Reverse engineering the face space: Discovering the critical features for face identification

Naphtali Abudarham
School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel

Galit Yovel
School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel
Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel

How do we identify people? What are the critical facial features that define an identity and determine whether two faces belong to the same person or different people? To answer these questions, we applied the face space framework, according to which faces are represented as points in a multidimensional feature space, such that face space distances are correlated with perceptual similarities between faces. In particular, we developed a novel method that allowed us to reveal the critical dimensions (i.e., critical features) of the face space. To that end, we constructed a concrete face space, which included 20 facial features of natural face images, and asked human observers to evaluate feature values (e.g., how thick are the lips). Next, we systematically and quantitatively changed facial features, and measured the perceptual effects of these manipulations. We found that critical features were those for which participants have high perceptual sensitivity (PS) for detecting differences across identities (e.g., which of two faces has thicker lips). Furthermore, these high PS features vary minimally across different views of the same identity, suggesting high PS features support face recognition across different images of the same face. The methods described here set an infrastructure for discovering the critical features of other face categories not studied here (e.g., Asians, familiar) as well as other aspects of face processing, such as attractiveness or trait inferences.

Introduction

While observing the two pairs of faces presented in Figure 1A, most observers indicate that the two faces on the left are more similar and may belong to the same person whereas the face pair on the right are of two different people. However, in both pairs, we generated the right face by modifying a different subset of facial features of the left face. Why is the right face pair perceived to be more different? In other words, what are the critical features that define an identity and determine whether two faces belong to the same person or different people? In essence, we are asking: How do we identify people?

One prominent theory, suggested by Valentine (1991, 2001), which could have potentially revealed these critical features, is the face space theory. According to the face space theory, faces are represented in a multidimensional space in which each dimension corresponds to a feature in the face. Thus, each face is represented by a point in space, or a feature vector, in which each of the vector values indicates the magnitude of a feature on its unique scale. Distances between feature vectors ("face space distances") correspond to perceptual differences between faces. This theory further assumes that each identity takes up a subspace, which includes its different appearances, such as changes in head pose, illumination, aging, expression, and so on (Lewis, 2004; Tanaka, Giles, Kremen, & Simon, 1998; Valentine, 1991, 2001; see Figure 1B). Despite this comprehensive account of the representation of face identity and the many studies it inspired (Blank & Yovel, 2011; Lee, Byatt, & Rhodes, 2000; Leopold, Bondar, & Giese, 2006; Lewis, 2004; Rhodes & Jeffery, 2006; Rhodes & Leopold, 2011), neither the original theory nor later empirical work has revealed what the dimensions of the face space are, that is, which facial features are used for determining the identity of a face.

In the current study, we applied the face space framework to discover the critical features for face identification by using the following simple but crucial observation: Features that are critical for identification

Citation: Abudarham, N., & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*(3):40, 1–18, doi:10.1167/16.3.40.

doi: 10.1167/16.3.40 Received September 19, 2015; published February 29, 2016 ISSN 1534-7362

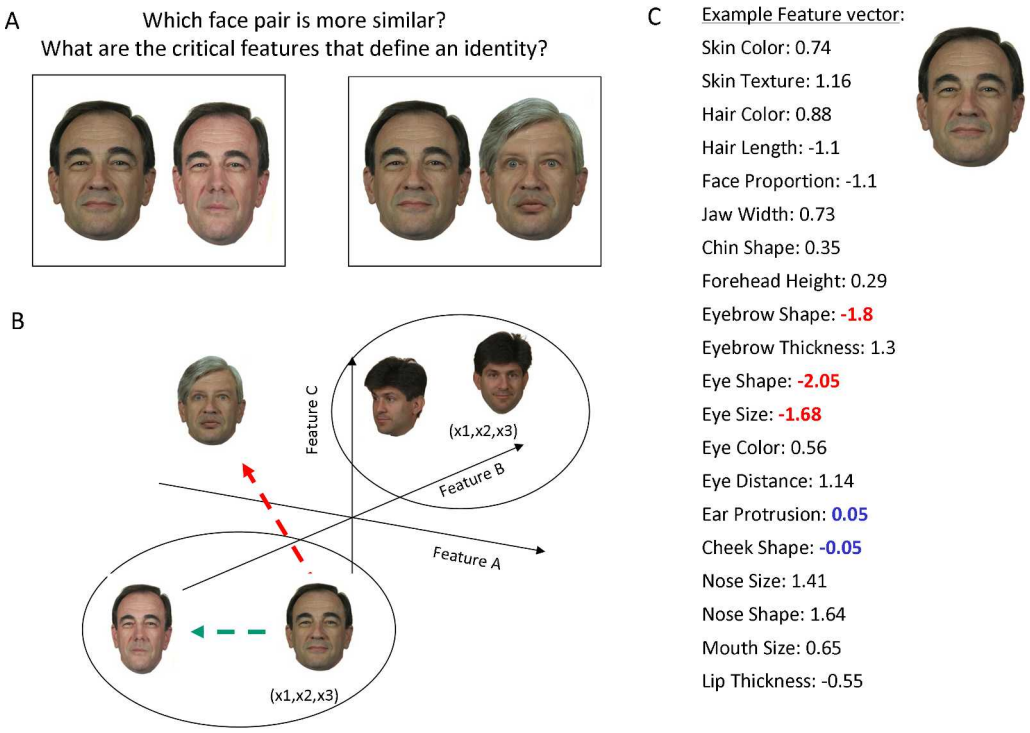

Figure 1. (A) Are these pairs of pictures of the same person or of different people? Which of the two pairs is more similar? In both pairs, we took an original picture (left face in each pair) and changed it in a different way, producing a different perceptual difference between the original and the changed face. (B) A schematic description of the face space theory (Valentine, 1991). The face space is a multidimensional space: Each dimension corresponds to a facial feature; therefore, each face may be described by a vector of values. Perceptual differences between faces are described as distances in the face space. Identities take up a subspace (black ellipse), in which all points belong to that same identity. The green arrow indicates the distance between an original and changed face in which feature changes did not change the identity of a face whereas the red arrow indicates feature changes that change the identity of a face and move it outside of its subspace. (C) An example feature vector of one of the faces in the database. Extreme values (in red) correspond to distinctive features, and average values (in blue) correspond to average features relative to features of the 100 faces in the database.

are those that when *changed* would move the face outside its subspace, causing a change in identity. Feature changes that do not move the face outside its subspace are features that do not change the identity of a face and therefore are not critical for face identification (Figure 1B). This approach is inspired by the reverse engineering technique known as "adversarial learning" (Lowd & Meek, 2005), which is typically used to discover what features a classifier uses by systematically changing the input to the classifier and examining the respective output. Because the face space theory makes an analogy between face identification and machine classification, our intent is to discover what features are critical for human face identification by deliberately changing features and testing the resulting perceptual effect. This idea also could be thought of as a process for dimensionality reduction in the (possibly infinite dimension) face space (Townsend, Solomon, & Smith, 2001; Townsend & Thomas, 1993).

To apply this theoretical approach, we constructed a face space using a concrete set of 20 features of natural faces. Such a face space has not been constructed before with natural face images. Next, we conducted a series of experiments based on human judgments to quantify feature differences between faces on one hand and to measure whole face similarity on the other hand. We then showed that distances in this space correspond with perceptual similarity judgments, thus validating the chosen dimensions of the face space. This enabled us to examine the relative importance of different features to perceptual face identity judgments, thereby discovering what the critical features for face identification are. Notably, in this study, we constrained our stimuli to unfamiliar faces of male Caucasian adults and used a simultaneous face-matching task for measuring perceptual similarity and identification. Nevertheless, the method we propose here can be applied to any other type of face (e.g., Asian faces, famous faces) and any other facial aspects (e.g., attractiveness, trait inferences).

To achieve this goal, our study included five experiments. In Experiment 1, we constructed and

Category	Feature name and description	Scale (ranging between)
General Appearance	Skin color	Light–dark
	Skin texture:	Smooth–textured
	Textured includes marks, scars, freckles, wrinkles	
Hair	Hair color	Light–dark
	Hair length	Bald–long hair
Face shape	Face proportion:	Wide & short – symmetrical – narrow & tall
	Ratio between length (top to bottom) and width	
	Jaw width	Narrow–wide
	Chin shape	Pointed–rounded–flat (square)
Forehead	Forehead height:	Short–long
	Distance between the eyebrows and the hairline	
Eyebrows	Eyebrow shape	Rounded–straight
	Eyebrow thickness	Thin–thick
Eyes	Eye shape	Narrow–round
	Eye size	Small–large
	Eye color	Light–dark
	Eye distance	Small–large
	The distance between eye centers (pupils)	
Ears	Ear protrusion:	Adjacent to the skull–protruding outward
	Flat on the skull or protruding outward	
Cheeks	Cheek shape	Sunken–puffy
	Sunken and skinny cheeks or full and puffy	
Nose	Nose size:	Small–Large
	Overall size	
	Nose shape:	Pointed–Flattened
	Pointed and thin or flat and wide	
Mouth	Mouth size:	Small–Large
	General size, width from left to right	
	Lip thickness	Thin–Thick

Table 1. The 20 features that were selected for the construction of the face space.

validated a 20-feature face space by showing that distances in face space are correlated with perceptual similarity judgments (Figure 2A). In Experiment 2, we further validated our choice of features and the metric of the face space by changing a subset of features in a quantitative manner and examining the effects of these changes on perceptual judgments of the whole face (Figure 2B). In Experiment 3, we assessed the discriminative power of the 20 features by measuring the perceptual sensitivity (PS) of humans to detect differences in each of these features across different faces (for example, which face has thicker lips). This enabled us to identify a subset of features with high PS that may be critical for face recognition (Figure 4). In Experiment 4, we showed that replacing features of high PS (and therefore of high discriminative power) changed the identity of a face more than replacing features of low discriminative power (Figures 5 and 6). Finally, in Experiment 5, we showed that features of high discriminative power vary minimally across different variations of the same identity (Figure 7), suggesting that they play an important role in our ability to recognize faces across different appearances.

