bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Cortical Sensitivity to Natural Scene Structure

Daniel Kaiser[1], Greta Häberle[1,2], Radoslaw M. Cichy[1,2,3,4]

[1]*Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany*

[2]*Einstein Center for Neurosciences Berlin, Berlin, Germany*

[3]*Berlin School of Mind and Brain, Humboldt-Universität Berlin, Berlin, Germany*

[4]*Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany*

Author Note

We thank Sina Schwarze for help in EEG data collection and manuscript preparation.

D.K. and R.M.C. are supported by Deutsche Forschungsgemeinschaft (DFG) grants (KA4683/2-1, CI241/1-1, CI241/3-1). R.M.C. is supported by a European Research Council Starting Grant (ERC-2018-StG 803370).

Correspondence should be addressed to Dr. Daniel Kaiser, Department of Education and Psychology, Freie Universität Berlin, Habelschwerdter Allee 45, 14195 Berlin, Germany. Email: danielkaiser.net@gmail.com

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Abstract

Natural scenes are inherently structured, with meaningful objects appearing in predictable locations. Human vision is tuned to this structure: When scene structure is purposefully disrupted (e.g., by jumbling scene images), perception is strongly impaired. Here, we tested how such perceptual effects are reflected in neural sensitivity to natural scene structure. During separate fMRI and EEG experiments, participants passively viewed scenes whose spatial structure (i.e., the position of the scene's parts) and categorical structure (i.e., the content of the scene's parts) could be intact or jumbled. Using multivariate decoding analyses, we show that spatial (but not categorical) scene structure has a profound impact on cortical processing: Scene-selective responses in the occipital and parahippocampal cortices (fMRI) and after 255ms (EEG) accurately differentiated between spatially intact and spatially jumbled scenes. Importantly, this differentiation was more pronounced for upright than for inverted scenes, indicating genuine sensitivity to spatial scene structure rather than sensitivity to low-level visual attributes. The cortical sensitivity to spatial structure may reflect perceptual adaptations to real-world statistics, which support efficient scene understanding in everyday situations.

Keywords: visual perception, scene representation, spatial structure, fMRI / EEG, multivariate decoding

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Cortical Sensitivity to Natural Scene Structure

Humans can understand natural scenes from just a single glance (Potter, 1975; Thorpe, Fize, & Marlot, 1996). One reason for this perceptual efficiency lies in the structure of natural scenes: for instance, a scene's spatial structure tells us where specific objects can be found and its categorical structure tells us which objects are typically encountered within the scene (Bar, 2004; Oliva & Torralba, 2007; Potter, 2012; Võ, Boettcher, & Draschkow, 2019).

The beneficial impact of scene structure on perception becomes apparent in jumbling paradigms, where the scene's structure is purposefully disrupted by shuffling blocks of information across the scene. Jumbling makes it harder to categorize scenes (Biederman, Rabinowitz, Glass, & Stacy, 1974), recognize objects within them (Biederman, 1972; Biederman, Glass, & Stacy, 1973) or to detect subtle visual changes (Varakin & Levin, 2008; Zimmermann, Schnier, & Lappe, 2010).

These perceptual effects prompt the hypothesis that scene structure impacts perceptual stages of cortical scene processing. However, while there is evidence that real-world structure impacts visual responses to everyday objects (Kim & Biederman, 2011; Kaiser & Cichy, 2018; Kaiser & Peelen, 2018; Roberts & Humphreys, 2010) and people (Bernstein, Oron, Sadah, & Yovel, 2010; Brandman & Yovel, 2016; Chan, Kravitz, Truong, Arizpe, & Baker, 2010; Sorisa Bauser & Suchan, 2015), it is unclear whether real-world structure has a similar impact on scene-selective responses.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Here, we used multivariate pattern analysis (MVPA) on fMRI and EEG responses to jumbled scenes to demonstrate that cortical scene processing is indeed sensitive to scene structure. We reveal three key characteristics of this sensitivity: (1) Cortical scene processing is primarily sensitive to the scene's spatial structure, rather than its categorical structure. (2) Spatial structure impacts the perceptual analysis of scenes, in occipital and parahippocampal cortices (Epstein, 2012) and shortly after 200ms (Harel, Groen, Kravitz, Deouell, & Baker, 2016). (3) Spatial structure impacts cortical responses more strongly for upright than inverted scenes, indicating robust sensitivity to scene structure that goes beyond sensitivity to low-level features.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Method

Participants

In the fMRI experiment, 20 healthy adults participated in session 1 (mean age 25.5, *SD*=4.0; 13 female) and 20 in session 2 (mean age 25.4, *SD*=4.0; 12 female). Seventeen participants completed both sessions, three participants only session 1 or session 2, respectively. In the EEG experiment, 20 healthy adults (mean age 26.6, *SD*=5.8; 9 female) participated in a single session. All participants had normal or corrected-to-normal vision. Participants provided informed consent and received monetary reimbursement or course credits. All procedures were approved by the local ethical committee and were in accordance with the Declaration of Helsinki.

