Critical Features for Face Recognition in Humans and Machines

Naphtali Abudarham\textsuperscript{1}, Lior Shkiller\textsuperscript{1}, Galit Yovel\textsuperscript{1,2}

\textsuperscript{1}School of Psychological Sciences, \textsuperscript{2}Sagol School of Neuroscience
Tel Aviv University, Tel Aviv, Israel

Correspondence regarding this manuscript should be addressed to:

Galit Yovel
School of Psychological Sciences & Sagol School of Neuroscience
Tel Aviv University
Tel Aviv, 69978, Israel
Email: gality@post.tau.ac.il,
Abstract

Face recognition is a computationally challenging task that humans perform effortlessly. Nonetheless, this remarkable ability is limited to familiar faces and does not generalize to unfamiliar faces. To account for humans superior ability to recognize familiar faces, current theories suggest that familiar and unfamiliar faces have different perceptual representations. In the current study, we applied a reverse engineering approach to reveal which facial features are critical for familiar face recognition. In contrast to current views, we discovered that the same subset of features that are used for matching unfamiliar faces, are also used for matching as well as recognition of familiar faces. We further show that these features are also used by a deep neural network face recognition algorithm. We therefore propose a new framework that assumes similar perceptual representation for all faces, and integrates cognition and perception to account for humans’ superior recognition of familiar faces.
Introduction

Face recognition is a computationally challenging task that requires fine discrimination between similarly looking images of different identities, as well as generalization across different images of the same individual. Although humans are considered experts in face recognition, studies have shown that this expertise is limited to faces we are familiar with, whereas our ability to match unfamiliar faces is error-prone \(^1\). These findings led to the suggestion that familiar face recognition depends on a different set of facial features, based on the extensive experience that we have with them, than those used for unfamiliar faces. For example, it has been suggested that familiar face recognition is primarily based on internal facial features whereas unfamiliar face matching is primarily based on external facial features \(^3\)\(^\text{-}\)\(^6\). According to another view, the representation of familiar faces is based on the average of their different appearances, which excludes superficial image-based information that may dominate the representation of unfamiliar faces. This view further posits that throughout our experience with familiar faces we learn their idiosyncratic features that are unique for each identity, therefore suggesting that a different set of features is used to recognize different familiar faces \(^1\),\(^7\).

In a recent study, we used a novel reverse engineering approach to reveal which facial features are critical for face identity. We found a subset of features for which humans have high perceptual sensitivity to detect differences between different faces (high-PS features) \(^8\). We then showed that changing high-PS features changes the identity of faces, whereas changing features for which humans have low perceptual sensitivity (low-PS features)
did not change the identity of faces (see Figure 1). Importantly, these high-PS features remain invariant across different head views, making them useful not only for discrimination between identities but also for generalizing across different appearances of the same identity.

Nevertheless, this subset of features was shown to be critical for unfamiliar faces and may not generalize to familiar faces, with which we have much greater experience. Thus, the goal of the current study was to use the same reverse engineering approach to reveal which features are critical for familiar face recognition. This will allow us to test the common view that different facial features are used for the identification of familiar and unfamiliar faces.
To that end, we first examined the role of high-PS vs. low-PS features in a familiar face matching task, using the same matching task that was used for unfamiliar faces in our previous study (Experiment 1). An important difference between familiar and unfamiliar faces is that familiar faces are represented in memory. Features that are used for matching two faces presented simultaneously, may not be used for matching a familiar face to its representation in memory. Therefore, we also examined whether these features are also used for face recognition (Experiments 2 & 3). Finally, the features that we found correspond to semantic descriptions of facial features (e.g., eyes, mouth), and may therefore overlook visual information that cannot be described by these labels. We therefore examined whether these features are also used by a face recognition algorithm, that is not bound to these semantic meanings. Recently, Deep Neural Network (DNN) algorithms have reached human level performance on unconstrained ("wild") facial images, in
which faces appear in various poses, expressions and illuminations\textsuperscript{9}. These advances are the result of the capability of DNNs to extract the invariant information through supervised learning with many different images of the same identity. We therefore hypothesize that a DNN may be tuned to the same invariant, high-PS features that humans use for face recognition (Experiment 4).