Experiment 1: Constructing and validating a face space

Experiment 1A: Constructing a face space

To construct a face space, we selected a set of 20 features that could be used to describe faces as feature vectors (see Table 1, Figure 1C; see also Catz, Kampf, Nachson, & Babkoff, 2009; Freiwald, Tsao, & Livingstone, 2009, for examples of feature sets). Participants were asked to assign values to each of the 20 features, allowing us to describe each face as a point in a face space.

Methods

Participants

A total of 55 subjects participated in the feature value assignment procedure. Each subject tagged four features selected randomly out of the total 20 features across all

100 faces (the limited number of features was to avoid fatigue). This resulted in an average of 11 subjects per feature. The subjects were first-year psychology students from Tel Aviv University, performing the experiment for course credit. The experiment was approved by the ethics committee of Tel Aviv University.

Stimuli

The stimuli used for face feature evaluation were 100 pictures of unfamiliar faces taken from the color-FERET database (Phillips, Moon, Rizvi, & Rauss, 2000; Phillips, Wechsler, Huang, & Rauss, 1998). Out of the 1,199 different persons in the color-FERET database, we randomly selected 100 persons whose images met the following criteria: adult male Caucasians who had two different frontal view images with uniform lighting, neutral expression, no glasses, and no facial hair. These facial images were then cropped from below the chin and up (including hair and ears) and placed on a white background.

Face-tagging procedure

To tag faces and to create feature vectors for faces, subjects were presented with faces on a computer screen, and under each face was a scale, ranging from −5 to +5, for marking the magnitude of the currently measured feature (see Figure 3A and also Catz et al., 2009). Subjects were asked to tag all faces in the data set according to one feature before moving on to the next feature. All 100 faces were available for viewing on the screen per feature, and subjects were asked to make their judgments based on all faces. They could scroll up and down, viewing all faces and adjusting the values as they pleased until they were done and ready to move on to the next feature. To assure that subjects understood the scale of each feature, we provided, prior to displaying the faces, schematic faces that portrayed extreme values of the feature that was currently measured (these faces were generated by the FaceGen software; Inversions, 2006). The order of the faces on the screen was random for every feature, and the type and order of features were randomized between subjects.

Calculation of face feature vectors

To calculate the feature vector of a face, we first calculated the average value for each feature and then used the z score of that value with respect to all the values that other faces received for each feature. The result of this particular method for calculating the feature vector, together with the fact that all faces were on the screen when subjects determined their evaluation, was that feature values were determined relative to the whole face data set.

Results

An example feature vector is shown in Figure 1C. In this example, it is possible to see that high negative or positive values correspond to extreme features whereas average (near 0) values correspond to average features. Supplementary Figure S1 shows the correlation among all feature values. Some feature values are highly correlated (for example, face proportion, jaw width, chin shape, and cheek shape as well as eye shape and size and nose shape and size) whereas other features are not correlated. These results were taken into account when we decided which features to change (see Experiment 4).

Experiment 1B: Validating the face space dimensions

The next step was to assess whether the 20 features that we selected to construct the face space conformed with the core theoretical requirement from a face space: that perceptual similarity between faces is correlated with distances in the face space. We hypothesized that face space distances will be inversely correlated with similarity, i.e., that faces that are far from each other in the face space will be perceived as less similar than faces that are close to each other in the face space. To this end, we defined distances in the face space as L1 distance between feature vectors (i.e., the sum of absolute differences or what is known as the "city block" metric). The L1 norm was chosen for convenience to "sum up the differences" between faces. Using the more popular L2 norm was also tested and yielded very similar results and therefore had no advantage over the L1 norm. In addition, the L2 norm gives more weight to large differences and diminishes the contribution of small differences, and we had no theoretical justification for that. We then assessed the perceptual similarity between pairs of faces with the 10 smallest or 10 largest face space distances between them.

Methods

Participants

Twenty subjects participated in this face similarity experiment. The subjects were first-year psychology students from Tel Aviv University, performing the experiment for course credit. The experiment was approved by the ethics committee of Tel Aviv University. None of these subjects participated in Experiment 1A.

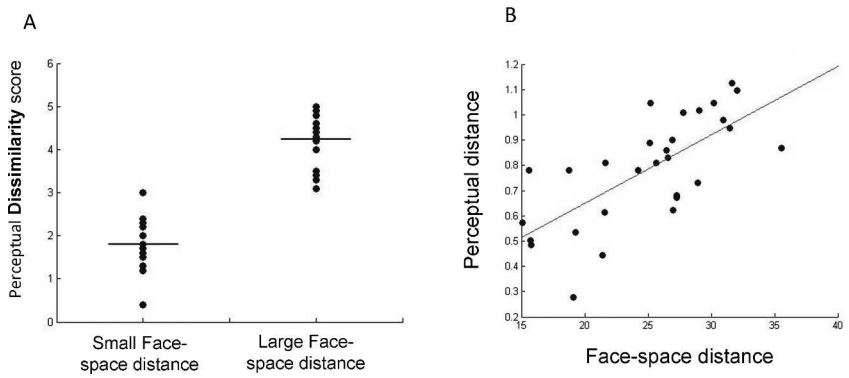

Figure 2. (A) Experiment 1B: Face dissimilarity scores for large and small face space distances: In this experiment, subjects were asked to rank, on a scale of 1 to 6, the similarity between 20 face pairs that had either small or large face space distances between them. The dots in the scatter plot indicate the dissimilarity scores for each face pair, and the horizontal lines indicate the mean dissimilarity score for each condition. Dissimilarity is significantly greater for faces with large face space distances between them, suggesting that the face space that we constructed meets the theoretical requirements described in Figure 1. (B) Experiment 2: The correlation between face space distances and perceptual distances for changed faces: The *x*-axis indicates the face space distances between pairs of faces before and after change (based on all 20 features), and the *y*-axis indicates the perceptual distances between the same face pairs. The correlation between the face space distances and the perceptual judgments is high and significant, again validating our face space and also demonstrating that we can systematically manipulate faces and "move them" in the space in a perceptually meaningful way.

Stimuli

To construct face pairs for similarity measurements, we calculated the face space distances between all possible pairs of the 100 faces used in Experiment 1A. We then took the 10 pairs with the largest face space distances between them and the 10 pairs with the smallest face space distances between them (see examples in Supplementary Figure S2), thereby creating 20 face pairs for similarity measurements.

Procedure

To measure face similarity, we simultaneously presented pairs of faces for an unlimited time and asked subjects to rank on a scale of 1 to 6 the similarity between the faces (1 = very different, 6 = very similar). The order of presentation of the face pairs was randomized between subjects and so was the right/left position of the faces within each pair. Because none of the subjects participated in Experiment 1A, the faces were unfamiliar to them.

Results and discussion

We found that similarity scores were significantly higher for pairs with small face space distance ($M = 4.2$, $SD = 0.62$, 95% CI [3.9, 4.5]) than for pairs with large face space distance ($M = 1.76$, $SD = 0.52$, 95% CI [1.5, 2.0]), $t(19) = 16.95$, $p < 0.0001$, Cohen's $d = 4.36$. Figure 2A shows the mean *dissimilarity* scores for each of the types of face pairs (dissimilarity was calculated as the max score [6] minus the similarity score), indicating that dissimilarity is greater when face space distances are larger. (See also Supplementary Figure S2 for example pairs of similar and different faces.) This means that the features that we chose and the face space distance that we defined can describe similarity between faces, and therefore, the face space that we constructed conforms with the theoretical definition of a face space. Our results are consistent with a previous report by Catz et al. (2009) that used a similar set of features and face-tagging procedure to that we employed here but with cropped faces with no hair, which may be important for identification (e.g., Sinha & Poggio, 1996, 2002). To validate the face space they created, Catz et al. showed that faces with extreme feature values were those that were judged as distinctive faces in a subsequent perceptual test. In the current experiment, we went further and showed that face space distances can account for face similarity judgments regardless of face distinctiveness measures.

Experiment 2: Moving faces in the face space

The results of Experiment 1 show that we are able to construct a face space using concrete features and that distances between faces in this space are inversely correlated with perceptual similarity. These results were

obtained for natural faces, the original faces that exist in our database. Nevertheless, when studying natural faces, we have little control over the feature differences between faces, making it difficult to test the effect of specific feature differences. Given our premise that features that are critical for identification are those that changing them would change the identity of a face, in the following experiment, we tested the perceptual effects of deliberate feature changes.

