Using CNNs to understand how bottom-up and top-down processes shape human face detection

Sule Tasliyurt-Celebi (sule.tasliyurt-celebi@psychol.uni-giessen.de)
Department of Psychology, Justus Liebig University Giessen
Giessen, 35394, Germany

Benjamin de Haas (benjamin.de-haas@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

Melissa L.-H. Võ (mlvo@psych.uni-frankfurt.de)
Department of Psychology, Goethe University Frankfurt
Frankfurt, 60323, 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

Understanding the interplay between bottom-up and top-down processing remains a crucial challenge in human perception and cognition. The success of feedforward deep convolutional neural networks (CNNs) in mirroring aspects of human visual perception offers support for the role of bottom-up processing. Here, we leverage the feedforward characteristics of CNNs to differentiate between bottom-up and top-down processes in a core visual task–the rapid detection of faces. By manipulating the presence of scene previews in human face detection tasks, we examine the influence of top-down processing. We found that scene preview enhances face detection, supporting the role of top-down processing in this condition. Encoding model analyses show that while basic visual features, such as face eccentricity, and high-level features extracted from CNNs predict face detection latency, scene preview selectively alters their predictivity, revealing a dynamic and context-dependent contribution. Our results offer a novel approach for understanding the complex dynamics between bottom-up and top-down influences in human visual perception.

Keywords: Face perception; Scene perception; Convolutional Neural Networks; Top-down processing; Bottom-up processing; Predictive processing

Introduction

Visual perception and cognition involve complex processes that allow us to understand and interact with our environment, through bottom-up and top-down processing. Bottom-up processing is driven by sensory input, whereas top-down processing relies on expectations, tasks at hand, knowledge, and previous experiences. The sophisticated interplay between these two processes has been central to the study of visual perception, from visual illusions (Gregory, 1970) to attention (Treisman & Gelade, 1980) and their neural mechanisms (Gilbert & Li, 2013)and continues to drive ongoing debates within the field (Peters et al., 2024).

In the realm of face perception, research has predominantly focused bottom-up mechanisms. This focus is supported by findings that humans can rapidly and accurately detect faces across diverse scenes (Bindemann & Lewis, 2013), with saccades to faces occurring as fast as 100 ms after stimulus onset (Crouzet et al., 2010; Martin et al., 2018), suggesting that face detection might be minimally influenced by top-down processes, such as scene context (Crouzet & Thorpe, 2011). However, the strong impact of scene context on object perception (Aldegheri et al., 2023; Lauer & Võ, 2022) raises questions about its potential role in face detection (Lewis & Edmonds, 2003).

Recently, deep convolutional neural networks (CNNs) have emerged as powerful tools for understanding human perceptual processes, including face recognition (Dobs et al., 2022) and scene categorization (Groen et al., 2018; Karapetian et al., 2023). Critically, feedforward CNNs do not incorporate recurrent or top-down connections, offering unique opportunities to study the interplay between sensory input and scene context. Here, we aim to dissect how top-down processing provided by scene previews shapes face detection, focusing on the role of basic visual features (e.g., face eccentricity) and high-level features derived from task-optimized CNNs.

Methods and Results

Scene Preview Facilitates Face Detection

Methods To test the role of scene information on face detection latency, we curated a dataset of 120 natural scene images from 12 indoor scene categories (e.g., bathroom, kitchen), each containing a single face. For each target scene, we produced a corresponding faceless version by manually editing out the face and body, serving as scene preview. In the experiment, participants (n=38) were briefly (250 ms) shown either a scene preview (preview condition) or a gray screen (no-preview condition) prior to the presentation of the target scene (**Figure 1A**). Using eye tracking, we measured face detection latency as the time to first fixation on the face in the target scene.

