Face recognition depends on specialized mechanisms tuned to view-invariant facial features: Insights from deep neural networks optimized for face or object recognition.

.

Naphtali Abudarham[1] Idan Grosbard[2] & Galit Yovel[1,2]

[1]School of Psychological Sciences, [2]Sagol School of Neuroscience

Tel Aviv University, Tel Aviv, Israel

Corresponding author:
Galit Yovel
gality@tauex.tau.ac.il
School of Psychological Sciences
Sagol School of Neuroscience
Tel Aviv University, Tel Aviv
Israel, 69987

Abstract

Face recognition is a computationally challenging task. Deep convolutional neural networks (DCNNs) are brain-inspired algorithms that have recently reached human-level performance in face and object recognition. However, it is not clear to what extent DCNNs generate a human-like representation of face identity. We have recently revealed a subset of facial features that are used by humans for face recognition. This enables us now to ask whether DCNNs rely on the same facial information and whether this human-like representation depends on a system that is optimized for face identification. In the current study, we examined the representation of DCNNs of faces that differ in features that are critical or non-critical for human face recognition. Our findings show that DCNNs optimized for face identification are tuned to the same facial features used by humans for face recognition. Sensitivity to these features was highly correlated with performance of the DCNN on a benchmark face recognition task. Moreover, sensitivity to these features and a view-invariant face representation emerged at higher layers of a DCNN optimized for face recognition but not for object recognition. This finding parallels the division to a face and an object system in high-level visual cortex. Taken together, these findings validate human perceptual models of face recognition, enable us to use DCNNs to test predictions about human face and object recognition as well as contribute to the interpretability of DCNNs.

Keywords: Face Recognition, Deep Convolutional Neural Networks, Artificial Intelligence, High-level visual system, Face Space, Explainability

Highlights:

- Deep convolutional neural networks (DCNNs) use human-like facial features.
- Sensitivity to these features is larger in face- than object-trained DCNNs.
- Sensitivity to these features is larger in higher than lower layers of DCNN.
- Sensitivity to these features is highly correlated with DCNN performance.
- Face-trained DCNNs are valid computational models of human face recognition.

1. Introduction

Recent advances in artificial intelligence (AI) enable machines to solve complex tasks at the level of human performance. This remarkable achievement can potentially offer a computational model for human intelligence. However, in most cases, these algorithms and the nature of the representations that they generate are too complex to provide an interpretable solution that can be tested in humans (Voosen, 2017). To overcome this challenge, we can more simply ask whether these algorithms rely on the same representation that humans use to perform the task. In case they do, this will both validate models that are based on human behavior as well as advance algorithm interpretability. Here, we use this approach to explore one of the most well-known achievements of AI, the ability to recognize faces.

Face recognition is a computationally challenging task that requires discrimination of numerous images to different identities and at the same time identifying the same person across highly variable appearances. Discovering the solution for this taxing task has been an ongoing effort of both cognitive and computer scientists. The goal of research in both fields is to uncover the nature of the representation that determines the identity of a face (O'Toole, Castillo, Parde, Hill, & Chellappa, 2018; Taigman, Yang, Ranzato, & Wolf, 2014; Valentine, 2001). But do humans and machines reach a similar solution? Do they rely on the same facial information for face recognition? Answering this question by integrating the complementary approaches used in these two disciplines, will advance our understanding of both human and machine face recognition, and ultimately offer a possible solution to this unresolved problem.

To assess whether DCNNs generate a human-like face representation, we employed our recent findings based on human psychophysical experiments, which revealed a subset of view-invariant facial features that are critical for human face recognition (Abudarham & Yovel, 2018; Abudarham, Shkiller, & Yovel, 2019; Abudarham & Yovel, 2016). To demonstrate that these features are critical for human face recognition, we used a reverse-engineering approach and defined critical features as features that changing them changes the identity of a face. Fig. 1 shows the effect of changing critical and non-critical features on the identity of George Clooney. Our findings showed that none of the celebrity faces that we tested was recognized after replacing 4-5 critical features (Abudarham, Shkiller, & Yovel, 2019). Similar findings were also reported in an identity matching task of pairs of unfamiliar faces (Abudarham & Yovel, 2016). In other words, pairs of unfamiliar faces that differed in 5 critical features were judged as different identity faces.

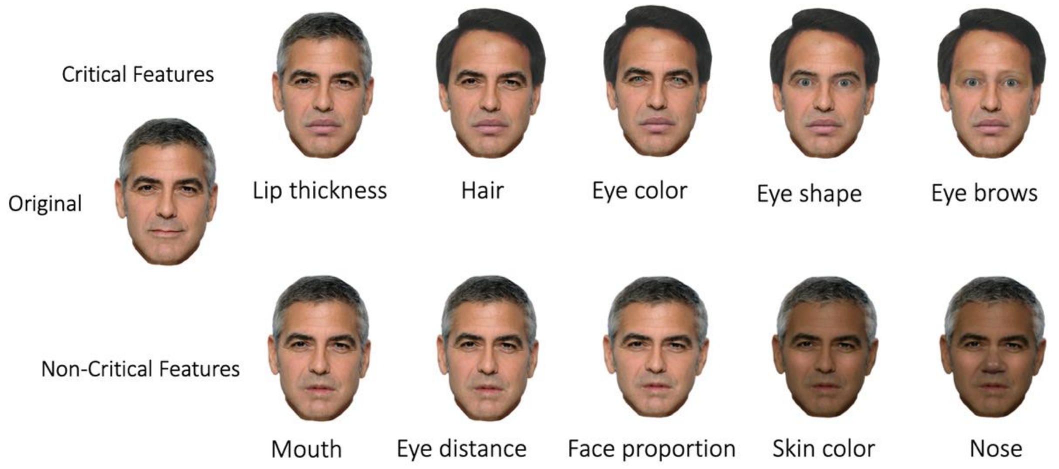

Figure 1: An example demonstrating the effect of feature changes on face identity (George Clooney). Changing five critical features (top row) changes the identity of a face, whereas changing five non-critical features (bottom row) did not change the identity of Clooney (for more details see Abudarham & Yovel, 2016; Abudarham et al., 2019).

These findings, however, were based on perceptual judgments performed by humans, on a limited set of pre-selected, namable facial features, such as eye shape or lip thickness. Thus, it is necessary to validate the relevance of these features for face recognition with a system that has reached human-level performance but is agnostic to the semantic meaning of these features. Deep convolutional neural networks (DCNNs) optimized for face recognition have recently reached human-level performance (Phillips et al., 2018; Taigman et al., 2014) and are therefore ideal to validate models of human perception.

There are two main advantages for using DCNNs to validate models of human visual processing, on top of their human-level performance: First, DCNNs have a brain-inspired hierarchical architecture. Recent studies have shown that earlier layers of DCNNs represent low-level visual features, similar to primates' early visual cortex; intermediate layers correspond to representations in mid ventral temporal cortex (Grossman et al., 2019) and high-level layers represent high-level visual features that support object/face recognition (Khaligh-Razavi & Kriegeskorte, 2014; Kubilius, Bracci, & Op de Beeck, 2016; Yamins & DiCarlo, 2016). We therefore predict that sensitivity to the critical facial features used by humans will emerge at higher layers of the network. Second, it is well-established that faces engage specialized neural mechanisms that diverge from a general object processing system at higher levels of visual processing. With DCNNs, we can fully control the type of stimuli that these models are trained with, and this way separately model a face recognition and an object recognition system. We predict that sensitivity to the critical facial features will be found in higher layers of a system that is optimized for face recognition but not for object recognition. Support for these predictions will not only advance our understanding of human face recognition but can also inform face-trained DCNNs, which are based on millions of parameters and were therefore criticized for providing non-interpretable solutions (Marcus, 2018; Voosen, 2017).

