Cognition 208 (2021) 104484

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short Communication

Deep learning of shared perceptual representations for familiar and unfamiliar faces: Reply to commentaries

Nicholas M. Blauch[a,b,*], Marlene Behrmann[b,c], David C. Plaut[b,c]

[a] Program in Neural Computation, Carnegie Mellon University, United States
[b] Neuroscience Institute, Carnegie Mellon University, United States
[c] Department of Psychology, Carnegie Mellon University, United States

ARTICLE INFO

Keywords:
Face recognition
Perceptual representations
Deep convolutional neural networks
Familiarity
Expertise

ABSTRACT

We recently argued that human unfamiliar face identity perception reflects substantial perceptual expertise, and that the advantage for familiar over unfamiliar face identity matching reflects a learned mapping between generic high-level perceptual features and a unique identity representation of each individual (Blauch, Behrmann and Plaut, 2020). Here we respond to two commentaries by Young and Burton (2020) and Yovel and Abudarham (2020), clarifying and elaborating our stance on various theoretical issues, and discussing topics for future research in human face recognition and the learning of perceptual representations.

1. The role of idiosyncratic within-identity variability in face recognition

A major point of agreement between us and both Young and Burton (2020), and Yovel and Abudarham (2020), hereafter Y&B and Y&A, respectively, is the fact that faces exhibit a large degree of idiosyncratic within-identity variability, which places fundamental constraints on face recognition performance (Young & Burton, 2018; Kramer, Young & Burton, 2018). Until this within-identity variability is learned for each individual, face verification involving that individual will be more error-prone. This is why even the highly trained generic perceptual face representations learned by a face-trained deep convolutional neural network (DCNN) provide good, but imperfect verification performance, as also pointed out by Y&A.

2. Deep versus image-based mechanisms for unfamiliar face perception

Y&B make extensive reference to work by Burton, Kramer, Ritchie, and Jenkins (2016) who, like Kramer et al. (2018), utilized an Active Appearance Model (AAM) of human face recognition, which first aligns pixel representations of faces to a common template of several (<100) face landmark positions (the positions of which yield a "shape" representation), and then performs a linear reweighting of these post-aligned pixel or "texture" representations, using Principal Components Analysis (PCA) for dimensionality reduction. However, differently from Kramer et al. (2018), Burton et al. (2016) performed PCA on different faces separately, and found that faces differ along fundamentally different texture-based PCs. This is an interesting finding; however, the question is whether it supports the associated claims, drawn from Burton et al. (2016) that: 1) "the dimensions of variability in one face do not generalize well to another," 2) "the expertise that comes with learning faces is not expertise for faces as a class of objects…(but) expertise for the individual faces that have been learned," and 3) familiarization allows human perceivers to move from a "simple image-dependent recognition strategy to a more sophisticated, abstractive recognition strategy that generalizes to novel instances of the person".

These claims are similar to many of the original claims (Young & Burton, 2018; Kramer et al., 2018) we addressed in the target article (Blauch et al., 2020), in which we asked whether human unfamiliar face recognition is well described as a simple image-dependent recognition process, or rather, is demonstrative of substantial perceptual expertise. Several of our results highlight the perceptual expertise account over the image-dependent account: 1) a simple (linear, post-alignment) image-based active-appearance model (as in Kramer et al., 2018) performs substantially more poorly than humans in unfamiliar face recognition, a detail which becomes more apparent when utilizing an unbounded metric such as *d'* rather than accuracy, suggesting that its representations do not capture the complexity of the human mechanism; 2) a deep neural network trained for face recognition performs on par with humans in unfamiliar face recognition, and this effect emerges relatively deep in the network, after several nonlinear

[*] Corresponding author at: Baker Hall 342C4825 Frew St., Pittsburgh PA 15213 United States.
E-mail address: blauch@cmu.edu (N.M. Blauch).

