be the case in arrays of greater homogeneity. In particular, in the practical situations where the spatial distance measure may be of great usefulness (e.g., lineup and identikit construction) the target population is typically fairly homogenous. The use of heterogenous collections of faces may explain the difference in the size of the correlations between the present study and those reported by Hancock et al. (1998). However, it should also be noted that the use of homogenous collections of faces runs the risk of introducing range restrictions. If researchers select sets of faces that are very similar to each other, they would curtail facial variation, and this range restriction would probably evidence itself in lower correlations between human ratings of similarity and those based on distance in PC spaces. Nevertheless, in order to use the spatial distance measure in the practical ways suggested in Studies 4 and 5, researchers would need to test it with sets of faces that are more like those they would encounter in practice, and those would be more homogeneous than the sets used in the studies reported in this article.

An important further piece of research would thus be to examine the relation between the spatial distance measure and human judgments of facial similarity in homogeneous subpopulations of face images. One would need to be very careful in such research, however, to separate the limitations of the spatial distance measure and the effects of variability in human judgments of facial similarity. A striking finding in all the empirical research reported in this article was the presence of considerable variability in judgments of similarity. Although the variability does not appear to be due to intrarater instability over time (as Study 3 showed), it does raise questions of some significance. In particular, it makes the task of validating substitute measures of similarity very difficult—what is there to correlate the substitute measure against?

There are also many questions that need to be explored before the spatial distance measure can be considered ready for practical applications. In the first instance, little is known about the performance of the PCA approach across diverse “face populations.” Although some authors claim that relatively few faces are needed to serve as a generative basis for relatively large sets of faces, it remains to be seen how well the approach works when faces from a particular population (say *Y*) are projected into a space formed from faces from a different population (say *X*). One possibility, in the absence of research on this question, is to quantify the difference between the projection of a face image into the space and the original face image (see, e.g., Kirby & Sirovich, 1990). If the difference is large, the face is poorly represented in the space.

Second, the present research used a somewhat cumbersome procedure for standardizing face images before applying PCA. Most practical applications will probably need an implementation that works in real time, and although this may presently seem far-fetched, there is a good deal of exciting work that points to the real possibility of automating the process of face identification, standardization, and segmentation (see especially Cootes & Taylor, 2001).

Despite these limitations and uncertainties, there are several potentially useful and exciting applications of the PCA facial similarity measure. Two are singled out for discussion here, and I am able to report that researchers have already made substantial progress in implementing one of these.

The first potential application is the development of a software tool for constructing face lineups of varying similarity. In this application, a large database of faces is collected and standardized according to the procedure outlined earlier in this article. The faces are then subjected to PCA, and the database is updated to contain coefficients for each face in the associated eigenspace. A new face (i.e., one which is not represented in the existing space) can be projected into the space and coefficients determined on each of the underlying eigenfaces. It is trivial then to identify the nearest neighbors to the face in the eigenspace and to treat these as the most similar faces in the database. Some experimentation would be necessary to avoid selecting faces that are too similar, but it may be a very useful tool for both researchers and law enforcement agencies. Researchers expect to soon have a prototype database ready for pilot experimentation, using a set of about 5,000 distinct faces. This initial database contains faces of varying photographic quality, and researchers recognize that it is important to collect high quality images of faces that are photographed under controlled lighting conditions, and my colleagues are cooperating with other researchers to achieve a large, high quality database.

The second potential application is to produce facial composite software. One of the greatest potential benefits of a PCA “face-space” is that it is possible to represent any face in such a space with a measurable amount of error, even if the face is not in the original set of faces. In other words, a face that is not in the original set can be constructed from a linear composite of the underlying eigenfaces, and it is easy to show that there exists an optimal set of component coefficients which will give the best possible reconstruction within the limitations of the underlying set of eigenfaces. The task is to find the optimal set of weights. When an image of the target face is available, this can be done by projecting the target face into the eigenspace, but when the target face is only available as a memory image, the task is somewhat more difficult. One approach researchers have explored is the application of a learning algorithm known as *population based incremental learning* (Rosenthal, de Jager, & Greene, 1998), and preliminary experiments have shown that a PCA-based system can be used in an interactive fashion to successfully reconstruct memory images of faces. Researchers believe that this prototype system holds great promise, particularly as an alternative to police identikit systems and hope to take a revised, working model into the field for experimentation in the near future. It is also a fact that researchers at Stirling University have developed a similar system and have recently reported first test results (Hancock, 2000).

References

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

Benson, P. J., & Perrett, D. I. (1991b). Synthesizing continuous tone caricatures. *Images & Vision Computing, 9,* 123–129.

Bowns, L., & Morgan, M. J. (1993). Facial features and axis of symmetry extracted using natural orientation information. *Biological Cybernetics, 70,* 137–144.

Bruce, V. (1979). Searching for politicians: An information-processing approach to face recognition. *Quarterly Journal of Experimental Psychology, 31,* 373–395.

Bruce, V. (1994). Stability from variation: The case of face recognition. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 47A,* 5–28.

Cootes, T. F., & Taylor, C. J. (2001). *Statistical models of appearance for computer vision.* Unpublished manuscript, Wolfson Image Analysis Unit, University of Manchester.

Craw, I., & Cameron, P. (1991). Parameterising images for recognition and reconstruction. *Proceedings of the British Machine Vision Conference*

BMCV '91 (pp. 367–370). New York: Turing Institute Press and Springer-Verlag.

