Behavioral/Cognitive

Rapid Extraction of the Spatial Distribution of Physical Saliency and Semantic Informativeness from Natural Scenes in the Human Brain

John E. Kiat, Taylor R. Hayes, John M. Henderson, and Steven J. Luck
Center for Mind & Brain and Department of Psychology, University of California–Davis, Davis, California 95618

Physically salient objects are thought to attract attention in natural scenes. However, research has shown that meaning maps, which capture the spatial distribution of semantically informative scene features, trump physical saliency in predicting the pattern of eye moments in natural scene viewing. Meaning maps even predict the fastest eye movements, suggesting that the brain extracts the spatial distribution of potentially meaningful scene regions very rapidly. To test this hypothesis, we applied representational similarity analysis to ERP data. The ERPs were obtained from human participants ($N = 32$, male and female) who viewed a series of 50 different natural scenes while performing a modified 1-back task. For each scene, we obtained a physical saliency map from a computational model and a meaning map from crowd-sourced ratings. We then used representational similarity analysis to assess the extent to which the representational geometry of physical saliency maps and meaning maps can predict the representational geometry of the neural response (the ERP scalp distribution) at each moment in time following scene onset. We found that a link between physical saliency and the ERPs emerged first (~78 ms after stimulus onset), with a link to semantic informativeness emerging soon afterward (~87 ms after stimulus onset). These findings are in line with previous evidence indicating that saliency is computed rapidly, while also indicating that information related to the spatial distribution of semantically informative scene elements is computed shortly thereafter, early enough to potentially exert an influence on eye movements.

Key words: **attention; EEG; ERP; meaning map; representational similarity analysis; saliency**

Significance Statement

Attention may be attracted by physically salient objects, such as flashing lights, but humans must also be able to direct their attention to meaningful parts of scenes. Understanding how we direct attention to meaningful scene regions will be important for developing treatments for disorders of attention and for designing roadways, cockpits, and computer user interfaces. Information about saliency appears to be extracted rapidly by the brain, but little is known about the mechanisms that determine the locations of meaningful information. To address this gap, we showed people photographs of real-world scenes and measured brain activity. We found that information related to the locations of meaningful scene elements was extracted rapidly, shortly after the emergence of saliency-related information.

Introduction

Visually guided behavior relies on rapid prioritization of incoming visual information. The precise mechanisms by which our brains perform this prioritization, however, remain unclear. Two distinct theoretical perspectives have emerged: one emphasizing physical saliency and the other emphasizing cognitive guidance. Physical saliency theories propose that attention is drawn to locations that differ from their surroundings in low-level features (Koch and Ullman, 1985; Itti et al., 1998; Itti and Koch, 2001; Harel et al., 2007; for review, see Veale et al., 2017). By contrast, cognitive guidance theories propose that, from the very earliest viewing moments, attention is instead guided by the distribution of semantic or task-relevant content within scenes (Wolfe, 1994; Henderson, 2003, 2017; Hayhoe and Ballard, 2005).

To distinguish between these possibilities, recent studies have compared physical saliency maps (maps indicating physical saliency at each location) with meaning maps (maps indicating semantic informativeness levels at each location) (Henderson and Hayes, 2017). Although saliency is thought to have largely

Received Mar. 17, 2021; revised Oct. 6, 2021; accepted Oct. 12, 2021.

Author contributions: J.E.K., T.R.H., J.M.H., and S.J.L. designed research; J.E.K. and S.J.L. performed research; J.E.K. analyzed data; J.E.K. wrote the first draft of the paper; J.E.K., T.R.H., J.M.H., and S.J.L. edited the paper; J.E.K. and S.J.L. wrote the paper; T.R.H. and J.M.H. contributed unpublished reagents/analytic tools.

This work was supported by National Institutes of Health Grants R01MH076226 and R01MH065034 to S.J.L. and Grant R01EY027792 to J.M.H. We thank Mazze L. Whitney for assistance with data collection.

The authors declare no competing financial interests.

Correspondence should be addressed to John E. Kiat at jekiat@ucdavis.edu.

https://doi.org/10.1523/JNEUROSCI.0602-21.2021

equivalent effects on covert and overt attention, most research on models of saliency have focused on overt shifts of gaze. Several recent studies have shown that eye movement patterns are predicted better by meaning maps than by physical saliency (Henderson and Hayes, 2017; Henderson et al., 2019). This advantage has been observed across multiple tasks, including visual search (Hayes and Henderson, 2019), simple free viewing (Peacock et al., 2019a), and scene and action description (Henderson and Hayes, 2018; Rehrig et al., 2020). The predictive advantage of meaning maps is present even when the task is to count the number of physically salient scene regions (Peacock et al., 2019b).

Although one might expect that information related to meaning would be extracted relatively slowly, the explanatory advantage of meaning maps is often present from the very first saccade (Hayes and Henderson, 2019; Peacock et al., 2020). This suggests that meaning-related computations arise quickly enough to overcome physical saliency in the control of attention.

Although the representation of physical saliency emerges very rapidly in nonhuman primates (e.g., superior colliculus: 65 ms, White et al., 2017; V1: 90-100 ms, Li et al., 2006), less is known about when the brain determines which regions are likely to contain meaningful objects. Here, we considered two competing possibilities. First, information related to semantic informativeness might be extracted substantially later than physical saliency-related information, reflecting the additional computations involved in computing meaning. For example, previous research suggests that it takes the human brain ~150 ms to complete the extraction of meaning from complex scenes (Thorpe et al., 1996; Fabre-Thorpe et al., 2001; Gordon, 2004). However, it may be sufficient to determine that a location is likely to contain meaningful information before shifting covert or overt attention to that location, which is presumably faster than computing the meaning itself at that location. This raises the alternative hypothesis that brain extracts the locations of semantically informative regions almost as rapidly as it extracts saliency-related information.

To distinguish between these alternatives, we assessed the onset of information related to physical saliency maps and meaning maps in neural responses elicited by photographs of real-world scenes, leveraging the high temporal precision of ERPs. Subjects viewed a series of 50 different scenes while performing a modified 1-back task (Fig. 1). Because we were examining the processes that precede covert and overt shifts of attention, and because eye movements create large electrical artifacts, subjects maintained central fixation throughout the task. We used representational similarity analysis (RSA) to link the ERP scalp distribution at each moment in time with computationally generated maps of physical saliency (Harel et al., 2007) and crowd-sourced meaning maps (Henderson and Hayes, 2017) (Fig. 2). We predicted that representational similarity between the ERPs and meaning maps would arise shortly after representational similarity between the ERPs and physical saliency maps, consistent with the fact that the fastest eye movements are better explained by semantic informativeness than by physical saliency.

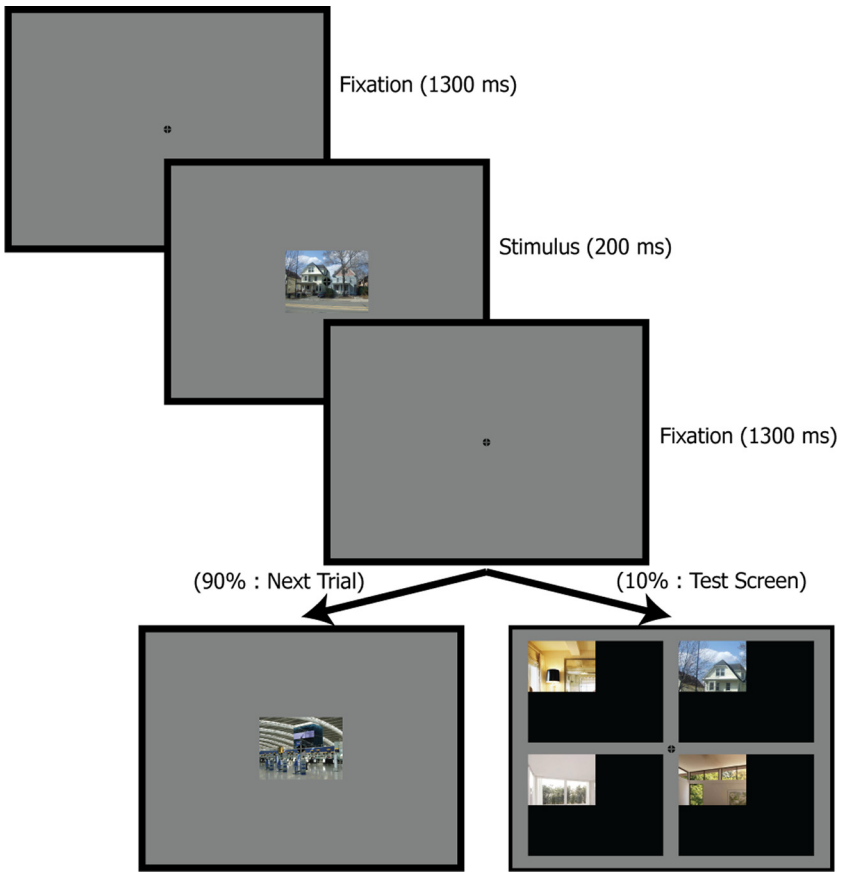

Figure 1. Example stimulus sequence from the experimental task. The target stimuli were photographs of real-world scenes presented at the center of the screen. Subjects were tested on a randomly selected 10% of trials. In each test, four quadrants from four different images (one matching the immediately preceding scene, three selected at random from the other task scenes).

