How Exposure to Diverse Faces Shapes the Computational Mechanism of Face Perception

Elaheh Akbarifathkouhi (elaheh.akbarifathkouhi@psychol.uni-giessen.de)
Department of Psychology, Justus Liebig University Giessen
Giessen, 35394, Germany
Center for Mind, Brain and Behavior, Universities of Marburg, Giessen, and Darmstadt
Marburg, 35032, Germany

Katharina Dobs (katharina.dobs@psychol.uni-giessen.de)
Department of Psychology, Justus Liebig University Giessen
Giessen, 35394, Germany
Center for Mind, Brain and Behavior, Universities of Marburg, Giessen, and Darmstadt
Marburg, 35032, Germany

Abstract

The Other-Race Effect (ORE) refers to the difficulty humans experience when recognizing faces from races less familiar to them. Prior research has linked the ORE with limited exposure to diverse faces, yet the precise nature of this relationship remains unclear. Here, we use deep convolutional neural networks (CNNs) to investigate how racially varied exposure affects face perception. We trained three CNNs: one on white faces, another on Asian faces, and a Dual CNN on both. While the single-trained CNNs exhibited an ORE on the untrained race, the Dual CNN showed less bias and performed well across both races. Surprisingly, in a target-matching task, the Dual CNN most closely matched both white and Asian participants' choices, despite their own ORE. Furthermore, only the Dual CNN developed a unified representational space for both races. When testing on an unfamiliar third race, the Dual CNN outperformed single-trained models, highlighting its feature space's generalizability. Our results show that racially diverse exposure not only reduces biases in CNNs but also results in a unified, more generalized representational geometry, thereby offering new insights into how experience may shape the computational mechanisms of face perception.

Keywords: Face Perception, Deep Convolutional Neural Networks, Other-Race Effect, Representational Geometry

Introduction

Despite humans' proficiency in recognizing faces, their perception often carries inherent biases. Among the most studied of these biases is the Other-Race Effect (ORE), which refers to the difficulties humans experience when identifying faces from races less familiar to them. Previous research has established a link between the ORE and limited exposure to racially diverse faces (O'Toole et al.,1991; Walker et al., 2008). However, the precise impact of facial exposure on human facial representations remains unclear: Does diverse exposure result in an integrated feature space or multiple race-specific feature spaces? Moreover, does experience with diverse faces enhance generalizability beyond familiar races? Addressing these questions is challenging due to the inherent difficulty of controlling visual experience in humans.

Recently, task-optimized deep convolutional neural networks (CNNs) have emerged as promising computational models for studying perceptual representations in humans (Kriegeskorte, 2015; Kanwisher, Khosla, & Dobs, 2023). Notably, CNNs exhibit behavioral characteristics similar to those observed in human face perception, thereby enabling inquiries into the computational mechanism governing their origins (Dobs et al., 2023). An intriguing advantage of these models lies in the full control over their 'training diet'. Here, we trained CNNs on either race-specific or racially diverse datasets to explore the computational mechanisms underlying the ORE. Our primary focus was to investigate how exposure to racially varied faces affects the CNNs' representational geometry and its generalizability to unfamiliar races, as well as their consistency with human face perception behavior.

Methods and Results

Training CNNs on Diverse Datasets

To assess the impact of racially varied training on CNNs, we trained three CNNs, all based on the VGG16 architecture, for face recognition tasks: one on 1,654 Asian identities (Asian CNN), another on 1,654 white identities (White CNN), and a third combining both sets of identities (Dual CNN) (**Figure 1A**).

Racial Biases in CNNs

Would racially diverse training lead to reduced racial biases? We evaluated each CNN's performance using a target-matching task on a completely novel dataset comprising 80 female identities (40 Asian and 40 white, 5 images each), none of which were included in the training phase. Performance was quantified by the minimum distance (1 – cosine similarity) in activation patterns (fc7 layer) between two matching images and a target image. The ORE index was computed as

$$ORE\ Index = \frac{Accuracy_{trained-race} - Accuracy_{untrained-race}}{Accuracy_{trained-race}}$$

For the Dual CNN, we refined this index for both races by weighting the difficulty of each race's task, using the performance of the single-trained CNNs as a baseline.

