Facial expression is retained in deep networks trained for face identification

Y. Ivette Colón Behavioral and Brain Sciences, The University of Texas at Dallas, TX, USA

Carlos D. Castillo University of Maryland Institute for Advanced Computer Studies, MD, USA

Alice J. O'Toole Behavioral and Brain Sciences, The University of Texas at Dallas, TX, USA

Facial expressions distort visual cues for identification in two-dimensional images. Face processing systems in the brain must decouple image-based information from multiple sources to operate in the social world. Deep convolutional neural networks (DCNN) trained for face identification retain identity-irrelevant, image-based information (e.g., viewpoint). We asked whether a DCNN trained for identity also retains expression information that generalizes over viewpoint change. DCNN representations were generated for a controlled dataset containing images of 70 actors posing 7 facial expressions (happy, sad, angry, surprised, fearful, disgusted, neutral), from 5 viewpoints (frontal, 90° and 45° left and right profiles). Two-dimensional visualizations of the DCNN representations revealed hierarchical groupings by identity, followed by viewpoint, and then by facial expression. Linear discriminant analysis of full-dimensional representations predicted expressions accurately, mean 76.8% correct for happiness, followed by surprise, disgust, anger, neutral, sad, and fearful at 42.0%; chance ≈14.3%. Expression classification was stable across viewpoints. Representational similarity heatmaps indicated that image similarities within identities varied more by viewpoint than by expression. We conclude that an identity-trained, deep network retains shape-deformable information about expression and viewpoint, along with identity, in a unified form—consistent with a recent hypothesis for ventral visual stream processing.

Introduction

Historically, the neural processing of facial expression and identity were considered wholly separate. This conclusion was based on early neuropsychological case studies of patients who could recognize faces, but not facial expressions (e.g., Kurucz & Feldmar, 1979), and conversely, on patients who could recognize facial expressions, but not identify faces (e.g., Bruyer et al., 1983). This double dissociation led researchers to conclude that facial expression and face identity are perceived and represented independently. Accordingly, in Bruce and Young's 1986 model (Bruce & Young, 1986), view-dependent descriptions of faces feed into specialized processes for the analysis of facial speech, expression, and identification.

With the benefit of additional evidence from functional neuroimaging, Haxby, Hoffman, and Gobbini (2000) proposed a distributed model of face processing. In this model, invariant aspects of a face (e.g., identity) are processed in the fusiform face area (FFA) in the ventral visual stream, whereas changeable aspects of a face (e.g., expression, gaze, viewpoint/pose) are processed in the posterior superior temporal sulcus, in the dorsal visual stream. This separated processing of facial identity from expression is functionally consistent with earlier work (Bruce & Young, 1986).

Since then, the assumption that facial expression and identity processing are carried out independently has been reassessed. Calder and Young (2001), for example, showed that expression and identity can be modeled in a unified visual representation, with only partial dissociation. Their model was based on a principal component analysis (PCA) of face images that varied in identity and expression. PCA, applied to these images, generated PCs that coded identity, expression, or both. In the 1990s, image-based PCA was considered a highly effective computational model for face recognition and was used widely in commercial face recognition systems. However, image-based PCA operates in a viewpoint- and illumination-dependent way. It is, therefore, limited as a model of human face perception.

More recently, Duchaine and Yovel (2015) proposed modifications of the Haxby model (Haxby et al., 2000)

Citation: Colón, Y. I., Castillo, C. D., & O'Toole, A. J. (2021). Facial expression is retained in deep networks trained for face identification. *Journal of Vision*, *21*(4):4, 1–10, https://doi.org/10.1167/jov.21.4.4.

https://doi.org/10.1167/jov.21.4.4 Received May 7, 2020; published April 6, 2021 ISSN 1534-7362

to suggest that the ventral stream processes shape and form information (including identity and some aspects of expression), whereas the dorsal stream processes dynamic information from faces. This revision is based on findings suggesting that the ventral stream's FFA may contribute to the perception of changeable, as well as invariant, aspects of a face—for example, changes in the "shape" of a face that might be due to facial expressions.

Evidence for the influence of shape information in the FFA comes from studies measuring neural activity in response to facial expression (Fox et al., 2009; Ganel et al., 2005). Their findings indicate that FFA responses differentiate for distinct facial expressions. Further, neural activity in the FFA increases along with the intensity of facial expressions (Surguladze et al., 2003). This may be caused by changes to the shape of a face as increase expressions become more intense. According to Duchaine and Yovel's (2015) proposal, expression changes cause shape changes that are processed in the FFA.