Materials and Methods

Participants. Thirty-two college students (17 female, 15 male; 18-30 years of age) with normal or corrected-to-normal visual acuity participated in this study for monetary compensation. Given the small number of prior ERP studies using RSA, and the lack of any ERP RSA studies of this specific issue, it was difficult to conduct a conventional power analysis to determine an appropriate sample size. Instead, we made an initial choice of $N = 32$ by doubling (out of an abundance of caution) the $N = 16$ used in prior ERP studies from our laboratory using other multivariate pattern analysis methods (Bae and Luck, 2018, 2019). We then conducted simulations (10,000 runs; for code, see https://osf.io/zg7ue/) using this $N = 32$ sample size and found that we could detect 80% of the time points exhibiting a significant effect in our target analysis with this target N. Specifically, we created simulated data with a small representational similarity effect ($r = 0.05$) that extended for 100 ms within a 500 ms analysis window. On average, our method was able to detect significant effects for 80% of the time points within this 100 ms period after correcting for multiple comparisons over the 500 ms analysis window (using the analytic approach described in Statistical analysis). From this, we concluded that $N = 32$ was sufficient for the present study. All study procedures were approved by the University of California-Davis Institutional Review Board.

Stimuli and task. All task elements were presented in MATLAB (The MathWorks) using PsychToolbox (Brainard, 1997; Pelli, 1997; Kleiner et

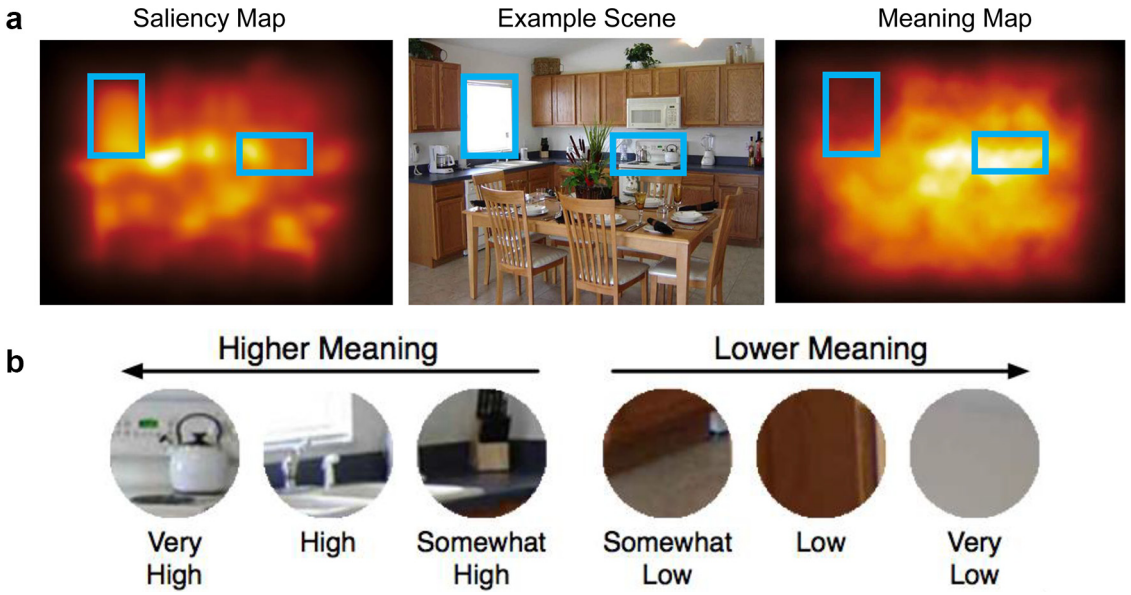

Figure 2. ***a***, Example scene along with its corresponding physical saliency map and meaning map. The blue rectangles were not present in the scene but were added here to highlight specific regions in each map type. In this example, the region highlighted on the left is high in physical saliency (being brighter than the surrounding regions) but low in meaning (being largely homogeneous). By contrast, the region highlighted on the right is high in semantic saliency (as it contains easily identifiable objects) while being relatively low in physical saliency. ***b***, Examples of the patches used to construct the meaning maps and their ratings. Subjects viewed and rated the meaningfulness of each individual patch in isolation.

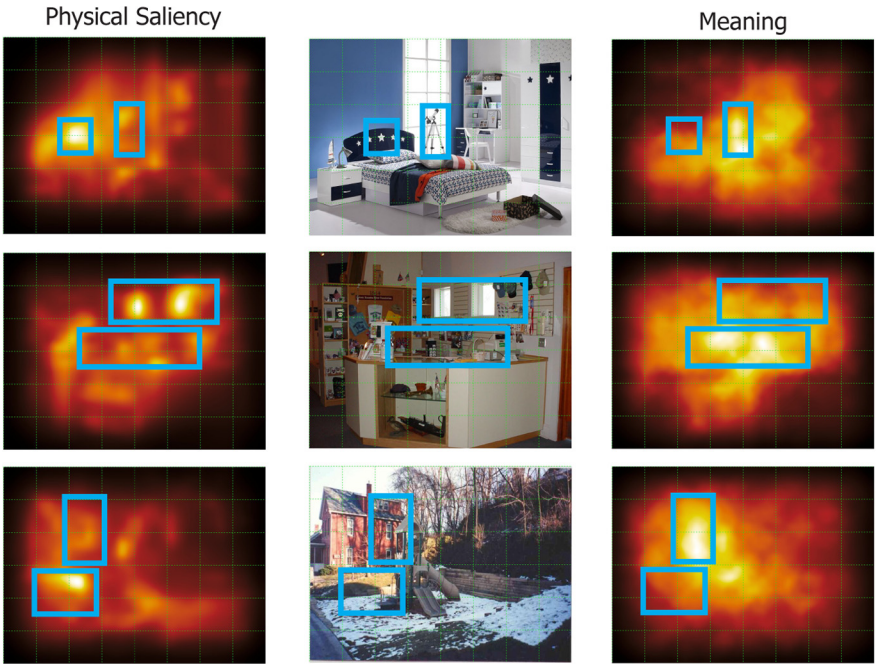

Figure 3. Examples of three scenes used in the present study and their corresponding physical saliency maps and meaning maps. The blue rectangles and grid were not present in actual scenes but were added here to highlight correspondences between the maps and the scenes.

al., 2007) (all scene images are available at https://osf.io/ptsvm/). The stimuli were presented on an LCD monitor (HP ZR2440W) with a gray background (31.2 cd/m^2) at a viewing distance of 100 cm. The monitor presentation delay was measured (24 ms), and all timing values were adjusted accordingly.

The experimental task is illustrated in Figure 1. Throughout the task, an empirically optimized fixation symbol (Thaler et al., 2013) was continuously present in the middle of the screen. Because the goal of the present study was to examine processes that precede both covert and overt shifts of attention, subjects were instructed to maintain central fixation on this symbol throughout the task.

The primary stimuli consisted of 50 digitized photographs of real-world scenes (Henderson and Hayes, 2017, 2018; Henderson et al., 2020). Sample images are presented in Figures 2 and 3, and all 50 of the images are available in a public repository (https://osf.io/ptsvm/). Mean local contrast energy and spatial coherence statistics (Groen et al., 2013) for these images were also calculated and are provided in the repository. Each image subtended 8 × 6 degrees of visual angle. On each trial of the task, a scene was presented for 200 ms, followed by a 1300 ms interstimulus interval. The brief stimulus duration was designed to discourage eye movements.

To promote general attentiveness, subjects were instructed to remember the most recent scene, and their memory was tested after a randomly selected 10% of trials. As illustrated in Figure 1, each test display contained four options: one matching the immediately preceding scene and the others selected at random from the other 49 scenes. Each option consisted of one quadrant of a given scene so that subjects could not perform the task by focusing on a narrow region when encoding the scenes. For example, if participants focused narrowly on the center of the house scene in Figure 1, they would have difficulty determining which of the four test options matched this scene. The position of the selected quadrant was selected at random, with all four options being extracted from the same quadrant (e.g., all four from the upper left quadrant of the scenes). Each option subtended 3.2 × 2.4° and was embedded within a 12.9 × 9.7° black region. The position of a given option relative to the black background corresponded with the quadrant's position in the original scene. Subjects were instructed to indicate which of the four options matched the immediately preceding scene by pressing one of four trigger buttons on a gamepad with the index and middle fingers of the left and right hands, mapped to the corresponding four locations in the test display. This task was designed to minimize task-based categorization or response-related

activity until the test display so that these factors would not influence the RSA results.

Before the main task began, subjects were required to achieve at least 75% accuracy in a 50 trial practice block. This block was repeated until the required performance level was achieved. The images in this practice block were not used in the main task.