We found that single-trained CNNs showed decreased performance on untrained races, indicating significant racial bias (**Figure 1B**). In contrast, the Dual CNN performed well across both trained races, outperforming the single-trained CNNs. This finding was quantitatively supported by the ORE indices, with single-trained CNNs displaying biases towards their trained races (ORE indices: ~6-9%). Meanwhile, the Dual CNN showed substantially lower bias (difficulty-weighted ORE index: ~3%), highlighting the effectiveness of diverse training in mitigating racial biases. These findings align with prior research in human development, suggesting that early exposure to racially diverse faces can reduce the ORE in human perception (McKone et al., 2019; Spangler et al., 2013; Suhrke et al., 2014).

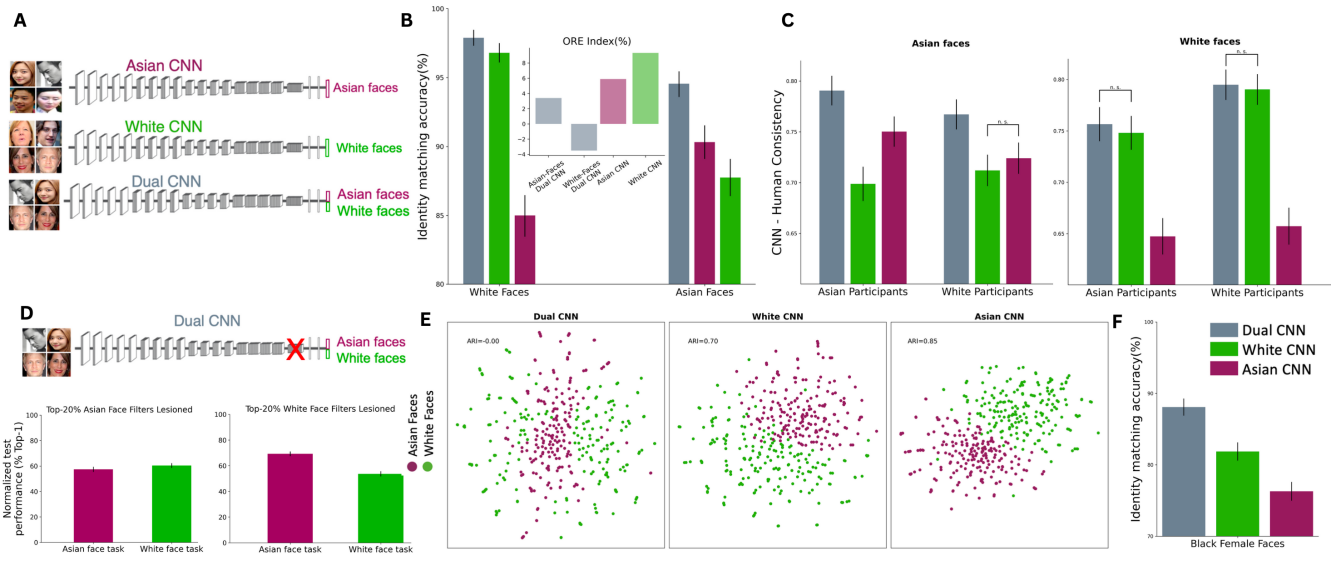

Figure 1: **A.** Three task-optimized CNNs, **B.** CNNs' performance in a target-matching task with their ORE index, **C.** CNN-human consistency on Asian (left) and white (right) faces. All pairwise differences were significant ($p<0.01$), unless marked otherwise (n.s). **D.** Lesioning results in the Dual CNN, **E.** t-SNE visualization of Asian and white faces in each CNN, **F.** CNNs' performance in a target-matching task on a third unfamiliar race – black faces.