To this end, we created a set of modified faces, which were generated from original faces by replacing some facial features with features taken from "donor" faces (see Figure 5 for an example of facial feature changing). Feature vectors were then calculated for the new faces, using the same procedure as in Experiment 1A, enabling us to compute the distances between the original and modified faces. In "face space terminology," this manipulation "moved" faces from their original place in the space. Thus, the goal of this experiment was to test what type of changes (i.e., what directions and what distance in the face space) cause a change in identity or "move faces out of their identity subspace" (Figure 1B).

We created a set of changed face pairs with variable distances between the original and changed face, measured the face space distance between them, and measured the perceptual difference between original and changed faces and the correlation between these two measurements. Similar to Experiment 1B, we hypothesized that there would be a high correlation between *face space distances* between pairs of original and modified faces and the *perceptual distances* between these pairs.

Methods

Participants

Twenty-seven subjects participated in measuring perceptual difference between original and changed faces. The subjects were first-year psychology students from Tel Aviv University, performing the experiment for course credit. The experiment was approved by the ethics committee of Tel Aviv University. None of these subjects had participated in Experiment 1.

Stimuli

Thirty faces from the 100 tagged faces used in Experiment 1 were randomly selected to create changed faces. To test the effect of different facial feature changes, we developed a feature-changing method, using Adobe Photoshop©, which included copying features from donor faces based on feature values obtained during the tagging procedure (see Figure 5 for the face-changing procedure). To decide which features to change, we sorted the facial features of each face by the absolute magnitudes of their values and started replacing features from the largest values and downward to smaller values (in other words, we started by replacing the most distinctive features). Each feature was replaced with a feature with as far away a value as possible (from the existing database of tagged faces), thus making the largest possible change in the feature vector. Our goal was to create a variety of changes with a variety of face space distances, so we changed each of the 30 faces by replacing a different number of features. The stopping criterion was the number of features at which a face started to look unnatural or "Photoshopped," according to judgments made by a group of subjects not participating in the main experiment. This created a set of 30 pairs of original and changed faces that varied in the number of features that were changed (ranging between two and 11) and in the resulting face space distances among them (ranging between 15 and 37). We could therefore measure if face space distances covaried with perceptual differences between the original and changed face across the 30 face pairs. (See Supplementary Materials for more details on the feature-changing method.)

To measure perceptual differences between the original and changed face, based on these 30 changed faces, we created 90 face pairs in three conditions: Same, Changed, and Different face pairs. In the Same condition, we used two slightly different pictures of the same face (calling one of them "reference" and the other one "base"), belonging to the 30 faces that we manipulated. These two pictures were physically different, i.e., taken at different times (either consecutively in the same session or with some larger time difference) but were taken under similar lighting/pose/camera conditions (see Supplementary Figure S3 for an example of a Same pair). In the Changed condition, we used the reference pictures used in the Same condition and the changed picture, which was created from the base picture. The result was that in the Changed condition the pictures did not only differ in the features that were changed in our manipulation, but also in the low-level pixel information. This was done to ensure that subjects would not rely on this low-level information when they made their similarity judgments and would perform face matching rather than image matching. Finally, in the Different condition, we used two pictures of different identities, being either original or changed pictures (see Supplementary Figure S3).

Procedure

Measuring the perceptual effect of changing facial features: The goal of this procedure was to measure the perceptual difference that was caused by our feature replacement manipulation relative to the score that

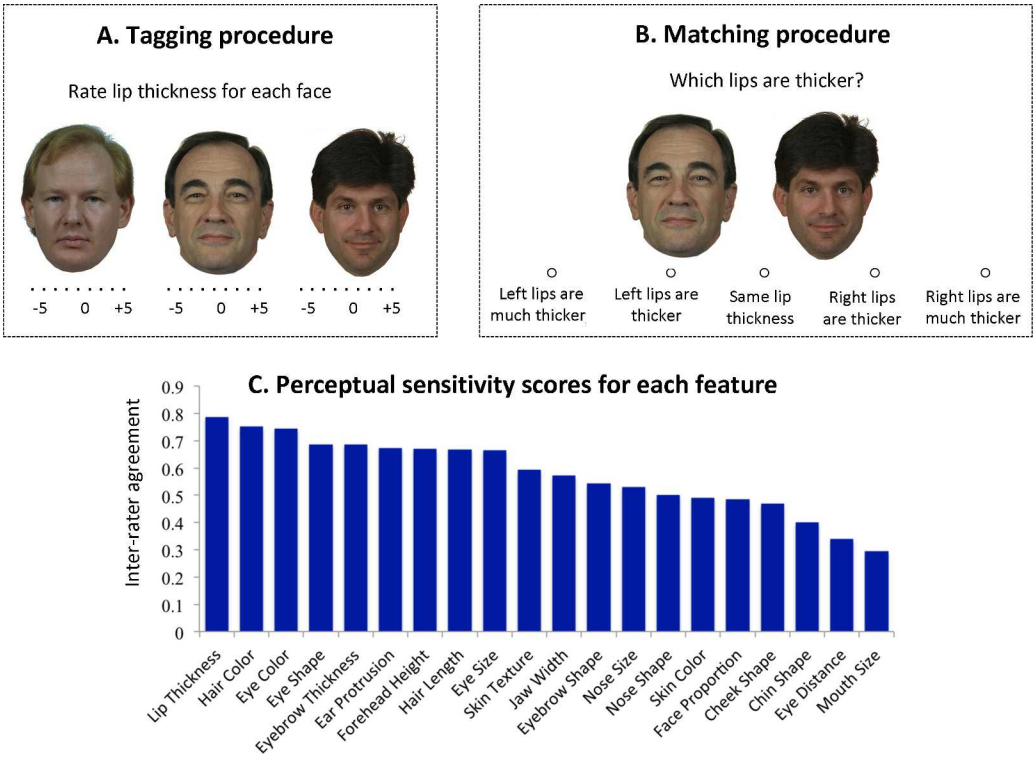

Figure 3. Two methods were used to measure perceptual sensitivity to detect differences in each feature across different faces: (A) In the face-tagging task (Experiment 1A), subjects are asked to indicate a value for each of the 20 selected features on a predefined scale (see Table 1) for 100 faces. (B) In a feature-matching task (Experiment 3), two faces were presented simultaneously. Subjects were asked to compare 50 pairs of faces for each of the 20 different features. For example, it is easier to indicate which of the two faces has thicker eyebrows than which has larger eye distance. (C) The inter-rater agreement was calculated based on the tagging of the 100 faces and the ranks that subjects gave the 50 pairs of faces per feature. Because the results were very similar across the two methods (Table 2) we combined them to obtain a more stable measure of PS.

would be given to Same and Different face pairs for each of the 30 faces (See Supplementary Figure S3). To this end, we presented subjects with 30 face pairs of each of the three conditions (Same, Changed, and Different) and asked them to mark, on a scale of 1 to 6, whether the two pictures belong to the same person or not (1 being "definitely the same person" and 6 being "definitely different people"). The two pictures in each pair were shown simultaneously until the subject responded. The order of the face pairs and the right/left position in each pair were randomized among subjects.

To analyze the results, the Same and Different conditions served as a baseline for calculating the similarity between changed pairs. This baseline was used in three ways: First, it provided the subjects with reference conditions to tune themselves and to understand how same and different people look. Second, we measured the baseline of identification in natural faces in both Same and Different conditions on the same set of pictures. Previous studies (Bruce et al., 1999; Burton, White, & McNeill, 2010) have shown that human performance is far from perfect on this task; therefore, we needed to tune our scale. Third, we used the baseline score for each Same pair for calculating the score for the Changed pair. Same pairs consisted of two pictures, and one of these was then modified to create the Changed pair. Therefore, we needed to subtract the score for the Same pair from the score for the Changed pair to compensate for the basic differences between the two pictures. Accordingly, the formula for calculating the perceptual distance between a Changed pair was (mean score for Changed pair – mean score for the Same pair of that face)/(mean score for all Different pairs – mean score for the Same pair of that face). The resulting perceptual distance scores ranged between ~0 and ~1, where 0 is the case in which the Changed pair got the same score as the Same pair of that face, and 1 is the case in which the Changed pair got a score that is close to the mean score for the Different pairs (see Supplementary Figure S3). To relate these perceptual distance scores with face identification, we regard perceptual distances closer to 1 as a change in identity, meaning that these faces were misidentified (i.e., perceived as different people), and scores that are closer to 0 as "same identity," meaning that these faces were identified correctly (i.e., perceived as the same person).

Measuring the face space distances between original and changed faces: To measure the face space distances, i.e., the distances between the feature vectors of faces before and after change, the changed faces were tagged again, using the same tagging procedure as in Experiment 1A. Forty-four subjects tagged the 30 changed faces (each subject tagging seven out of the 20 features to avoid fatigue, resulting in an average of 15 raters per feature) to obtain feature vectors for the changed faces. These subjects did not participate in the face-matching procedure to avoid familiarity effects. Face space distances were calculated by taking the sum of the absolute differences between the feature vectors (based on all 20 features) before and after change (L1 norm).