The reevaluation of the two-stream hypothesis raised the possibility that the ventral stream contributes more to the perception of facial expression than previously thought. Moreover, experimental results have suggested that the neural representations of identity and expression might be linked in ventral stream processing—potentially mediated by the viewpoint of the face. Expression and viewpoint perception are considered dorsal stream processes because both expression and viewpoint *change* the shape of a face, though in different ways. Expression deforms the face itself, whereas viewpoint changes the shape of the two-dimensional projection onto the retina, which is experienced by a viewer.

Although Calder and Young (2001) showed that image-based PCA can encompass expression and identity information in a unified representation, this representation cannot support recognition over changes in viewpoint. By contrast, current DCNN-based computational models of face recognition support face recognition across changes in viewpoint. The face representations generated from these networks might be used to address the question of whether identity and expression information can co-exist in a unified and generalizable representation of the face. DCNNs operate using cascaded convolutional layers, with millions of non-linear computations. These layers first expand the representation, and then condense it into a compact face descriptor. The representation that emerges from a DCNN trained for face identity retains both invariant (identity, gender) and changeable (viewpoint, illumination) aspects of faces (Hill et al., 2019; O'Toole et al., 2018; Parde et al., 2017).

Given that the deep convolutional neural network (DCNN) codes retain aspects of image information, examining whether expression information is likewise retained in these codes is pertinent. To date, the representation of expression in facial identification DCNNs has not been explored. Moreover, DCNNs have an added benefit of identifying faces accurately over image variation (e.g., viewpoint). This advantage allows us to probe expression representations over changes in viewpoint in the context of an identity-trained network. We expect that a DCNN will retain information about facial expression for two reasons. First, research has shown that, in addition to identity information, DCNN face representations also retain data about subject characteristics, including sex (Hill et al., 2019) and social traits (Song et al., 2017, Parde et al., 2019), as well as image characteristics, such as illumination and viewpoint (Parde et al., 2017). Second, in a visual sense, facial expressions are not specific to individuals. This makes expression a source of face variation that the identity-trained network has to manage, analogous to viewpoint or illumination.

To explore the representation of expression and viewpoint in a neural network trained for identity, we analyzed the network's output. The output was defined as the unit responses at the penultimate layer of the network. We will refer to this output as the "face-image representation." This representation is analogous to an ensemble of neural unit responses for identity. We analyze these top-layer unit responses in three steps.

First, we visualized the expression and view information present in a *face space* (Valentine, 1991). If face images that vary on particular dimensions (e.g., expression) are represented more similarly (i.e., closer together in the space defined by the full-dimensional descriptor vectors), it will be reflected by their proximity in a two-dimensional face space. Second, because face space visualization only provides information about the most salient elements of the space, next we considered a method for visualizing the full-dimensional face space. We used a representation similarity map to evaluate the influence of particular expression and viewpoint classes on the face-image representation (Kriegeskorte et al., 2008). If expression information is found reliably in the identity representation, it is worth exploring whether, and how, this expression representation is mediated by the viewpoint of a face in an image.

Third, we used linear classifiers, applied to the face-image representations in the high-dimensional space, to predict seven facial expressions from viewpoint-variable face images. This method provides the strongest and most quantitatively grounded estimate of how well expression is represented in the network-generated face-image descriptor. Above-chance expression classification would indicate that DCNN-generated face-image representations retain accurate information about expression. Generalization of expression classification over viewpoint would indicate that the

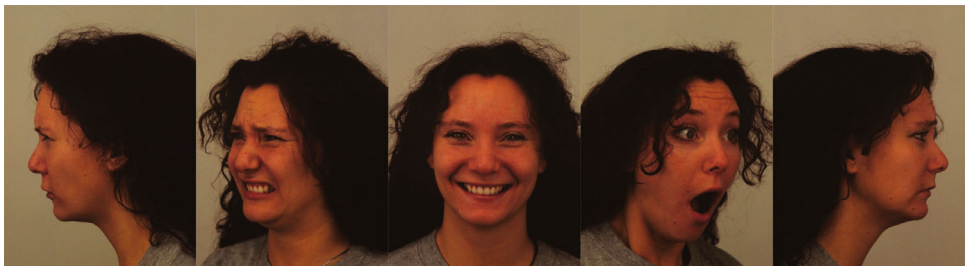

Figure 1. An example of image variation for one identity in the KDEF dataset. Image IDs, from left to right: F02ANFL, F02DIHL, F02HAS, F02SUHR, F02SAFR.

network retains this information in viewpoint-variable images.

As we shall see, a DCNN trained for face identification retains expression information. The quality of expression information varies across expressions in a manner comparable to human expression classification results (Matsumoto & Hwang, 2019). Expression information also generalizes over viewpoint.