\textbf{Experiment 1 – Critical features for matching familiar faces.}

To determine whether changing high-PS features, but not low-PS features, changes the identity of a familiar face, we used a matching task similar to the one we used in a previous study with unfamiliar faces\textsuperscript{8}. Familiar faces were modified by either changing five high-PS features or five low-PS features (Figure 2, Figure S3). We presented participants with pairs of celebrity faces, before and after feature change, and asked them to rate whether the two pictures belong to the same person or to different people (Figure 3A, top). Pairs of same identity and different identity faces were also presented to obtain baseline performance to which matching abilities for low-PS and high-PS pairs can be compared.

\textbf{Results & Discussion}

Figure 3A shows a histogram of the dissimilarity ratings for the four types of pairs (Figure 3A, middle row) and an average dissimilarity rating score for each of the four conditions (Figure 3A bottom row). A repeated measures ANOVA on the averaged rating revealed a significant difference between the 4 conditions ($F(3,111) = 257.75$, $p << 0.001$, $\eta^2 = 0.87$). Post-hoc comparisons of the four different conditions with p-value corrected for six comparisons ($p = 0.008$) revealed a significant difference between all four conditions ($t(37) > 8.43$, $p < .0001$), except the difference between high-PS and different faces ($t(37)= 2.2$, $p = .03$). These findings indicate that changing high-PS features changed
the identity of the face. Low-PS changes were not perceived as the same person but also did not make faces look like different people. Thus, low-PS features are noticeable changes but do not change the identity of a face. These results are similar to results we obtained in our previous study with unfamiliar faces (see Figure 3B based on data taken from Abudarham & Yovel, 2016).

We also compared the feature vector distance (L1-norm: the sum of absolute differences between 20 feature values) for the original faces and faces that underwent high and low-PS changes, based on the feature tagging data. Overall, the distances between the original and changed faces were not significantly different for High-PS changes (M = 24.05, STD = 1.4) than Low-PS changes (M = 22.05, STD = 4.3) (t(9) = 1.37, p = 0.2). Yet, as reported above, the perceptual distance (i.e. how different the changed faces were perceived) was larger for High-PS changes. These results indicate that although the face-space distance based on all 20 features was the same for both types of changes, different features contributed differently to the perceptual distance between faces. In other words, the direction of change in space is more critical than its distance.

In summary, our results suggest that the same features that were critical for unfamiliar face matching are also critical for familiar face matching. These findings are therefore inconsistent with the suggestion that different features are used to determine the identity of familiar and unfamiliar faces. Instead, our data show that the same subset of features are used to determine the identity of all faces regardless of familiarity or their specific identity.
Figure 3: Results of the face matching task. A: Familiar faces; B: Unfamiliar faces. Top: The four types of face pairs presented in the experiment. Middle: histogram of the distribution of same-different identity rating from 1 – definitely same identity to 6 – definitely different identity. Bottom: The average similarity rating scores for each pair type shows that matching the original and a changed high-PS face yielded scores that are similar to different pairs.

**Experiment 2 – Critical features for familiar face recognition**

In Experiment 1 we found that high-PS features are critical for familiar face matching. However, matching two images presented simultaneously may not necessarily depend on the same features used to match an incoming face to its representation in memory. Furthermore, in real life we hardly ever need to match different images of familiar faces, but often need to recognize familiar faces by matching them to their representation in memory. In Experiment 2, we therefore examined whether high-PS, but not low-PS features are also critical for familiar face recognition. In addition, we asked how many feature changes are needed to change the identity of a face. To that
end, we presented faces in a face recognition task, and measured recognition performance as a function of the number of high-PS or low-PS feature changes (see Figure 4A for an example of the face of George Clooney).

Results & Discussion

Figure 4B shows recognition rates for high and low-PS changes as a function of the number of changes. A mixed two-way ANOVA with PS type (high/low) as a between participants factor, and number of changes (0-5) as a within subject factor, revealed a significant effect of PS type indicating better recognition for low-PS than high-PS changed faces. A significant effect of number of features indicates that recognition decreases with the increase in the number of changes ($F(5,190) = 76.26, p < .00001, \eta^2 = .67$). An interaction
between the two variables indicates that recognition dropped much more steeply for high-PS than for low-PS features (F(5,190) = 21.13, p < .00001, $\eta^2$ = .36). Post-hoc comparisons revealed a significant difference between high and low-PS changes for 5, 4 and 3 changes (t(38) > 3.06, p < .004) but not for 2 and 1 changes (t(38) < .7, p > .37). Thus, by changing only 3 high-PS changes we caused a significant drop in recognition of familiar faces.