Behavioral Consistency Between Human Participants and CNNs

To directly assess the findings' relevance to human behavior, we examined the consistency between CNNs' decisions and those of Asian (n=102) and white (n=269) participants, using behavioral responses obtained from the same target-matching task (Dobs et al., 2023). We analyzed the correlation of decision patterns between participants and CNNs on a trial-by-trial basis (**Figure 1C**). Our analysis revealed that the decisions of the Dual CNN most closely matched the behavioral choices of all participants (white Participants: mean r=0.78, Asian Participants: mean r=0.77), despite their own ORE. This suggests that training CNNs on racially diverse datasets results in a more generalized understanding of facial features across different races.

Properties of the Representational Feature Space

Next, we asked whether diverse exposure leads to an integrated feature space or multiple race-specific feature spaces within the Dual CNN. To this end, we used lesion techniques and t-SNE visualization.

Lesion Technique We conducted lesion experiments on the last convolutional layer of the Dual CNN. We ablated individual filters, assessed their impact on recognition performance for each racial task, and ranked them based on their impact. We found that lesioning the top 20% most impactful filters for either race had a similar effect on the CNN's performance across both racial tasks (**Figure 1D**). This suggests an integrated feature space for white and Asian faces within the Dual CNN, rather than separate race-specific feature spaces.

t-SNE To further explore each CNN's feature space, we visualized their representational geometries using t-SNE (**Figure 1E**). We quantified the degree of segregation between races within these spaces using the Adjusted Rand Index (ARI). We found that the single-trained CNNs showed distinct clusters for white and Asian faces (white CNN: ARI=0.7; Asian CNN: ARI=0.85), while the Dual CNN demonstrated a fully integrated feature space (ARI=0).

Benefits of the Integrated Representational Feature Space

Does experience with racially diverse faces enhance out-of-distribution generalization? To address this question, we tested all CNNs on a target-matching task featuring faces from an entirely unfamiliar third race – black faces. We found that the Dual CNN outperformed the single-trained CNNs, showing enhanced generalization to this new race (**Figure 1F**).

Conclusion

We investigated the computational mechanisms underlying the ORE, offering testable hypotheses on the impact of diverse facial exposure on human perception. Our results show that such exposure not only reduces biases in CNNs but also enhances their generalizability and alignment with human perceptual behavior, thereby deepening our understanding of how experience shapes human face perception.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)-project number 222641018-SFB/TRR 135 TP C9, “The Adaptive Mind”, funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art, and the European Research Council (ERC Starting Grant DEEPFUNC, ERC-2023-STG-101117441) to K.D..

References

Dobs, K., Yuan, J., Martinez, J., & Kanwisher, N. (2023). Behavioral signatures of face perception emerge in deep neural networks optimized for face recognition. *Proceedings of the National Academy of Sciences*, *120*(32), e2220642120.

Kanwisher, N., Khosla, M., & Dobs, K. (2023). Using artificial neural networks to ask ‘why’ questions of minds and brains. *Trends in Neurosciences*, *46*(3), 240-254.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, *1*, 417-446.

McKone, E., Wan, L., Pidcock, M., Crookes, K., Reynolds, K., Dawel, A., ... & Fiorentini, C. (2019). A critical period for faces: Other-race face recognition is improved by childhood but not adult social contact. *Scientific reports*, *9*(1), 12820.

O'Toole, A. J., Deffenbacher, K., Abdi, H., & Bartlett, J. C. (1991). Simulating the ‘other-race effect’ as a problem in perceptual learning. *Connection Science*, *3*(2), 163-178.

Spangler, S. M., Schwarzer, G., Freitag, C., Vierhaus, M., Teubert, M., Fassbender, I., ... & Keller, H. (2013). The other-race effect in a longitudinal sample of 3-, 6- and 9-month-old infants: Evidence of a training effect. *Infancy*, *18*(4), 516-533.

Suhrke, J., Freitag, C., Lamm, B., Teiser, J., Fassbender, I., Poloczek, S., ... & Schwarzer, G. (2014). The other-race effect in 3-year-old German and Cameroonian children. *Frontiers in psychology*, *5*, 73289.

Walker, P. M., Silvert, L., Hewstone, M., & Nobre, A. C. (2008). Social contact and other-race face processing in the human brain. *Social Cognitive and Affective Neuroscience*, *3*(1), 16-25.