Results and discussion

Figure 2B shows the high and significant correlation (Spearman's $r = .72$, $p < 0.001$) between face space distances (calculated based on all 20 features) between Changed pairs of pictures and the perceptual distances between these faces. Similar to Experiment 1B, this significant correlation shows again that the features that we selected as well as the distance function that we defined can serve as a face space and to account for face similarity. We are able to manipulate faces and "move them around" in the face space in a way that is correlated with perceptual similarity measures. A similar correlation was found between the *number* of changed features and perceptual distance ($r = .69$, $p < 0.001$). Because of our feature-changing method in which we replace features with donor features that are as far away as possible, there was a high correlation between the number of changed features and the resulting face space distance ($r = .72$, $p < 0.001$), so based on this particular experiment, we cannot determine which parameter is more important: the distance that faces moved in the space or the number of features that were changed—a point that will be tested in Experiment 4.

Experiment 3: Assessing the discriminative power of different facial features

Experiment 2 showed that the face space that we constructed and the face space distances that we measured between faces could account for perceptual similarity between faces. Yet our goal was to find out which subset of these features is more important for determining face identity, i.e., to analyze the relative contribution of different features to face similarity. The standard approach to this question is to use some tool of regression analysis on the feature vector values collected in Experiment 2. However, our 20-dimensional space is quite large, and the number of data points that we have is relatively small, causing any attempt to make such analysis prone to overfitting (Todorov & Oosterhof, 2011). Therefore, we took a different approach, and that is to measure the discriminative power of the different features. The logic is that if a feature is useful in discriminating between faces, i.e., it is easy to tell the difference in a feature across different faces, then perhaps it would be used not only to discriminate among features but also to discriminate among faces. For example, if it is easy to tell which face has thicker lips, i.e., we have high PS to this feature, but it is difficult to tell which face has larger eye distance (low PS), then the former would be more important than the latter for face identification. We therefore hypothesized that features with high PS will have more discriminative power and will account for a more significant part of the variance in perceptual distances obtained in Experiment 2 whereas features with low PS will account for a smaller part of that variance.

But how do we measure PS? Normally, PS is measured by showing subjects different pictures of the same face with slight changes in a given feature and testing if subjects can tell the difference in that feature (for review, see McKone & Yovel, 2009). However, this type of measurement is not useful for studying face identification because face identification is about comparing features across *different* faces. Thus, to measure PS for each of the 20 features across faces, we used inter-rater agreement on feature values. The idea is that if most subjects agree on the values of some feature, it means that it is easy to measure it, and therefore it is useful for identification whereas, if subjects disagree on the values of some feature, it means that it is not useful for identification.

To this end, we measured inter-rater agreement in two different methods: (a) inter-rater agreement in feature values in Experiment 1A (Figure 3A) and (b) inter-rater agreement when subjects are asked to compare pairs of faces (Figure 3B). This latter method is more direct: Subjects are asked to directly compare two faces based on a feature, for example, which face has larger eye distance.

Methods

For inter-rater agreement in face tagging (Experiment 1A), we calculated the median value of all the pair-wise correlations across all subjects' ranks per feature across all 100 faces. To further validate the PS measures we obtained in Experiment 1A, we asked a

Feature name	Experiment 1 feature tagging	Experiment 3 feature comparison
Lip thickness	0.80	0.75
Hair color	0.77	0.71
Eye color	0.75	0.73
Eye shape	0.72	0.61
Eyebrow thickness	0.71	0.63
Ear protrusion	0.71	0.60
Forehead height	0.65	0.69
Hair length	0.71	0.57
Eye size	0.65	0.68
Skin texture	0.68	0.42
Jaw width	0.58	0.54
Eyebrow shape	0.58	0.47
Nose size	0.49	0.60
Nose shape	0.56	0.37
Skin color	0.45	0.57
Face proportion	0.53	0.38
Cheek shape	0.41	0.57
Chin shape	0.54	0.12
Eye distance	0.37	0.27
Mouth size	0.34	0.21

Table 2. Perceptual sensitivity (PS) of features as measured in Experiments 1A and 3. *Notes*: The inter-rater agreement values that were measured in Experiment 1A, facial feature-tagging task, and in Experiment 3, feature comparison task between pairs of faces. These inter-rater agreement values indicate the PS for each feature and were highly correlated across the two methods.

different group of subjects to explicitly compare the magnitude of each feature across two simultaneously presented faces.

Participants

A total of 36 subjects who did not participate in Experiment 1 participated in this experiment. Subjects were Amazon Mechanical Turk workers, participating in the experiment for money (approximately $1 per 15 min).

Procedure

In this experiment, PS was measured by comparing features between faces as described in Figure 3B. Subjects were presented with 50 pairs of faces in a random order (the left/right position of the faces was also randomized) and were asked to judge the differences in features between faces. Similar to the tagging procedure, the definition of each feature and its scale were explained to the subjects along with example schematic images where it was applicable. Each subject was asked to rank six randomly selected features (so each feature was evaluated by an average of 11 subjects), and subjects first evaluated all pairs in one feature before moving on to the next feature. Feature comparison was ranked on a 5-point scale, and the specific scale wording was matched to the scale of the feature. For example, for comparing eye distance, the scale was 1 = left face eye distance much larger than right face eye distance, 2 = left face eye distance larger than right face eye distance, 3 = both faces have similar eye distance, 4 = right face eye distance larger than left face eye distance, 5 = right face eye distance much larger than left face eye distance.

PS per feature was calculated as the inter-rater agreement on comparison ranks. It was calculated by taking median value of all the pair-wise correlations between all subjects' ranks per feature.

Results and discussion

Inter-rater agreement for each of the features in the face-tagging procedure (Experiment 1A) is shown in the left column of Table 2. It is evident that there is high agreement to determine the magnitude of some features, such as lip thickness or hair color, indicating that human observers have high PS for detecting differences in these features. For other features, such as mouth size or eye distance, there is low inter-rater agreement, indicating it is harder to judge the relative magnitude of these features across faces, reflecting low PS to detect differences in the magnitude of these features.

The results for inter-rater agreement in the face-matching experiment are shown in the right column of Table 2. These measures were highly correlated with the inter-rater agreement we obtained in the tagging procedure used in Experiment 1A (Spearman's correlation: $r = .71$, $p < 0.01$), and we therefore combined them to obtain a more stable measure of perceptual sensitivity as shown in Figure 3C.

Finally, we assessed whether a subset of features of the highest PS may account for the perceptual similarity scores across faces obtained in Experiment 2. The way we did this was to calculate the face space distances based not on all 20 features but on a subset of features, starting from one feature, the highest PS feature, and increasing the number of features used to calculate the distance up to all 20 features, according to a descending PS order. For each set of these modified face space distances, we calculated the correlation between the face space distances and the perceptual distances. We performed the same analysis, calculating face space distances using an increasing number of features but starting from the lowest PS feature and adding features in an ascending PS order. Figure 4A shows the results of these calcula-

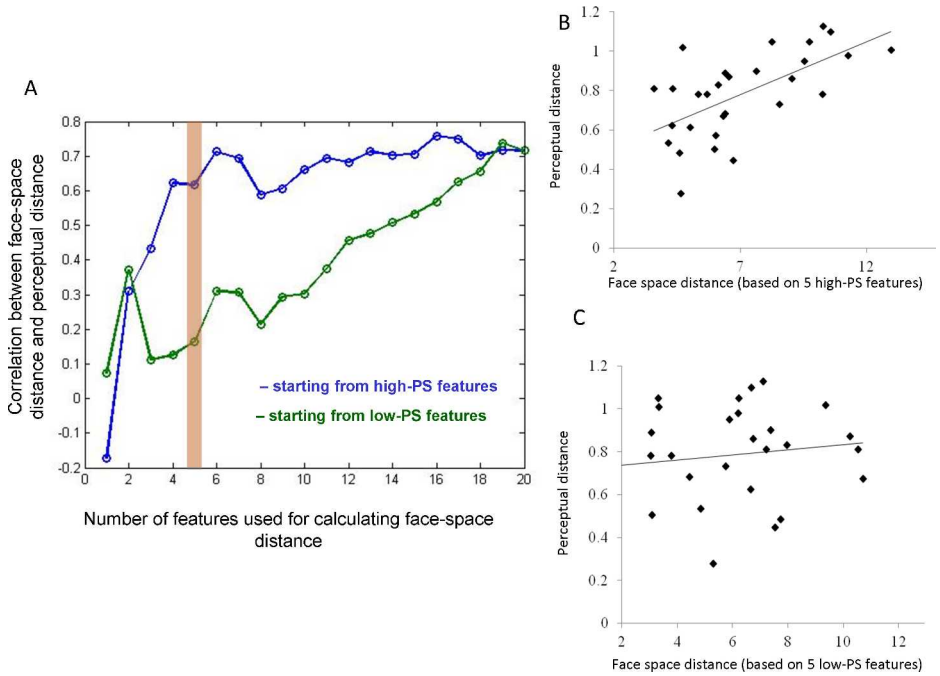

Figure 4. (A) Correlations between face space distances and perceptual distances using a different number of features to calculate the face space distance. The *x*-axis denotes how many features were used to calculate the face space distance. The blue line shows the correlations when we use an increasing number of features starting with high PS features, and the green line shows the correlations when we use an increasing number of features, starting with low PS features. It takes four to six high PS features to reach significant correlation, similar to using all 20 features, but it takes 16 low PS features to reach a similar level of correlation. B and C show the scatterplots of the correlations corresponding to the points highlighted by the red vertical line. (B) The scatterplot of the correlation between face space distances and perceptual distances using the five highest PS features is high and significant ($r = .62$, $p < 0.01$), similar to using all 20 features, thus accounting for most of the variance in perceptual-distances. (C) The scatterplot of the correlation between face space distances and perceptual distances using the five lowest PS features is insignificant.

