Supporting Information for

Saccade Onset, Not Fixation Onset, Best Explains Early Sensory Responses Across the Human Visual Cortex During Naturalistic Vision

Carmen Amme, Philip Sulewski, Eelke Spaak, Martin N. Hebart, Peter König, Tim C. Kietzmann

* Corresponding author: Carmen Amme

Email: carmenamme@outlook.de

This PDF file includes:

Extended Methods

The MEG - Eye-Tracking Data Set

Procedure. Each of the 5 participants completed 10 measurement sessions over a period between 2 and 5 weeks. At the beginning of the first session, all participants were given instructions in German to freely explore the image visually. All participants were further presented with written instructions in German about giving a verbal caption, which they were indicated to do after 25% of the trials. The participants were then shown 10 example images with 5 example captions each. During the experiment, the caption and non-caption images were presented in random order. The participants were seated with a viewing distance of 70 cm to the screen and the stimuli were presented with a visual angle of 28,54° x 21,61°. At the beginning of each session, the eye tracker was calibrated through a 9-point calibration where participants were asked to fixate subsequently appearing dots at 9 different locations on the screen. Every new trial then began with the central appearance of a white fixation cross on gray background for a minimum duration of 1s, which also served as a drift control for the fine-calibration of the eye-tracking system. If the fixation deviated more than 1.3° from the fixation cross, the trial was re-started. The stimulus image was displayed for 4 seconds, allowing participants to freely explore it through natural eye movements. On 75% of the trials, the presentation of the fixation cross indicated the start of the subsequent trial. On the remaining 25% of the trials, the stimulus presentation was followed by the appearance of a black microphone on a gray background for 1s in the center of the screen, prompting the participants to give a verbal description after the microphone had disappeared. The recording of the caption started with the disappearance of the microphone and ended after a duration of 8s with the start of the following trial. The stimuli were presented in blocks of 30 trials. The first session consisted of 10 blocks (300 trials in total), while the remaining 9 sessions consisted of 14 blocks each (420 trials/session). After every 3 to 4 blocks, the eye tracker was re-calibrated and the participants could take a small break. After the experimental sessions, the recordings of the captions were transcribed manually.

Eye Tracking Data Acquisition and Preprocessing. Eye tracking data were recorded using a SR Research EyeLink 1000 system (SR Research; https://www.sr-research.com) with a sampling rate of 1,000 Hz. The raw eye-tracking data were preprocessed by automatically detecting saccades, identifying samples that exceeded a predefined velocity threshold (Engbert & Mergenthaler, 2006). Blinks were automatically detected by the eye tracker. All samples that were not classified as a blink or a saccade were classified as a fixation. Because we were only interested in fixations that finished during the stimulus presentation, the last fixations on all images were excluded.

MEG Data Acquisition and Preprocessing. To minimize head movements during the measurement, we created individually fitted head stabilizers for each participant. We took each participant's head shape using a Structure Sensor Pro (XRPro, LLC; https://structure.io) and compared that to a 3D scan of the MEG head piece. The space between the MEG headpiece and the participant's head shape was milled into styrofoam. Before each measurement session, a Polhemus FASTRAK tool (Polhemus; https://polhemus.com) was used to digitize the participants' head shape. To track head positions inside the MEG, a total of five coils were placed above the nasion on the forehead, at the inions bilaterally, and in-between these points on the forehead. The experiment took place in a magnetically shielded room using a NeuroMag VectorView MEG system (Megin; https://megin.com) consisting of 306 channels (102 magnetometers, 204 planar gradiometers). The raw data were recorded with a sampling rate of 1,000 Hz, an online filter was applied between 0 and 330 Hz.

After the recording, the raw MEG data was bandpass filtered through a casual filter between 0.2 and 200 Hz. MEG channels that were visually observed to be noisy during the measurement were tracked and interpolated before analysis. The raw MEG data was then cleaned by applying a spatiotemporal *MaxFilter* (maxwell filter; Elektra), including a temporal Signal Space Separation (tSSS; Taulu et al., 2004), which also accounted for different head positions and head movements made during the measurement.