In the main task, the trials were divided into a series of 32 blocks, each containing one trial with each of the 50 scenes. Thus, each of the 50 scenes was presented 32 times for each subject, yielding a fully balanced within-subjects design. Image presentation order was randomized within each block with the restriction that the last image in one block could not be the first image in the next block. Participants were given a break after every block.

Generation of physical saliency maps and meaning maps. The term saliency can be defined in different ways. Following in the tradition of Koch and Ullman (1985) and Itti and Koch (2000, 2001), the present study uses the phrase physical saliency to refer to information about saliency that is computed by early visual cortex on the basis of low-level physical features in the sensory input. With this aim in mind, we selected the Graph-Based Visual Saliency (GBVS) Toolbox as our model of physical saliency given its biological plausibility (Harel et al., 2007) along with its track record of performance (Walther and Koch, 2006; Nuthmann et al., 2017).

We applied the GBVS algorithm to our scenes using the default parameter settings (saliency map size: 32; selected channels: color, orientation, intensity; Gabor angles: 0, 45, 90, 135; contrast width: 0.10; blur fraction: 0.02). The GBVS method first extracts low-level color, orientation, and contrast features, vectors from an image using biologically inspired filters. These features are then used to compute activation maps for each unique feature type. Subsequently, these activation maps are normalized and additively combined to form a single global saliency map. Finally, the resulting map is blurred using a Gaussian kernel.

Akin to how physical saliency maps represent physical saliency at each location in an image, meaning maps aim to quantify the extent to which meaningful information is present at each location. Meaning maps for the scenes used in the present study were previously generated by Henderson and Hayes (2017). In the map generation process, each scene (768 × 1024 pixels) was decomposed into a series of partially overlapping and tiled circular patches at two spatial scales (fine: patch diameter of 87 pixels, 300 patches per scene; coarse: patch diameter of 205 pixels, 108 patches per scene; for full details, see https://osf.io/suzex/). An example scene and patches from that scene are shown in Figure 2.

These patches were then evaluated by 204 Amazon Mechanical Turk subjects who rated how informative or recognizable each patch was on a 6 point Likert scale. The subjects viewed each patch in isolation, without any context (e.g., they never saw the intact scenes). The patches from multiple scenes were intermixed and presented in random order. Thus, the ratings reflect the extent to which a given patch contains meaningful information, not the specific meaning of that patch or the relationship of that patch to the rest of the scene. Each unique patch was then rated by three unique raters. Given the substantial spatial overlap between patches, any given point in a scene typically received dozens of ratings. A meaning map was generated for each scene by averaging the rating data at each spatial scale separately at the pixel level, then averaging the spatial scale maps together, and finally smoothing the average rating map with a Gaussian filter (i.e., σ = 10, FWHM= ~23 px; for the image processing code, see https://osf.io/654uh/).

Given that images in this study were centered on the target fixation point, it is important to account for the expected center bias in the processing of each scene in both the GBVS and meaning maps. Maps from the GBVS model are intrinsically center-biased, with the center-bias being an emergent property of the distribution of graph node locations used to compute the image maps (Harel et al., 2007). To implement the same center-bias weighting to the meaning maps, the center-bias weights included in the GBVS package were applied to the meaning maps via pointwise multiplication. Through this procedure, the weighting of map features across the two map types was effectively standardized. Examples of these maps are shown alongside the original scene images in Figure 3.

EEG recording and preprocessing. Continuous voltages were recorded from 64 electrodes using a Brain Products ActiCHamp recording system (Brain Products). Electrodes were located at a broad set of 59 scalp sites (AF3, AF4, AF7, AF8, FC1, FC2, FC3, FC4, FC5, FC6, FP1, FP2, F1, F2, F3, F4, F5, F6, F7, F8, C1, C2, C3, C4, C5, C6, CP1, CP2, CP3, CP4, CP5, CP6, P1, P2, P3, P4, P5, P6, P7, P8, P9, P10, PO3, PO4, PO7, PO8, T7, T7, TP7, TP8, O1, O2, Fz, FCz, Cz, CPz, Pz, POz, and Oz), at the left and right mastoids, and at three electrooculogram (EOG) sites. The two horizontal EOG electrodes were placed lateral to the external canthi and were used to record horizontal eye movements; the vertical EOG electrode was placed below the right eye and was used to record eyeblinks and vertical eye movements (for a comprehensive description of the electrode application and recording procedures, see Farrens et al., 2020). Electrode impedances were maintained at $<15\ k\Omega$. All signals were recorded single-ended with a customized version of the PyCorder EEG recording software and then referenced offline. The EEG was filtered online with a cascaded integrator-comb antialiasing filter (half-power cutoff at 130 Hz) and digitized at 500 Hz.

The EEG preprocessing began by referencing the scalp EEG to the average of the left and right mastoid sites. A bipolar horizontal EOG derivation was then computed as the difference between the two horizontal EOG electrodes, with a vertical EOG derivation computed as the difference between Fp2 and the vertical EOG electrode. All the signals were then bandpass filtered (noncausal Butterworth impulse response function, DC offset removed, half-amplitude cutoffs at 0.1 and 30 Hz, 12 dB/oct roll-off), and resampled at 250 Hz. Portions of EEG containing large muscle artifacts or extreme voltage offsets (identified by visual inspection) were removed.

Independent component analysis (ICA) was then performed on the retained continuous EEG for each subject to identify and remove components that were associated with blinks (Jung et al., 2007) and eye movements (Drisdelle et al., 2017). The criterion for excluding an ICA component was the consistency between the shape, timing, and spatial location of the component compared with the HEOG and VEOG signals. The data for each channel (excluding HEOG and VEOG) were then reconstructed from the other ICA components. Individual trials were rejected if the peak-to-peak voltage was $>200\ \mu V$ in any 200 ms window in any electrode, or if a blink or eye movement (defined as a step-like voltage change) (see Luck, 2014) was detected in the uncorrected HEOG or VEOG signals between 200 ms before stimulus and 200 ms after stimulus (and might therefore impact the perception of the stimulus).

The ICA-corrected EEG signals were then segmented for each trial from −500 to 1500 ms relative to the onset of the target scenes. Epochs preceded by test trials were discarded to reduce trial-by-trial variability. The retained trials were then baseline-corrected using the mean voltage from −500 to 0 ms, and the averaged ERP waveform was computed for each of the 50 scenes. The key experimental effects occurred within 200 ms of stimulus onset, minimizing any concern that residual EOG activity or secondary effects of eye movements might have impacted the results. Moreover, there was little motivation for subjects to move their eyes in a scene-dependent manner because the task required perceiving the entirety of each scene, the scenes were centered at the fixation point, and the scenes terminated after 200 ms. Additional analyses are provided in Results to demonstrate that eye movements had little or no impact on the RSA findings.

Statistical analysis. RSA (Kriegeskorte et al., 2008) was used to link the ERPs with the physical saliency and meaning maps. RSA makes it possible to compare multiple distinct measurement spaces for a set of stimuli, in this case ERP topographies and the spatial distribution of saliency and meaning. RSA is widely used in fMRI research to assess the correspondence between computational models and the pattern of activation across voxels, and it solves the fundamental problem of linking data modalities that have intrinsically different measurement spaces. Specifically, the RSA approach abstracts away from the activity patterns themselves to compute, for each measurement modality of interest, representational similarity matrices (RSMs) (or complementary representational dissimilarity matrices), which represent the overall pattern of similarity observed between the activity patterns produced by a set of stimuli.

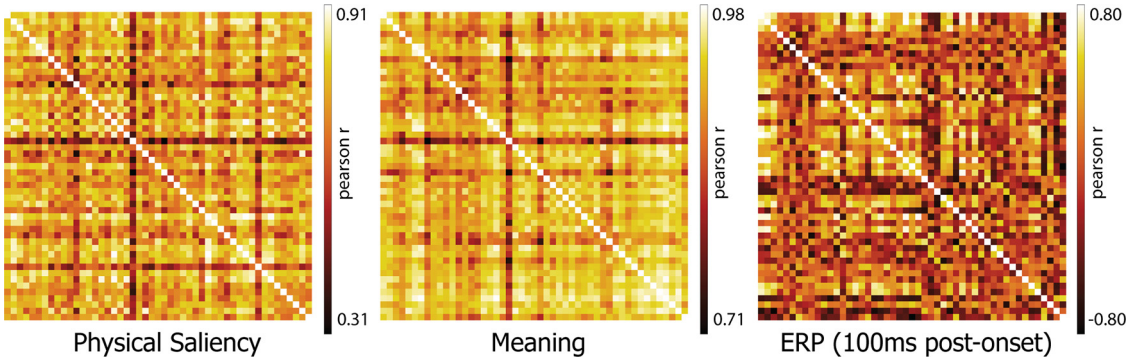

Figure 4. RSMs for the physical saliency and meaning maps, as well as an example RSM from the ERP data (drawn from a single subject from 100 ms after stimulus onset). The ordering of items in each matrix is identical and corresponds to the arbitrary numbering assigned to each scene in the experimental task. Similarity values are presented in Pearson *r* units. Shading in each cell represents the computed similarity between a pair of scenes.