Davies, G. M., Shepherd, J. W., & Ellis, H. D. (1979). Similarity effects in face recognition. *American Journal of Psychology, 92,* 507–523.

Doob, A. N., & Kirshenbaum, H. M. (1973). Bias in police lineups—Partial remembering. *Journal of Police Science and Administration, 1,* 287–293.

Hancock, P. J. B. (2000). Evolving faces from principal components. *Behavior Research Methods, Instruments and Computers, 32,* 327–333.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (1998). A comparison of two computer-based face identification systems with human perceptions of faces. *Vision Research, 38,* 2277–2288.

Hancock, P. J. B., Burton, A. M., & Bruce, V. (1996). Face processing: Human perception and principal components analysis. *Memory & Cognition, 24,* 26–40.

Harmon, L. D. (1973). The recognition of faces. *Scientific American, 229,* 70–82.

Hirschberg, N., Jones, L. E., & Haggerty, M. (1978). What's in a face: Individual differences in face perception. *Journal of Research in Personality, 12,* 488–499.

Kalocsai, P., Zhao, W., & Elagin, E. (1998). Face similarity space as perceived by humans and artificial systems. *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition,* Nara, Japan, 177–180.

Kirby, M., & Sirovich, L. (1990). Application of the Karhunen–Loeve procedure for the characterization of human faces. *IEEE: Transactions on Pattern Analysis and Machine Intelligence, 12,* 103–108.

Laughery, K. R., Fessler, P. K., Lenorovitz, D. R., & Yoblick, D. A. (1974). Time delay and similarity effects in facial recognition. *Journal of Applied Psychology, 39,* 490–496.

Li, H. Y., Qiao, Y., & Psaltis, D. (1993). Optical network for real-time face recognition. *Applied Optics, 32,* 5026–5035.

Lindsay, R. C. L. (1994). Biased lineups: Where do they come from? In D. Ross, J. D. Read, & M. P. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 182–200). New York: Cambridge University Press.

Lindsay, R. C. L. (Ed.). (1999). Measuring lineup fairness [Special issue]. *Applied Cognitive Psychology, S1.*

Lindsay, R. C. L., & Wells, G. L. (1985). Improving eyewitness identifications from lineups: Simultaneous versus sequential lineup presentation. *Journal of Applied Psychology, 70,* 556–564.

Malpass, R. S. (1981). Effective size and defendant bias in eyewitness identification lineups. *Law and Human Behaviour, 5,* 299–309.

Malpass, R. S., & Devine, P. G. (1983). Measuring the fairness of eyewitness identification lineups. In S. M. A. Lloyd-Bostock & B. R. Clifford (Eds.), *Evaluating witness evidence* (pp. 81–102). London: Wiley.

Malpass, R. S., & Devine, P. G. (1984). Research on suggestion in lineups and photospreads. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 12–37). Cambridge, England: Cambridge University Press.

Milord, J. T. (1978). Aesthetic aspects of faces: A (somewhat) phenomenological analysis using multidimensional scaling methods. *Journal of Personality and Social Psychology, 36,* 205–216.

O'Toole, A. J., Abdi, H., Deffenbacher, K. A., & Valentin, D. (1993). Low-dimensional representation of faces in higher dimensions of the face space. *Journal of the Optical Society of America, 10,* 405–411.

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 406–417.

Penev, P. S., & Sirovich, L. (2000). The global dimensionality of face space. *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition,* 264–270.

Rhodes, G. (1988). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception, 17,* 43–63.

Rosenthal, Y., de Jager, G., & Greene, J. (1998). *A computerised face recall system using eigenfaces.* Unpublished manuscript, University of Cape Town, Private Bag, Rondebosch, South Africa.

Shepherd, J. W., Ellis, H., & Davies, G. M. (1982). *Identification evidence: A psychological evaluation.* Aberdeen, England: Aberdeen University Press.

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

Tredoux, C. G. (1998). Statistical inference on measures of lineup fairness. *Law and Human Behavior, 22,* 217–237.

Valentine, T. (1991a). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A,* 161–204.

Valentine, T. (1991b). Representation and process in face recognition. In R. Watt (Ed.), *Vision and visual dysfunction. Vol. 14: Pattern recognition in man and machine.* London: MacMillan.

Valentine, T., & Endo, M. (1992). Towards an exemplar model of face processing: The effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 44A,* 671–703.

Valentine, T., & Ferrara, A. (1991). Typicality in categorization, recognition and identification: Evidence from face recognition. *British Journal of Psychology, 82,* 87–102.

Wells, G. L., Leippe, M. R., & Ostrom, T. M. (1979). Guidelines for empirically assessing the fairness of a lineup. *Law and Human Behaviour, 3,* 285–293.

Wells, G. L., & Lindsay, R. C. L. (1980). On estimating the diagnosticity of eyewitness nonidentifications. *Psychological Bulletin, 88,* 776–784.

Wells, G. L., Seelau, E. P., Rydell, S. M., & Luus, C. A. E. (1994). Recommendations for properly conducted lineup identification tasks. In D. Ross, J. D. Read, & Toglia, M. P. (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 223–244). New York: Cambridge University Press.

Young, M. P., & Yamane, S. (1992, May 29). Sparse population coding of faces in the inferotemporal cortex. *Science, 256,* 1327–1331.

Received November 2, 2001
Revision received April 15, 2002
Accepted April 18, 2002 ■