tions, indicating that four to six high PS features are sufficient to account for most of the variance in perceptual distances whereas the same correlation is reached only when using 16 low PS features. Figure 4B shows that the face space distances calculated based on the five features with highest PS (lip thickness, hair color, eye color, eye shape, eyebrow thickness) are highly correlated with perceptual distances ($r = .62$, $p < 0.01$), similar to the correlation we obtained using all 20 features to calculate the face space distance (compare Figures 2B and 4B). Figure 4C shows the correlation of face space distances with perceptual similarity judgments when only the five low PS features are used to calculate face space distances. This time, the correlation between these distances and perceptual distances is very low and insignificant.

These results show that PS is related to the discriminative power of different features. Importantly, a subset of features—high PS features—are sufficient to explain the variance in perceptual distances and therefore may be enough to account for face identification.

Experiment 4: The effect of changing high PS or low PS features on perceptual identity judgments

Experiment 3 indicated that high PS features may account for face identity similarity judgments, suggesting that high PS but not low PS features fit our definition of critical features; i.e., critical features are those that changing them would change the identity of the face. To directly assess this suggestion, we examined whether changing high PS features would result in a change of identity whereas changing low PS features would not change face identity. In other words, using face space terminology, we hypothesized that changing high PS features, but not low PS features, would move faces out of their identity subspace (Figure 1B). Based on the results of Experiment 3 (Figure 4), we decided to test our hypothesis by changing five high PS or five low PS features. In addition, we tested whether it is more important *which* features to change or *how far* to change them by adding a third condition of changing high PS features to a smaller face space distance than the pairs that differed in low PS features (see Figure 6).

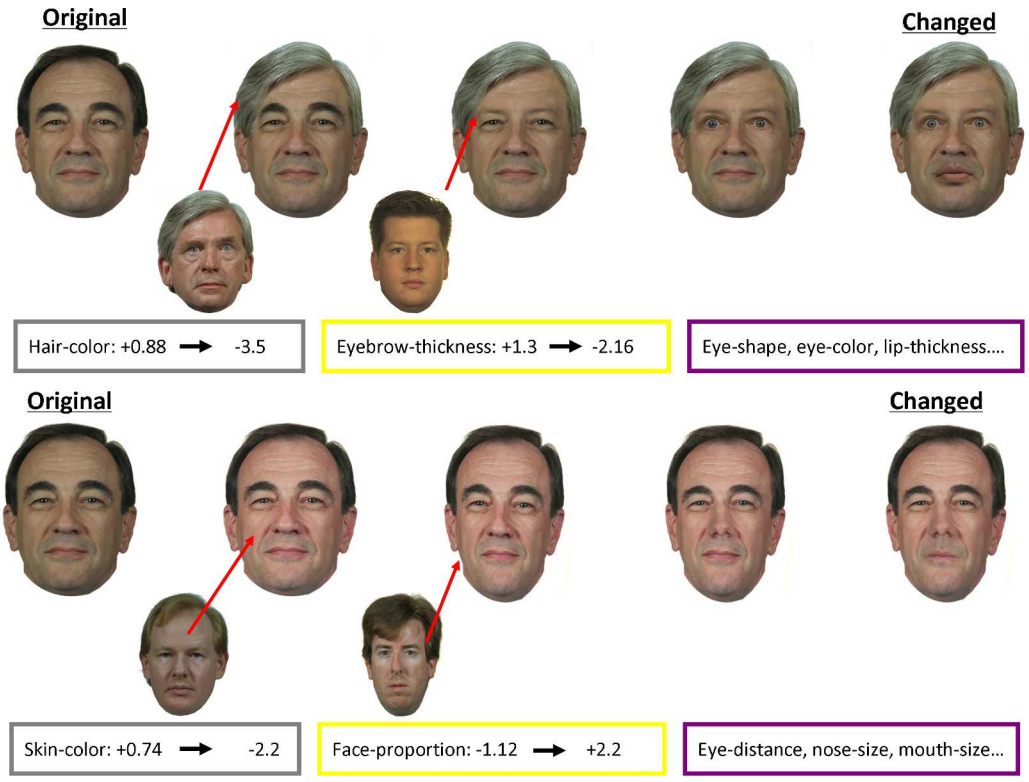

Figure 5. The procedure used to change faces in Experiments 2 and 4: Features were copied from donor faces (taken from the 100 faces in the database), based on the feature values obtained in the tagging procedure. The result is a new, natural-looking face with new features. In Experiment 2, the features that were changed were the ones with the most extreme values whereas in Experiment 4 the changed features were either high PS or low PS features. To achieve maximal face space distances, we copied features with values that are opposite from the original value—this was the method used in Experiment 2, in the far high PS condition (top row), and in the far low PS condition (bottom row). In the random high PS condition, the donor feature was selected from a randomly chosen face.

Method

Participants

Forty-eight subjects participated in the face-matching experiment (measuring perceptual distances between changed pictures). Seventy different subjects participated in tagging the changed faces (for measuring face space distances between original and changed faces). All were Amazon Mechanical Turk workers, participating in the experiment for money (approximately $1 per 15 min). To avoid familiarity effects, we monitored worker IDs to ensure that subjects did not participate in more than one task throughout this study.

Stimuli

Fifteen faces were randomly selected from the 100 faces that were used in Experiment 1 (these faces were different than the 30 faces used in Experiment 2). We used a similar feature-changing method as in Experiment 3, only this time, for each of the 15 faces, we changed either the five lowest or five highest PS features (see Figure 3C). In case of high correlations between features (see Supplementary Figure S1), we replaced only one of the features. Thus, for high PS changes, we changed lip thickness, hair color, eye color, eye shape, and eyebrow thickness. As a result of our feature-changing method, copying the hair from a donor face sometimes resulted in changing also the hair length and forehead height (these are also high PS features, and they are highly correlated between them). For low PS changes, we chose mouth size, eye distance, face proportion (which is highly correlated with chin shape and cheek shape), skin color, and nose size (correlated with nose shape). The results were natural-looking faces that were different from the original faces in the selected high or low PS features (see Figure 5).

Applying the face-changing procedure, the 15 faces were changed in three different ways (Figure 6A): (a) The high PS features were replaced with donor features that had opposite values as far as possible from the values of the original features, moving the face to the maximal distance that is possible in the face space when changing only these features. We call this feature manipulation "far high PS." (b) The same high PS features were replaced, but this time, each donor feature was selected randomly from one of the faces in the original 100-face

database. This resulted in a smaller face space distance than the "far high PS" change. We call this feature manipulation "random high PS." This condition was used to test whether it is important to merely change high PS features or if it is also important *how far* we change the features. (c) The low PS features were replaced with donor features of as far as possible values, again moving the face to the maximal distance in the face space. We call this group "far low PS." See Figure 6A for two examples of the three types of changes.

Procedure

Measuring the perceptual effect of changing facial features: We used a similar procedure as in Experiment 2 to determine perceptual distance between changed faces. Again we had Same, Different, and Changed conditions. This time, we had 15 original faces, each changed in three different ways (far high PS, random high PS, and far low PS), creating 45 Changed pairs for each face. In addition, we used 15 Same pairs and 15 Different pairs for each face (see Supplementary Figure S3). We presented each subject with the 15 Same pairs, 15 Different pairs, and 15 (out of the 45) Changed pairs, five for each type of change. Based on our previous studies of face matching that recruited about 12–15 subjects (e.g., Brandman & Yovel, 2012; Yovel & Kanwisher, 2004), we obtained data from 16 subjects for each of the changed face stimuli, which resulted in a sample of 48 subjects. The face pairs were presented in a random order (and the right/left picture location was also randomized), each pair presented until the subject's response. For each pair, subjects were asked to judge, on a scale of 1–6, whether this is a pair of pictures of the same person or of different people.
Measuring the face space distances between original and changed faces: To measure the face space distances, i.e., the distances between the feature vectors of faces before and after change, the changed faces were tagged again, using the same tagging procedure as in Experiment 1A. Seventy subjects tagged the 45 changed faces (each subject tagging four out of the 20 features to avoid fatigue, resulting in an average of 14 raters per feature) to obtain feature vectors for the changed faces. These subjects did not participate in the face-matching task to avoid familiarity effects. Face space distances were calculated by taking the sum of the absolute differences between the feature vectors (based on all 20 features) before and after change (L1 norm).