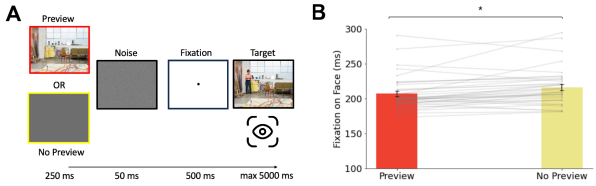

Figure 1: A. Task procedure (n=38). B. Face detection latencies in preview and no-preview conditions (*p<0.001). Gray lines show participant's data.

Results Does prior scene information enhance the rapid detection of faces? We found that participants' face detection latency was shorter in the preview than in the no-preview condition (207 vs. 216 ms; p<0.001) (**Figure 1B**). This suggests that top-down processing influences face detection, but how so?

Scene Preview Affects Features Driving Face Detection

Methods To test the role of top-down processes on visual features guiding face detection, we analyzed nine visual feature categories from our stimulus set. We measured physical characteristics, including face and body size, along with their eccentricity. To measure complex visual scene- and face-specific features, we used four VGG-16 networks each optimized for distinct tasks (**Figure 2A**): face discrimination (Face CNN), scene categorization (Scene CNN), combined scene categorization and face discrimination (SceneFace CNN), and combined scene categorization and face detection (SceneFaceDet CNN) (Dobs et al., 2022, 2023). The CNNs' effectiveness in face detection was measured by calculating pairwise cosine distances between fc7 layer activations for preview and target scenes (**Figure 2B**). As a control for low-level visual information, we analyzed pixel-level similarity between scenes. We hypothesized a reduced contribution from all these features in the preview condition due to their primarily feedforward nature.

Furthermore, we hypothesized that scene previews might enhance face detection by providing cues about the expected locations of faces. To explore this, we ran an additional experiment asking participants (n=127) to indicate the most likely location of a face within the faceless scene previews (**Figure 2C**). A face expectation measure was derived using spatial autocorrelation (Moran's I), quantifying the degree to which face predictions were clustered or dispersed.

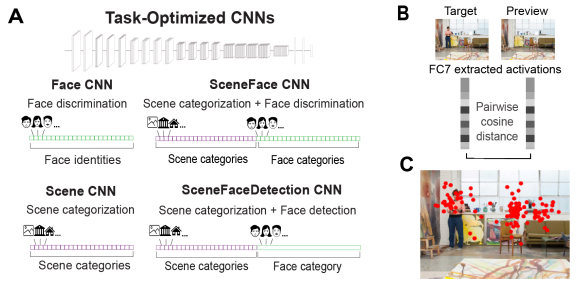

Figure 2: A. Task-optimized CNNs. B. Pairwise cosine distance between activations to preview and target scenes served as measure of face detection in CNNs. C. Example of face likelihood patterns in a scene.

To understand the contribution of each feature category to face detection across conditions, we employed an encoding model analysis. For each feature category, an Ordinary Least Squares (OLS) regression model was fitted using data split across participants by randomly dividing participants into halves (100 times). Explained variance was calculated as squared Spearman's *r* between predicted and true latencies in the test set. To isolate the unique contribution of each feature, we used variance partitioning by comparing the explained variance of a full regression model including all features with a reduced model that excluded the feature of interest.

Results We found that all feature categories, except for pixel features, significantly predicted human face detection latency across preview conditions (**Figure 3A**). Notably, CNN features, particularly those optimized for scene and combined scene and face recognition, were the most predictive ($p<0.001$). We further found lower predictivity of all features in the preview compared to the no-preview condition ($p<0.001$), with the notable exception of face expectation, suggesting that top-down processes decrease the reliance on bottom-up visual features.