General methods

Dataset

Our stimuli consisted of 4,060 images of 58 actors from the Karolinska Directed Emotional Faces (KDEF) (Lundqvist et al., 1998) database. Each actor posed seven unique expressions (happiness, fear, anger, sadness, surprise, disgust, and neutral), which were photographed over five viewpoints (frontal, 90° left and right, 45° left and right). This resulted in 35 images per identity. All viewpoint and expression conditions were captured in two sessions for a total of 70 images per identity. Figure 1 shows an example of the views and expressions. All available images of the following identities from KDEF were used for analyses: F02 to F09, F12, F14 to F19, F21 to F29, F32 to F35, M01 to M20, M23, M26 to M33, and M35.

DCNN specifications

All simulations were conducted on a ResNet101-based DCNN (Ranjan et al., 2017). The network was trained on the Universe dataset, which is a mixture of three datasets (UMDFaces Bansal et al. 2017; UMDVideos Bansal et al., 2017; and MS1M Guo et al., 2016). The dataset includes in-the-wild images and video frames that vary widely in imaging condition (pose, illumination, etc.). In total, 5,714,444 images were used in training. A Crystal Loss (L2 Softmax) normalization function was used in the network (Ranjan et al., 2018).

DCNN face representation

The KDEF images were processed by the DCNN to produce a 512-dimensional "top-layer" feature vector for each image. All analyses were conducted on these top-level representations.

Visualizing expression and viewpoint in the face space

The goal of this visualization was to get a first look at the structure of a face space that contains information about identity, expression, and viewpoint. To that end, we visualized the face space to highlight how these face attributes are organized in the space.

Procedure

Unit-normalized face-image representations of all images in the dataset were visualized using t-Distributed Stochastic Neighbor Embedding (t-SNE; Maaten & Hinton, 2008). This dimensionality reduction technique projects the 512-dimensional feature vectors into a lower-dimensional space, while preserving the relative distances between points in the full dimensionality. The output of the t-SNE was then plotted to visualize the face space created by the face-image representations. We used a Barnes-Hut implementation of t-SNE with perplexity set to 30, following the advice of Hill et al. (2019).

Results

Figure 2 shows a t-SNE visualization of the DCNN face-image representations, color-coded by

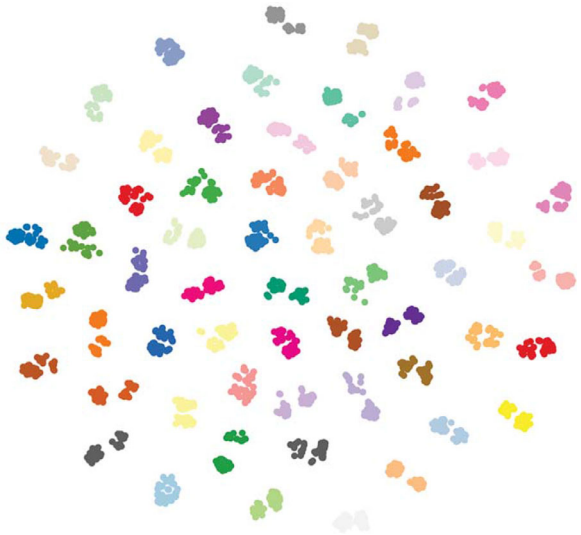

Figure 2. A visualization of the two-dimensional t-Distributed Stochastic Neighbor Embedding (t-SNE) projections of image representations for the KDEF dataset (color-coded by identity) shows that identities are well-separated by the network. Note: because there were more identities than colors, some colors were used for two identities.

identity.[1] The DCNN clusters face images by identity (color-constant groups) accurately, despite viewpoint and expression variability. Within each identity cluster, Figure 2 also shows two groups of images, which we consider next.

To examine the organization of expression and viewpoint within an identity cluster, we zoomed-in on individual identities. Two typical example identity clusters are depicted in Figure 3A and B. For each of these identity clusters, the space cleanly divides face images into two viewpoint groups: near-frontal images, which include frontal and 45° profile images, and 90° profile images. This bifurcation can be seen in Figure 2 as the cleaving of images within an identity cluster into two subclusters. The figure shows consistent bifurcation across all identities in the dataset.

Combined, the resulting structure of clusters in the two-dimensional projection reveals three things. First, the face images in the space clearly cluster by identity. Second, within individual identity clusters, near-frontal images are separated from profile images. Third, within the viewpoint bifurcation groups, images loosely cluster by facial expression (Figure 3).