For example, if we had conducted an fMRI experiment in which we showed an observer 50 different scenes, we could take the pattern of BOLD activation across the voxels in visual cortex for each scene and then compute the correlation between the pattern of activation for each pair of scenes. This would yield a 50×50 correlation matrix. A matrix calculated in this manner is termed an RSM because each cell of the matrix indicates the similarity between the representations of a given pair of stimuli. We could then take those same scenes and construct another 50×50 correlation matrix in which each cell contains the correlation between the maps from a given pair of images (correlated at the pixel-by-pixel level). Each of these RSMs would provide information about the representational geometry of the system that produced it. In other words, each RSM indicates how similar or dissimilar different scenes are with regard to (1) the fMRI voxel patterns they evoke and (2) the outputs they produced in the model.

By assessing the relationship between these two RSMs (using a rank-order correlation to avoid assuming linearity), we could assess the relationship between the representational geometry of the two measurement spaces. In other words, the correlation between the RSMs indicates the degree to which stimuli evaluated as being similar/dissimilar in one system are considered similar/dissimilar in the other. Only the lower or upper triangle of each RSM is used in computing this correlation (because the upper and lower triangles are mirror images of each other, and the cells along the diagonal always have values of 1). A rank-order correlation was selected here as it provides a robust method that does not depend on the assumption of a linear relationship between the true similarities produced by the systems underlying the RSMs (Diedrichsen and Kriegeskorte, 2017). Furthermore, there is reason to believe that a monotonic transform best accounts for the expected effect of the activity-pattern noise of given system on its RSM (Kriegeskorte et al., 2008).

In the present study, we used the voltage pattern across electrode sites at a given latency rather than the pattern of BOLD activation across voxels to construct the neural RSMs. This provides much better temporal resolution because the EEG is a measurement of the actual extracellular potentials produced by the neurons, with zero delay (but spatially blurred by the brain, meninges, skull, and scalp). We applied RSA to the averaged ERPs in a three-step process. First, for the ERP data from a given subject, a separate RSM was computed at each moment in time relative to stimulus onset. Each cell in one of these RSMs represents the similarity in scalp distribution between the ERPs elicited by two of the scenes at that moment in time. Second, RSMs were computed for the saliency and meaning maps; these RSMs were identical across subjects. Third, the relationship between the ERP and saliency/meaning RSMs was estimated using rank regression, separately for each subject. All reported *p* values are two-tailed unless otherwise specified.

To take full advantage of the ERP technique's temporal resolution, we computed a separate ERP RSM at each time point for each subject. We began by taking the scalp distribution of the averaged ERP for a given scene and storing it as a vector of 59 voltages (i.e., a list with one voltage for each electrode). To represent the similarity between the scalp distributions for two scenes at a given time point, we computed the Pearson *r* correlation between the ERP scalp distribution vectors for those two scenes. This was done separately for each pair of scenes, yielding a 50×50 RSM for each time point. These computations were performed independently at each time point for each subject to avoid representational geometry distortions associated with the averaging of distance data (Ashby et al., 1994). We repeated this process for each time point in the 2000 ms ERP epoch (500 time points at 4 ms per sample), producing 500 different 50×50 ERP RSMs for each subject. We used the Pearson *r* correlation between scalp distributions as our measure of similarity because it quantifies similarity in the spatial pattern of the scalp distribution, disregarding differences in overall amplitude.

For the saliency map RSMs, we began by reshaping each two-dimensional saliency map (one saliency value per pixel) into a single 1-dimensional vector (list) of saliency pixel values. We then computed the Pearson *r* correlation between the vectors for a given pair of scenes to represent the saliency-based similarity between those scenes. This yielded a 50×50 saliency RSM. This process was repeated for the meaning maps to produce a 50×50 meaning map RSM. Figure 4 presents the physical saliency and meaning map RSMs alongside an example ERP RSM from one recording time point from a single subject.

When choosing the 50 scenes for this study, we intentionally selected scenes in which the physical saliency and meaning maps were not overly similar (Pearson $r < 0.50$). To assess the degree of correlation between the saliency and meaning RSMs, we computed the Spearman ρ rank-order correlation between them, with a permutation test (10,000 iterations) to assess statistical significance. Consistent with our selection criteria, we found that the physical saliency and meaning map RSMs were only modestly correlated ($\rho = 0.243$, $p = 0.007$).

To assess the link between the ERPs, physical saliency maps, and the meaning maps, a rank regression procedure (Iman and Conover, 1979) was used to regress the ERP RSMs onto the physical saliency and meaning map RSMs. The estimated parameters of interest were the semipartial correlations between the rank-ordered ERP RSMs and the rank-ordered saliency and meaning RSMs. These semipartial correlations quantify how much of the variance in the representational geometry of the neural activity (the ERP RSMs) is uniquely accounted for by the representational geometry of the saliency and meaning maps (the saliency and meaning RSMs). In other words, variance in the ERP RSM that was explained by the physical saliency map RSM was partialled out when examining the correlation between the ERP RSM and the meaning map RSM, and vice versa. This approach allows us to examine the unique representational contribution of each source of information with regard to the ERP response. This procedure was repeated independently for each of the 500 time points for each of the 32 unique subjects, resulting in two sets of separate 32 semipartial correlations (one for each subject) at each time point for physical saliency maps as well as for meaning maps.

We used parametric (Pearson) correlations to assess the similarity between the scalp distributions when constructing the RSMs because different scalp distributions from the same subject can be directly related to each other. However, we used nonparametric (rank-order) correlations

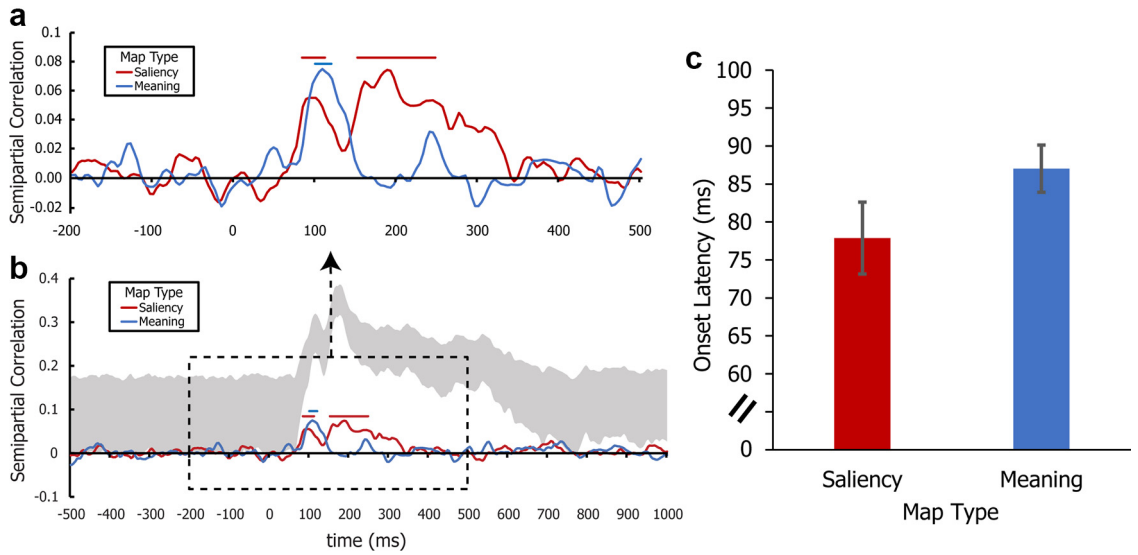

Figure 5. ***a***, Representational similarity time course between the ERP data (i.e., the ERP representational dissimilarity matrices computed from all scalp electrodes) and each of the two map types (saliency and meaning) from −200 ms before stimulus to 500 ms after stimulus. Representational similarity was computed separately for each participant, with the mean across participants being shown here. Horizontal line segments across the top indicate time periods in which the representational similarity values were significantly >0 ($p < 0.05$ correcting for the FDR). ***b***, Full time course of the representational link shown in ***a***. The upper and lower edges of the gray region denote the upper and lower bound estimates of the estimated noise ceiling (i.e., the highest expected observed correlation) of the ERP data. ***c***, Jackknifed mean onset latency for physical and semantic saliency in the ERP data. Error bars indicate the jackknife-corrected SE.

to assess the similarity between the ERP and saliency RSMs because these RSMs come from different sources of data that may not be linearly related.

Given that our hypotheses focused on sensory/perceptual activity, our analytic window focused on the 500 ms period following stimulus onset. Negative RSA correlations are typically uninterpretable and were treated as noise. We therefore used a one-tailed Wilcoxon sign-rank test against zero to determine whether the average of the 32 single-subject semipartial correlations at a given time point within the analytic window was significantly >0. This was done separately for the physical saliency and meaning map RSMs. A FDR correction ($q = 0.05$) was then applied to each set of saliency and meaning p values as an adjustment for multiple comparisons (Benjamini and Yekutieli, 2001).