Stimuli and design

Stimuli were 24 scenes from four different categories (church, house, road, supermarket; Figure 1a), taken from an online resource (Konkle, Brady, Alvarez, & Oliva, 2010). We split each image into quadrants and systematically recombined the resulting parts in a 2×2 design, where both the scenes' spatial structure and their categorical structure could be either intact or jumbled (Figure 1b/c). This yielded four conditions: (1) In the "spatially intact & categorically intact" condition, parts from four scenes of the same category were combined in their correct locations. (2) In the "spatially intact & categorically jumbled" condition, parts from four scenes from different categories were combined in their correct locations. (3) In the "spatially jumbled & categorically intact" condition, parts from four scenes of the same category

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

were combined, and their locations were exchanged in a crisscrossed way. (4) In the "spatially jumbled & categorically jumbled" condition, parts from four scenes from different categories were combined, and their locations were exchanged in a crisscrossed way. For each participant separately, 24 unique stimuli were generated for each condition by randomly drawing suitable fragments from different scenes[1]. During the experiment, all scenes were presented both upright and inverted.

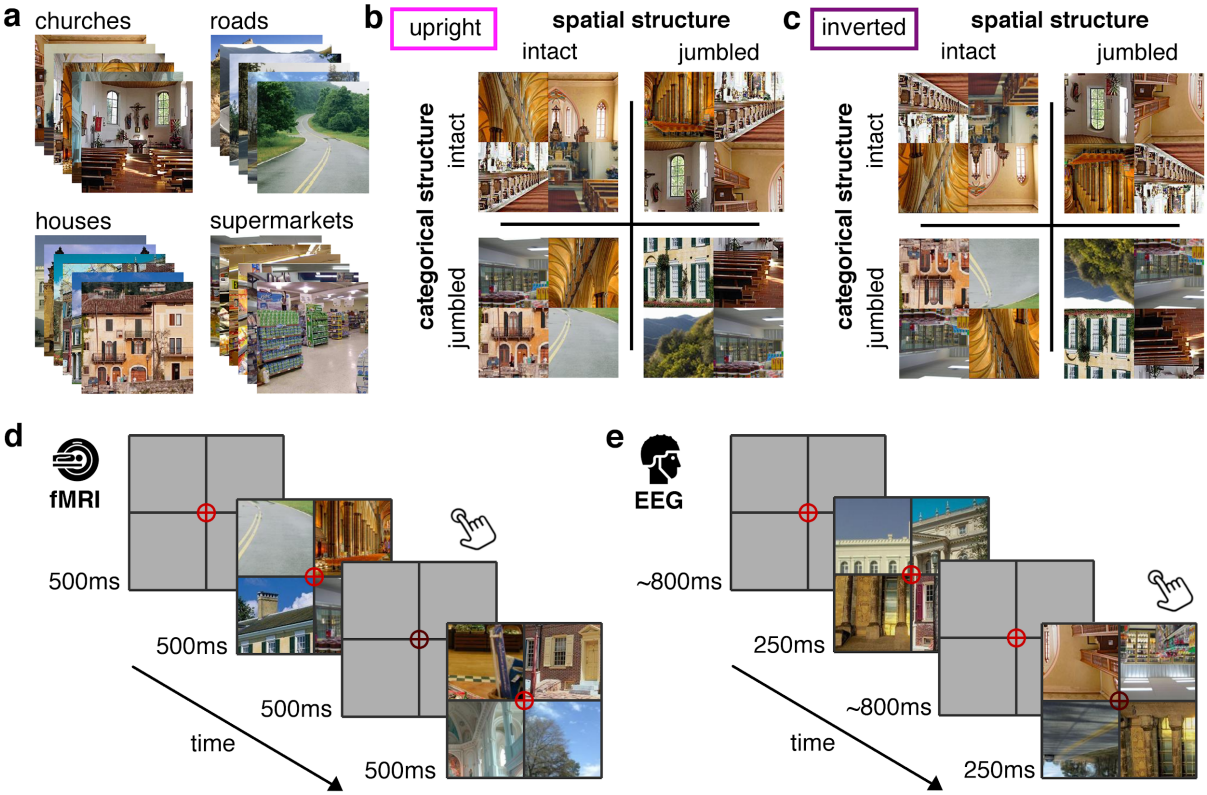

Figure 1. Stimuli and Paradigm. We combined parts from 24 scene images from four categories (a) to create a stimulus set where the scenes' structural (e.g. the spatial arrangements of the parts) and their categorical structure (e.g., the category of the parts) was orthogonally manipulated; all scenes were presented both upright and