2. Study 1

In a series of studies, we discovered a subset of view-invariant facial features that are used by humans to recognize faces (Abudarham & Yovel, 2018; Abudarham et al., 2019; Abudarham & Yovel, 2016). To discover these features, we asked human participants to compare pairs of faces presented from the same view or different views on a list of 20 facial features. For example, participants were asked to indicate which of two faces has thicker lips, thicker eyebrows, a larger nose and so on. Results showed that humans show high perceptual sensitivity for a subset of these features for both same view and different view faces (for more details see Abudarham & Yovel, 2016). This subset of facial features includes the hair, lip-thickness, eye-color and shape, and eyebrow-thickness. We further found that when these features are modified, faces cannot be identified and are judged as different identities (see Figs. 2A and 1, for an example of George Clooney). We therefore named these view-invariant features, *critical features*, as they are critical for the identity of a face. In contrast, we discovered another set of features that were not well discriminated across faces presented from the same or different head-views. These features include eye-distance, face-proportion, mouth-size and nose-shape. We then found that changing faces by modifying these features did not change the identity of a face. These features were therefore named *non-critical* features (Figs. 2A and 1).

To examine the sensitivity of DCNNs to critical and non-critical facial features, we computed the Euclidian distance between face-representations of two types of face-pairs: an original face vs. the same identity in which we modified critical features, and an original face vs. the same identity in which we modified non-critical features. As a reference, we also measured the distance between same identity and different identity face pairs (see Fig. 2A). If the face-trained DCNN is sensitive to the same critical/view-invariant features as humans, the distance between faces that differ in critical features will be larger than faces that differ in non-critical features. Furthermore, the distance between faces that differ in critical features will be similar to the distance between different identity faces. We further examined how the sensitivity to these features evolves from low-level to high-level layers of the network and whether it is specific to a DCNN that is trained for face but not for object recognition.

3. Methods:

3.1. **Stimuli:**

25 face identities were used to generate image pairs. For each of the 25 identities we had an original image, an image in which we replaced critical features (modified from the original image), and an image in which we replaced non-critical features (also modified from the same original image) (for more information about the creation of the face images, see (Abudarham & Yovel, 2016)). In addition, we had a different non-modified image of that person, which we used as a reference image. Thus, we created four image pairs: *Same*– the reference image vs. the original image, *Different*– the reference image vs. a reference image of a different identity, *Critical features*– the reference image vs. the original image with different critical features, and *Non-Critical features*– the reference image vs. the original image with different non-critical features (See Fig. 2A for example face pairs).

3.2. Face-trained and Object-trained DCNNs:

For the object-trained DCNN we used the pre-trained inception_v3 DCNN from https://pytorch.org/docs/stable/torchvision/models.html, that was trained to classify the 1000 categories of the ImageNet Large Scale Visual Recognition Challenge (ILSVRC, http://image-net.org/challenges/LSVRC/). This object-trained DCNN was not trained for face identification but only for object classification. For face training we took the same inception_v3 DCNN (defined in https://github.com/pytorch/vision/blob/master/torchvision/models/inception.py), and trained it on a subset of the VGGFace2 face image dataset (http://www.robots.ox.ac.uk/~vgg/data/vgg_face2/), by randomly selecting 1000 identities from the dataset, taking the first 300 images for each identity as train data, and 50 images per identity as validation data. We started with random weights, using the default training parameters of https://github.com/pytorch/examples/blob/master/imagenet/main.py and trained the network for 120 epochs of 1000 iterations each.

To measure the DCNN level of performance on a face verification task, we used the standard Labeled Faces in the Wild (LFW) benchmark (http://vis-www.cs.umass.edu/lfw/). This test was performed by running a forward pass (inference) for LFW images and extracting the representations from the penultimate fully connected (fc) layer of each DCNN. We then measured the Euclidian distance between LFW image pairs and calculated the best accuracy measure ($Accuracy = \frac{True\ Positive + True\ Negative}{Number\ Positive + Number\ Negative}$) across all possible distance thresholds (between the minimum and maximum distance). Accuracy of the face-trained DCNN on LFW was 96.5%, whereas the accuracy of the object-trained DCNN was 69.8%. The performance of the face-trained DCNN on LFW (96.5%) is somewhat lower than the current state-of-the-art (http://vis-www.cs.umass.edu/lfw/results.html), because the training set that we used was limited to 300K images, but is still very close to human performance level (97.53%) (http://vis-www.cs.umass.edu/lfw/results.html#Human).

To assure that results are generalized to other DCNN, the same training set and procedure was used to train the DCNN VGG-16 (https://github.com/pytorch/vision/blob/master/torchvision/models/vgg.py). Performance of VGG-16, trained on the same 300K images, on LFW was 94.3%.

3.3. Extracting representations from DCNNs:

To extract the representations that were generated by the DCNNs, we ran the trained models in evaluation mode on a predefined set of image stimuli (see Stimuli section above). The face images were first aligned using the MTCNN face alignment algorithm (Xiang & Zhu, 2017). Following alignment, the images were normalized with the standard ImageNet normalization (mean=[0.485,0.456,0.406], std=[0.229,0.224,0.225]).

We first measured the pixel-based representations of all face images. We then examined the representations at the penultimate, fully connected (fc) layer. This is the layer that generates the final representation that is transformed to the output probability layer. We then examined the representations across the different layers of the network: "Conv2d_1a_3x3", "Conv2d_2a_3x3", "Conv2d_2b_3x3", "Conv2d_3b_1x1", "Conv2d_4a_3x3", "Mixed_5b", "Mixed_5c", "Mixed_5d", "Mixed_6a", "Mixed_6b", "Mixed_6c", "Mixed_6d", "Mixed_6e", "Mixed_7a", "Mixed_7b", "Mixed_7c".

To measure the distances between representations we computed the Euclidian distance between pairs of faces (python's numpy linalg.norm method). Because this distance is influenced by the size of each layer, the absolute values cannot be compared directly across layers. To compare between dissimilarity measures across different layers and between DCNNs and humans we normalized the distance scores by dividing the measured distances by the maximal distance value in each layer across all stimuli and conditions. This yielded a normalized score that ranged from 0-1 (see Figure S3 for absolute distance scores).