The results of this experiment clearly show that high-PS features are critical for face recognition. Changing more than three high-PS features completely modified the identity of a familiar face such that none of the participants were able to recognize it. In contrast, only 50% of the celebrity faces in which five low-PS features were replaced could be recognized. These results indicate that high-PS features are not only used for matching unfamiliar or familiar faces presented simultaneously, but are also central to the representation of the identity of a familiar face in memory. In the next set of experiments, we examined the relative contribution of different high-PS features by changing the order in which they were modified.

**Experiment 3: The relative contribution of different High-PS features for face recognition**

To assess the relative contribution of individual feature changes, we altered the order in which features were changed. Since there are 120 different orders of 5 feature changes, we had to select a few of interest. We started with changes in the reverse order to the one that we used in Experiment 2.

**Experiment 3A**

**Results & Discussion**

Figure 5 shows recognition rates for the new order of the high-PS changes.
relative to the low and high-PS changes used in Experiment 2. A mixed ANOVA with Type (high-PS original order, high-PS reversed order) as a between subject factor, and number of changes as a within subject factor, revealed a main effect of no. of changes ($F(5,190) = 112.96, p < .0001, \eta^2 = .75$) and an interaction of no. of changes and order of changes ($F(5,190) = 3.08, p < .01, \eta^2 = .07$). The interaction reflects the difference in recognition for 3 feature changes in which the hair changed in the original order but not in the reversed order. As a result, performance level was high when the original hair was present than when it was changed ($t(38) = 3.08, p < .004$).

Results therefore suggest that hair has important contribution for face recognition. We therefore conducted another experiment in which we changed the hair first or last and examined recognition rate for the same faces.

**Experiment 3B**

**Results & Discussion**
A mixed ANOVA with Type (Hair First, Hair Last) as a between subject factor, and number of changes as a within subject factor, revealed a main effect of no. of changes ($F(5,185) = 176.37, p < .0001, \eta^2 = .83$) and an interaction of no. of changes and order of changes ($F(5,185) = 9.53, p < .001, \eta^2 = .20$). Independent sample t-tests corrected for 6 multiple comparisons ($p = .008$) revealed better performance for hair-last than hair-first changes for 2 and 4 changes ($t(38) > 3.84 p < .0001$) and a marginally significant difference for 1 and 3 changes ($t(38) > 2.6 p < .01$). There was no significant difference between no changes and five changes, which were based on identical images in the two versions of the experiment ($p = .35$).

These findings show a dramatic effect of hair on familiar face recognition. When the hair is changed last, recognition level for high-PS features drops below low-PS features changes after 3-4 feature changes. Whereas when the hair is changed first, recognition level drops below low-PS changes already at 2 feature changes.
Hair is often cropped in face recognition experiments that focus on the role of the internal facial features for face recognition. Here we show that face recognition is highly influenced by hair even in celebrities that tend to modify their hair styles. These findings suggest that despite the fact that hair is a relatively variable facial feature, it is an integral and critical part of the representation of the identity of a person.

**Experiment 4—Critical features for machine face recognition**

Our results so far reveal a subset of features that are used to determine the identity of a familiar and unfamiliar faces in matching and recognition tasks. However, these features correspond to semantic descriptions of facial features (eyes, mouth). It was therefore important to assess whether these features are also used by a face recognition algorithm that is not bound to these semantic meanings. To this end we tested the similarity scores for the feature vectors of the original and changed pictures, obtained using a face-recognition DNN.

**Results & Discussion**
Figure 7 shows the mean similarity scores for the familiar and unfamiliar face images used in the human matching task described in Experiment 1. The results for machine face matching are very similar to the results of human face matching. A repeated measure ANOVA on the four face conditions with face stimuli as a random variable revealed a main effect of condition ($F(3,72) = 117.63, p < .0001, \eta^2 = .36$). Post hoc t-tests corrected for 6 comparisons ($p = .008$) revealed a significant difference among all conditions ($t(24) > 3.5, p < .002$) in the same direction as was found for Humans.

A Mixed ANOVA with DNN/Human as a between-groups factor and Condition (Same, Low-PS, High-PS, Different) as repeated measures reveal a significant main effect of Condition ($F(3,144) = 388.83, p < .0001, \eta^2 = .89$) and an interaction between the two factors ($F(3,144) = 3.68, p < .02, \eta^2 = .07$), which reflects the larger difference between high-PS and different faces for the algorithm than humans ($F(1,48) = 10.33, p < .005, \eta^2 = .18$). This may result from the fact that that algorithm training is performed on faces with cropped hair, which has a major effect on human face recognition (see Figure 6). We
therefore further examined the effect of each feature on the representation of faces by the DNN by comparing the feature vector distance of faces that differed in only one feature. This analysis revealed that out of the high-PS features, hair and eye color are not used as much by the algorithm, and out of the low-PS features, the nose is used by the algorithm (see figure S3).