These two-dimensional projections offer a first look at the structure of the face space created by the DCNN. Next, we turn our attention to a visualization analysis of full-dimensional representation to examine the organization of expression and viewpoint variation in the similarity space.

Organized representational similarity map

We examined single identities using representational similarity heatmaps, organized by expression and viewpoint (Kriegeskorte et al., 2008). If the two-dimensional visualizations (Figure 3) accurately reflect the image representations in the full-dimensional space, we would expect the following. The most similar images of an identity would be those taken from near-frontal viewpoints. Differences between discrete expressions would be smaller, relative to differences between viewpoint classes. The heatmaps would also indicate whether this principle of representational similarity (i.e., viewpoint differences more dissimilar than expression differences) operates consistently across expressions.

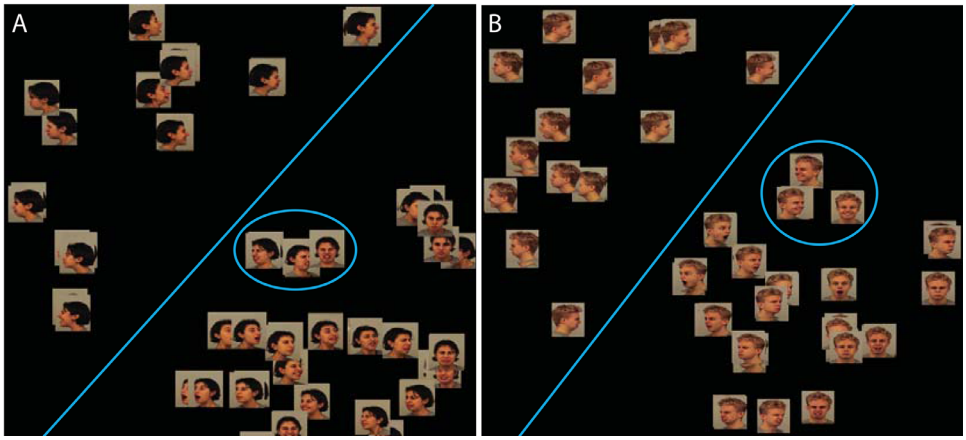

Figure 3. Two example identities in the t-SNE projection. Each panel (A and B) shows one identity. A hand-drawn blue line shows that the near-frontal images can be separated from profile images of the identity in the face space. Circles illustrate an example of expression clustering within viewpoint groups.

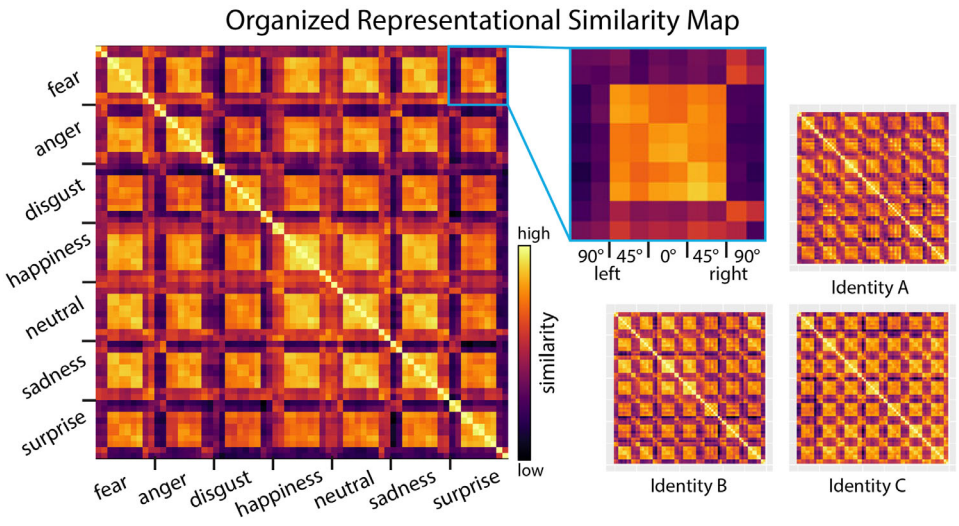

Figure 4. Representational similarity maps comparing representations of 70 images of 4 randomly selected, individual identities. Heatmaps were organized first by expression, then by viewpoint within each expression. The pattern of similarity indicates that, for all identities, and all expressions, representations of images in near-frontal viewpoint groups are represented more similarly than full-profile viewpoint images.

Procedure

We computed the similarity of all possible pairs of full-dimensional representations of images for all identities in the KDEF dataset. Similarity was defined as the cosine between representation vectors. Because the DCNN produces a normalized unit-length output, cosine was an appropriate distance measure. Each heatmap is made from the DCNN's representations of 70 images of a single identity (7 expressions taken from 5 viewpoints in 2 sessions). The heatmaps are organized first by expression, then by viewpoint condition within each expression.