All further analyses were implemented in Python by making use of the toolbox MNE-Python (Gramfort et al., 2013) and several custom written scripts. We applied independent component analysis (ICA) separately for each participant and session to identify and remove noisy components. To identify and remove eye movement-related components, we correlated the epoched fixation-based data of each component with the corresponding x-/y-coordinates of the eye position. We median-centered the data per channel and session to account for different sessions-wise offsets in the MEG data.

MEG-ET Alignment. To investigate fixation-related neural processes, we mapped the MEG data to the eye tracking data by aligning MEG and eye tracking data to scene onset, which was annotated in both MEG and eye tracking data, and then mapped fixation annotations onto MEG data based on fixation/saccade onset latency and duration within each scene presentation. The MEG data was then epoched based on fixation or saccade onset. We corrected the scene onset MEG trigger by 20 ms due to a systematic delay in the trigger sending.

Participants. We recorded MEG data from 5 right-handed participants (3 female), all aged between 26 and 33 years ($\overline{m}_{\text{age}}$ = 27.8) . All participants were native German speakers, reported to be healthy and had normal or corrected-to-normal vision. Everyone gave written informed consent at the beginning of each experimental session, and both at separate days for scanning and fitting the personalized head stabilizers. The study was approved by the ethics committee of the University Medical Center Leipzig and adhered to the Declaration of Helsinki. All participants received monetary reimbursement for their participation.

Stimuli. We included 4,080 images from the Natural Scenes Dataset (NSD; Allen et al., 2022), which consists of 73,000 images of natural scenes from the Microsoft Common Objects in Context (MS COCO) image dataset (Lin et al., 2015). 25% of the sampled images (n = 1,020) were caption-images, where participants were asked to provide a verbal description. All images, as well as whether an image belonged to the caption images, were pseudo-randomly selected with the following conditions: (i) To make sure that the semantic content of the images is equally represented, we performed k-means clustering on all NSD scene captions and identified 60 different semantic clusters. We selected an equal number of images from each cluster. (ii) Allen et al. (2022) defined a set of 100 images that maximally span the semantic space (‘special100’). We included the ‘special100’ as caption images, and the ‘shared1000‘, which are images presented to all participants that were part of NSD, as preferably caption images, but only if the equal distribution of clusters allowed it. All participants were presented with the same caption and non-caption images, which were randomly shuffled for each participant.

Source Reconstruction. We collected a T1-weighted anatomical MRI scan before the start of the MEG experiment. The brain surface reconstruction was done using the *freesurfer* toolbox (Fischl, 2012). The brain surface reconstruction was utilized to compute the boundary element model (BEM), which was then used to generate an individualized forward model for each participant. The MRI and MEG coordinate systems were aligned using the MNE-Python coregistration GUI (Gramfort et al., 2013). The MEG data were source reconstructed using Dynamic Statistical Parameter Mapping (dSPM; Dale et al., 2000). The noise covariance used for the source reconstruction was computed by merging two empty-rooms recordings per participant (~5 minutes of empty room recording each, recorded before and after the experiment). The selected empty room recording sessions per participant were: P1: 3, P2: 3, P3: 2, P4: 2, P5: 1. The source activations were projected onto the cortical surface normal. Spatiotemporal source activations were obtained for each time point and cortical source.

Data Analysis

Peak Latency Analysis. The fixation-based epochs were binned into equally sized bins (n_{bins} = 160) based on the duration of the preceding saccade. For each saccade duration bin, we obtained a median-based event-related field (ERF), which was visualized in a heatmap showing the ERF of a single sensor sorted by the duration of the preceding saccade, with each row corresponding to one saccade duration bin (Fig. 1A, B). The participant-wise sensor selection was based on the strongest positive gradiometer activation between a time window from 90 to 110 ms of either the fixation-locked ERF (Fig. 1A; P1: MEG1923, P2: MEG1733, P3: MEG1933, P4: MEG2123, P5: MEG2512) or the saccade-locked ERF (Fig. 1B; P1: MEG2343, P2: MEG2512, P3: MEG1933, P4: MEG2123, P5: MEG2512). All further analyses were based on the sensor with maximum activation for saccade-locked ERFs.