https://doi.org/10.1016/j.cognition.2020.104484
Received 14 August 2020; Received in revised form 1 October 2020; Accepted 3 October 2020
Available online 24 October 2020

processing stages; 3) the same network trained on object recognition performs much more poorly than humans in unfamiliar face recognition, and fails to efficiently learn to recognize familiar faces from a small number of examples, in contrast to the face-trained network; and 4) unfamiliar recognition performance improves consistently with further prior experience with faces, along with the ability to learn new familiar faces robustly. In addition to highlighting the perceptual expertise necessary to achieve human-level performance on unfamiliar faces, these results show that performance on unfamiliar and familiar faces are linked—learning generic features helps to perceptually discriminate unfamiliar faces, and provides a basis for rapidly learning the idiosyncratic variability of familiar faces. Thus, we argue that Y&B's theoretical account neglects the complexity of the generic perceptual face representations that must be learned by humans to master face recognition, and thus overstates the case for idiosyncratic variability (see also Rossion, 2018; Sunday & Gauthier, 2018).

3. The relevance of model complexity and absolute performance level to explaining expertise in human face perception

Citing Roberts and Pashler (2001), Y&B argued that it is no surprise that DCNNs are able to better model human face recognition when compared to linear image-based models, as their greater number of parameters and model complexity gives them a trivial advantage in model fitting. However, whereas Roberts and Pashler cautioned against the use of a good fit as evidence in favor of a theory, they did so in reference to models which adjust free parameters directly in order to maximize the fit to human data. In our application of deep learning models to face recognition, the models 1) are fit to maximize performance on facial recognition, not to reproduce any details of human performance; and 2) are always tested on unseen images, thereby evaluating generalization rather than model fit. The only factors that are varied with reference to human performance are the domain (e.g. objects vs. faces) and the extent of visual experience, which have a direct bearing on our interpretation. Moreover, how these factors affect the match to human data reveals an interesting point: model complexity alone is insufficient to yield an accurate model of human face recognition performance—a large amount of training on the specific domain of faces is also needed to learn the perceptual transformations that can capture human-like performance. Thus, the criticism that deep networks are better models simply by virtue of complexity is invalid.

Y&B also claim that absolute performance level is not a critical aspect of models of face perception. We agree that absolute performance level is not always the most critical factor and certainly not the only important factor; simpler models that perform poorly on some real-world tasks may still offer many merits for interpretability in certain situations. In this specific case, however, where the complexity and expertise of the system is the question of interest, we hold that a good match to absolute performance level is a critical aspect of a good model. Only by reaching this performance level can we reason about the level of expertise necessary for the task at hand. In attempting to reach this performance level with DCNNs, we find that a large degree of face-specific experience and multiple levels of nonlinear processing are necessary, thus advancing our view that the human skill is expert.

4. From unfamiliar to familiar: Linking perceptual variation with conceptual constancy

We agree with Y&B and Y&A that the process of familiarization involves linking multiple, variable perceptual instances of a face with some conceptual information—be it environmental context (e.g. your barista) or a name—to serve as a supervisory signal to bind the perceptual variation to a conceptual constant. As stated in Section 1 above, learning this relationship between perceptual variation and conceptual constancy is required for optimal facial identification. This point of agreement highlights a similarity between the AAM and DCNN models, both of which utilize supervised learning of perceptual variation with the conceptual constant of identity labels (names). However, a point which we discovered using deep learning based models, discussed in Section 2 above, is that such familiarization can be done very efficiently in highly variable and potentially non-frontal naturalistic views only if high-level facial-identity perceptual features are available to be bound with the conceptual information. Thus, the expert perceptual mechanism is useful not only for discriminating between unfamiliar faces, but also for grouping together a person's perceptual variation in terms of reliable, high-level perceptual features rather than image-based or non-face specific features which fail to generalize in this scenario.

5. Is face expertise “for” familiar faces?

Both Y&B and Y&A argue that perceptual face expertise is “for” familiar faces, roughly based on three considerations:

1) the greater robustness in familiar versus unfamiliar face identity matching for humans and DCNNs;
2) the fact that the DCNNs trained using supervised learning on a set of familiar face identities learn representations which successfully model both unfamiliar and familiar human face recognition, claiming that the benefit for unfamiliar faces is merely coincidental; and
3) The intuition that familiar face identity perception is much more important than unfamiliar face identity perception.