In summary, we found that the same high-PS features that are critical for human face matching, are also used by DNNs trained with unconstrained faces. The fact that the same features are critical also for machine face recognition, indicates that these features have visual importance beyond their semantic meaning. We do not argue that the feature-vector representation of the DNN that we used specifically has eyebrow thickness or lip thickness encoded in it, but that high-PS features are highly correlated with the information represented in the network. We suggest that these features are important because they are invariant across different appearances of the same identity, whereas low-PS features, such as skin color and eye distance, vary across different appearances, such as lighting and head pose.

**General Discussion**

We discovered a subset of facial features that are critical for human familiar and unfamiliar face matching, human familiar face recognition and machine face recognition. To reveal these features we used a novel reverse-engineering approach, in which we replaced different facial features until faces were perceived as different identities. We found that faces that differed in 5 or even 3 high-PS features but not those that differed in low-PS features, were perceived as different identities. An important difference between high and low-PS features is that the latter tend to vary under variations in appearance caused by changes in pose ⁸, in illumination or expression, thus making them
Our findings have important implications for several major topics in the study of face recognition. First, they suggest a novel categorization of facial features instead of the commonly used external-internal and part-configuration categorization; Second, in contrast to current theories that suggest that different facial features are used for familiar and unfamiliar faces, they suggest that the same features are used for familiar and unfamiliar faces. Third, they go beyond current comparisons between human and machine face recognition, which primarily focus on performance level, and examine the similarity of the internal face representations in humans and machines. We will discuss each of these topics in turn.

The face recognition literature has long been interested in discovering which facial features determine the identity of the face. Two prominent categorizations of facial features have been extensively examined in previous studies: external vs. internal features, and configural vs. part-based categorization. In particular, it has been suggested that internal features are more critical for the identification of familiar faces but not for unfamiliar faces \cite{3,6}. These studies typically crop external facial features and present either the external part, which includes the ears, face outline and the hair or only the internal features. Our method, which provided us with a much more detailed examination of the role of 20 different facial features, each was measured independently, and relies on human perceptual sensitivity \cite{8}, revealed a new categorization that is inconsistent with the external-internal distinction. Out of the 5 high-PS features that we changed, 4 are internal features (lip-thickness, eye-color, eye-shape and eyebrow-thickness), and the fifth is the hair, which had a large effect on both familiar and unfamiliar face recognition (Figure 5,
see also\textsuperscript{10}. Furthermore, low-PS changes such as eye-distance and mouth-size are also internal features but were found to be non-critical for face identity. Thus, our fine distinction allowed us to separately test features such as lip thickness vs. mouth size or eyebrow thickness vs. eyebrow shape, showing that the former but not the latter in these two examples, are important for the identity of the face. These findings suggest that a gross categorization to internal and external features is inept to reveal which features are used for face recognition.

A second frequently used categorization of facial features is the part vs. configural (i.e., spacing) distinction. Whereas earlier studies indicated that configural processing is more important than part-based processing for face recognition\textsuperscript{11} later studies have shown that both spacing and part-based information are used for face recognition\textsuperscript{12–14} for review see\textsuperscript{16}. The special role of configural processing in face recognition has been also challenged by studies showing that compressed familiar faces can be easily recognized despite the fact that their eye distance and face proportions are largely distorted\textsuperscript{17,18}. Our findings that eye distance and face proportion are not used for face recognition are also consistent with these results. On the other hand, eye shape and lip thickness, which are considered local, part-based features, were found to be important for face recognition. These data are therefore inconsistent with the common notion that face recognition does not depend on face parts. Instead, we suggest an alternative categorization, according to which facial features that are critical for face recognition are those that remain invariant across different appearances of the same individual and can be easily discriminated across different identities.

A second important finding our study reveals is that the same critical features are used for familiar and unfamiliar faces. Recent studies have
convincingly shown that our ability to match different images of the same individual is relatively poor for unfamiliar faces, but is superior for familiar faces. These findings have led to the suggestion that the perceptual representation that we acquire during our experience with familiar faces is robust to these different variations, whereas the lack of experience with unfamiliar faces results in a picture-based representation that does not support generalization across such different images of the same individual. However, our findings show that face identity of familiar and unfamiliar faces is based on the same facial features. We therefore suggest that the perceptual representation of unfamiliar faces is based on our experience in the world with familiar faces. Our system learns which features are suitable for both discrimination and generalization of familiar faces, and then applies the same features for matching unfamiliar faces.