[1] Note that all scenes were jumbled to some extent, as also in the categorically intact scenes four different exemplars were intermixed.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

inverted (b/c). In the fMRI experiment, scenes were presented in a block design, where each block of 24s exclusively contained scenes of a single condition (d). In the EEG experiment, all conditions were randomly intermixed (e). During both experiments, participants responded to color changes of the central crosshair.

fMRI paradigm

The fMRI experiment (Figure 1d) comprised two sessions. In the first session, upright scenes were shown, in the second session inverted scenes were shown; the sessions were otherwise identical. Each session consisted of five runs of 10min. Each run consisted of 25 blocks of 24 seconds. In 20 blocks, scene stimuli were shown with a frequency of 1Hz (0.5s stimulus, 0.5s blank). Each block contained all 24 stimuli of a single condition. In 5 additional fixation-only blocks, no scenes were shown. Block order was randomized within every five consecutive blocks, which contained each condition (four scene conditions and fixation-only) exactly once.

Scene stimuli appeared in a black grid (4.5° visual angle), which served to mask visual discontinuities between quadrants. Participants were monitoring a central red crosshair, which twice per block (at random times) darkened for 50ms; participants had to press a button when they detected a change. Participants on average detected 80.0% (*SE*=2.5)[2] of the changes. Stimulus presentation was controlled using the Psychtoolbox (Brainard, 1997).

In addition to the experimental runs, each participant completed a functional localizer run of 13min, during which they viewed images of scenes, objects, and scrambled scenes. The scenes were new exemplars of the four scene categories

[2] For two participants, due to technical problems, no responses were recorded.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

used in the experimental runs; objects were also selected from four categories (car, jacket, lamp, sandwich). Participants completed 32 blocks (24 scene/object/scrambled blocks and 8 fixation-only blocks), with parameters identical to the experimental runs (24s block duration, 1Hz stimulation frequency, color change task).

EEG paradigm

In the EEG experiment (Figure 1e), all conditions were randomly intermixed within a single session of 75min (split into 16 runs). During each trial, a scene appeared for 250ms, followed by an inter-trial interval randomly varying between 700ms and 900ms. In total, there were 3072 trials (384 per condition), and an additional 1152 target trials (see below).

As in the fMRI, stimuli appeared in a black grid (4.5° visual angle) with a central red crosshair. In target trials, the crosshair darkened during the scene presentation; participants had to press a button and blink when detecting this change. Participants on average detected 78.1% (*SE*=3.6) of the changes. Target trials were not included in subsequent analyses.

fMRI recording and preprocessing

MRI data was acquired using a 3T Siemens Tim Trio Scanner equipped with a 12-channel head coil. T2*-weighted gradient-echo echo-planar images were collected as functional volumes (*TR*=2s, *TE*=30ms, 70° flip angle, 3mm^3 voxel size, 37 slices, 20% gap, 192mm FOV, 64×64 matrix size, interleaved acquisition). Additionally, a T1-weighted anatomical image (MPRAGE; 1mm^3 voxel size) was

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

obtained. Preprocessing was performed using SPM12 (www.fil.ion.ucl.ac.uk/spm/). Functional volumes were realigned, coregistered to the anatomical image, and normalized into MNI-305 space. Images from the localizer run were additionally smoothed using a 6mm full-width-half-maximum Gaussian kernel.

EEG recording and preprocessing

EEG signals were recorded using an EASYCAP 64-electrode[3] system and a Brainvision actiCHamp amplifier. Electrodes were arranged in accordance with the 10-10 system. EEG data was recorded at 1000Hz sampling rate and filtered online between 0.03Hz and 100Hz. All electrodes were referenced online to the Fz electrode. Offline preprocessing was performed using FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011). EEG data were epoched from -200ms to 800ms relative to stimulus onset, and baseline-corrected by subtracting the mean pre-stimulus signal. Channels and trials containing excessive noise were removed based on visual inspection. Blinks and eye movement artifacts were removed using independent component analysis and visual inspection of the resulting components. The epoched data were down-sampled to 200Hz.