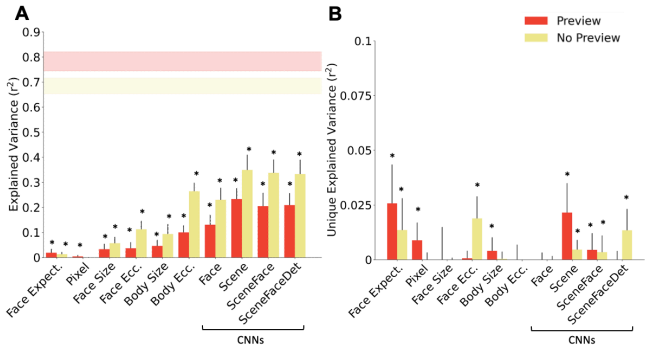

Figure 3: A. Explained variance and B. unique explained variance in preview and no-preview conditions. Shaded horizontal bars represent noise ceiling. Error bars are SEMs across split-halves. (*p<0.01; sign permutation test, Bonferroni-corrected).

When examining the unique contribution of each feature category, distinct patterns emerged between conditions (**Figure 3B**). In the no-preview condition, face eccentricity and features from the SceneFaceDet CNN contributed most, suggesting that bottom-up face detection primarily relies on features optimized for both face detection and scene categorization, as well as the face's location. In contrast, in the preview condition, face expectation and features from the Scene CNN were predominant, indicating that top-down face detection is largely driven by information about faces' probable location and general scene–rather than face-specific–features. Notably, face eccentricity no longer influenced detection latency, suggesting that scene previews facilitate detecting faces in the periphery.

Conclusion

Our findings highlight the significance of top-down processes in face detection, showing how they interact with bottom-up signals to enhance visual cognition. These results further point towards developing more advanced AI models that better mimic human cognitive processes.

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

Aldegheri, G., Gayet, S., & Peelen, M. V. (2023). Scene context automatically drives predictions of object transformations. *Cognition*, *238*.

Bindemann, M., & Lewis, M. B. (2013). Face detection differs from categorization: Evidence from visual search in natural scenes. *Psychonomic Bulletin and Review*, *20*(6), 1140–1145.

Crouzet, S. M., Kirchner, H., & Thorpe, S. J. (2010). Fast saccades toward faces: Face detection in just 100 ms. *Journal of Vision*, *10*(4), 1–17.

Crouzet, S. M., & Thorpe, S. J. (2011). Low-level cues and ultra-fast face detection. *Frontiers in Psychology*, *2*(11).

Dobs, K., Martinez, J., Kell, A. J. E., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science Advance* (8).

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 of the United States of America*, *120*(32).

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Gregory, R. (1970). *The Intelligent Eye*. Weidenfeld and Nicolson.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife,* 1-26.

Karapetian, A., Boyanova, A., Pandaram, M., Obermayer, K., Kietzmann, T. C., & Cichy, R. M. (2023). Empirically Identifying and Computationally Modeling the Brain–Behavior Relationship for Human Scene Categorization. *Journal of Cognitive Neuroscience*, *35*(11), 1879–1897.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11).

Lauer, T., & Võ, M. L. H. (2022). The Ingredients of Scenes that Affect Object Search and Perception. In Ionescu, B., Bainbridge, W. A., & Murray, N. (Eds.), *Human Perception of Visual Information: Psychological and Computational Perspectives*, 1-32. Springer.

Lewis, M. B., & Edmonds, A. J. (2003). Face detection: Mapping human performance. *Perception*, *32*(8), 903–920.

Martin, J. G., Davis, C. E., Riesenhuber, M., & Thorpe, S. J. (2018). Zapping 500 faces in less than 100 seconds: Evidence for extremely fast and sustained continuous visual search. *Scientific Reports*, *8*(1).

Peters, B., DiCarlo, J. J., Gureckis, T., Haefner, R., Isik, L., Tenenbaum, J., Konkle, T., Naselaris, T., Stachenfeld, K., Tavares, Z., Tsao, D., Yildirim, I., & Kriegeskorte, N. (2024). *How does the primate brain combine generative and discriminative computations in vision?*

Treisman, A. M., & Gelade, G. (1980). A Feature-Integration Theory of Attention. Cognitive Psychology (12).