4. Results

4.1. The Representation of Critical Facial Features in Face-trained and Object-trained DCNNs

Figure 2A shows an example of the four types of face pairs. We performed an ANOVA with Face Type (Same, Non-Critical, Critical, Different) as a repeated measure and normalized distance between face pairs as the dependent variable. Figure 2B (left) shows performance on an identity similarity rating task performed by humans (reported in Abudarham et al., 2019). A repeated measure ANOVA reveals a main effect of Face Type (F(3,72) = 538.78, p < .001, η^2_p = .95). Post hoc comparisons (corrected for multiple comparisons) showed that all face pairs differ significantly from one another (t(24) > 17, p < .0001, Cohen's d > 3.3), except no difference between faces that differ in critical features and different identity faces (p = .76). These findings indicate that changing critical features changes the identity of a face. Figure 2B (right) shows the distances between face pairs based on pixel-based representation. A significant effect of Face Type (F(3,72) = 7.22, p < .001, η^2_p = .23) reflects a smaller distance between same identity face pairs than all other face pairs (t(24) > 3.2, p < .001, Cohen's d range 0.66-0.84), but no difference between faces that differ in non-critical features, critical features and different identity faces (t(24) < 1, p > .65, Cohen's d range 0.08-0.17). Thus, perceptual differences between faces that differ in critical and non-critical features are not due to image-based differences.

We then examined the average distances across all face identities of each of the face pairs, based on the representation in the penultimate, fully connected (fc) layer of the face-trained and the object-trained DCNNs (Fig. 2C). A mixed ANOVA with Training Type (Face, Object) as a between groups factor and Face Type (Same, Non-Critical, Critical, Different) as repeated measures on dissimilarity scores of all face pairs revealed a significant interaction between the two factors (F(3,144) = 57.37, p < .001, η^2_p = .54). The difference between critical and non-critical features was larger in the face-trained than the object-trained networks (F(1,48) = 10.65, p < .002, η^2_p = .18). The difference between Same and Different pairs was also larger in the face-trained than object-trained networks (F(1,48) = 116.55, p < .001, η^2_p = .71).

We then compared the representation of the fc layer of the object or face-trained DCNNs to human's similarity ratings of the same stimuli (Fig. 2B, left). A mixed ANOVA with System (Human, DCNN) as a between-subject factor and Face Type (Same, Non-Critical, Critical, Different) on dissimilarity scores of all face identities, was performed separately for the face-trained and the object-trained DCNNs. These ANOVAs revealed a significant interaction for humans and object-trained DCNN (F(3,144) = 135.58, p < .001, η^2_p = .74) and a much smaller difference between humans and face-trained DCNN (F(3,144) = 9.84, p < .001, η^2_p = .17). As can be seen in Figure 2 (B, C), distances between the different types of face pairs were relatively similar for the object-trained

DCNN, whereas both humans and the face-trained DCNN showed a larger difference between same and different identity pairs, as well as a larger difference between critical and non-critical feature changes relative to the object-trained DCNN (see Fig. S1 for the contribution of individual facial features to the distance between face pairs).

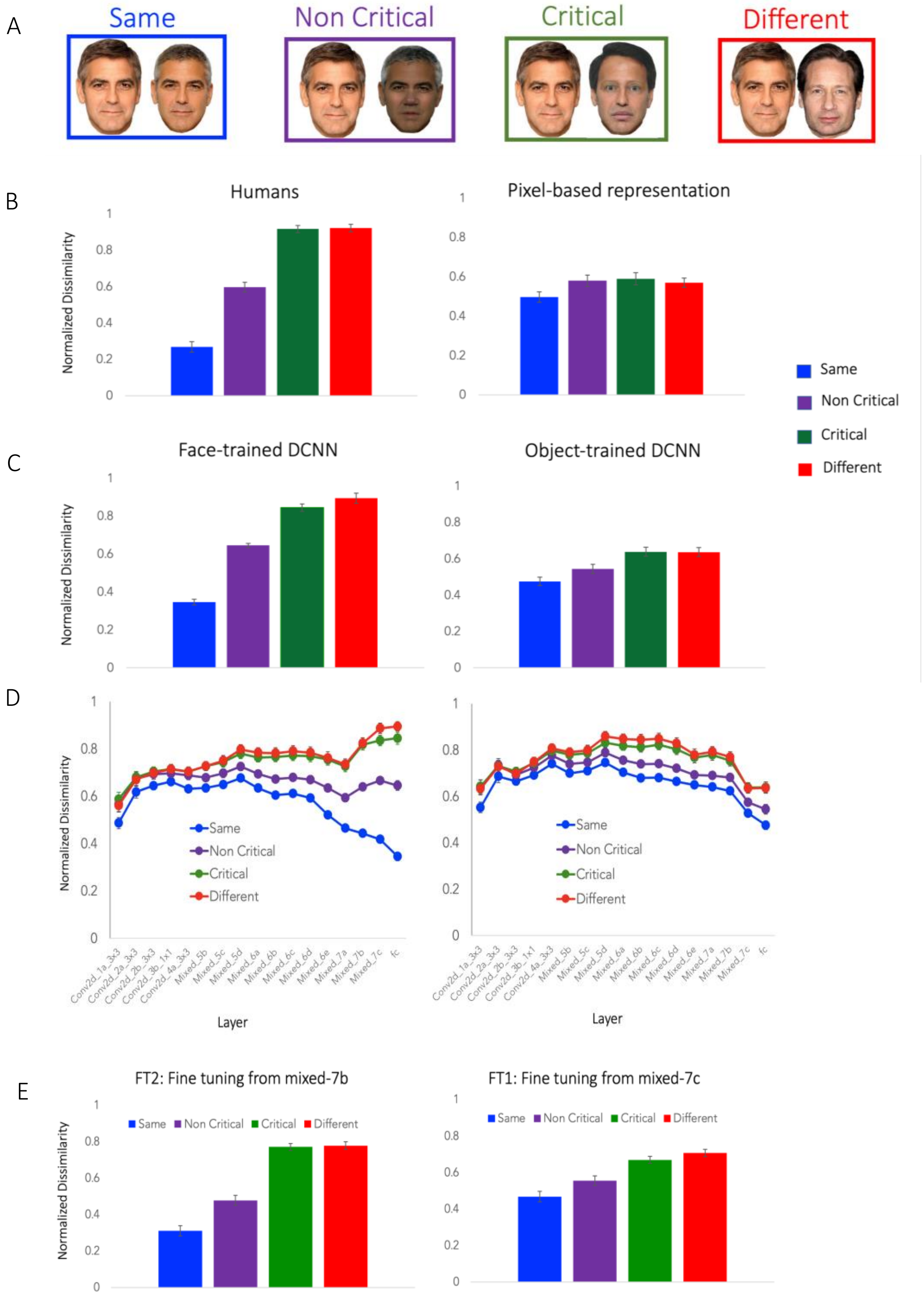

Figure 2: A. An example of the four conditions with the face of George Clooney, including a face pair of same identity faces, a face pair in which non-critical features were modified, a face pair in which critical features were modified and different identity faces. The effects are similar for unfamiliar faces B. left: Normalized human similarity ratings indicate no difference between faces that differ in critical features and different identity faces. Right: Distances between the pixel-based representations of the four face pairs indicate no difference between faces that differ in critical features non-critical features and different identity faces. C. The representation at the penultimate, fc layer of a face-trained DCNN (left) and object-trained DCNN (left). D. The representations across the different DCNN layers show that sensitivity to critical features emerges at higher layers of the face-trained DCNN. Low-level layers of both DCNNs and high-level layers of the object-trained DCNN were less sensitive to critical features (Critical > Non Critical) as well as to face identity (Different > Same). E. Training the final layers of the object-trained DCNN on face identification (fine-tuning) starting from layer mixed_7b (left) generated a representation that was similar to the fully-trained face DCNN, whereas training that started from layer mixed_7c (FT1) generated a representation that was more similar to an object-trained network. Error bars indicate the standard error of the mean dissimilarity across image pairs.