Results

Figure 6B shows the perceptual distance scores, and Figure 6C shows the face space distances for each of the three types of face changes.

To compute the results of the perceptual distances between changed pictures, we calculated for each subject the average perceptual distance scores in each of the three types of changed faces. A repeated-measures ANOVA revealed significant differences between perceptual distances for far high PS changes ($M = 1.03$, $SD = 0.14$, 95% CI [0.99, 1.06]), random high PS changes ($M = 0.92$, $SD = 0.21$, 95% CI [0.87, 0.97]), and far low PS changes ($M = 0.47$, $SD = 0.58$, 95% CI [0.33, 0.61]), $F(2, 141) = 31.44$, $p < 0.001$, $\eta^2 = .31$. Post hoc analysis revealed that the perceptual distance for far low PS changes was significantly lower than that of the random ($p < 0.01$, Cohen's $d = 1.04$) and far high PS changes ($p < 0.01$, $d = 1.33$) whereas the difference between the two types of high PS feature changes was not significant.

A repeated-measures ANOVA for face space distance scores revealed significant differences between face space distances for far high PS changes ($M = 24.11$, $SD = 3.48$, 95% CI [22.53, 25.69]), random high PS changes ($M = 16.29$, $SD = 3.01$, 95% CI [14.93, 17.66]), and far low PS changes ($M = 20.79$, $SD = 4.37$, 95% CI [18.8, 22.78]), $F(2, 42) = 17.2$, $p < 0.001$, $\eta^2 = .45$. Post hoc analysis revealed that random high PS changes resulted in significantly smaller face space distances compared with both far high PS ($p < 0.01$, $d = 2.48$) and far low PS changes ($p < 0.01$, $d = 1.24$), but there was no significant difference between the face space distances of the far low PS and far high PS changes.

We also found that the simple face space distance function that we defined (L1 norm) based on all 20 features was not correlated with perceptual distances between changed face pairs (Supplementary Figure S4A). Nevertheless, modifying the distance function by taking into account only the high PS features (i.e., giving high PS features a weight of 1 and all other features a weight of 0) resulted in a very high correlation ($r = .77$, $p < 0.001$) between face space distances and perceptual distances (see Supplementary Figure S4B and compare with Figure 4).

Discussion

The results clearly show that changing high PS but not low PS features resulted in a change in identity as defined by our perceptual distance measure: The average perceptual distance score for far high PS changes was 1, meaning that these faces were perceived as pairs of different faces. This suggests that high PS features are critical for face identification. In addition, comparing perceptual distances and face space distances between face pairs shows that the *type* of features that were changed was more critical than *how much* the features were changed. Namely, high PS

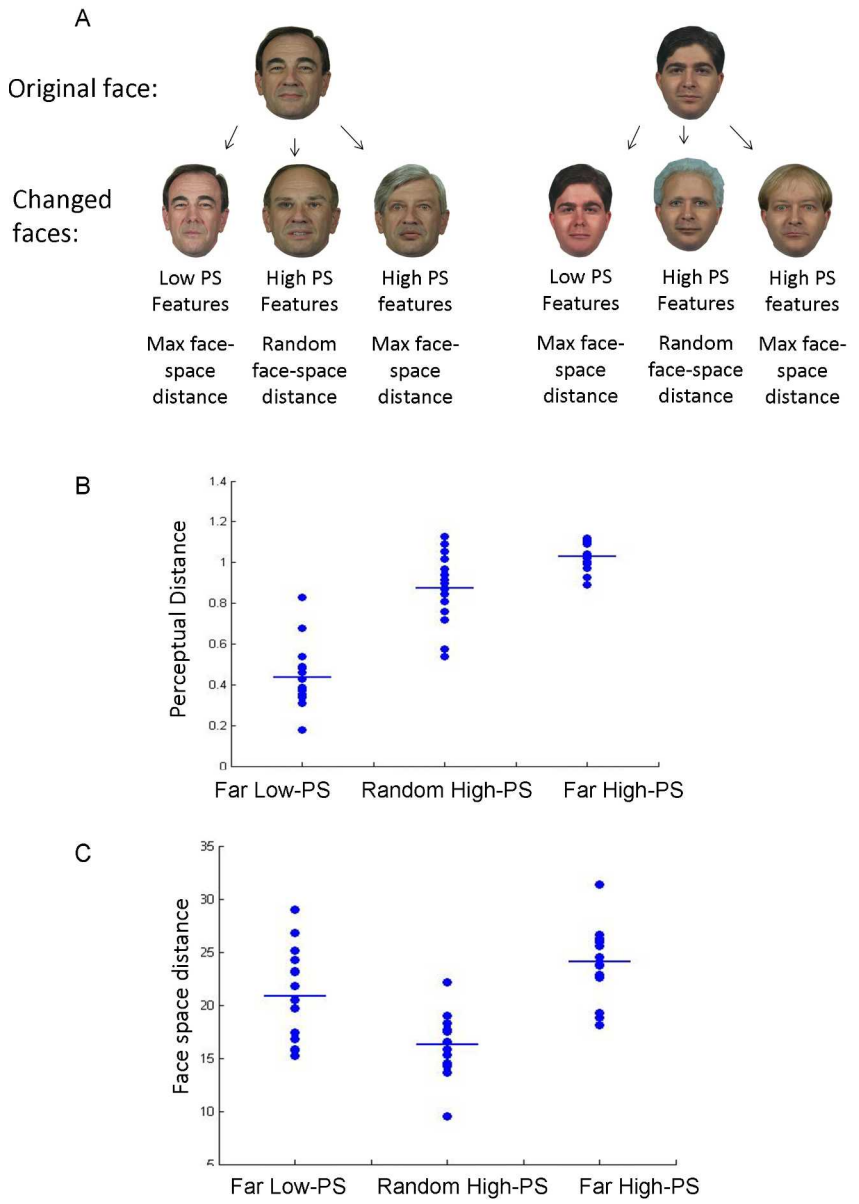

Figure 6. (A) Examples for the three types of changes for two example faces. Far high PS: changing high PS features with an opposite feature (i.e., max face space distance). Random high PS: changing high PS features with a random feature. Far low PS: changing low PS features with an opposite feature (i.e., max face space distance). (B) Perceptual distances for the three types of feature changes (the dots indicate the perceptual distances between original and changed pictures in each condition, and the horizontal lines indicates the means): Perceptual distance scores were larger following changing high PS features than low PS features regardless of whether the change was random or maximal. (C) The face space distances for the three types of feature changes: Face space distance was larger for the far high and far low PS changes than the random changes and did not correspond to perceptual distances.

changes caused a much larger perceptual change even when changes in low PS features resulted in a larger or similar face space distance (i.e., distance between feature vectors) relative to random or far high PS features, respectively (Figure 6). Interestingly, despite the fact that human observers can detect large changes in low PS features (as indicated by the face space distance), they *ignore* these detected changes when coming to determine the identity of a face.

Experiment 5: Are critical features invariant to changes in face appearance?

Results of Experiments 3 and 4 demonstrate that high PS features but not low PS features are critical for face identity. However, we are still left with the question of *why* some features are more important than others. As