We addressed point 1 in arguing that familiar and unfamiliar faces are perceived by largely overlapping perceptual processes that capture the generic variability in faces (see Section 2), and that the idiosyncratic variability of faces mandates that an additional stage of processing link the perceptual variability of an individual face with the constant conceptual/identity information for maximal performance (see Section 1), in agreement with Y&A. In our view, this extra stage of processing available for familiar faces does not negate the role of perceptual representations in perceiving identity in unfamiliar faces, due to the large overlap and mutual dependence of the perceptual representations supporting face identity perception regardless of familiarity.

Point 2 requires further elaboration. While our computational model learned perceptual representations through supervised learning over many natural images of a set of familiar identities with associated identity labels, we do not maintain a strong theoretical commitment to either purely supervised learning or the application of learning to only conceptually familiar faces, in contrast to the position outlined by Y&A. Indeed, one of the earlier and most influential approaches in deep face representation learning (FaceNet; Schroff, Kalenichenko, & Philbin, 2015) utilizes supervision with a same/different identity label per pair of triplets of images, rather than a unique classification label per image for each familiar identity. This approach, which minimizes a contrastive loss to encourage “same” pairs to be close and “different” pairs to be far in representational space, can be scaled to video-based unsupervised learning on unfamiliar faces, where the same/different label for pairs of images is inferred continuously from a trajectory of image frames (Sharma, Tapaswi, Sarfraz, & Stiefelhagen, 2019). Moreover, self-supervised learning approaches that emphasize prediction of future or augmented perceptual representations without labelled supervision have achieved broad success in object recognition recently (Chen, Kornblith, Norouzi, & Hinton, 2020; Hénaff et al., 2020; van den Oord, Li, & Vinyals, 2019). Thus, perceptual face representations might also be learned through more perceptual forms of learning utilizing environmental cues to conceptual constancy without access to identity-specific conceptual information, such as a name, even if this conceptual information does assist learning when available. Moreover, it is, of course, true that learning could not proceed on wholly unfamiliar faces, as, by definition, they are not part of one's experience and, thus, provide no learning signals. We agree with Y&A that familiar faces elicit very strong learning signals through important and extensive social interactions. However, we maintain that perceptual learning might still proceed on faces even when semantic information is unavailable, or when the experience with an identity is so transient as to leave that person largely unfamiliar. In our view, perceptual variation linked with conceptual constancy and a sense of task-relevance are the crucial ingredients driving the learning of perceptual face representations, rather than the specific nature of the conceptually constant signal (e.g., a labelled name vs. environmental cues) or level of familiarity.

There remains the question of whether unfamiliar faces are sufficiently relevant to daily life to drive perceptual learning. This brings us to point 3—the belief, grounded in an observer's personal experience, that only familiar face identity perception is important. Note, however, that in the course of daily life, humans can never be sure whether individuals whom they encounter will ultimately become familiar. Indeed, familiarization can only occur when an unfamiliar face is identified over multiple perceptual instances. Outside of the lab, cues to identification and names are not always available, and learning a face might more generally involve aggregating a conceptual representation of a person over multiple encounters through episodic memory. As argued in Section 4, expert generic face descriptors appear to be necessary for such identity perception, thus providing an important purpose for these representations as applied to identity perception of unfamiliar faces. Thus, while the task-relevance of familiar face recognition is obvious, the task-relevance of unfamiliar face identity perception may also be noteworthy, and could vary across individuals and influence differences in face recognition performance.