We then examined the sensitivity to critical features across the different layers of the face-trained and object-trained DCNNs. Inspection of Figure 2D shows that the sensitivity to face identity (difference between Same and Different pairs) and to critical features in particular (difference between critical and non-critical features), increases gradually across the layers, reaching its maximal value at higher layers of the face-trained DCNN. The face representations at earlier layers of both DCNNs, and the higher layers of the object-trained DCNN, were less sensitive to face identity as well as to critical features. Indeed, a mixed ANOVA with Training Type (Face, Object) as a between groups factor and Layer (all 17 layers) and Face Type (Same, Non-Critical, Critical, Different) as repeated measures, on the dissimilarity scores of 25 face identities, revealed a significant interaction of the three factors (F(48,2304) = 30.53, p < .001, η^2_p = .39).

4.2 Face training for specific layers (i.e. Fine Tuning) of the object-trained DCNN:

Comparison of the face-trained and object-trained DCNNs indicates that the representation of critical and non-critical images at early layers of the networks is similar. These findings indicate that selective training of the final layers of the object-trained network on face recognition may suffice to generate a representation that is sensitive to critical features. As can be seen in Figure 2D, the stage in the hierarchy of processing where the representations of the two DCNNs diverge is at 3-5 last layers of the DCNNs. To further examine the exact point of divergence, we took the object-trained DCNN and re-trained it with the same face images that were used to train the face-DCNN (i.e. 1000 identities with 300 images each), updating the weights of each layer at a time, until the representation was similar to the fully face-trained network (a procedure known as "fine-tuning") (see Fig. 2E).

First, we trained the weights between layer mixed_7c and layer fc (we name this condition fine-tuning 1 (FT1)). Performance on LFW was 83.3%. A repeated measure ANOVA with Training type (Full, FT1) and Face Type (Same, Non-Critical, Critical, Different) on the representation of the last layer, revealed a significant interaction ($F(3,72) = 29.88$, $p < .001$, $\eta^2_p = .55$), indicating highly different face representations for the two DCNNs. Next, we repeated the same procedure, for the weights between layer mixed_7b and Mixed_7c (we name this condition fine-tuning 2 (FT2)). Performance on LFW was 95.7%, similar to the performance of the fully trained face network. Also, the difference between the representations of the fine-tuned and fully-trained networks was not significant ($F(3,72) = 1.05$, $p = .38$, $\eta^2_p = .04$). As can be seen in Figure 2, the representations of a fully-trained DCNN (Fig. 2C) and FT2 (Fig. 2E) are nearly identical.

Finally, to assess whether sensitivity to critical features is associated with better performance on a face identification task, we measured the performance of each of the layers of the face-trained DCNN on the benchmark face verification task (LFW), as explained in the Methods section of Study 1. We then computed a measure of sensitivity to critical features, by subtracting the distances between the representations of pairs that differ in critical features and pairs that differ in non-critical features for each layer. A higher difference indicates greater sensitivity to critical features. We then computed the correlations between the two measures. Figure 3 shows a very strong correlation (r = 0.98) indicating that

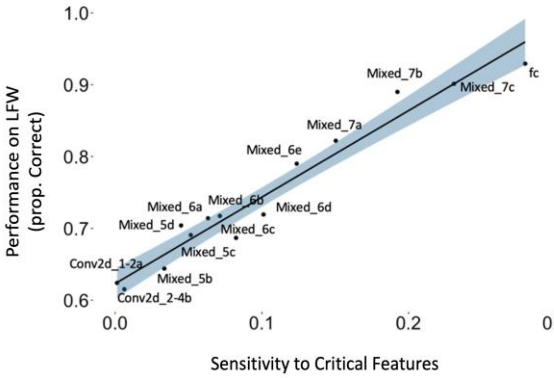

Figure 3: Performance on a benchmark face verification task (LFW) for each of the 17 layers of the network is highly correlated with sensitivity of each layer to critical features. Higher sensitivity to critical features is associated with higher performance on a face matching task, highlighting their importance for face identification. The shaded area indicates 95% confidence interval for the linear model.

increased sensitivity to critical features used by humans, is associated with improved performance of the DCNN on a benchmark face identity task (LFW).

To examine whether the pattern of findings that was found with Inception-V3, can be generalized to other DCNNs, we trained another commonly used network, VGG-16 (Simonyan & Zisserman, 2014), with the same face and object data sets, and performed the same analysis. Figure 4 shows that results of VGG-16 were very similar to Inception-V3 (see Fig. 2D), with a larger difference between Different and Same identity faces, and between critical than non-critical features, in higher layers of a face-trained DCNN than an object-trained DCNN, and a similar pattern of results in lower layers of both networks. A mixed ANOVA with Training Type (Face, Object) as between groups factor and Face Type (Same, Non-Critical, Critical, Different) as repeated measures on dissimilarity scores of 25 face pairs revealed a significant interaction between the two factors (F(18,864) = 14.49, p < .001, η^2_p = .23), indicating different patterns of results in a face and object-trained network.

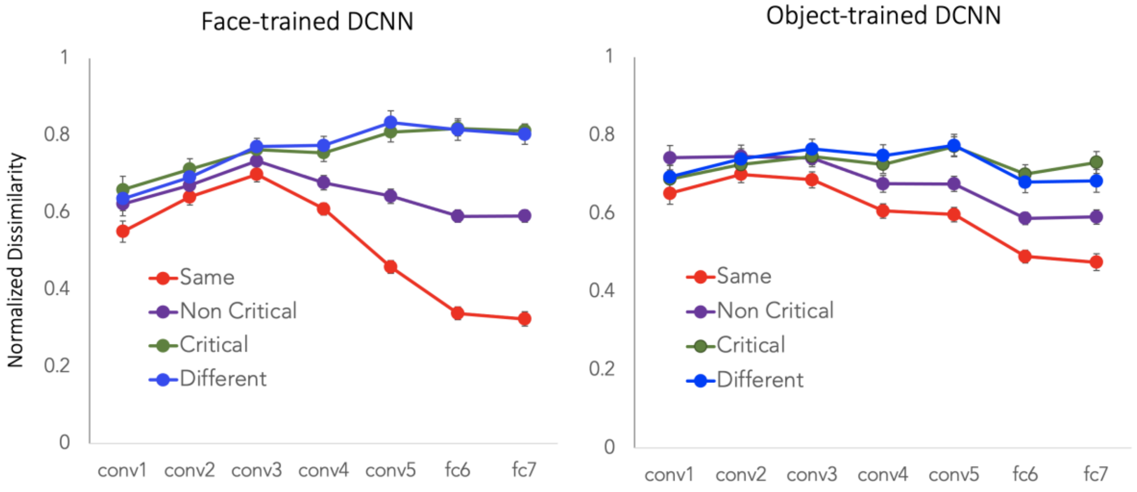

Figure 4: Results with VGG-16 trained with faces (left) or objects (right) were similar to results revealed with Inception-V3 (see Fig. 2D). Error bars indicate the standard error of the mean dissimilarity across image pairs.