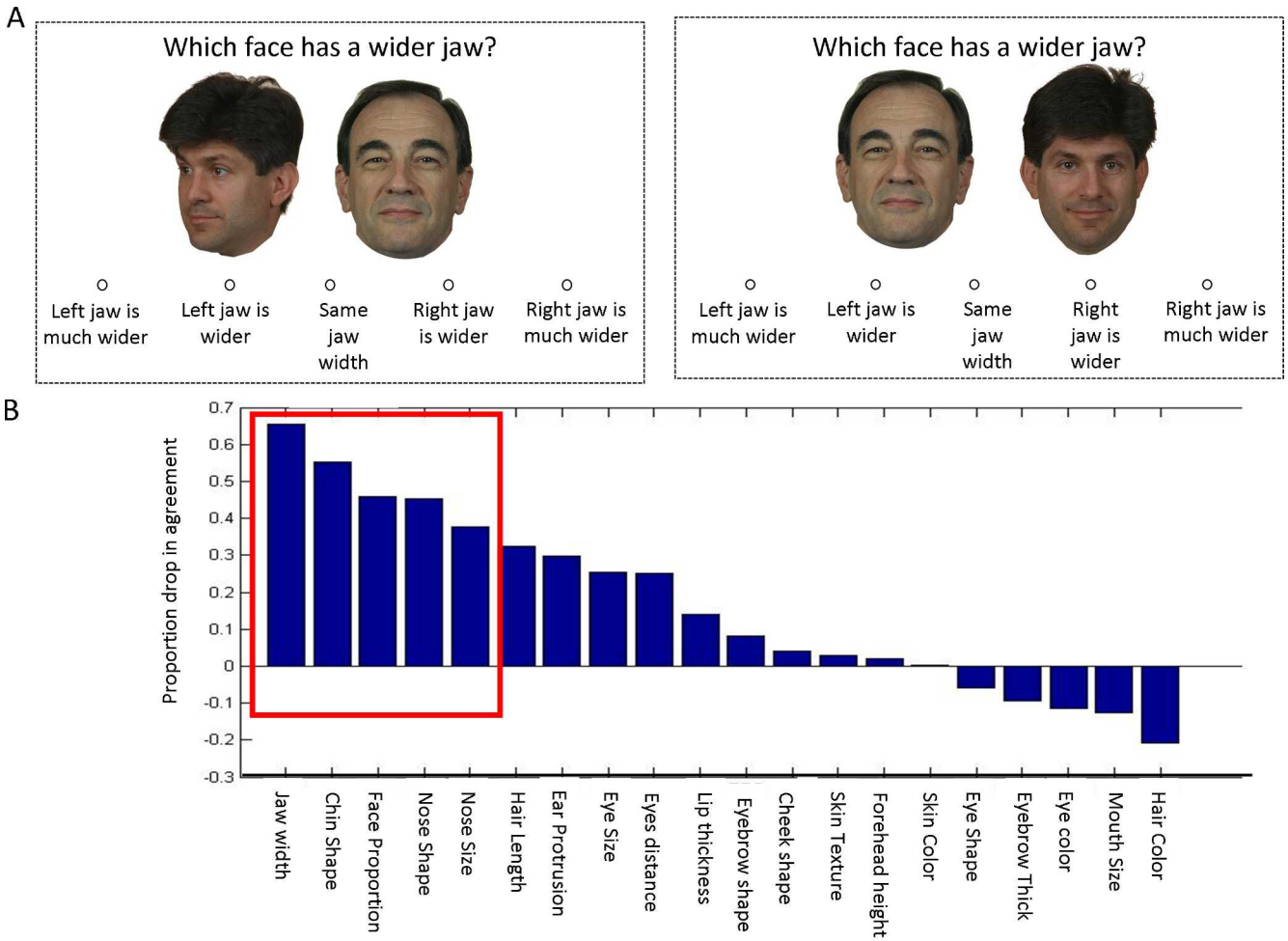

Figure 7. (A) An example of a matching trial used to measure PS for each feature across faces of different views (left) and same view (right). (B) The relative change in PS for each feature when comparing faces across views relative to same view faces. The five features that suffered the greatest drop in agreement for different relative to same view faces (inside the red rectangle) are low PS features, suggesting a correlation between feature invariance to different views and its importance to face identification.

explained above, the face space theory assumes that each identity takes up a subspace, in which different points represent different appearances of the same identity (Figure 1B). We therefore hypothesized that low PS features are those that may vary within the identity's subspace across different appearances (such as pose, aging, weight change, etc.). These low PS features are "ignored" in the face identification process because they cannot be used reliably to identify faces under "allowed" variations. For example, whereas eye color and eyebrow thickness remain the same across different head views of the same identity, eye distance and face proportion vary across different views of the same identity, making the former but not the latter features more effective for correct identification under varying conditions.

Using the same logic that we used to estimate PS, we hypothesized that when human observers are asked to compare a given feature across two faces in *different poses*, some low PS features will suffer a drop in inter-rater agreement, and high PS features will be resistant or invariant to changes in viewpoint.

Method

Participants

Sixty Amazon Mechanical Turk workers participated in this experiment for payment (approximately $1 per 15 min), none of whom participated in previous experiments in this study.

Procedure

To test PS for each feature across face views, we asked subjects to compare facial features between two faces as before; only this time each feature was shown in two conditions: Same view condition: both faces were in frontal view as in Experiment 2 and Different view condition: one face was in frontal view and the

other face in left three-quarters view (see Figure 7A). Each subject rated 30 pairs of faces in six features out of the total 20 (to avoid fatigue), three of which were in frontal view and the other three in left three-quarters view. An average of nine subjects rated each feature in each condition.

The inter-rater agreement was calculated by correlating between ratings for all pairs across judges. The *drop* in inter-rater agreement between frontal and left three-quarters views was calculated by dividing the difference between the two inter-rater agreement values by the inter-rater agreement value for the frontal view.

Results

Figure 7B shows the change in inter-rater agreement for Different view relative to Same view face pairs, sorted from the largest to the smallest drop in agreement. The five features that suffered the largest drop in inter-rater agreement due to change in viewpoint were jaw width, chin shape, face proportion, and nose shape and size, all of which are low PS features (see Figure 3C). In contrast, most high PS features showed a smaller or no drop in inter-rater agreement (for example eye shape, eyebrow thickness, or eye color). Some low PS features, such as eye distance, also showed no drop in inter-rater agreement, but the initial inter-rater agreement of this feature was very low to begin with (see Table 2).

Discussion

Results of Experiment 5 show that there is a correlation between which features are important for identification and which are invariant to "allowable" changes in face appearance. Critical features may be invariant under *most* changes in appearance whereas the less critical features that are less used for face identity will change their appearance in *at least one* of these types of changes (e.g., skin color varies across different lighting conditions, mouth size varies across different expressions, and eye distance varies across different head views), making them generally less useful for identification. Here we set the methodological groundwork for testing all these possible changes in future studies.

General discussion

We have identified a subset of features that are critical for face identification by showing that changing them changes the perceived identity of a face, whereas changing a different set of features does not change the perceived identity of a face (Experiment 4, Figure 6). These findings are novel both conceptually and methodologically: Conceptually, we have shown that critical features are those for which we have high PS to detect differences across different faces (Experiment 3, Figures 3 and 4). Furthermore, these features vary minimally across variations of the same identity (Experiment 5). Thus, these findings address one of the most fundamental queries in the study of face recognition—known as the invariance problem—of how faces are recognized across their different variations (expression, pose, lighting). Methodologically, we developed a novel procedure that allows us to construct a concrete face space and can be now applied to ask many different questions about critical features of other facial aspects (e.g., gender, attractiveness) or face types (e.g., Asian faces).

In this study, we examined the question of which features are critical for processing unfamiliar faces, a question that has been extensively discussed in the literature. Some researchers hold a holistic view on face processing, suggesting that face identity is determined by an interactive processing among face parts (Le Grand, Mondloch, Maurer, & Brent, 2004; Rossion, 2013). In addition, there is abundant literature on the debate of whether face recognition is based on "featural" or "configural" information (Maurer, Grand, & Mondloch, 2002; McKone & Yovel, 2009; Schwaninger, Lobmaier, & Collishaw, 2002; Shin, Jang, & Kwon, 2011). Others have emphasized the importance of the hair (Sinha & Poggio, 1996; Toseeb, Keeble, & Bryant, 2012), the internal features, or the external features (Andrews, Davies-Thompson, Kingstone, & Young, 2010; Clutterbuck & Johnston, 2002; Ellis, Shepherd, & Davies, 1979). Nevertheless, none of these numerous studies have explicitly pointed to the critical features that define an identity. Our findings suggest a new categorization of facial features based on their discriminative power, which is measured by PS to detect differences *across identities*. We also suggest that PS is related to the variance or invariance of features across different appearances of the face. We suggest that these two criteria (high PS across identities and low variance within identities) better characterize the importance of features for face identification than the currently prevailing classifications to featural versus configural information or to external versus internal features.

PS for facial features has been used in previous studies to assess which features are important for face recognition. This has been typically done by using a change detection task in which two images of the same person differed on one specific feature (e.g., shape of the eyes or distance between the eyes (Le Grand, Mondloch, Maurer, & Brent, 2001; McKone & Yovel,

2009; Yovel & Kanwisher, 2004) and were presented either upright or inverted. However, detecting differences between features for images of the *same identity* may not reflect our ability to detect these changes *across identities*. Indeed, our study clearly shows low PS for comparing eye distances across different identities, a feature that is easily detected when manipulated and compared within the same face and has been therefore considered in many studies to be important for upright face recognition (Le Grand et al., 2004; Maurer et al., 2002). Consistent with our findings, recent studies show intact face recognition for compressed faces in which the eye distance was significantly distorted (Andrews et al., 2013; Hole, George, Eaves, & Rasek, 2002) suggesting that eye distance may play a little role in face recognition.

Previous studies attempted to construct face spaces and to use them for measuring perceptual effects but have not indicated what are the dimensions of the face space (Leopold et al., 2006; Leopold, O'Toole, Vetter, & Blanz, 2001; Rhodes, 1988; Rhodes & Jeffery, 2006; Rhodes & Leopold, 2011). Some studies used computer-generated faces to systematically change and quantitatively measure facial features (Freiwald et al., 2009; Gao & Wilson, 2013; Oosterhof & Todorov, 2009). Here we show that the same can be achieved also with natural faces, which are more ecologically valid. A study conducted by Rhodes (1988) used multidimensional scaling to discover which features contribute more to face identity and reported that the eyes, eyebrows, and mouth as well as eye position, spatial relations between features, and chin shape were correlated with the dimensions of the space. Rhodes' measurements were performed by asking subjects to measure feature sizes and distances (in millimeters) and angles (in degrees) on pictures of natural faces. In contrast, the current study used subjective perceptual estimations, which may better reflect the perceptual comparisons that we normally perform in natural face processing. Lewis (2004) performed a mathematical analysis of the face space and concluded that the face space should consist of 15 to 22 dimensions to enable recognition of the thousands of faces people are able to memorize and recognize. The constraint that leads to this number of dimensions is the finding of Benson and Perrett (1991) that 4.4% of exaggeration away from the prototypical face results in the best likeness of the person, which gives an estimate of the size of the subspace each identity takes up in space. In the current study, we used 20 features (a number which is consistent with Lewis's estimate), and we further show the relative importance of different features in this set. A more recent study by Nestor Vettel, and Tarr (2013) used schematic faces and noise-based images to try and discover the visual structures underlying face processing from BOLD responses to such images. This is an interesting approach to discover neural representation of facial features, and is similar to the study reported by Freiwald et al. (2009) on macaques. Nevertheless, unlike our study that used natural images and examined face identification, Nestor et al. studied face detection rather than recognition and used schematic faces rather than natural images.