We defined ERF latency as the halfway point during increase leading to the maximum deflection between 30 and 130 ms after fixation onset. The relationship between saccade duration and ERF peak latency was investigated by fitting a linear regression to the fixation onset-locked ERF latencies, binned based on preceding saccade duration (n_{bins} = 160), for each of the 10 measurement sessions (dashed lines; Fig. 1C). We obtained the correlation coefficient (*r*) and the slope (*m*) of the fitted line (reported only for session 1), as well as an average across all sessions.

To illustrate the shift of the ERF as a function of (preceding) saccade duration, we binned the fixation and saccade ERFs into 10 equally sized bins defined by the duration of the (previous) saccade and visualized the median-based ERF, either fixation onset-locked (Fig. 1D) or saccade onset-locked (Fig. 1E).

Mixing Factor Analysis. To get a data-driven estimate of the optimal event onset for investigating ERFs in natural vision, we investigated whether either fixation, saccade onset or any hypothetical event onset between or beyond those events leads to an ERF that is most powerful in explaining the single epoch variability. We therefore linearly inter- and extrapolated event onsets for each epoch. The simulated event onset can be defined as:

$$t_{alpha} = t_{fix} + \alpha \cdot (t_{fix} - t_{sacc}),$$

where t_{alpha} represents the simulated event onset, α the temporal distance with respect to fixation onset, and t_{fix} and t_{sacc} the fixation and saccade onset time, respectively. A mixing factor of $\alpha = 0$ corresponds to fixation onset (t_{fix}) of a given epoch, whereas a mixing factor of $\alpha = -1$ indicates saccade onset (t_{sacc}). A mixing factor of $-1 < \alpha < 0$ corresponds to an interpolated event onset between saccade and fixation onset. Accordingly, $-2 > \alpha > -1$ indicates an extrapolated event onset up to twice the duration of the saccade before $t_{sacc} \triangleq \alpha = -1$, and $0 > \alpha > 1$ indicates an event onset up to twice the saccade duration after $t_{fix} \triangleq \alpha = 0$.

As a measure of how well any α-based ERF could explain the single epochs, we evaluated the mean correlation (*r*) of the α-based ERF with the epoch-wise activation between ± 40 ms around each epoch's activation peak. Epochs that peaked earlier than 10 ms before, or later than 260 ms after t_{alpha}, were excluded. The correlation values between the overall ERF and each epoch were Fisher z-transformed before obtaining the average, and inverse-transformed afterwards.

We ran the mixing factor analysis separately for all gradiometers. Fig. 1F shows the averaged correlation values per α for channel MEG2123 and a weighted average over all channels (participant 4). The weights are computed as the absolute mean correlation value per channel. The density of the peak α value per channel is given in Fig. 1G (participant 4).

Saccade ERFs and Grand Average ERFs. Saccade onset-locked epochs were obtained from the MEG data. Per participant, time point and channel, we took the median across epochs (participant 4: $n_{saccades}$ = 41,756) to obtain saccade onset-locked ERFs. The grand average ERFs were obtained by taking the median ERF per channel across participants.

Residual-Locking. For each participant and sensor, the saccade onset-locked ERF for the previously selected sensor (Fig. 1J, upper row) was subtracted from the epoch-wise saccade onset-locked data. The remaining residuals were then shifted forward in time by the duration of the saccade, aligning the data with fixation onset. We obtained a residual fixation ERF by taking the median across the fixation onset-shifted residuals.

Zero-Duration Saccade ERF Extrapolation. For each participant and sensor, we extrapolated a hypothetical ERF for a saccade with a duration of zero ms. We took the shortest 25% of the 160 saccade duration bin ERFs. For each timepoint within the ERF, we trained an Ordinary Least Squares (OLS) regression model to predict the sensor activation as a function of saccade duration. The resulting model is used to predict the ERF for a hypothetical saccade with a duration of 0 ms (Fig. 2A).