What may therefore underlie the better generalization across images of the same identity for familiar than unfamiliar faces? An important difference between familiar and unfamiliar faces is that familiar faces are represented in memory. Thus, unlike unfamiliar faces that can only be matched perceptually, familiar faces are also matched to their representation in memory. In contrast to the suggestion that it is the averaged representation of the different appearances of a familiar face that supports this generalization across its different appearances, we suggest that images of the same individual that are highly dissimilar, such that a person who is unfamiliar with that face perceive them as different individuals, may be stored as separate exemplars under the same conceptual representation. An example of such conceptual representation are the concept cells revealed in the medial temporal lobe that respond to very different images of the same individual as well as their name. If each of two very different images of a familiar face can be each
matched with their stored representations, they can still be matched correctly as they are linked to the same concept of that familiar person. Thus, matching of two different images of the same familiar face does may not depend only on perceptual matching but also on conceptual matching. This suggestion should be further explored in future computational and experimental studies.

A third finding reported in our studies is that the same features that were found to be important for human face recognition are also used by a DNN. Comparing human and machine face recognition adds further support for the validity of our suggested feature categorization. Even though the features that we used to tag faces are "nameable" features, i.e. features with semantic meaning, the fact that machine face recognition is also sensitive to the same features indicates that high-PS features capture the essence of the identity of a face.

Previous studies that compared human and machine face recognition have focused on performance level, aiming to develop an algorithm that reaches the level of humans or beyond. Our novel method allowed us to go beyond performance level and compare the nature of the representation of humans and machines. In particular, we showed that humans and DNNs are tuned to the same facial information and overlook other facial information that may not be useful for face recognition. We believe that this similarity in representations is because both humans and DNNs are trained on high variability of labeled images of the same and different individuals and need to learn which features remain invariant across different images of the same individual and at the same time vary across different identities.

Finally, it is noteworthy that the features that we discovered are useful for the identity of Caucasian male faces. Features such as eye color and hair color that are critical for Caucasian faces are unlikely to be used for Asian or
African faces, which show different variations in these features. Similarly, hair and eye-brow thickness that are also used for Caucasian faces are unlikely to be used to discriminate infant faces that hardly have hair or eyebrows, and are all look very similar to one another. Our findings may therefore account also for the well-established other-race effect, indicating that lower performance for other-race faces may be due to the usage of set of facial features that are not diagnostic for the race we had no experience with. Our methods can be therefore used to determine the critical features of any category of faces based on the same principle of perceptual sensitivity and the reverse engineering approach. These features are typically learned in real life through our experiences with different categories of faces. As mentioned above, we suggest that the training needed to extract these features is based on our experience with familiar faces, rather than passive exposure to unfamiliar faces. This claim is based on previous studies that show that it is the association of faces with person-related information that improves face recognition rather than the pure perceptual exposure to a large number of familiar faces.

In summary, we discovered a subset of features that determine the identity of familiar and unfamiliar faces in humans and a DNN. We suggest that these features are learned through our rich experience with familiar faces to obtain optimal generalization and discrimination. These features are then applied to determine the similarity between any incoming face stimuli and are therefore similar to both familiar and unfamiliar faces. Future studies are needed to further explore the mechanisms that account for the different abilities that humans show for familiar than unfamiliar faces, and the extent to which they are based on conceptual rather than perceptual matching as was suggested here.
Method

Experiment 1

Participants:

All participants were Amazon-Mechanical-Turk workers, participating in the experiment for payment (approximately 1$ per 15 minutes of work). A total of 38 participants (American residence, 18 females, 28 Caucasians, 6 East-Asians, 2 African-American, 1 Hispanic/Latin and 1 Middle-Eastern, ages 23-66 (mean 39.4, standard deviation (SD) 13.6) performed the experiment.

Stimuli:

Ten American celebrities - all adult Caucasian males - were selected for the experiment. For each identity, we downloaded from the internet two frontal neutral expression images, with no glasses, hat or facial hair, and with adequate lighting and quality. All pictures were cropped from the background, and cut below the chin, leaving just the face, including the hair and ears. One of these images was selected as a "base" picture, a picture that was later modified, and the other designated as a "reference" picture, which was left unchanged. Additional 100 pictures of Caucasian male faces, with no glasses or facial hair, were taken from the Color FERET database, and cropped in the same way.