Results

Figure 4 shows the similarity heatmap for images of four example identities. The heatmaps confirm the organization seen in the t-SNE face space. First, representations of individual identities remain consistent across variations in expression. The example identities are typical of all identities in the KDEF dataset.

Second, viewpoint variation dominates both within and across expression classes, as evidenced by the "ring" of dissimilarity in the comparison of extreme profiles to near profiles in every expression class.

Predicting facial expressions from identity representations

The primacy of viewpoint in organizing images in the DCNN space makes it difficult to interpret the structure of expression within individual identities, and across the entire KDEF dataset. Therefore, we used a pattern classification approach to determine the quality of expression information retained in the face-image representation.

Procedure

Expression predictions were made by applying linear discriminant analysis (LDA) on the 512-dimensional face-image representations. Specifically, linear classifiers were trained to identify seven expressions across all five viewpoints, using the face-image representation. Separate linear discriminant analysis classifiers were trained for each of the five viewpoints, holding out the other viewpoints. The linear discriminant analyses were tested on all expressions and remaining viewpoints. In each case, classification was cross-validated by identity, whereby a classifier was trained on images of all-but-one identity, and was tested on the held-out identity. This process was repeated for each viewpoint classifier, holding out one identity at a time, until all identities were tested. Expression classification was evaluated as percent correct (chance performance ≈ 14.3%).

Results

All expressions in the dataset were classified at levels well above chance (Figure 5). These computational results mimic the most consistently found features of human expression recognition. Specifically, happiness has been found to be the most recognizable expression

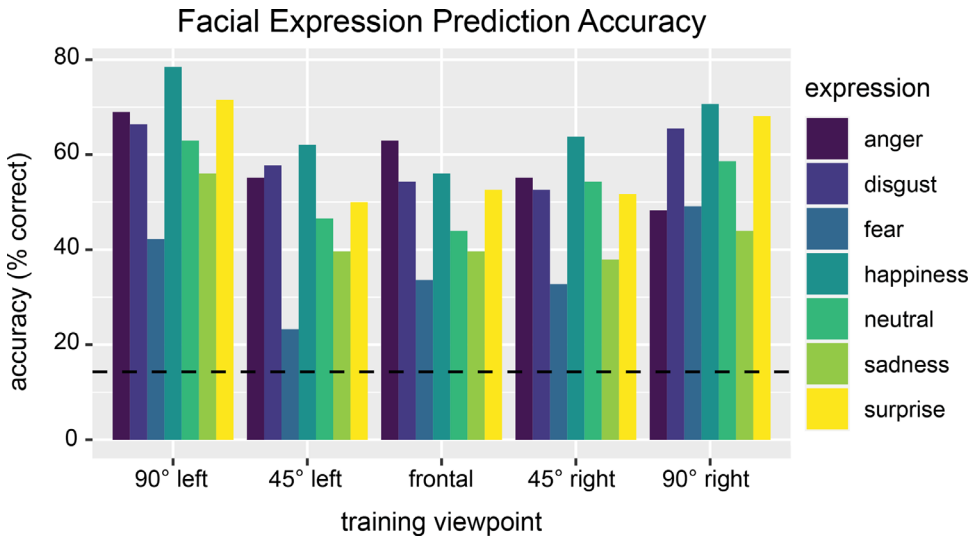

Figure 5. Expression classification results for the KDEF dataset using deep features shown by viewpoint and expression. All expressions are classified above chance. Note that chance performance, indicated by the dashed line on the figure, is approximately 14.3%.

and fear is often found to be the least recognizable expression Mancini et al., 2018; Matsumoto & Hwang, 2019; Wells et al., 2016. The mean standard error of predictions across expressions was 0.031.

Although human facial expression perception has been studied extensively over the last few decades, surprisingly little is known about the robustness of expression perception over changes in viewpoint. Here, we examined the ability of the DCNN to generalize expression perception over viewpoint change. Figure 5 shows that expression classification accuracy was roughly equivalent across viewpoint variation. The mean standard error of predictions across viewpoints was 0.048, indicating that there is no real difference between classification performance with any particular viewpoint class. Note again that all classification conditions were above chance.