5. Discussion

Results of Study 1 indicate that face-trained DCNNs are sensitive to the same critical/view-invariant features used by humans for face recognition. Importantly, the sensitivity to these facial features emerges only at the higher layers of the face-trained DCNNs. Object-trained DCNNs were significantly less sensitive to critical features in higher-layers of the network. Interestingly, the representations of faces in low-level layers were similar in the face and object-trained DCNNs. Consistent with these findings, retraining of the last two layers of an object-trained network with faces, generated similar sensitivity to critical features and performance level on a face verification task, as a DCNN that was fully trained with faces. These patterns of results are consistent with the architecture of the primate visual system, which extracts similar features from faces and objects in low-level visual regions but diverge to a face and an object system only at high-level visual processing. Thus, face recognition depends on a system that is specifically tuned to face-specific features. Indeed, the sensitivity to these view-invariant, critical facial features was strongly correlated across the DCNN layers with performance on a face verification task that requires matching faces across variable appearances (Fig. 3). Higher layers showed increased sensitivity to critical features than lower layers and also improved performance on a benchmark face verification task (LFW).

It is important to note that we do not suggest that face-trained DCNNs are specifically tuned to measure lip-thickness or eye-shape, nor do we suggest that face neurons are tuned that way. We do propose, however, that the type of information that humans and DCNNs rely on for face recognition is correlated with the critical features that we discovered. Our previous psychophysical studies indicated that critical features are useful for face identification because they remain invariant across different head-views (Abudarham & Yovel, 2016). These findings imply that the sensitivity to critical features that we reveal in higher layers of the network will correspond with the emergence of a view invariant representation of face identity. We examined this hypothesis in Study 2.

6. Study 2

Results of Study 1 show that high-performance face-recognition in a face-trained DCNN, is correlated with sensitivity to the subset of view-invariant, critical facial features used by humans to recognize faces. Face recognition depends on the generation of a view-invariant representation of faces, to enable generalization and discrimination of faces across different appearances. Therefore, the sensitivity of the face-trained but not object-trained systems to critical facial features, predicts that the face-trained DCNN will generate a view-invariant face representation, whereas the object-trained DCNN will generate a view-specific representation. The hierarchical architecture of DCNNs enables us also to examine at what stage of processing the view-invariant representation is generated. Single unit recording studies of the face areas of the macaque have shown that a view-invariant representation is generated in the anterior face area (AM), whereas posterior face areas show a view specific representation (ML) (Freiwald & Tsao, 2010). Accordingly, we expect a view-specific representation in earlier layers of the DCNN and a view-invariant representation in the higher layers. Freiwald & Tsao (2010) also showed evidence for a mirror-symmetric representation (i.e. similar response to left-right head-view faces) at intermediate stages of face processing, consistent with human fMRI findings (Axelrod & Yovel, 2012; Kietzmann, Swisher, König, & Tong, 2012).

To examine whether, and at what stage of processing, a view-invariant representation is generated in a face-trained and an object-trained network, we measured the distances between representations generated for images of the same identity in different head views, relative to pairs of different identity same view faces (Fig. 5A). A view-invariant representation is indicated by a larger distance between different identity faces in the same head view, than same identity faces presented from different views. A view-specific representation is indicated by a larger distance between same identity faces presented from different views, than different identity faces presented from the same view.

7. Methods

7.1. Stimuli:

To quantify view-invariance we used images of 15 identities from the color FERET face-image dataset (Phillips, Wechsler, Huang, & Rauss, 1998). For each identity we took four images: a frontal image, hereby referred to as the “reference” image; a second frontal image, different from the “reference” image, hereby referred to as the “frontal” image; a quarter-left image, and a half-left image. All face images were of adult Caucasian males, had adequate lighting, with no glasses, hats or facial hair. The images were cropped just below the chin to leave only the face, including the hair and ears. Four types of face pairs were generated. Same Frontal, Same quarter view, Same half view and Different Frontal (Fig. 5A).

7.2. Quantifying view-invariance of face-representations in DCNNs:

We computed the Euclidian distance between the representations of the following pairs of faces for the 15 different identities: Same identity faces- same view, Same identity faces – quarter view, Same identity faces – half view, Different identity faces – same view (see Fig. 5A). The face alignment procedure failed to detect 4 of the half-view faces, for this reason we only had 11 face pairs in the frontal – half-view condition. These distances were computed across the different layers of the face-trained, and object-trained DCNNs. The face images used in the current study enable us to also examine whether and at what stage of processing face-trained DCNNs generate a mirror-symmetric representation. We expand on this topic in supplementary material.

8. Results

8.1 A view-invariant representation in Face-trained but not Object-trained DCNNs

Figure 5 depicts the average normalized Euclidian distances between the representations generated by the face-trained and object-trained DCNNs, for pairs of images of the three same-identity conditions – same identity frontal view, same identity quarter-left view, same identity half-left view as well as the different identity-same frontal view (see Fig. 5A), in the penultimate fc layer (Fig. 5B) and across all layers (Fig. 5C). A view-invariant representation is indicated by smaller distances for same-identity pairs across different views, compared with images of different identities from the same view. A view-specific representation is indicated by similar distance scores of same and different identity faces from the same view and higher distance for same identity faces from different views.

A repeated measure ANOVA with DCNN training (Face, Object) and Head-view (Same Frontal, Same quarter left, Same half left and Different frontal) on distances between face pairs in the penultimate, fc layer of the two DCNNs, revealed a significant interaction ($F(3,63) = 40.49$, $p < .0001$, $\eta^2_p = .67$). This interaction reflects the view-invariant representation in the face-trained and view-specific representation in the object-trained networks (Fig. 5B). In the face-trained DCNN, the distances between different identity faces were the largest and significantly different from distances between same identity faces across different head views ($p < .001$, corrected for multiple comparisons, Cohen's d = 2.3 – 7.49). In contrast, for the object-trained DCNN the distances between different identity faces were significantly smaller than the distances for same identity different-view faces ($p < .005$, Cohen's d = (-0.68) – (-1.8)).

We also examined the representations across the different layers (Fig. 5C). A significant interaction between DCNN training, Head-view and Layer ($F(48,1008) = 29.15$, $p < .0001$, η^2_p = .58), indicates different representations for identity and head-view across layers in the two networks. Inspection of Fig. 5C shows that the representations in the face and object-trained DCNNs were similar at the initial layers, which showed a view-specific representation, but different for higher layers, which remained view-specific for the object-trained DCNN but became view-invariant for the face-trained DCNN. This was indicated by higher dissimilarity for same identity faces across head views than different identity, and same identity faces within the same head-view in lower layers of the face and object-trained DCNNs. In the higher layers of the DCNNs, starting at mixed-7c, we see a view invariant representation of face identity for the face-trained but not the object-trained DCNN. (See supplementary material for report of a mirror-symmetric representation).

Next we examined whether, similar to sensitivity to critical features, fine tuning of the last two layers of an object-trained DCNN on face identification, will generate a view invariant representation. Fig. 5D (right) shows that fine tuning of the last layer (FT1 - weights between layers Mixed_7c and fc) still generates a view-specific representation. A repeated measure ANOVA with Training type (Full, FT1) and Face Type (Same, quarter-left, half-left, Different) on the representation of the last layer, revealed a significant interaction ($F(3,30) = 101.23$, $p < .001$, η^2_p = .91), indicating different face representations for the a fully face-trained and the FT1 DCNN. Next, we repeated the same procedure, after fine-tuning of another layer (FT-2 - the weights between layer mixed_7b and Mixed_7c). Here, the interaction was only marginally significant ($F(3,30)$ = 3.05, $p = .04$ (uncorrected), η^2_p = .23) overall showing a view-invariant representation. As can be seen in Figure 5, the representations of a fully-trained face DCNN (Fig. 5C-middle) and FT2 (Fig. 5D, left) both generate a view-invariant representation of face identity.