Face tagging: converting faces into feature vectors, and measuring face-space distances

In our previous study we demonstrated that it is possible to describe faces as feature-vectors embedded in a multidimensional feature space. We showed that by perceptually assigning values to a set of 20 features, we can measure distances between faces, and these distances were correlated with perceptual face similarity. In this study, we repeated this procedure with familiar faces and converted each one of the faces in our database into a
For the 100 faces from the color-FERET dataset we used the feature-vector representations obtained in our previous study. For tagging the ten celebrity faces we ran a face-tagging procedure. To provide participants with a large enough dataset for tagging, allowing them to judge facial features with respect to a variance of feature sizes and shapes, we created a dataset of 60 face images. These 60 images included the selected 10 celebrity faces, 20 pictures of other celebrities of similar characteristics as the original 10, and 30 randomly selected pictures from the 100 color FERET dataset. In the tagging procedure, participants were asked to rank each of the 20 features for each of the sixty faces on a scale of -5 and +5 (for example: how bright is the skin? how large are the eyes?). We then normalized these values to get a relative rating score (Figure S1). A total of sixty-eight participants participated in the tagging experiment, each participant rated 7 features (to avoid fatigue), such that an average of 10 participants tagged each feature in each face.

Facial feature substitution:

Based on the feature vectors that we obtained for each face, we were able to select features from other faces, to replace the original features (see Figures 2). The features were taken from faces in our original database of 100 tagged faces used in the previous study. Feature replacement was done by copying and pasting a feature with an opposite and as far away value as possible. For example, to replace thin lips, we took one of the three thickest lips available in the database (we did not always take the top thickest, to avoid repeating the same feature in too many faces). This face modification process was performed by a professional graphic designer using Adobe Photoshop. We created two sets of changed pictures: in one set we changed high-PS...
features, namely the lip-thickness, the hair, eye-color, eye-shape and eyebrow-thickness. In the second set, we changed low-PS features, namely mouth-size, eye-distance, face-proportion, skin-color and nose. (See Figure 2 for an example of George Clooney) (see Figure S3 for 5 high and low-PS feature changes for the 10 celebrity faces). Following feature substitution, we ran again the face tagging procedure, on different participants, and computed feature vectors for the changed faces. This allowed us to calculate the face-space distances (the sum of absolute differences between feature values, or L1-norm) between faces before and after change.

**Face matching task:**

Following the feature substitution, we had 4 images for each celebrity face: two original images ("base" and "reference"), one image with high-PS feature changes, and one with low-PS feature changes. As mentioned above, we had two images for each identity, a “base” image and a “reference” image. Only the “base” image of each identity was modified. The “reference” image was left unchanged so matching the modified and the original face is not based on low-level image based similarity.

We then created 4 types of pairs for the matching task for each celebrity: *Same pair* — "reference" and "base" images of the same identity, *Different* pair— "reference" image and the "reference" image of another celebrity, *High-PS* pair— the "reference" image and the image with high-PS changes, and *Low-PS* pair – the "reference" image and the image with low-PS changes. We used these 40 pairs to create two versions of a matching task, each version with 30 face pairs: Version 1 – the 10 Same pairs, the 10 Different pairs, 5 High-PS pairs and 5 Low-PS pairs, and Version 2 – the 10 Same pairs, the 10 Different pairs, the 5 High-PS pairs not included in Version 1, and the 5 Low-PS pairs not included in Version 1. Each subject participated in one of two
versions of the experiment, in which half of the faces were changed in their high-PS features, and the other half in their low-PS features.

To determine perceptual similarity between face pairs, the two face images were presented simultaneously, side by side on the computer screen, and participants were asked to judge whether they belonged to the same person or to different people. The participants' response was on a scale of 1 to 6: ‘1’ indicating "definitely the same person", ‘2’ indicating “same person”, ‘3’ indication “possibly the same person”, ‘4’ indicating “possibly different people”, ‘5’ indicating “different people” and ‘6’ indicating "definitely different people". The two pictures were presented on the screen until response, after which the next two faces were presented. The order of the pairs, as well as the right-left positions of the images within each pair were randomized across participants.

**Experiment 2**

**Participants:**

Participants were 40 Amazon-Mechanical-Turk workers (American residents, 22 females, 34 Caucasians, 3 Hispanic/Latins, 1 African American, 1 South-Asian and 1 Native American, Ages 21-51, mean age 32.7, STD = 7.9), participating in the experiment for payment (approximately 1$ per 15 minutes of work). They were randomly assigned to two groups: 20 participants were presented with faces that were modified by changing high-PS features, and the other 20 participants were presented with faces that were modified by changing low-PS features (see Figure 4A).