This study reveals the critical features for identification of unfamiliar Caucasian faces. The fact that we used different faces (from our database) for Experiments 2 and 4 shows that the high PS features that were identified are generalized across different data sets of Caucasian male faces. These features, however, may not generalize to other types of faces, such as Asian or African faces. For example, hair color has high discriminative power in Caucasian faces, but there is little variation in hair color among Asian or African faces, and therefore, hair color may not be a high PS feature in these faces. However, the method we propose here can be used to reveal the crucial features of other categories of faces. Importantly, the principle that high PS features are those that are critical for face identification is expected to account for identification of any category of faces.

Our study sets an infrastructure for studying many phenomena in natural face processing. The tools used in this study—a multidimensional face space, measurement of PS, systematic feature changes, and measurement of perceptual differences between natural faces—may be used for understanding many basic questions in face processing, such as which features are critical for determining face attractiveness or various facial expressions and which features are critical for identification across different races. Whereas in the current study we applied this method to study the role of a subset of 20 features, future studies may focus on other/additional features, perhaps even manipulating different scales of multiple features. We hope that this study will inspire future studies to apply this approach to answer the many fundamental but still open questions in human face processing.

Keywords: face processing, face identification, face space theory, face features

Acknowledgments

Portions of the research in this manuscript use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office. The said images were processed by the author for this specific experiment. You may not use any of the images in this experiment without written permission from NIST and from the author. We wish to thank Dr. Vadim Axelrod

and Michal Bernstein for their helpful comments on earlier versions of this manuscript. We would also like to thank our gifted graphic artist, Maya Goldstein, for creating the changed face stimuli for this study. Special thanks go to Prof. Yonatan Goshen and Prof. Alex Bronstein for their advice and discussion.

Commercial relationships: none.
Corresponding author: Naphtali Abudarham.
Email: naphtalia@post.tau.ac.il.
Address: School of Psychological Sciences, Tel Aviv University, Tel Aviv, Israel.

References

Andrews, T. J., Baseler, H. A., Harris, R. J., Jenkins, R., Burton, A. M., & Young, A. W. (2013). Invariance to linear but not non-linear changes in the spatial configuration of faces in human visual cortex. *Journal of Vision, 13*(9): 168, doi:10.1167/13.9.168. [Abstract]

Andrews, T. J., Davies-Thompson, J., Kingstone, A., & Young, A. W. (2010). Internal and external features of the face are represented holistically in face-selective regions of visual cortex. *The Journal of Neuroscience, 30*(9), 3544–3552.

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

Blank, I., & Yovel, G. (2011). The structure of face—space is tolerant to lighting and viewpoint transformations. *Journal of Vision, 11*(8):15, 1–13, doi: 10.1167/11.8.15. [PubMed] [Article]

Brandman, T., & Yovel, G. (2012). A face inversion effect without a face. *Cognition, 125*(3), 365–372.

Bruce, V., Henderson, Z., Greenwood, K., Hancock, P. J., Burton, A. M., & Miller, P. (1999). Verification of face identities from images captured on video. *Journal of Experimental Psychology: Applied, 5*(4), 339.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow face matching test. *Behavior Research Methods, 42*(1), 286–291.

Catz, O., Kampf, M., Nachson, I., & Babkoff, H. (2009). From theory to implementation: Building a multidimensional space for face recognition. *Acta Psychologica, 131*(2), 143–152.

Clutterbuck, R., & Johnston, R. A. (2002). Exploring levels of face familiarity by using an indirect face-matching measure. *Perception-London, 31*(8), 985–994.

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

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009). A face feature space in the macaque temporal lobe. *Nature Neuroscience, 12*(9), 1187–1196.

Gao, X., & Wilson, H. R. (2013). The neural representation of face space dimensions. *Neuropsychologia, 51*(10), 1787–1793.

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception-London, 31*(10), 1221–1240.

Inversions, S. (2006). FaceGen 3.1 full software development kit documentation. Retrieved October 1, 2010.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2001, Apr 19). Neuroperception: Early visual experience and face processing. *Nature, 410*(6831), 890.

Le Grand, R., Mondloch, C. J., Maurer, D., & Brent, H. P. (2004). Impairment in holistic face processing following early visual deprivation. *Psychological Science, 15*(11), 762–768.

Lee, K., Byatt, G., & Rhodes, G. (2000). Caricature effects, distinctiveness, and identification: Testing the face-space framework. *Psychological Science, 11*(5), 379–385.

Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006, Aug 3). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature, 442*(7102), 572–575.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience, 4*(1), 89–94.

Lewis, M. (2004). Face-space-R: Towards a unified account of face recognition. *Visual Cognition, 11*(1), 29–69.

Lowd, D., & Meek, C. (2005, August). Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on knowledge discovery in data mining* (pp. 641–647). New York: ACM.

Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences, 6*(6), 255–260.

McKone, E., & Yovel, G. (2009). Why does picture-plane inversion sometimes dissociate perception of features and spacing in faces, and sometimes not? Toward a new theory of holistic processing. *Psychonomic Bulletin & Review, 16*(5), 778–797.

Nestor, A., Vettel, J. M., & Tarr, M. J. (2013). Internal representations for face detection: An application of noise-based image classification to BOLD responses. *Human Brain Mapping, 34*(11), 3101–3115.

Oosterhof, N. N., & Todorov, A. (2009). Shared perceptual basis of emotional expressions and trustworthiness impressions from faces. *Emotion, 9*(1), 128.

Phillips, P. J., Moon, H., Rizvi, S. A., & Rauss, P. J. (2000). The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*, 1090–1104.

Phillips, P. J., Wechsler, H., Huang, J., & Rauss, P. J. (1998). The FERET database and evaluation procedure for face-recognition algorithms. *Image and Vision Computing, 16*(5), 295–306.

Rhodes, G. (1988). Looking at faces: First-order and second-order features as determinants of facial appearance. *Perception, 17*(1), 43–63.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision Research, 46*(18), 2977–2987.

Rhodes, G., & Leopold, D. A. (2011). Adaptive norm-based coding of face identity. In A. Calder, G. Rhodes, M. Johnson, & L. Haxby (Eds.), *The Oxford handbook of face perception* (pp. 263–286). Oxford, UK: Oxford University Press.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition, 21*(2), 139–253.

Schwaninger, A., Lobmaier, J. S., & Collishaw, S. M. (2002). Role of featural and configural information in familiar and unfamiliar face recognition. *Lecture Notes in Computer Science*, *2525*(2002), 643–650.

Shin, N. Y., Jang, J. H., & Kwon, J. S. (2011). Face recognition in human: The roles of featural and configurational processing. *Face Analysis, Modelling and Recognition Systems, 9*, 133–148.

Sinha, P., & Poggio, T. (1996). I think I know that face. *Nature*, *384*, 404.

Sinha, P., & Poggio, T. (2002). Last but not least. *Perception, 31*, 133.

Tanaka, J., Giles, M., Kremen, S., & Simon, V. (1998). Mapping attractor fields in face space: The atypicality bias in face recognition. *Cognition, 68*(3), 199–220.

Todorov, A., & Oosterhof, N. N. (2011). Modeling social perception of faces [social sciences]. *Signal Processing Magazine, IEEE, 28*(2), 117–122.

Toseeb, U., Keeble, D. R., & Bryant, E. J. (2012). The significance of hair for face recognition. *PloS One, 7*(3), e34144.

Townsend, J. T., Solomon, B., & Smith, J. S. (2001). The perfect Gestalt: Infinite dimensional Riemannian face spaces and other aspects of face perception. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives of facial cognition: Contexts and challenges* (pp. 39–82). Mahwah, NJ: Lawrence Erlbaum Associates.

Townsend, J. T., & Thomas, R. D. (1993). On the need for a general quantitative theory of pattern similarity. *Advances in Psychology, 99*, 297–368.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. *The Quarterly Journal of Experimental Psychology, 43*(2), 161–204.

Valentine, T. (2001). Face-space models of face recognition. In M. J. Wenger & J. T. Townsend (Eds.), *Computational, geometric, and process perspectives on facial cognition: Contexts and challenges* (pp. 83–113). Mahwah, NJ: Lawrence Erlbaum Associates.

Yovel, G., & Kanwisher, N. (2004). Face perception: Domain specific, not process specific. *Neuron, 44*(5), 889–898.