Data and code accessibility. All EEG preprocessing methods were implemented in MATLAB using the open-source EEGLAB (Delorme and Makeig, 2004) and ERPLAB (Lopez-Calderon and Luck, 2014) toolboxes. The EEG data, experimental control scripts, EEG preprocessing scripts, and in-house custom MATLAB functions that implement the RSA analyses are available at https://osf.io/zg7ue/ whereas the meaning maps used in this study are available at https://osf.io/ptsvm/.

Results

Behavioral results

The mean accuracy across subjects for the behavioral task was 86% (SD = 9.68), with a mean response time of 2.22 s (SD = 0.72).

Representational similarity time course analyses

Figure 5*a* shows the representational similarity (semipartial rank correlations) between the ERP data, the physical saliency map RSM, and the meaning map RSM within the analytic window. Values that were significantly greater than chance (after correction for multiple comparisons) are indicated using horizontal lines. In general, the representational similarity values rose above chance rapidly after scene onset, with a slightly later onset for meaning than for saliency. The representational similarity was significantly above chance from 84 to 112 ms for saliency and from 100 to 120 ms for meaning. Saliency also exhibited a second period of significant representational similarity from 152 to 248 ms. We would like to stress that these are semipartial correlations, in which variance explained by saliency maps was partialled out of the meaning map values and vice versa. Thus, the data in Figure 5 reflect the unique contribution of each map type.

Figure 5*b* presents the RSA results for the full epoch, along with the noise ceiling, which reflects the highest representational similarity values that would be expected given the noise in the ERP data. The lower and upper bounds of the noise ceiling were estimated independently for each time point using the technique described by Nili et al. (2014). Specifically, the upper bound was estimated by computing the correlation between a given subject's ERP RSM at a specific time point and the grand average of the ERP RSMs across all subjects at that time point and then averaging the correlations across subjects. The lower bound was estimated using a similar approach, except that the grand average RSM used for the correlation with a given subject excluded that subject.

Onset latencies for the RSA waveforms were estimated using the fractional onset latency technique (Hansen and Hillyard, 1980; Luck, 2014) in ERPLAB. In this approach, the peak value is first determined to normalize for differences in magnitude. Then, the onset latency is defined as the time point at which the value reaches 50% of the peak value. Simulations have shown that this approach provides an accurate and precise metric of onset latency (Kiesel et al., 2008). These measurements were obtained using the jackknife approach (Miller et al., 1998; Ulrich and Miller, 2001; Kiesel et al., 2008). Because jackknifing increases precision, a spline interpolation algorithm was applied to the RSA waveforms to provide a 1 ms measurement precision (Luck, 2014).

Figure 5*c* presents the estimated onset latencies for the physical saliency map RSA waveform (mean = 77.88 ms, SE = 3.10) and the meaning map RSA waveform (mean = 87.02 ms, SE = 4.25). We then compared the onset latencies of these peaks using a jackknife-adjusted paired t test (Miller et al., 1998). The observed difference in latency (mean difference = 9.14 ms, SEM = 3.96) between these onset latencies was statistically

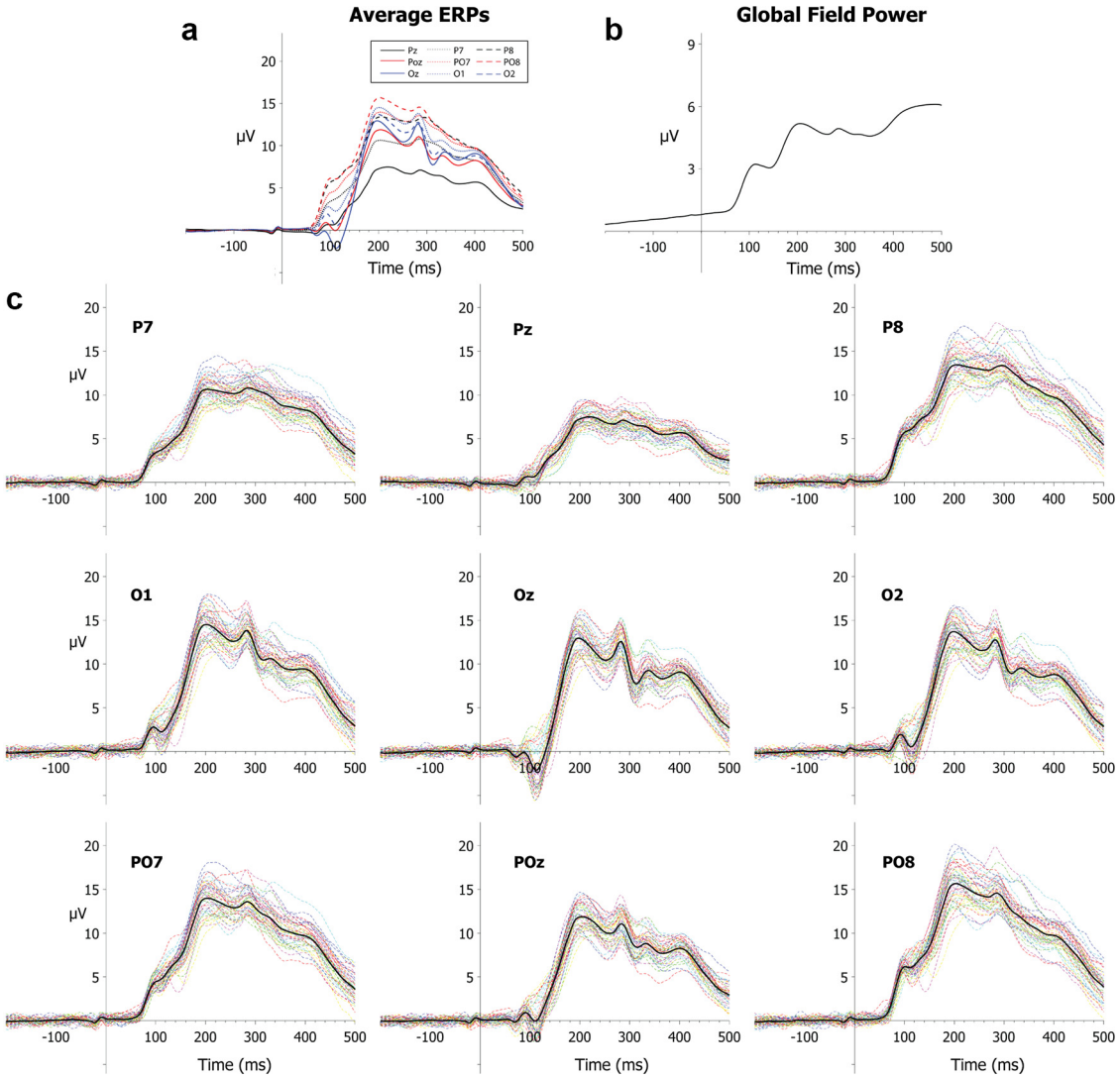

Figure 6. ***a***, Grand average ERPs, collapsed across scenes and subjects, at nine different electrode sites. ***b***, Mean GFP (Skrandies, 1990) collapsed across scenes and subjects. Time zero is the onset of the scene in all waveforms. ***c***, Grand average ERPs for all scenes, collapsed across subjects at nine different electrode sites. The grand average waveform is highlighted in black in each panel.

significant (jackknifed-adjusted $t = 2.31$, $p = 0.027$). Thus, information related to the spatial distribution of physical saliency within the scenes was present in the neural responses quite early, followed ~10 ms later by information related to the spatial distribution of meaningful scene features.

To put these latencies into context, Figure 6 shows the ERP waveforms from a set of representative electrode sites over visual cortex. The earliest ERP responses began at ~75 ms, which was close to the onset time of the representational similarity waveform for physical saliency. Thus, physical saliency-related information was present from near the beginning of the cortical activity that could be detected on the scalp, with meaningfulness-related information following rapidly.

Ruling out eye movement confounds

Although subjects were instructed to maintain central fixation, and we rejected trials with clear eye movements and used ICA to correct for any remaining eye movements, it is possible that some small eye movements escaped rejection and correction, varying systematically across scenes. The initial RSA effects were too early to have been a result of such scene-driven changes in eye position, but the later effects may have been influenced by eye movements.

To assess this possibility, we repeated our analyses with ERP RSMs computed in two ways. First, we computed RSMs using the ICA-corrected HEOG and VEOG channels that were excluded during the construction of the original ERP RSMs (Fig. 7*a*). If residual EOG activity that survived correction drove the main RSA results, then these RSA results should be even clearer if we limit the RSA analyses to the channels where these signals are largest. Second, we computed RSMs from ERPs reconstructed using only the ICA components associated with ocular activity (i.e., the components that were removed from the data during the preprocessing phase), eliminating all sources of brain activity captured by the other ICA components (Fig. 7*b*). These ocular ICA components should isolate eye movement signals, allowing us to see if they contain information that can be predicted from the physical saliency and meaning maps. As shown in Figure 7, neither of these ocular RSMs exhibited a statistically significant relationship to the physical saliency or meaning map RSMs during the analytic time window. Thus, it is unlikely that the main ERP RSA results were substantially influenced by eye movements.