Scene Onset ERF. The MEG data was epoched based on the onset of the scene (n_{scene} = 4,080). To make sure we didn't include eye movement-related data to the scene onset ERFs, the epochs were masked from the time point of the first saccade onset on. The scene onset ERF was computed by taking the median of the masked scene onset-locked epochs (Fig. 2B).

Topographic Comparison of Different Event Type ERFs. We compared the zero-duration saccade ERF, residual fixation ERF and scene onset ERF with respect to their topographic similarity. The magnetometer topographies were averaged between 90 to 100 ms after event onset (Fig. 2C, D, E), and the ERFs for the 3 different events were projected into source space (SI The MEG Eye-Tracking Dataset - Source Reconstruction). Additionally, we show the absolute source activation per source on the cortical surface. To achieve this, the individual source reconstruction was transformed into *fsaverage* (Fischl, 2012) coordinates and then projected onto the *fsaverage* flattened cortical surface map.

Correlations of Source Activations for Different Events. We correlated the source-wise surface normal activations for each pairwise comparison between zero-duration saccade ERF, residual fixation ERF and scene onset ERF. This was done for five different regions of interests (ROIs): early visual, midventral and ventral, midparietal and parietal, lateral and midlateral (as defined in Allen et al. (2022), and posterior cingulate cortex (combination of the following ROIs, as defined in Glasser et al. (2016): POS1, POS2, RSC, d23ab, v23ab, 7m, 31pv, 31pd, 31a, 23d, 23c, and PCV). The ROIs were projected into *fsaverage* coordinates and then fitted to the individual source space using spherical averaging. We averaged the source activations within a time window between 90 to 100 ms after event onset. For each ROI, the averaged source activations were correlated using Pearson correlation. An OLS regression model with bootstrapped confidence intervals was fitted for each comparison. Fig. 2F visualizes the spread of surface activations between two events for all comparisons, along with the fitted lines.

To calculate the grand average correlation values across all participants (Fig. 2G), we computed the Pearson correlation of the temporally averaged surface normal activations for each ROI, comparing each pair of the three different events individually for each participant. The correlation values were Fisher z-transformed before obtaining the average across participants, and inverse-transformed afterwards.

SI References

Allen, E. J., St-Yves, G., Wu, Y., Breedlove, J. L., Prince, J. S., Dowdle, L. T., Nau, M., Caron, B., Pestilli, F., Charest, I., Hutchinson, J. B., Naselaris, T., & Kay, K. (2022). A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nature Neuroscience*, *25*(1), 116–126. https://doi.org/10.1038/s41593-021-00962-x

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic Statistical Parametric Mapping: Combining fMRI and MEG for High-Resolution Imaging of Cortical Activity. *Neuron*, *26*(1), 55–67. https://doi.org/10.1016/S0896-6273(00)81138-1

Engbert, R., & Mergenthaler, K. (2006). Microsaccades Are Triggered by Low Retinal Image Slip. *Proceedings of the National Academy of Sciences*, *103*(18), 7192–7197. https://doi.org/10.1073/pnas.0509557103

Fischl, B. (2012). FreeSurfer. *NeuroImage*, *62*(2), 774–781. https://doi.org/10.1016/j.neuroimage.2012.01.021

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Essen, D. C. V. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615). https://doi.org/10.1038/nature18933

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG Data Analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(267), 1–13. https://doi.org/10.3389/fnins.2013.00267

Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., & Dollár, P. (2015). *Microsoft COCO: Common Objects in Context*. https://doi.org/10.48550/arXiv.1405.0312

Taulu, S., Kajola, M., & Simola, J. (2004). Suppression of Interference and Artifacts by the

Signal Space Separation Method. *Brain Topography*, *16*(4), 269–275.

https://doi.org/10.1023/B:BRAT.0000032864.93890.f9