**Stimuli:**

We used the same original pictures, of ten celebrities, as in Experiment 1. For each identity, we generated 10 different modified faces in which we gradually changed either 5 high-PS features or 5 low-PS features. The order of
changes corresponded to the descending order of the 5 top PS scores of features (see Figure S2). For high-PS changes, faces were gradually changed in the following order: lip-thickness, hair, eye-color, eye-shape, and eyebrow-thickness. For low-PS changes we had the following order: mouth-size, eye-distance, face-proportion, skin-color and nose, matching the ascending order of the low-PS features. Thus, a total of 10 identities with 20 modification for each (5 High-PS and 5 Low-PS changes) were used in the experiment. Figure 4A shows the 10 different modifications for the one identity.

**Procedure:**

Participants were presented with faces that were changed in high-PS or in low-PS features. The images were presented one at a time, in the following order: first the 10 faces with 5 feature changes (the order within these 10 faces was randomized across participants), then the 10 faces with 4 feature changes, and so on until finally the 10 original celebrity images were presented. Thus, each participant was presented with a total of 60 faces. For each image, participants were asked to write the name of the person in the image. If they did not recall the name they were asked to describe him as best as they could. If they could not recognize the face they marked "I cannot recognize this face". After they typed the response they pressed a key, which initiated the presentation of the next face.

**Data Analysis:**

All responses were analyzed manually to determine whether the participant correctly identified the face in the image. For example, replies like "Mark Zuckerberg", "Zuckerberg" or "The Facebook guy" were all accepted as correct recognition of the picture of Mark Zuckerberg. For each participant, each reply was scored "1" for correct recognition, and "0" for an incorrect recognition or an "I cannot recognize this face" response. Recognition rate was
calculated for each of the face manipulations by dividing the correct scores for each participant by the total number of 10 identities (Figure 4B).

**Experiment 3A**

**Participants:**

Twenty Amazon-Mechanical-Turk workers (American residents, 10 females, 12 Caucasians, 3 African-Americans, 3 Hispanic/Latins, 1 East-Asian and 1 Native American, ages 25-53, mean age 33.75, STD = 7.7), participated in the experiment for payment (approximately 1$ per 15 minutes of work).

**Stimuli:**

In the previous experiment the order of feature changes started with features with the highest PS score (lip thickness, hair, eye-shape, eye-color, eyebrow-thickness) (see Figure S2). Here we reversed the order of feature changes (i.e. the order was: eyebrow-thickness, eye-shape, eye-color, hair and lip-thickness).

**Experiment 3B**

**Participants:**

Thirty-nine Amazon-Mechanical-Turk workers (American residents, 20 females, 31 Caucasians, 4 East-Asians, 2 Hispanic/Latins, 1 South-Asian and 1 African-American, ages 18-56, mean age 35, STD = 9.1) participated in the experiment for payment (approximately 1$ per 15 minutes of work). 20 of them were randomly allocated to the hair-first version of the experiment and the rest to the hair-last version.

**Stimuli:**

We used the same stimuli but changed the features by either changing the hair first or last (see Figure 6). The order of the other features was similar to the order used in Experiment 2.

**Experiment 4**
Stimuli

We used the 10 celebrity face images that were used in Experiment 2 and 15 unfamiliar faces that were manipulated in the same way. All images were pre-processed using the DLIB image processing library: the faces were aligned to match eyes and nose locations across all faces, and then were cropped to a size of 96 by 96 pixels (cropping out the hair).

Face recognition algorithm:

We used the OpenFace DNN (https://cmusatyalab.github.io/openface/) for face recognition, and the pretrained model nn4.small2.v1.t7 available for download from the OpenFace site. OpenFace is trained with 500,000 images from labeled face recognition datasets, CASIA-WebFace and FaceScrub. The algorithm learns a mapping from face images to an embedding space where distances in the embedding space directly correspond to a measure of face similarity (faces of the same person have small distances and faces of different people have large distances). OpenFace trains its output to be a feature vector of length 128 using a triplet-based loss function. The triplets consist of two matching face images and one non-matching face images and the loss aims to separate the positive pair from the negative by a distance margin.