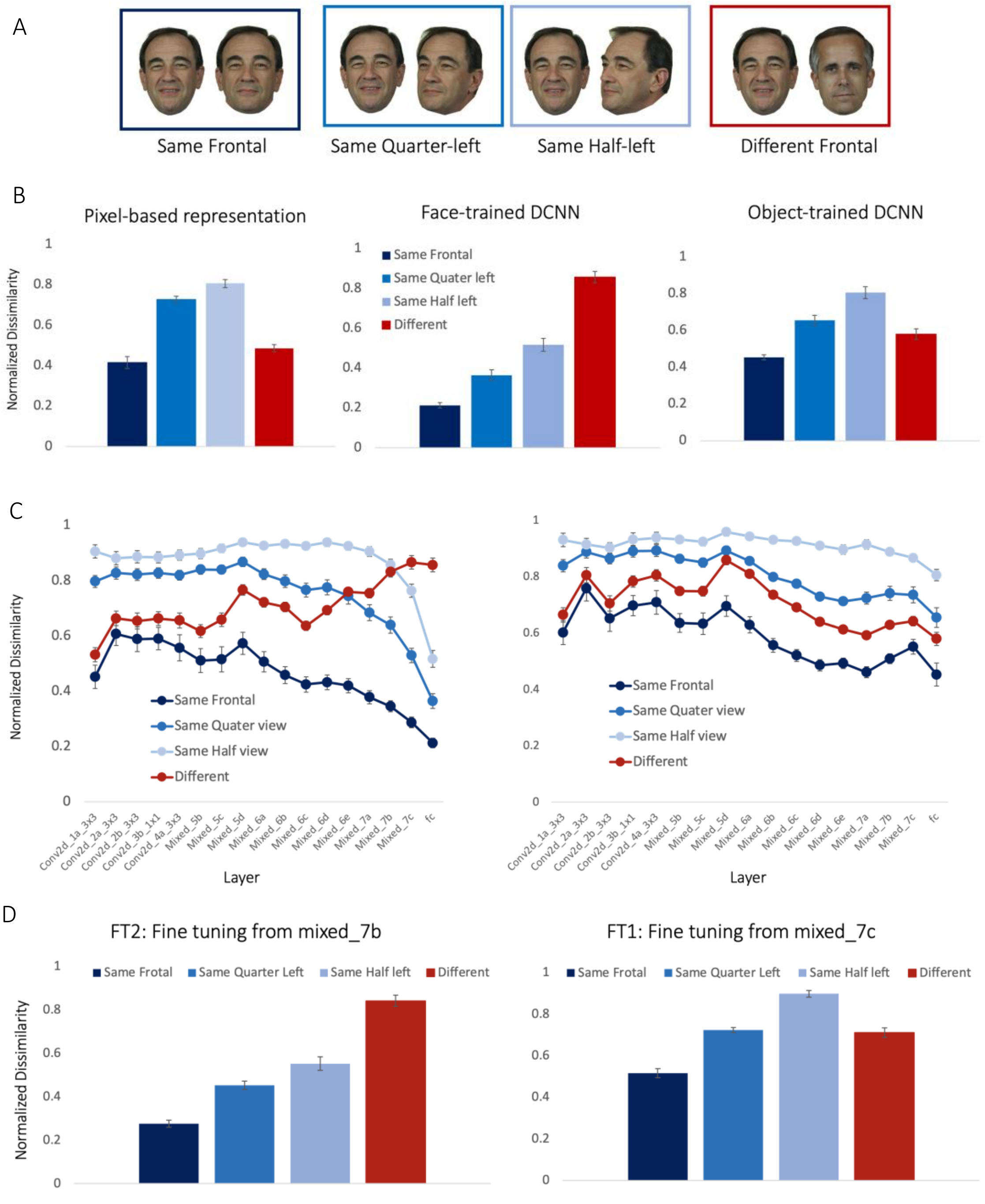

Figure 5: A. To quantify view invariance, pairs of same identity faces from the same frontal view or different views as well as same-view different identity faces were used. B. The normalized Euclidian distances between the representations of each pair were computed based on pixel-based measures (left) and for the penultimate fc layer of the face-trained (middle) and object-trained (right) DCNNs. Results show higher dissimilarity for different identity same-view faces than same identity faces across head views - indicating an identity-based, view-invariant representation in the face-trained DCNNs. The object-trained DCNN shows higher dissimilarity for same identity different-view faces, than for different identity same-view faces, indicating a view-specific, identity-independent representation. C. The representations across the layers indicate a view-specific representation in both the face and object-trained networks for low-level and mid-level layers, but a view-invariant representation only for the higher-layers of the face-trained network. D. Re-training the final layers of object-trained DCNN with faces (fine-tuning) starting from layer mixed_7b (FT2) generated a view-invariant representation that was similar to the fully-trained face DCNN, whereas training that stated from layer mixed_7c (FT1) generated a view-specific representation that was more similar to an object-trained network. Error bars indicate the standard error of the mean dissimilarity across image pairs.

To examine whether results generalize to other DCNN, we examined the representations of the four face pairs in VGG-16 that was trained with faces or objects, similar to Study 1. Figure 6 shows that results were similar to the findings we revealed with Inception-v3 (Fig. 5C). A significant interaction between DCNN training (Face, Object), Face Type and Layer ($F(18,360) = 70.65$, $p < .0001$, $\eta^2_p = .78$), indicating different representations for identity and head-view across the layers of the face and object-trained networks.

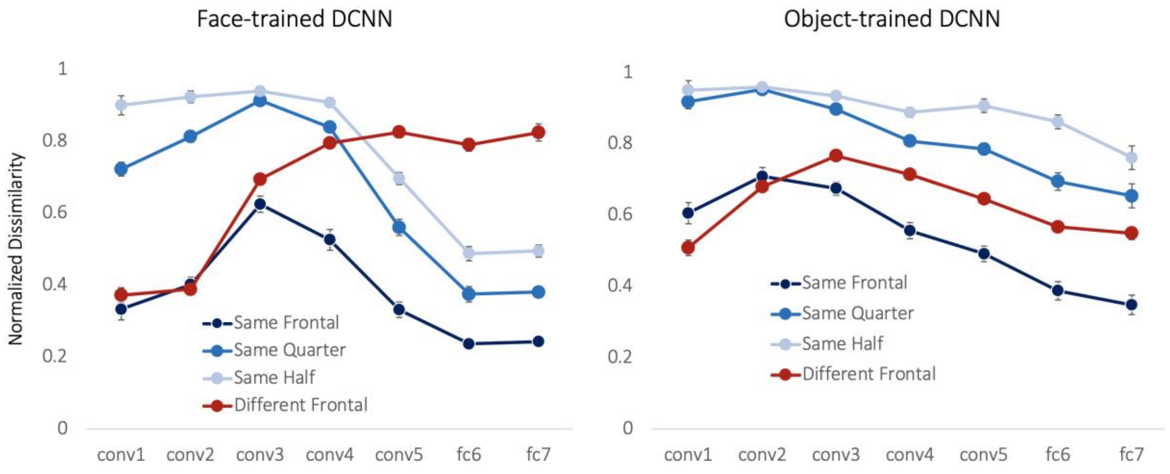

Figure 6: Results with VGG-16 that was trained with faces (left) or objects (right) were similar to results revealed with Inception-V3 (see Fig 5C). Error bars indicate the standard error of the mean dissimilarity across image pairs.

