Neural oscillations encode context-based informativeness during naturalistic free viewing

Songyun Bai (songyun.bai@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 EN Nijmegen, Netherlands

Philip Sulewski (psulewski@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49074 Osnabrück, Germany

Carmen Amme (camme@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49074 Osnabrück, Germany

Peter König (peter.koenig@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49074 Osnabrück, Germany

Tim C. Kietzmann (tim.kietzmann@uni-osnabrueck.de)
Institute of Cognitive Science, University of Osnabrück, 49074 Osnabrück, Germany

Marius Peelen (marius.peelen@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 EN Nijmegen, Netherlands

Eelke Spaak (eelke.spaak@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, 6525 EN Nijmegen, Netherlands

Abstract

In everyday vision, humans actively sample the environment by moving their eyes several times per second. While predictive processing accounts have been successful in explaining neural oscillatory activity during fixated viewing in non-human primates, whether these accounts extend to free viewing and humans is still unknown. To address this, we developed a novel analysis pipeline combining large-sample, head-fixed MEG, eye-tracking, and a generative deep neural network model to investigate how the brain encodes visual input with varying levels of contextual predictability. Our results show that the informativeness of the current fixation is positively associated with the occurrence of alpha-beta oscillations across large parts of posterior cortex. It is furthermore positively associated with gamma-band activity, though more specifically localized to central posterior regions corresponding to the foveal representation. In conclusion, contextual predictability is rapidly and transiently encoded in neural oscillations in different frequency bands during free viewing.

Keywords: naturalistic viewing; neural oscillations; deep neural networks

Introduction

Rather than passively receiving visual input, humans actively explore the world by moving our eyes several times per second. This process of sampling is well explained from the perspective of predictive processing (Henderson, 2017): observers predict where the most informative visual information is likely to be found, then acquire new information with each fixation.

Crucial questions in this framework are: (1) What is the contribution to such behaviour of different forms of informativeness/predictability (Spaak, Peelen, & de Lange, 2022)? (2) How are these sources of predictability encoded in and processed by (visual) cortex? A recent monkey study with a fixation design (Uran et al., 2022) has shown that V1 firing rates decrease with the predictability of high-level image features, while gamma-synchronization increases with the predictability of low-level image features. However, it remains unclear whether these neural signatures generalize to human observers, especially under conditions of unconstrained, free-viewing behavior.

To tackle this problem, we developed an analysis pipeline that leverages a unique large-sample, head-fixed human MEG dataset, and integrates it with eye tracking and generative deep neural networks (DNNs). This approach allows us to quantify the informativeness and predictability of each fixation during naturalistic image viewing and explore the corresponding neural correlates of predictive processing in gaze behaviour.

Method

Stimuli and Paradigm

Five healthy adults (3 female; mean age = 27.8 years, SD = 2.6) with normal or corrected-to-normal vision participated in the study. Participants freely viewed 4,080 natural scenes from the Natural Scenes Dataset (Allen et al., 2022) with their eye movement recorded using EyeLink 1000 system (SR Research Ltd., Ottawa, Canada). The scenes were semantically balanced across categories and presented at a resolution of 1024 × 768 pixels on a screen positioned 70 cm from the participants. Each of the 5 participants completed 10 measurement sessions. Stimuli were presented in blocks of 30 trials, with each trial lasting 4 seconds.. First session had 10 blocks (300 trials), and the remaining sessions had 14 blocks each (420 trials/session).

To promote active engagement, 25% of trials included a scene description task, where participates verbally described the content of the preceding scene. Brain activity was recorded using a 306-channel MEG system (Elekta Neuromag TRIUX) at 1000 Hz, with continuous head tracking and foam-stabilized support to allow for natural viewing while minimizing head movement artifacts. Results presented here reflect analyses of 10 recording sessions from one participant (final analyses will reflect 50 sessions in total).

Quantifying contextual predictability of fixations

The visual input of each fixation was defined as a 224 × 224 pixel image patch centered on the fixation point (≈6.83° visual angle). Patches that spatially overlapped with previous fixations or extended beyond image boundaries were excluded, yielding approximately 1,200 valid fixation patches per session.

We implemented a stable diffusion inpainting model (stable-diffusion-v1-5) to quantify the inherent contextual predictability of the stimulus input at each fixation. This was done by removing the fixation patch from the full scene, inpainting the missing region based on the remaining context, and comparing the inpainted patch to the ground truth. Specifically, both the original and inpainted patches were passed through AlexNet, and cosine distances between activation patterns at each convolutional and fully connected layer were computed. This provides a standardized way to quantify the contextual predictability of any region in a naturalistic scene across different levels of visual features. A high similarity between the inpainted and ground-truth patch means that the patch was of low informativeness, given the context; and vice versa.

MEG processing

MEG data were processed using MNE-Python (Gramfort et al., 2013). Temporal signal space separation (tSSS) with movement compensation (MaxFilter, Elekta Oy) was applied to suppress external noise. The data were band-pass filtered between 0.2 and 200 Hz and resampled to 500 Hz. Independent component analysis (ICA) was used to remove ocular artifacts.