To illustrate this last point, imagine a socially curious baby who looks at nearly every passing person while in the stroller on a walk around the neighborhood, and moreover, who, like the system of Sharma et al. (2019), learns from these short “perceptual video” experiences as well as from experience at home with familiar faces. Now imagine a second baby who learns from only the familiar faces of family and friends. Based on our analyses showing that face recognition performance improves with more previously learned identities (Blauch, Behrmann, Plaut 2020), we would expect that the curious baby with additional perceptual experience with unfamiliar faces would have more developed and socially advantageous face representations than would the baby who learned only from the familiar faces. Given machine-learning demonstrations of unfamiliar face identity-based learning, it is thus improbable to us that babies, and people more broadly, perceptually learn from only conceptually familiar faces. It remains an interesting and important goal for psychology to determine the extent to which humans actually do learn from experience with unfamiliar faces, and how this might change throughout development.

6. Concluding remarks

Our view is that generic facial expertise enables expert but variation-limited identity perception of unfamiliar faces, along with the rapid acquisition of robustly separable familiar face identity representations through the binding of perceptual and conceptual information. We believe that deep learning offers a promising framework within which to study human face recognition and that this constitutes an advance beyond simple linear image-based models. We look forward to future investigations that uncover the precise details of how humans develop perceptual face representations and the nature of the binding of perceptual and conceptual information across a network of important brain areas. Moreover, we anticipate that future computational modeling work using tools from deep learning will play a strong role in providing a more detailed account of human face recognition.

Acknowledgements

This research was supported by NIH grant RO1 EY027018 (to M.B. and D.C.P.) and a Presidential Fellowship from Carnegie Mellon University to N.M.B.

References

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2020). Computational insights into perceptual representations for familiar and unfamiliar face recognition. *Cognition*. https://doi.org/10.1016/j.cognition.2020.104341.

Burton, A. M., Kramer, R. S. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science, 40*(1), 202–223. https://doi.org/10.1111/cogs.12231.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *ArXiv:2002.05709 [Cs, Stat]*. http://arxiv.org/abs/2002.05709.

Hénaff, O. J., Srinivas, A., De Fauw, J., Razavi, A., Doersch, C., Eslami, S. M. A., & van den Oord, A. (2020). Data-efficient image recognition with contrastive predictive coding. *ArXiv:1905.09272 [Cs]*. http://arxiv.org/abs/1905.09272.

Kramer, K., Young, A. W., & Burton, A. M. (2018). Understanding face familiarity. *Cognition, 172*, 46–587. https://doi.org/10.1016/j.cognition.2017.12.005.

van den Oord, A., Li, Y., & Vinyals, O. (2019). Representation learning with contrastive predictive coding. *ArXiv:1807.03748 [Cs, Stat]*. http://arxiv.org/abs/1807.03748.

Roberts, S., & Pashler, H. (2001). How persuasive is a good fit? *A Comment on Theory Testing. 10*.

Rossion, B. (2018). Humans are visual experts at unfamiliar face recognition. *Trends in Cognitive Sciences, 22*(6), 471–472.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 815–823). . https://doi.org/10.1109/CVPR.2015.7298682.

Sharma, V., Tapaswi, M., Sarfraz, M. S., & Stiefelhagen, R. (2019). Self-supervised learning of face representations for video face clustering. *ArXiv:1903.01000 [Cs]*. http://arxiv.org/abs/1903.01000.

Sunday, M. A., & Gauthier, I. (2018). Face expertise for unfamiliar faces: A commentary on Young and Burton’s “Are we face experts?”. *Journal of Expertise, x*(x), 1–7.

Young, A. W., & Burton, A. M. (2018). Are we face experts? *Trends in Cognitive Sciences, 22*(2), 100–110. https://doi.org/10.1016/j.tics.2017.11.007.

Young, A. W., & Burton, A. M. (2020). Insights from computational models of face recognition: A reply to Blauch, Behrmann and Plaut. *Cognition*. https://doi.org/10.1016/j.cognition.2020.104422.

Yovel, G., & Abudarham, N. (2020). From concepts to percepts in human and machine face recognition: A reply to Blauch, Behrmann & Plaut. *Cognition*. https://doi.org/10.1016/j.cognition.2020.104424.