9. Discussion

Results of Study 2 show that a view-invariant representation of face identity emerges at higher levels of processing of a system that is specifically tuned to faces. The face-trained DCNN was trained on different appearances of many different identities and this way learned which features are useful for the generation of a view-invariant face representation. Importantly, a network that was trained with objects and was able to classify 1000 different categories of objects across their different appearances, did not generate a view-invariant representation of face identity. The representation of the object-trained network was view-specific and was similar across its different layers to the representation that was generated in lower-level layers of the face-trained network. Finally, consistent with the sensitivity to critical features (Fig. 1E), fine tuning the last two layers of an object-trained network for face identification, generated a view-invariant representation of face identity that was similar to the fully trained face network. These results show a correspondence between the generation of a view-invariant representation of face identity, the sensitivity to human-like critical features and performance level in face identity tasks, all emerge in high-level layers of a face-trained network.

In a recent study, Hill and colleagues (2019) examined the representation of the penultimate layer of a DCNN to face identity and head-view. Their findings show how information about both identity and head-view is preserved at the top layer of the network. These results are consistent with our findings that also show sensitivity to head view, as indicated by the differences between the same identity faces that differ in head views (the three blue bars in Fig. 5B), as well as to face identity, indicated by a larger distance between different identity and same identity faces (the red vs. the blue bars in Fig. 5B). Hills and colleagues examined only the top layer of a face-trained network, while our study shows that a view-invariant representation emerges at the top layer of a face-trained network, but not at its lower layers. We also show that a view-invariant representation does not emerge in an object-trained network that represent faces in a view-specific manner.

The emergence of a view-invariant representation at higher-level of face processing, following a view-specific representation at lower and mid-level face areas, is in line with findings reported in single unit recordings of face neurons in different face patches along the hierarchy of the face network of the macaque (Freiwald & Tsao, 2010). In particular, a view-specific representation that was not sensitive to face identity was found in the posterior face-area (area ML) whereas a view-invariant, identity-selective representation was found in the more anterior face-area (area AM). This pattern of response parallels the DCNN representations of the face-trained but not the object-trained network and indicates that the view-invariant representation depends on a system that is tuned to view-invariant, high-level facial features. One difference between the representation of head-view in DCNN and the primate face system was recently highlighted by Yildrim and colleagues (2020). Yildrim and colleagues (2020) examined the correspondence between the response of face neurons in different face areas of the macaque's brain with a DCNN (VGG-16) and an inverse graphic model (EIG). They found that an inverse graphic model, but not the DCNN, displayed a mirror-symmetric representation in its intermediate (f4) layer, before a view-invariant representation emerged, similar to the primate brain. Our findings also show that a mirror-symmetric representation and a full view-invariant representation emerged at the same layer of a DCNN (VGG-16) (Fig. S2), unlike the hierarchy in the primate face system. Yildrim and colleagues have also shown that EIG was better correlated with human performance than VGG, in two tasks which require 3D information: matching of faces with no texture information or with fish-eye style shape deformation, and the "hollow-face illusion". It will be interesting to test whether an EIG network, which explicitly codes 3D shape as well as texture information, codes the changes in the critical features described here, which humans use for face recognition.

10. General Discussion

The question of how humans recognize faces has been extensively studied in the past half-century (O'Toole et al., 2018; Young & Bruce, 2011). To answer this question, we need to unravel the nature of the representation that enables face identification in face images that vary greatly in illuminations, expressions and head-views. The same quest has also occupied computer scientists, who have aimed to generate algorithms that resolves this task (Taigman et al., 2014). Despite their recent success in reaching human-level performance, DCNNs have not provided us with an interpretable solution to this task. Here we combined our understanding of human face recognition, and the brain-inspired architecture of DCNNs, to unravel the nature of the representation that enables face recognition. Our findings show that both systems reach a similar solution. First, we found that DCNNs that were optimized for face recognition, without any explicit training of the features used by humans for face recognition, developed sensitivity to these features (Fig. 2B). Second, the sensitivity to these features emerged gradually in higher layers of the network (Fig. 2D) and was highly correlated with performance of each layer on a benchmark face verification task (Fig. 3), further stressing the importance of these features for face identification. Third, a system that was optimized for object recognition was less sensitive to these view-invariant facial features across all its layers (Fig. 2B, D) and did not perform well on the benchmark face identity task. Importantly, fine-tuning the last two layers of the object-trained network with faces, generated a representation that was similar to a fully trained face network (Fig. 2E). Similarly, a view-invariant representation of face identity was found only in higher layers of the face-trained network, whereas low-level layers and all layers of the object-trained network generated a view-specific face representation (Fig. 5B). Fine tuning the last two layers of the object trained network on faces, generated again a similar representation as the fully trained face network (Fig. 5E). These results parallel the division to a face and an object system at high-level visual cortex, where lower visual areas represent faces and objects similarly and diverge to separate systems only in high-level visual areas. Taken together these findings indicate that sensitivity to view-invariant facial features, that are critical for human face recognition, and a view-invariant representation of face identity, emerge at higher levels of processing of a system that is optimized for face identification.

Our findings are in line with a recent study that used a DCNN (VGG-16) to model human familiar face recognition (Blauch, Behrmann, & Plaut, 2020; see also Yovel & Abudarham, 2020). It is well established that human face identification is better for familiar than unfamiliar faces (Jenkins, White, Van Montfort, & Mike Burton, 2011; Ritchie et al., 2015). Blauch et al (2020) showed that this gap in performance is not mediated by the perceptual representation that is generated in the penultimate layer of the DCNN, but by the output, identification layer that is specifically tuned to the familiar, trained identities. Our findings that the penultimate layer of a DCNN is sensitive to critical features complement these results. We have recently shown that humans use the same critical features both for familiar and unfamiliar faces (Abudarham et al., 2019). Accordingly, we suggest that the penultimate perceptual layer of DCNNs extracts critical features from both familiar and unfamiliar faces. This representation is then used by the output layer to classify face images to different familiar identities. This classification operation provides an additional advantage, relative to performance that is merely based on perceptual distances between face images.

Our findings highlight the importance of specific training with faces for the generation of view-invariant facial features. Such training enables the system to learn which features are both invariant across different appearances of the same identity and are also useful for discrimination between identities. The features that we tested here are based on results of our previous studies that used faces of adult male Caucasian faces, and may not generalize to faces of other races. For example, hair and eye color, which are both invariant and discriminative for Caucasian faces, are invariant in Asian and African faces but may not be discriminative for these races. Indeed, it is well established that humans show poor recognition for races for which they have low experience with, an effect known as the Other Race Effect (Rhodes, Locke, Ewing, & Evangelista, 2009). Similarly, DCNNs were shown to be biased for the races that are included in their training set. State of the art and commercial algorithms that were developed in western countries show much lower performance for African and Asian faces than Caucasian faces (Phillips, Jiang, Narvekar, Ayyad, & O'Toole, 2011; Wang, Deng, Hu, Tao, & Huang, 2018). Thus, both human and DCNN representations indicate that the features that the face system is tuned to may not be selective merely to faces, but to facial features that are useful for the specific category of faces we have experience with. This further highlights the degree of specificity in visual experience that is required for intact face recognition.