Discussion

This study offers a proof-of-principle that expression, similar to other identity-irrelevant face information (illumination, viewpoint), can be coded in a single representation of a face image computed by neural-like operations in a deep network (Hill et al., 2019; Parde et al., 2017). The presence of this shape-deformable, non-identity information in a deep network trained for identification replicates, and expands on, two findings from previous studies. First, it shows that a high-level visual system, based on nonlinear combinations of low-level visual features, can co-represent different types of information in a single set of neural-like units (Parde et al., 2020). Second, it does so in a way that reinforces a hierarchical relationship between low-level visual information (viewpoint) and the categorical representation (identity) (Hong et al., 2016, 2019).

The DCNN's representation of identity is highly and consistently accurate. As with identity, the representation of expression in the DCNN is likewise robust to image-dependent viewpoint change (Parde et al., 2017). This combined representation of identity and expression emerges in artificial neurons. The question at the core of this work is whether the DCNN representation is relevant for understanding ventral visual stream processing. There are two arguments for why the DCNN might be relevant. First, as noted by Duchaine and Yovel (2015), there is evidence in the neuroscience literature for the co-existence of identity and expression information in the FFA in the ventral visual stream. More generally, the proposal of Duchaine and Yovel is based on the idea that the FFA supports aspects of dynamic face processing, because of its general role in processing face shape information. This shape information is relevant for accurate classification and prediction of both facial expression and viewpoint. The DCNN is able to encode identity, while maintaining and effectively managing information about facial expression and viewpoint. This is an advantage over image-based PCA, which is able to encode both identity and expression, but cannot operate over changes in viewpoint. Therefore, the unique representation that emerges from the DCNN allows us to simultaneously examine categorical face information (identity), as well as shape information (e.g., expression, viewpoint) in one set of neural-like units. This is not to say that the results presented here offer a direct analogy of visual processing in the ventral stream, but rather to suggest that because this category-trained system is able to simultaneously encode category-irrelevant information in its representation—as evidence has suggested for

ventral stream processing (Hong et al., 2016; Lindsay, 2020)—it is a hypothesis-grounded model of high-level visual processing in face areas of the ventral stream.

The second argument is that the pattern of facial expression classification in the DCNN mirrors human results. For both humans and the network, happiness is classified most accurately, and fear is classified the least accurately (Mancini et al., 2018; Matsumoto & Hwang, 2019; Wells et al., 2016), with other expressions in between. Expression classification accuracy for this DCNN is significantly lower in comparison to human accuracy (Hess et al., 2007; Matsumoto & Hwang, 2019). However, the pattern of expression classification matches that produced by a single, identity-driven representation of a face image. This accord between DCNNs and humans is surprising, because facial expressions are biologically important sources of information, and the network operates outside of this context using only visual information. The artificial ventral visual stream modeled by the DCNN retains accurate expression information across viewpoint changes, suggesting that perhaps some aspect of expression perception can be attributed to visual features of the input alone. Indeed, face images provide visual features useful for the perception of both identity and expression, and the availability of this information can be used by the ventral visual stream (and appropriate models) even beyond the identification and expression classification tasks. This does not imply that facial identity and expression perception rely solely on ventral stream processing. The DCNN merely demonstrates that a ventral-like identification system, operating alone, can support facial expression perception across realistic changes in viewing conditions, and can do so in a way that mirrors the pattern found in human expression perception.

Despite the importance of understanding "in-the-wild" expression perception, remarkably little is actually known about expression classification performance over changes in viewpoint. Studies examining this in humans have been sparse and have yielded contradictory results (Hess et al., 2007; Matsumoto & Hwang, 2019). In the DCNN, classification accuracy of facial identity and expression is remarkably stable across viewpoint—despite extreme viewpoint variation substantially altering an image of a face (e.g., visible surface area, availability of facial features, and expression cues). That does not mean, however, that information about viewpoint is lost in the representation. As noted elsewhere, DCNNs trained for identity retain several types of identity-irrelevant information in their final representations, including viewpoint (Hill et al., 2019; Parde et al., 2017, 2019).

In this study, we show that expression and viewpoint are preserved in a DCNN-generated representation of a face image. As demonstrated in the t-SNE visualizations, they are commingled and hierarchically organized in the representational space created by the DCNN. The representational similarity within images of single identities suggests that the underlying 'hierarchy' is based on the degree of alteration of the two-dimensional image by each type of image variation. Extreme viewpoint changes alter the DCNN representation more than expression changes. Notwithstanding, identity classification is not substantially affected by these two-dimensional deformations, because identity remains the most salient grouping principle in the representation (Hill et al., 2019; Parde et al., 2017). As such, there are no computational consequences for preserving viewpoint and expression information in the representation.