Image similarity measurements:

The OpenFace network processes image faces and outputs a feature vector representation of length 128 for each face. Image similarity scores are Euclidian distances between feature vectors. We computed similarity scores between original pictures and changed pictures for each type of change (high-PS or low-PS changes). We also computed similarity scores for Same pairs (the "base" and "reference" pictures for each celebrity), and for Different pairs (all pairs of "base" celebrity pictures). To compare human and machine distance
scores we normalized each of the ratings by dividing it with its largest score (6 for human dissimilarity scores and 2 for DNN dissimilarity scores)

Acknowledgements:

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.

References:

   doi:10.1177/0963721416688114


Supplementary Figures

Figure S1: Example of the feature vectors of two celebrity faces. The values of the vectors are based on standardized ratings of each of the features by human observers. For example, a value of -2.072 for the skin texture of Justin Bieber indicates a very smooth skin texture of this face. The rating of the same feature of -.036 reflects a more average skin texture for Ben Affleck.

```
<table>
<thead>
<tr>
<th>Feature</th>
<th>Ben Affleck</th>
<th>Justin Bieber</th>
</tr>
</thead>
<tbody>
<tr>
<td>skin color</td>
<td>-0.91577</td>
<td>-1.78666</td>
</tr>
<tr>
<td>skin texture</td>
<td>-0.36613</td>
<td>-2.0721</td>
</tr>
<tr>
<td>hair color</td>
<td>0.60688</td>
<td>-1.9045</td>
</tr>
<tr>
<td>face length</td>
<td>0.16192</td>
<td>1.2414</td>
</tr>
<tr>
<td>face proportion</td>
<td>0.60565</td>
<td>1.2643</td>
</tr>
<tr>
<td>jaw width</td>
<td>-1.2813</td>
<td>-1.5208</td>
</tr>
<tr>
<td>chin shape</td>
<td>0.13596</td>
<td>-0.76253</td>
</tr>
<tr>
<td>forehead height</td>
<td>-0.85322</td>
<td>-0.6144</td>
</tr>
<tr>
<td>eyebrow shape</td>
<td>1.9701</td>
<td>2.5753</td>
</tr>
<tr>
<td>eyebrow thickness</td>
<td>-0.60791</td>
<td>0.85477</td>
</tr>
<tr>
<td>eye shape</td>
<td>-1.0115</td>
<td>-0.89217</td>
</tr>
<tr>
<td>eye size</td>
<td>-1.4604</td>
<td>-0.84872</td>
</tr>
<tr>
<td>eye color</td>
<td>0.65791</td>
<td>1.1105</td>
</tr>
<tr>
<td>eye diameter</td>
<td>-0.11278</td>
<td>-0.96248</td>
</tr>
<tr>
<td>eye projection</td>
<td>0.56705</td>
<td>-1.1379</td>
</tr>
<tr>
<td>cheek shape</td>
<td>-1.0122</td>
<td>-0.90042</td>
</tr>
<tr>
<td>nose size</td>
<td>-1.3955</td>
<td>-1.3315</td>
</tr>
<tr>
<td>nose shape</td>
<td>-0.80298</td>
<td>-0.80464</td>
</tr>
<tr>
<td>mouth width</td>
<td>0.31155</td>
<td>-1.8782</td>
</tr>
<tr>
<td>lip thickness</td>
<td>-0.39614</td>
<td>1.9674</td>
</tr>
</tbody>
</table>
```

Figure S2: Figure is based on data taken from Abudarham & Yovel (2016) showing inter-rater agreement for rating each of the features. High inter-rater agreement reflects high perceptual sensitivity (high-PS) for detecting differences between faces (e.g., Lip Thickness, Hair). Low inter-rater agreement reflects low perceptual sensitivity (high-PS) for detecting differences between faces (e.g., Mouth size, Eye distance). The feature changes in Experiment 2 were made in the order of their perceptual sensitivity. The changes for the 5 high-PS features in Experiment 3A were made in the reverse order (starting with Eyebrow Thickness).
Figure S3: The 10 celebrity faces used in the experiments. The middle columns of the left and right three-column panel, show the original face, the left column 5 low-PS feature changes, and the right column 5 high-PS feature changes. As can be seen, changing 5 low-PS features have a much smaller effect on the identity of each face, whereas changing 5 high-PS features, completely changes the identity of a face.
Figure S4: The DNN feature vector distance between faces that differ in one feature indicates the contribution of each feature to the identity of a face. Whereas on average changing high-PS features results in higher dissimilarity scores than changing low-PS features, two high PS features, hair and eye color have little influence on DNN dissimilarity scores, and one low-PS feature, the nose, has high effect on DNN dissimilarity scores.