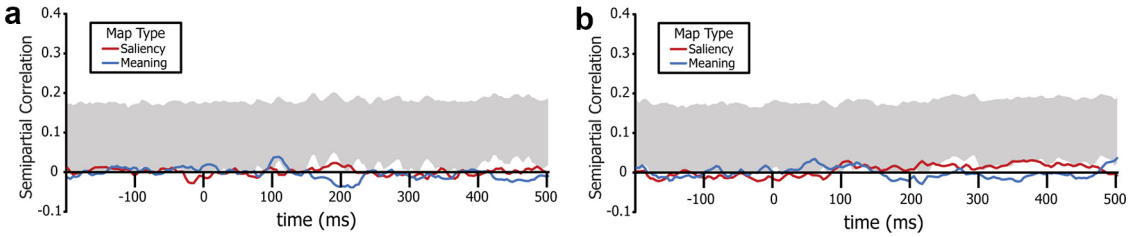

Figure 7. ***a***, Representational similarity time course from −200 ms before stimulus to 500 ms after stimulus between each of the map types (saliency and meaning) and the ERP data, but using only the horizontal and vertical EOG channels (after artifact correction). This makes it possible to estimate the effects of any residual eye movement activity. ***b***, Representational similarity time course when the ERP data at each electrode site were computed from the independent components flagged as being generated by ocular artifacts. This makes it possible to assess whether eye movements were systematically related to the physical saliency and meaning maps. ***a***, ***b***, Representational similarity was computed separately for each participant, and the mean across participants is shown here. The upper and lower edges of the gray region represent the upper and lower bound estimates of the estimated noise ceiling (i.e., the highest expected observed correlation).

Topographical assessment

The RSA time course analyses used all 59 scalp sites, providing no information about which sites were most important in producing the observed effects. To obtain information about scalp topography, we used a leave-one-electrode-out approach. Specifically, we repeated the rank regression procedure 59 times, each time leaving out one of the 59 electrodes when constructing the ERP RSMs. We then examined how much the representational similarity values dropped when a given electrode was excluded relative to when all 59 electrodes were included, using the magnitude of the drop as a metric of the contribution of that electrode to the representational similarity link. As before, these changes were computed at the single-subject level before being averaged across participants. Given the large number of electrodes and the high intercorrelations between them, we expected that the contribution from each individual electrode would be small but that the pattern across electrodes would be informative. These values were used to construct scalp topographies in which the value at each electrode was set to the magnitude of its unique contribution to the rank-order relationship between a given map type and the ERP RSM at a single point in time. A biharmonic spline interpolation was applied to all maps to facilitate visualization. This leave-one-out approach was used to produce topographic maps at each time point between 72 and 108 ms, encompassing the earliest ERP components as well as the initial portion of the physical saliency and meaning map RSMs.

As shown in Figure 8, the resulting topographies had a focus over visual cortex for both the saliency and meaning RSA data, with a slight lateralization to the right. Importantly, neither of these topographies included a substantial contribution from electrodes close to the eyes. Given the complex relationship between ERP generator locations and scalp electrodes (Nunez et al., 2006), and the large number of steps between the data and these topographic maps, it would be inadvisable to draw any strong conclusions from the topographies shown in Figure 8. Nonetheless, they do provide some descriptive information about the electrode sites that contributed most to the observed RSA effects, showing that electrodes over posterior midline and right lateral regions of visual cortex played a relatively strong role. Moreover, electrodes near the occipital pole showed a strong effect for the meaning map RSA data, suggesting that posterior visual areas may have played an important role in the meaning map RSA effects (Henderson et al., 2020). This is consistent with recent fMRI evidence indicating that information coded by GBVS maps is more strongly represented in occipital cortex and information related to meaning maps is more strongly represented in more anterior visual areas (Henderson et al., 2020).

Supplementary analyses

Our main analyses used semipartial correlations to assess the unique ability of the saliency maps to predict the ERP data after partialling out variance explained by the meaning maps and the unique ability of the meaning maps to predict the ERP data after partialling out variance explained by the saliency maps. However, it is possible that saliency and meaning also interact, which would not be captured by our primary analyses. To assess this possibility, we conducted an exploratory analysis in which we added an interaction term for the two map types to our rank regression analysis. This was calculated by assessing the effect size and significance of an interaction term computed via multiplying the centered meaning and saliency map RSM rank values. This analysis showed no evidence of an interaction between the two map types at any point in time, with near-zero values at all time points (peak $\rho = 0.017$) and no statistically significant effects after FDR correction.

In our main analyses, we used the Pearson r correlation coefficient to quantify the similarity in ERP scalp distributions for each pair of scenes, which assesses similarity in the pattern of activity over the scalp independent of the amplitude of the ERP response. We therefore conducted an additional analysis to determine whether the overall amplitude of the ERP response has a representational link to the saliency and meaning maps. In this analyses, we quantified the magnitude of the ERP response as the global field power (GFP), which is the SD of the voltage across electrode sites (Skrandies, 1990). This approach aggregates the data from all electrode sites into a single magnitude value at each time point. The difference in GFP between a given pair of scenes was used to quantify the (dis)similarity between the scenes. We used these values to construct a representational dissimilarity matrix, and then we reversed the ranks in this matrix to create an RSM. We then assessed the relationship between this RSM and the saliency and meaning RSMs at each time point for each participant using the same methods as in our main analyses. This metric of electrode-independent neural response magnitude was not clearly associated with either meaning or saliency, with no statistically significant effects at any time point after FDR adjustment. Thus, whereas the pattern of voltage over the scalp was clearly linked to both the saliency and meaning maps in the main analysis, we found no evidence for a link with the overall ERP magnitude.

Finally, we also conducted a supplementary analysis to assess the extent to which the spatial distribution of low versus high

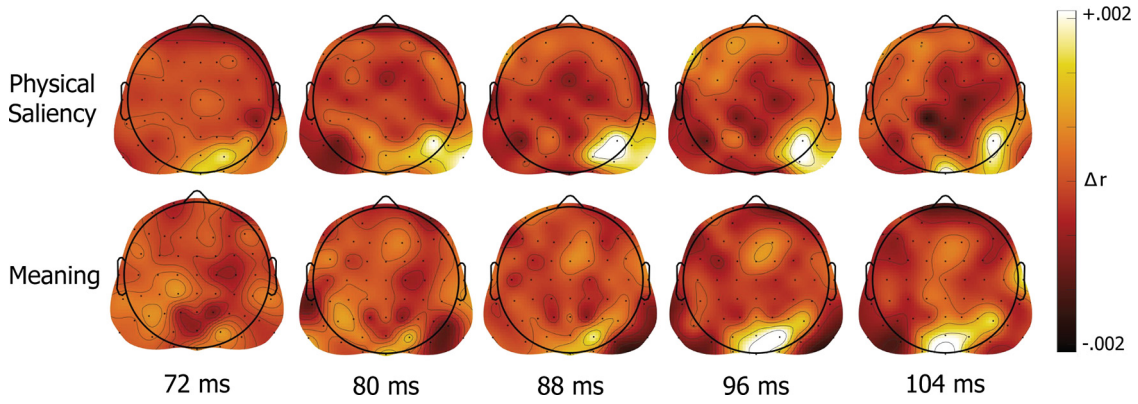

Figure 8. Scalp topographies derived from the leave-one-electrode-out analysis. Each scalp map presents the unique contribution (in rank correlation units) of each scalp electrode to the representational relationship between a given saliency map type and the neural RSM at a specific time point.

spatial frequencies played a role in the observed results. In this analysis, we created versions of the images that contained either only the low spatial frequencies or only the high spatial frequencies. Specifically, in line with the approach used by Dima et al. (2018), we converted the images to grayscale and applied either a low-pass Gaussian filter with a 3 cycles/degree cutoff (computed based on the visual angle of the scenes as presented in the task) or a high-pass filter with a 6 cycles/degree cutoff. Then, as was done with the saliency and meaning maps, we correlated the resulting maps at the pixel-by-pixel level to form a low spatial frequency RSM and a high spatial frequency RSM. We then compared these RSMs with the ERP RSMs using the same methods as in our primary analyses. This analysis yielded no link between these two RSMs and the ERP RSM at any point in time, with no significant time points after FDR correction (peak $r = 0.014$). Thus, similarity in the spatial distributions of low versus high spatial frequency information in the images does not appear to impact the similarity of ERP scalp distributions, at least for the set of images used in the present study.