Why might this hierarchy of identity-irrelevant, shape-related information occur in an identity representation? In a system trained for identification, the additional image-based information gleaned from expression and viewpoint might offer shape cues that are relevant for identification, albeit in different ways. For expression, identity-specific facial features can be accentuated or exaggerated by nonrigid facial deformations needed to produce an expression (e.g., a distinctive smile) (O'Toole et al., 2002; Yovel & O'Toole, 2016). For viewpoint, the two-dimensional shapes projected onto the retina change with viewpoints. This extra information has the potential to reveal new cues to face identity (e.g., a nose becoming more identifiable in profile). Viewpoint does not alter the intrinsic shape of the face, but rather reveals intrinsic face shape information as a result of specific image conditions. The network may be able to leverage the shape information available in an image to achieve a more robust identity representation. The network manages shape information from all of its sources, perhaps because it is a basic component of the visual stimulus that provides a richer context from which to extract useful identity cues.

A system designed for identity classification clearly does not compare with the complexity of processing in the ventral visual stream. The goal of this study was to examine whether an artificial visual system can accommodate several types of face information, including identity and expression. This could occur via the computational processing in the system, the richness of form information available in face images, or both. It remains an open question whether it is possible, by neuroscience-based techniques, to uncover what the ventral stream is optimized to do, if indeed it is optimized at all. This is why computational techniques can help to illustrate possible outcomes of various types of optimization schemes, and to note accord, or lack thereof, with human perception. The DCNN shows a hierarchy of face image information, and therefore generates a hypothesis about ventral visual stream representations of face information.

A remaining challenge for understanding facial expression perception, both computationally and psychologically, is that current facial expression datasets lack well-labeled, naturalistic (i.e., not acted) facial expressions. These datasets will likely be available in the near future, and will include richer visual information in face images that will contribute to our understanding of the relationship between human and machine face processing.

In the context of a global brain system for facial expression processing, the conclusions of this study in no way preclude representations of facial expression and viewpoint outside of the ventral visual stream. It is well-established that the dorsal visual stream and subcortical brain areas are involved in facial expression processing. Subcortical areas (e.g., the amygdala) process basic emotions and directed facial expressions that signal fear, disgust, and other emotional expressions associated with well-being (Gorno-Tempini et al., 2001; Karow et al., 2001). Concomitantly, the dorsal visual stream processes the visual motion signals associated with facial expression and viewpoint. There is important contextual information in emotional faces, and we do not discount the contributions these different brain areas make to the perception of facial expression. Instead, we propose that it is possible for a "ventral-like" identity system to form a single, hierarchical representation that encodes identity, expression, and viewpoint information.

What these findings imply for ventral stream processing is the idea that low-level features are maintained in a high-level categorical representation, possibly because low-level information provides context for understanding the categorical representation. They suggest that some aspect of the what depends on the how and where of the visual stimulus as a whole. Whether this is specific to faces is unclear, but it supports the consideration of shape information (including expression and viewpoint) when investigating face perception in the ventral visual stream.

Keywords: faces, expression, viewpoint, neural network, ventral stream

Acknowledgments

Funding provided by National Eye Institute Grant R01EY029692-01 and by the Intelligence Advanced Research Projects Activity (IARPA) to AOT.

Supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via IARPA R&D Contract No. 2014-14071600012 and 2019-022600002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation thereon.

Commercial relationships: none.
Corresponding author: Y. Ivette Colón.
Email: ycolon@wisc.edu.
Address: University of Wisconsin - Madison, Department of Psychology, 1202 West Johnson St. Madison, WI 53706-1611, USA.

Footnote

[1] Because there were more identities than colors, some colors are used for two identities.

References

Bansal, A., Castillo, C., Ranjan, R., & Chellappa, R. (2017). The do's and don'ts for cnn-based face verification. In: *Proceedings of the IEEE International Conference on Computer Vision*, 2545–2554.

Bansal, A., Nanduri, A., Castillo, C. D., Ranjan, R., & Chellappa, R. (2017). Umdfaces: An annotated face dataset for training deep networks. In: *2017 IEEE International Joint Conference on Biometrics (IJCB)*, IEEE, 464–473.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*(3), 305–327.

Bruyer, R., Laterre, C., Seron, X., Feyereisen, P., Strypstein, E., Pierrard, E., . . . Rectem, D. (1983). A case of prosopagnosia with some preserved covert remembrance of familiar faces. *Brain and Cognition, 2*(3), 257–284.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research, 41*(9), 1179–1208.

Duchaine, B., & Yovel, G. (2015). A revised neural framework for face processing. *Annual Review of Vision Science, 1*, 393–416.