fMRI region of interest definition

We restricted fMRI analyses to three regions of interest (ROIs): early visual cortex (V1), scene-selective occipital place area (OPA), and scene-selective parahippocampal place area (PPA). V1 was defined based on a functional group atlas (Wang, Mruczek, Arcaro, & Kastner, 2015). Scene-selective ROIs were defined using

[3] For two participants, due to technical problems, only data from 32 electrodes was recorded.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

the localizer data, which were modelled in a general linear model (GLM) with 9 predictors (3 regressors for the scene/object/scrambled blocks and 6 movement regressors). Scene-selective ROI definition was constrained by group-level activation masks for OPA and PPA (Julian, Fedorenko, Webster, & Kanwisher, 2012). Within these masks, we first identified the voxel exhibiting the greatest *t*-value in a scene>object contrast, separately for each hemisphere, and then defined the ROI as a 125-voxel sphere around this voxel. Left- and right-hemispheric ROIs were concatenated for further analysis.

fMRI decoding

fMRI response patterns for each ROI were extracted directly from the volumes recorded during each block. After shifting the activation time course by three TRs (i.e., 6s) to account for the hemodynamic delay, we extracted voxel-wise activation values from the 12 TRs corresponding to each block of 24s. Activation values for these 12 TRs were then averaged, yielding a single response pattern across voxels for each block. To account for activation differences between runs, the mean activation across all blocks was subtracted from each voxel's values, separately for each run. Decoding analyses were performed using CoSMoMVPA (Oosterhof, Connolly, & Haxby, 2016), and were carried out separately for each ROI and participant. We used data from four runs to train linear discriminant analysis (LDA) classifiers to discriminate multi-voxel response patterns (i.e., patterns of voxel activations across all voxels of an ROI) and response patterns from the left-out, fifth run to test these classifiers. This was done repeatedly until every run was left out once and decoding accuracy was averaged across these repetitions.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

EEG decoding

EEG decoding was performed separately for each time point (i.e., every 5ms) from -200ms to 800ms relative to stimulus onset, using CoSMoMVPA (Oosterhof et al., 2016). We used data from all-but-one trials for two conditions to train LDA classifiers to discriminate topographical response patterns (i.e., patterns across electrodes) and data from the left-out trials to test these classifiers. This was done repeatedly until each trial was left out once and decoding accuracy was averaged across these repetitions. Classification time series for individual participants were smoothed using a running average of five time points (i.e., 25ms).

Decoding sensitivity to scene structure

For both the fMRI and EEG data, we performed two complimentary decoding analyses. In the first analysis, we tested sensitivity for spatial structure by decoding spatially intact from spatially jumbled scenes (Figure 2a). In the second analysis, we tested sensitivity for categorical structure by decoding categorically intact from categorically jumbled scenes (Figure 2d). To investigate whether successful decoding indeed reflected sensitivity to scene structure, we performed both analyses separately for the upright and inverted scenes. Critically, inversion effects (i.e., better decoding in the upright than in the inverted condition) indicate genuine sensitivity to natural scene structure that goes beyond purely visual differences.

Statistical testing

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

For the fMRI data, we used *t*-tests to compare decoding against chance and between conditions. To Bonferroni-correct for comparisons across ROIs, all *p*-values were multiplied by 3. For the EEG data, given the larger number of comparisons, we used a threshold-free cluster enhancement procedure (Smith & Nichols, 2009). Multiple-comparison correction was based on a sign-permutation test (with null distributions created from 10,000 bootstrapping iterations) as implemented in CoSMoMVPA (Oosterhof et al., 2016). The resulting statistical maps were thresholded at $z>1.96$ (i.e., $p_{corr}<.05$).

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Results

Sensitivity to spatial scene structure

To uncover where and when cortical processing is sensitive to spatial structure, we decoded between scenes whose spatial structure was intact or jumbled (Figure 2a).

For the fMRI data (Figure 2b), we found highly significant decoding between spatially intact and spatially jumbled scenes. For upright scenes, significant decoding emerged in V1, $t(19)=13.03$, $p_{corr}<.001$, OPA, $t(19)=7.61$, $p_{corr}<.001$, and PPA, $t(19)=5.92$, $p_{corr}=.002$, and for inverted scenes in V1, $t(19)=9.92$, $p_{corr}<.001$, but not in OPA, $t(19)=2.08$, $p_{corr}=.16$, and PPA, $t(19)=0.85$, $p_{corr}>1$. Critically, we observed inversion effects (i.e., better decoding for the upright scenes) in the OPA, $t(16)=4.41$, $p_{corr}=.001$[4], and PPA, $t(16)=3.67$, $p_{corr}=.006$, but not in V1, $t(16)=1.32$, $p_{corr}=.62$. Therefore, decoding in V1 solely reflects visual differences, whereas OPA and PPA exhibit genuine sensitivity to the spatial scene structure. This result was confirmed by further ROI analyses and a spatially unconstrained searchlight analysis (see Supplementary Information).