Discussion

The results of this study indicate that the human brain rapidly extracts information associated with the spatial distribution of meaningful scene features shortly after the onset of cortical activity, almost as rapidly as it extracts information associated with physical saliency. These findings are consistent with the hypothesis that physical saliency is available first, but information about the locations of potentially meaningful scene regions is available soon after. To the best of our knowledge, the current study is the first to compare the processing time course of physical saliency and meaning-related information for natural scenes. This result extends prior work showing that the spatial distribution of physical saliency is rapidly represented in the frontal eye fields, in early visual cortex, and in parietal cortex (Gottlieb et al., 1998; Bogler et al., 2011; Henderson et al., 2020).

The rapid extraction of meaning-related information is consistent with behavioral work showing that the spatial distribution of meaningful scene features exerts an influence on even the initial shift of overt attention in real-world scenes (Henderson and Hayes, 2017, 2018; Hayes and Henderson, 2019; Peacock et al., 2020; Rehrig et al., 2020). Specifically, the 86 ms onset latency of the meaning-related activity observed in the present study is sufficiently fast to potentially influence even the earliest shifts of overt attention (Thorpe et al., 1996; Fabre-Thorpe et al., 2001; Gordon, 2004).

However, it is important to note that participants maintained central fixation throughout the present task (to avoid electrooculographic artifacts). This makes it impossible to determine whether the observed ERP effects play a causal role in shifts of overt attention (which would be difficult to ascertain even if eye movements were allowed). Nonetheless, these results clearly demonstrate that the brain extracts information that is predictive of semantic features sufficiently rapidly to guide scene-related eye movements.

More broadly, these RSA results indicated that similarities in scalp voltage patterns across scenes are associated with similarities in physical saliency maps and in meaning maps of these scenes. This implies that the neural representations of the spatial distribution of both saliency and potentially meaningful scene elements are mapped at a sufficiently large cortical scale in the brain that they can be detected even after the substantial spatial filtering that occurs when electrical potentials are recorded from the scalp.

Finer-grained information was provided by the topographical analysis shown in Figure 8, in which both physical saliency and meaning-related RSA effects were primarily accounted for by signals overlying visual cortex, with some indication of a right hemisphere lateralization for physical saliency. This right lateralization for physical saliency is interesting given prior transcranial magnetic stimulation work showing evidence for a right lateralization in posterior regions involved in maintaining physical saliency maps across saccades (van Koningsbruggen et al., 2010). While these results should be taken with caution given the complex dynamics underlying the generation of observed scalp-level EEG topographies (Nunez et al., 2006), it is not unreasonable to hypothesize that the effects observed in this study arise from retinotopic activations in visual cortex (DeYoe et al., 1996; Brewer et al., 2005) or other topographically mapped cortical regions (Arcaro et al., 2009; Silver and Kastner, 2009; Arcaro and Livingstone, 2017).

Finally, these results also draw on and contribute to an extensive body of work on the neurophysiological processes underlying scene perception. Of particular interest, previous work (Harel et al., 2007; Groen et al., 2012, 2013; Cichy et al., 2017; Henriksson et al., 2019; Kaiser et al., 2020) indicates that the time course of processing for low-level global statistics and scene geometries is similar to the time courses observed for the spatial maps of meaning and saliency in the present study. Further research into how these various factors interact, particularly with regard to spatial and nonspatial features, has significant potential for expanding our

understanding of the perceptual processing of scenes and how those processes drive shifts of overt attention.

Varieties of saliency

The original model of visual saliency by Koch and Ullman (1985) defines physical saliency based on a model of biologically plausible features that mimic the response of early visual processing regions (e.g., V1/LGN). Since then, computational models of saliency have been developed that instead rely on deep neural networks, such as AlexNet (Krizhevsky et al., 2017), VGG16 (Simonyan and Zisserman, 2015), or Resnet (He et al., 2016). These models therefore go well beyond the response properties of early visual processing regions. As a result, the models include more abstract, higher-level representations of the visual input, so they are not pure models of physical saliency. They do not contain semantics per se, but they are trained on human response data (e.g., human classification judgments or visual fixations from large datasets). As a result, they may be influenced by both physical saliency and the higher-level computations that presumably underlie the processing of meaning (Damiano et al., 2019). This makes it difficult to isolate physical saliency from semantic features in these models, so they were not relevant for the present study's goal of assessing the time courses of these two factors.

It is worth noting, however, that previous research has found links between the activation outputs of such models with patterns of EEG/MEG activity (Dima et al., 2018; Greene and Hansen, 2018). Such findings, taken in conjunction with the continued development of more biologically inspired neural network models (Schrimpf and Kubilius, 2018; Dapello et al., 2020), indicate that this line of research holds considerable promise for shedding light on other mechanisms of human vision.

Limitations and future directions

Although the present results provide strong evidence for the rapid emergence of information about the spatial distribution of semantic information in the human brain, some limitations must be considered. First, we examined only a single task and a limited number of scenes, and it is possible that the time course of physical saliency and meaning-related information may differ across tasks and scenes. The majority of behavioral work in this area suggests that meaning maps override physical saliency maps in the control of attention across a broad set of scenes and tasks (Henderson and Hayes, 2017, 2018; Hayes and Henderson, 2019; Peacock et al., 2019b), but it will be important for future research to explore a broader range of tasks and scenes. It will also be important for future research to assess the potential moderating influences of other scene-related features, such as naturalness and openness.

A second limitation is that these RSA results are, by definition, correlational. Thus, we cannot conclude that the brain was extracting physical saliency and meaningfulness per se, but only that the brain was extracting information that is associated with physical saliency and meaningfulness. This point is particularly important to note with regard to how quickly the representational effect of the meaning maps arose. That is, the fast onset of the meaning-related effects may indicate that neurons in visual cortex are tuned to features that are likely to be associated with meaningful objects either directly or indirectly (the former being more likely given the rapidity of the effect) via feedback from scene/object-selective regions. The causal direction of this effect could potentially be assessed with recently developed approaches using transcranial magnetic stimulation (Wischnewski and Peelen, 2021). Furthermore, recent work by Kiat et al. (2021) suggests that this tuning arises as a product of real-world experience and/or other developmental processes, presenting additional directions for further research.

Third, given the nature of scalp-based EEG, it is difficult to draw firm conclusions regarding the specific neural generators and systems underlying these effects. Future investigations involving representational similarity-based fusion of EEG/MEG and fMRI data (Cichy and Oliva, 2020) could shed light on this issue.

Fourth, the meaning maps used in this study likely do not fully represent all stages of semantic activity related to visual processing. Specifically, the meaning maps largely represent the context-free semantic density of local scene regions, excluding contextualized elements, such as object-scene semantic relations. These maps were selected as they are currently the best candidate available for representing the earliest stages of semantic feature processing. However, as scene processing progresses, representations of meaning are likely to become more context-dependent and less spatiotopically precise. As a result, these representations may no longer match the context-independent, spatially precise meaning maps, leading to low representational similarity between the ERPs and the meaning maps at later time points. Given prior work regarding the time course of contextual and semantic processing (Mudrik et al., 2010; Demiral et al., 2012), it is likely that a later, more sustained, link for contextually relevant semantic features would be obtained if we used maps that capture actual concepts and/or more contextualized aspects of semantic feature processing (Hayes and Henderson, 2021).

Finally, it is worth noting that the meaning maps used in the present study do not represent a theory of the processes underlying scene semantics. These maps instead provide an operational tool to quantify the spatial distribution of semantically informative scene elements, setting the stage for future investigations focused on disentangling how low-level image features are processed and integrated to give rise to semantic informativeness.

References

Arcaro MJ, Livingstone MS (2017) Retinotopic organization of scene areas in macaque inferior temporal cortex. J Neurosci 37:7373–7389.

Arcaro MJ, McMains SA, Singer BD, Kastner S (2009) Retinotopic organization of human ventral visual cortex. J Neurosci 29:10638–10652.

Ashby FG, Maddox WT, Lee WW (1994) On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. Psychol Sci 5:144–151.

Bae GY, Luck SJ (2018) Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. J Neurosci 38:409–422.

Bae GY, Luck SJ (2019) Decoding motion direction using the topography of sustained ERPs and alpha oscillations. Neuroimage 184:242–255.

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Statist 29:1165–1188.

Bogler C, Bode S, Haynes JD (2011) Decoding successive computational stages of saliency processing. Curr Biol 21:1667–1671.

Brainard DH (1997) The Psychophysics Toolbox. Spat Vis 10:433–436.

Brewer AA, Liu J, Wade AR, Wandell BA (2005) Visual field maps and stimulus selectivity in human ventral occipital cortex. Nat Neurosci 8:1102–1109.

Cichy RM, Oliva A (2020) A M/EEG-fMRI fusion primer: resolving human brain responses in space and time. Neuron 107:772–781.

Cichy RM, Khosla A, Pantazis D, Oliva A (2017) Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. Neuroimage 153:346–358.

Damiano C, Wilder J, Walther DB (2019) Mid-level feature contributions to category-specific gaze guidance. Atten Percept Psychophys 81:35–46.

Dapello J, Marques T, Schrimpf M, Geiger F, Cox DD, DiCarlo JJ (2020) Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. bioRxiv. Available at https://doi.org/10.1101/2020.06.16.154542.

Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods 134:9–21.

Demiral SB, Malcolm GL, Henderson JM (2012) ERP correlates of spatially incongruent object identification during scene viewing: contextual expectancy versus simultaneous processing. Neuropsychologia 50:1271–1285.

DeYoe EA, Carman GJ, Bandettini P, Glickman S, Wieser J, Cox R, Miller D, Neitz J (1996) Mapping striate and extrastriate visual areas in human cerebral cortex. Proc Natl Acad Sci USA 93:2382–2386.

Diedrichsen J, Kriegeskorte N (2017) Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput Biol 13: e1005508.

Dima DC, Perry G, Singh KD (2018) Spatial frequency supports the emergence of categorical representations in visual cortex during natural scene perception. Neuroimage 179:102–116.

Drisdelle BL, Aubin S, Jolicoeur P (2017) Dealing with ocular artifacts on lateralized ERPs in studies of visual-spatial attention and memory: ICA correction versus epoch rejection. Psychophysiology 54:83–99.

Fabre-Thorpe M, Delorme A, Marlot C, Thorpe S (2001) A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. J Cogn Neurosci 13:171–180.

Farrens JL, Simmons AM, Luck SJ, Kappe ES (2020) Electroencephalogram (EEG) recording protocol for cognitive and affective human neuroscience research. Protocol Exchange. Available at https://protocolexchange.researchsquare.com/article/pex-779/v2.

Gordon RD (2004) Attentional allocation during the perception of scenes. J Exp Psychol Hum Percept Perform 30:760–777.

Gottlieb JP, Kusunoki M, Goldberg ME (1998) The representation of visual salience in monkey parietal cortex. Nature 391:481–484.

Greene MR, Hansen BC (2018) Shared spatiotemporal category representations in biological and artificial deep neural networks. PLoS Comput Biol 14:e1006327.

Groen II, Ghebreab S, Lamme VA, Scholte HS (2012) Spatially pooled contrast responses predict neural and perceptual similarity of naturalistic image categories. PLoS Comput Biol 8:e1002726.

Groen II, Ghebreab S, Prins H, Lamme VA, Scholte HS (2013) From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. J Neurosci 33:18814–18824.

Hansen JC, Hillyard SA (1980) Endogenous brain potentials associated with selective auditory attention. Electroencephalogr Clin Neurophysiol 49:277–290.

Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: Advances in neural information processing systems, pp 545–552. Cambridge, MA: Massachusetts Institute of Technology.

Hayes TR, Henderson JM (2019) Scene semantics involuntarily guide attention during visual search. Psychon Bull Rev 26:1683–1689.

Hayes TR, Henderson JM (2021) Looking for semantic similarity: what a vector-space model of semantics can tell us about attention in real-world scenes. Psychol Sci 32:1262–1270.

Hayhoe M, Ballard D (2005) Eye movements in natural behavior. Trends Cogn Sci 9:188–194.

He K, Zhang X, Shaoqing R, Sun J (2016) Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778. Las Vegas, NV:IEEE.

Henderson JM (2003) Human gaze control during real-world scene perception. Trends Cogn Sci 7:498–504.

Henderson JM (2017) Gaze control as prediction. Trends Cogn Sci 21:15–23.

Henderson JM, Hayes TR (2017) Meaning-based guidance of attention in scenes as revealed by meaning maps. Nat Hum Behav 1:743–747.

Henderson JM, Hayes TR (2018) Meaning guides attention in real-world scene images: evidence from eye movements and meaning maps. J Vis 18:10.

Henderson JM, Hayes TR, Peacock CE, Rehrig G (2019) Meaning and attentional guidance in scenes: a review of the meaning map approach. Vision 3:19.

Henderson JM, Goold JE, Choi W, Hayes TR (2020) Neural correlates of fixated low- and high-level scene properties during active scene viewing. J Cogn Neurosci 32:2013–2023.

Henriksson L, Mur M, Kriegeskorte N (2019) Rapid invariant encoding of scene layout in human OPA. Neuron 103:161–171.e3.

Iman RL, Conover WJ (1979) The use of the rank transform in regression. Technometrics 21:499–509.

Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. Vision Res 40:1489–1506.

Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2:194–203.

Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Machine Intell 20:1254–1259.

Jung KY, Seo DW, Na DL, Chung CS, Lee IK, Oh K, Im CH, Jung HK (2007) Source localization of periodic sharp wave complexes using independent component analysis in sporadic Creutzfeldt-Jakob disease. Brain Res 1143:228–237.

Kaiser D, Häberle G, Cichy RM (2020) Cortical sensitivity to natural scene structure. Hum Brain Mapp 41:1286–1295.

Kiat JE, Luck SJ, Beckner AG, Hayes TR, Pomaranski KI, Henderson JM, Oakes LM (2021) Linking patterns of infant eye movements to a neural network model of the ventral stream using representational similarity analysis. Dev Sci 25:e13155.

Kiesel A, Miller J, Jolicoeur P, Brisson B (2008) Measurement of ERP latency differences: a comparison of single-participant and jackknife-based scoring methods. Psychophysiology 45:250–274.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in Psychtoolbox-3. Perception 36:1–16.

Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4:219–227.

Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60:84–90.

Li W, Piëch V, Gilbert CD (2006) Contour saliency in primary visual cortex. Neuron 50:951–962.

Lopez-Calderon J, Luck SJ (2014) ERPLAB: an open-source toolbox for the analysis of event-related potentials. Front Hum Neurosci 8:213.

Luck SJ (2014) An introduction to the event-related potential technique, Ed 2. Cambridge, MA: Massachusetts Institute of Technology.

Miller J, Patterson T, Ulrich R (1998) Jackknife-based method for measuring LRP onset latency differences. Psychophysiology 35:99–115.

Mudrik L, Lamy D, Deouell LY (2010) ERP evidence for context congruity effects during simultaneous object-scene processing. Neuropsychologia 48:507–517.

Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N (2014) A toolbox for representational similarity analysis. PLoS Comput Biol 10:e1003553.

Nunez PL, Nunez EP, Srinivasan R, Srinivasan AP (2006) Electric fields of the brain: the neurophysics of EEG. Oxford: Oxford UP.

Nuthmann A, Einhäuser W, Schütz I (2017) How well can saliency models predict fixation selection in scenes beyond central bias? A new approach to model evaluation using generalized linear mixed models. Front Hum Neurosci 11:491.

Peacock CE, Hayes TR, Henderson JM (2019a) The role of meaning in attentional guidance during free viewing of real-world scenes. Acta Psychol (Amst) 198:102889.

Peacock CE, Hayes TR, Henderson JM (2019b) Meaning guides attention during scene viewing even when it is irrelevant. Atten Percept Psychophys 81:20–34.

Peacock CE, Hayes TR, Henderson JM (2020) Center bias does not account for the advantage of meaning over salience in attentional guidance during scene viewing. Front Psychol 11:1877.

Pelli DG (1997) The VideoToolbox software for visual psychophysics: transforming numbers into movies. Spat Vis 10:437–442.

Rehrig G, Peacock CE, Hayes TR, Henderson JM, Ferreira F (2020) Where the action could be: speakers look at graspable objects and meaningful

scene regions when describing potential actions. J Exp Psychol Learn Mem Cogn 46:1659–1681.

Schrimpf M, Kubilius J (2018) Brain-Score: which artificial neural network for object recognition is most brain-like? bioRxiv. Available at https://doi.org/10.1101/407007.

Silver MA, Kastner S (2009) Topographic maps in human frontal and parietal cortex. Trends Cogn Sci 13:488–495.

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv 140915506:1–14.

Skrandies W (1990) Global field power and topographic similarity. Brain Topogr 3:137–141.

Thaler L, Schutz AC, Goodale MA, Gegenfurtner KR (2013) What is the best fixation target? The effect of target shape on stability of fixational eye movements. Vision Res 76:31–42.

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520–522.

Ulrich R, Miller J (2001) Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. Psychophysiology 38:816–827.

van Koningsbruggen MG, Gabay S, Sapir A, Henik A, Rafal RD (2010) Hemispheric asymmetry in the remapping and maintenance of visual saliency maps: a TMS study. J Cogn Neurosci 22:1730–1738.

Veale R, Hafed ZM, Yoshida M (2017) How is visual salience computed in the brain? Insights from behaviour, neurobiology and modelling. Philos Trans R Soc Lond B Biol Sci 372:20160113.

Walther D, Koch C (2006) Modeling attention to salient proto-objects. Neural Netw 19:1395–1407.

White BJ, Kan JY, Levy R, Itti L, Munoz DP (2017) Superior colliculus encodes visual saliency before the primary visual cortex. Proc Natl Acad Sci USA 114:9451–9456.

Wischnewski M, Peelen MV (2021) Causal neural mechanisms of context-based object recognition. eLife 10:e69736.

Wolfe JM (1994) Guided Search 2.0: a revised model of visual search. Psychon Bull Rev 1:202–238.