Fox, C. J., Moon, S. Y., Iaria, G., & Barton, J. J. (2009). The correlates of subjective perception of identity and expression in the face network: an fMRI adaptation study. *Neuroimage, 44*(2), 569–580.

Ganel, T., Valyear, K. F., Goshen-Gottstein, Y., & Goodale, M. A. (2005). The involvement of the

fusiform face area in processing facial expression. *Neuropsychologia, 43*(11), 1645–1654.

Gorno-Tempini, M. L., Pradelli, S., Serafini, M., Pagnoni, G., Baraldi, P., Porro, C., Nicoletti, R., Umita, C., . . . Nichelli, P. (2001). Explicit and incidental facial expression processing: an fMRI study. *Neuroimage, 14*(2), 465–473.

Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: *European Conference on Computer Vision*, Springer, 87–102.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences, 4*(6), 223–233.

Hess, U., Adams, R. B., & Kleck, R. E. (2007). Looking at you or looking elsewhere: The influence of head orientation on the signal value of emotional facial expressions. *Motivation and Emotion, 31*(2), 137–144.

Hill, M. Q., Parde, C. J., Castillo, C. D., Colón, Y. I., Ranjan, R., Chen, J.-C., Blanz, V., . . . O'Toole, A. J. (2019). Deep convolutional neural networks in the face of caricature. *Nature Machine Intelligence, 1*(11), 522–529.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience, 19*(4), 613.

Karow, C. M., Marquardt, T. P., & Marshall, R. C. (2001). Affective processing in left and right hemisphere brain-damaged subjects with and without subcortical involvement. *Aphasiology, 15*(8), 715–729.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Kurucz, J., & Feldmar, G. (1979). Prosopo-affective agnosia as a symptom of cerebral organic disease. *Journal of the American Geriatrics Society, 27*(5), 225–230.

Lindsay, G. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, 1–15, https://direct.mit.edu/jocn/article/doi/10.1162/jocn_a_01544/97402/Convolutional-Neural-Networks-as-a-Model-of-the.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). The Karolinska directed emotional faces (KDEF), CD ROM from Department of Clinical Neuroscience, Psychology section. *Karolinska Institutet, 91*(630), 2–2.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(Nov), 2579–2605.

Mancini, G., Biolcati, R., Agnoli, S., Andrei, F., & Trombini, E. (2018). Recognition of facial emotional expressions among Italian pre-adolescents, and their affective reactions. *Frontiers in Psychology, 9*, 1303.

Matsumoto, D., & Hwang, H. S. C. (2019). Culture, emotion, and expression. *Cross-Cultural Psychology: Contemporary Themes and Perspectives*, 501–515.

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, 22*(9), 794–809.

O'Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological and neural synthesis. *Trends in Cognitive Sciences, 6*(6), 261–266.

Parde, C. J., Castillo, C., Hill, M. Q., Colón, Y. I., Sankaranarayanan, S., Chen, J.-C., . . . O'Toole, A. J. (2017). Face and image representation in deep CNN features. In: *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, IEEE, 673–680.

Parde, C. J., Colón, Y. I., Hill, M. Q., Castillo, C. D., Dhar, P., & O'Toole, A. J. Single unit status in deep convolutional neural network codes for face identification: Sparseness redefined. arXiv preprint arXiv:2002.06274.

Parde, C. J., Hu, Y., Castillo, C., Sankaranarayanan, S., & O'Toole, A. J. (2019). Social trait information in deep convolutional neural networks trained for face identification. *Cognitive Science, 43*(6), e12729.

Ranjan, R., Bansal, A., Xu, H., Sankaranarayanan, S., Chen, J.-C., Castillo, C. D., . . . Chellappa, R. Crystal loss and quality pooling for unconstrained face verification and recognition. arXiv preprint arXiv:1804.01159.

Ranjan, R., Sankaranarayanan, S., Castillo, C. D., & Chellappa, R. (2017). An all-in-one convolutional neural network for face analysis. In: *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, IEEE, 17–24.

Song, A., Li, L., Atalla, C., & Cottrell, G. Learning to see people like people. arXiv preprint arXiv:1705.04282.

Surguladze, S. A., Brammer, M. J., Young, A. W., Andrew, C., Travis, M. J., Williams, S. C., . . . Phillips, M. L. (2003). A preferential increase in the extrastriate response to signals of danger. *Neuroimage, 19*(4), 1317–1328.

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

Wells, L. J., Gillespie, S. M., & Rotshtein, P. (2016). Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PloS One, 11*(12), e0168307.

Yovel, G., & O'Toole, A. J. (2016). Recognizing people in motion. *Trends in Cognitive Sciences, 20*(5), 383–395.