An important difference between the face and object-trained DCNNs is that the object DCNN is trained to classify among many different categories, but not within different exemplars of the same category, whereas the face-trained DCNN is trained to classify different exemplars within the category of faces. The goal of our study was to compare an object-general system, similar to the lateral occipital complex (Malach et al., 1995), to a face-selective system, similar to the FFA (Kanwisher & Yovel, 2006). A comparison between the representations that are generated for faces and a specific category of objects was beyond the scope of the current study but is worthwhile pursuing in future studies. Such an investigation may test the *expertise hypothesis* (e.g., Gauthier, Skudlarski, Gore, & Anderson, 2000; Tarr & Gauthier, 2000) by comparing the performance and the representations generated by the a face-trained DCNN and DCNNs optimized for the recognition of specific object categories (Dailey, Cottrell, & Padgett, 1997).

Finally, DCNNs have been criticized for being "black box" machines that are based on millions of parameters, and therefore reverse engineering their underlying feature representation is a great challenge (Cichy & Kaiser, 2019; Marcus, 2018). Here we show that insights from reverse engineering of the human mind, and the discovery of features that are used by humans, can shed light on the type of information used by DCNNs to accomplish their human-level performance. This similarity between the perceptual representations of humans and DCNNs is not trivial, given the many differences between the architecture and computations performed by the human brain and a feed-forward DCNN (Marcus, 2018). However, our findings indicate that feed-forward DCNNs are sensitive to the same features used by humans, and can be therefore used to test predictions on human visual processing.

In summary, with recent advances in artificial intelligence, humans and machines are now performing tasks of similar complexity. The discovery that they generate similar representations and reach a similar solution can advance our understanding of both the biological and the artificial systems (Ma & Peters, 2020). The approach we used here for the study of face recognition can be similarity applied to other cognitive tasks to improve our understanding of human cognition and the interpretability of artificial neural networks.

Acknowledgment

This work was supported by Joint China-Israel, NSFC-ISF Research Grant Application no. 2383/18.

11. References:

Abudarham, N., & Yovel, G. (2019). Same critical features are used for identification of familiarized and unfamiliar faces. *Vision Research*, *157*.

https://doi.org/10.1016/j.visres.2018.01.002

Abudarham, Naphtali, Shkiller, L., & Yovel, G. (2019). Critical features for face recognition. *Cognition*, *182*, 73–83. https://doi.org/10.1016/j.cognition.2018.09.002

Abudarham, Naphtali, & Yovel, G. (2016). Reverse engineering the face space: Discovering the critical features for face identification. *Journal of Vision*, *16*(3). https://doi.org/10.1167/16.3.40

Axelrod, V., & Yovel, G. (2012). Hierarchical processing of face viewpoint in human visual cortex. *Journal of Neuroscience*, *32*(7). https://doi.org/10.1523/JNEUROSCI.4770-11.2012

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2020). Deep learning of shared perceptual representations for familiar and unfamiliar faces : Reply to commentaries. *Cognition*, (August), 104484. https://doi.org/10.1016/j.cognition.2020.104484

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/J.TICS.2019.01.009

Dailey, M. N., Cottrell, G. W., & Padgett, C. (1997). A mixture of experts model exhibiting prosopagnosia. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*.

Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*. https://doi.org/10.1126/science.1194908

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature Neuroscience*. https://doi.org/10.1038/72140

Grossman, S., Gaziv, G., Yeagle, E. M., Harel, M., Mégevand, P., Groppe, D. M., … Malach, R. (2019). Convergent evolution of face spaces across human face-selective neuronal groups and deep convolutional networks. *Nature Communications*. https://doi.org/10.1038/s41467-019-12623-6

Jenkins, R., White, D., Van Montfort, X., & Mike Burton, A. (2011). Variability in photos of the same face. *Cognition*. https://doi.org/10.1016/j.cognition.2011.08.001

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476). https://doi.org/10.1098/rstb.2006.1934

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1003915

Kietzmann, T. C., Swisher, J. D., König, P., & Tong, F. (2012). Prevalence of selectivity for mirror-symmetric views of faces in the ventral and dorsal visual pathways. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.0126-12.2012

Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004896

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: towards using deep nets as models for human behavior*. 1–39. Retrieved from http://arxiv.org/abs/2005.02181

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., … Tootell, R. B. H. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.92.18.8135

Marcus, G. (2018). *Deep Learning: A Critical Appraisal*. 1–27. Retrieved from http://arxiv.org/abs/1801.00631

O'Toole, A. J., Castillo, C. D., Parde, C. J., Hill, M. Q., & Chellappa, R. (2018). Face Space Representations in Deep Convolutional Neural Networks. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2018.06.006

Phillips, P. J., Jiang, F., Narvekar, A., Ayyad, J., & O'Toole, A. J. (2011). An other-race effect for face recognition algorithms. *ACM Transactions on Applied Perception*. https://doi.org/10.1145/1870076.1870082

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.

Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., … O'Toole, A. J. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.1721355115

Rhodes, G., Locke, V., Ewing, L., & Evangelista, E. (2009). Race Coding and the Other-Race Effect in Face Recognition. *Perception*, *38*(2), 232–241. https://doi.org/10.1068/p6110

Ritchie, K. L., Smith, F. G., Jenkins, R., Bindemann, M., White, D., & Burton, A. M. (2015). Viewers base estimates of face matching accuracy on their own familiarity: Explaining the photo-ID paradox. *Cognition*, *141*, 161–169. https://doi.org/10.1016/j.cognition.2015.05.002

Simonyan, K., & Zisserman, A. (2014). VGG-16. *ArXiv Preprint*. https://doi.org/10.1016/j.infsof.2008.09.005

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2014.220

Tarr, M. J., & Gauthier, I. (2000). FFA: A flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience*. https://doi.org/10.1038/77666

Valentine, T. (2001). Face-space models of face recognition. *Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges*, 83–113.

Voosen, P. (2017). How AI detectives are cracking open the black box of deep learning. *Science*. https://doi.org/10.1126/science.aan7059

Wang, M., Deng, W., Hu, J., Tao, X., & Huang, Y. (2018). *Racial Faces in-the-Wild: Reducing Racial Bias by Information Maximization Adaptation Network*.

Xiang, J., & Zhu, G. (2017). Joint face detection and facial expression recognition with MTCNN. *Proceedings - 2017 4th International Conference on Information Science and Control Engineering, ICISCE 2017*. https://doi.org/10.1109/ICISCE.2017.95

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*. https://doi.org/10.1038/nn.4244

Young, A. W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*(4), 959–974. https://doi.org/10.1111/j.2044-8295.2011.02045.x

Yovel, G., & Abudarham, N. (2020). From Concepts to Percepts in Human and Machine Face Recgonition: A reply to Blauch, Behrmann & Plaut. *Cognition*, (July), 104424.

https://doi.org/10.1016/j.cognition.2020.104424