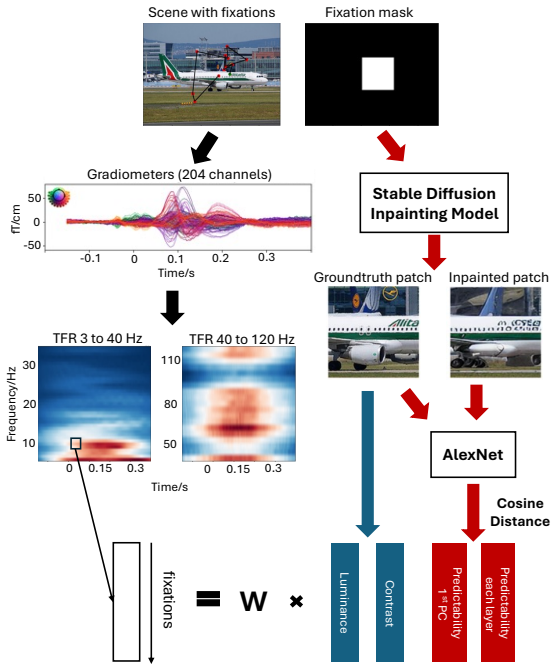

Figure 1: The pipeline

MEG recordings were segmented into epochs from 150ms before each fixation onset until 400ms after. To examine how different frequency bands represent visual scene context, time-frequency representations (TFRs) were computed for each epoch across all gradiometer sensors. For frequencies between 3 and 40 Hz, TFRs were computed using a single Hanning taper with 1 Hz steps. For the 40–120 Hz range, TFRs were calculated using a multitaper method with 16 Hz spectral smoothing to better capture broadband gamma activity.

Neural correlates of contextual predictability

To investigate how the brain encodes contextual predictability, we performed cross-validated linear regression analyses to predict power across the TFRs at each time point and frequency. As a baseline model, we included the brightness and contrast of the ground truth fixation crops to account for low-level visual features. We then added the similarity measures as an additional regressor. By measuring how much the performance of regression improves over the base model (Δr^2), we assess the neural encoding of contextual predictability.

Contextual predictability measures across AlexNet layers were highly correlated with each other, so we first applied principal component analysis (PCA) on the cosine distances for each CCN layers, and used the first principal component as a summary index of overall predictability, which explained around 80 percent of total variance. Additionally, to further examine how different levels of visual features are represented in neural dynamics, we then added the layer-specific predictability scores as individual regressors on top of the baseline + PC1 model.

Results

Context-based overall visual informativeness (i.e., inverse predictability) of fixated patches was associated with increased low frequency power, especially at alpha frequencies, across broad posterior regions (Figure 2A), compatible with widespread visual cortical engagement. Informativeness was likewise positively associated with gamma-band activity (40–80 Hz), with a topography more specifically localized to the central posterior region. Additionally, we observed enhanced high-gamma activity (100–120 Hz) with increased informativeness, which may reflect increased local neuronal firing associated with processing informative visual input. DNN-layer-specific analyses further revealed that these effects were primarily driven by predictability estimates from early to middle layers, corresponding to low-level and textural visual features (not shown).

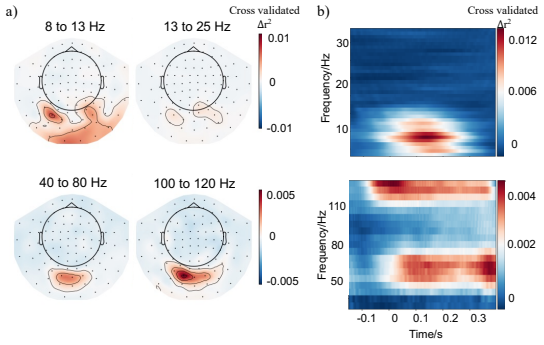

Figure 2: Results: cross-validated Δr^2. (a) Topographical map averaged between 0 and 200 ms. (b) Time-frequency map averaged across sensors that show top 10% effect.

Discussion

We analyzed (parts of) a large-sample, head-fixed MEG dataset acquired in humans freely viewing natural images, through the lens of context-based informativeness as quantified by a novel analysis pipeline. We found neural signatures of visual informativeness that in part align with previous fixated monkey work (alpha/beta, high-frequency broadband), yet in part are in direct opposition (gamma). This highlights the importance of studying visual perception and its neural correlates in naturalistic settings, while underscoring the key role played by various sources of predictions in guiding such gaze behaviour. We conclude that the pipeline we developed is a powerful tool to study dynamic, active vision.

References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., ... others (2022). A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, *25*(1), 116–126.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., ... others (2013). Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, *7*, 267.

Henderson, J. M. (2017). Gaze control as prediction. *Trends in cognitive sciences*, *21*(1), 15–23.

Spaak, E., Peelen, M. V., & de Lange, F. P. (2022). Scene context impairs perception of semantically congruent objects. *Psychological Science*, *33*(2), 299–313.

Uran, C., Peter, A., Lazar, A., Barnes, W., Klon-Lipok, J., Shapcott, K. A., ... Vinck, M. (2022). Predictive coding of natural images by v1 firing rates and rhythmic synchronization. *Neuron*, *110*(7), 1240–1257.