For the EEG data (Figure 2c), we also found strong decoding between spatially intact and jumbled scenes. For upright scenes, this decoding emerged between 55ms and 465ms, between 505ms and 565ms, and between 740ms and 785ms, peak $z>3.29$, $p_{corr}<.001$, and for inverted scenes between 65ms and 245ms, peak $z>3.29$, $p_{corr}<.001$. As in scene-selective cortex, we observed inversion effects, indexing

[4] Statistics for fMRI inversion effects are based on the 17 participants who completed both sessions.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

stronger sensitivity to spatial structure in upright scenes, between 255ms and 300ms and between 340ms and 395ms, peak z=2.78, p_{corr}=.005.

Together, these results show that in scene-selective OPA and PPA, and after 255ms, cortical activations are sensitive to the spatial structure of natural scenes.

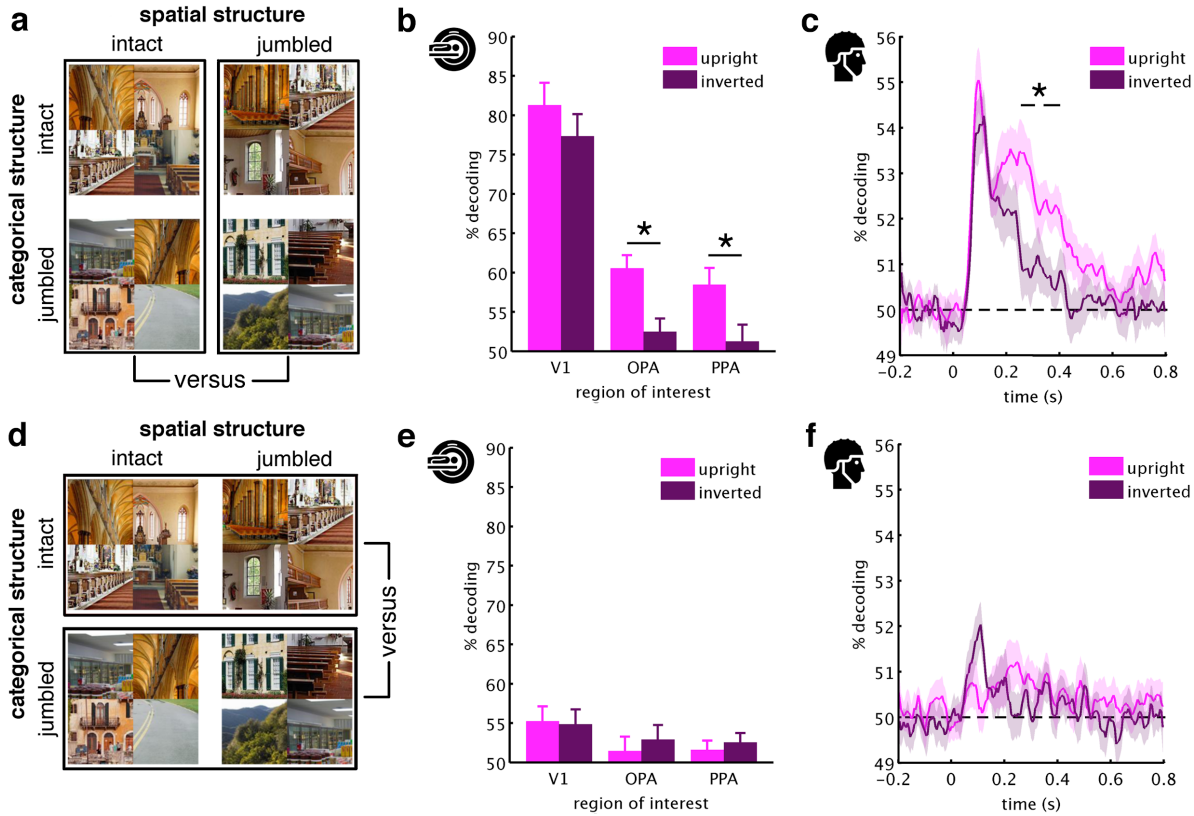

Figure 2. MVPA results. To reveal sensitivity to spatial scene structure, we decoded between scenes with spatially intact and spatially jumbled parts (a). Already during early processing (in V1 and before 200ms) spatially intact and jumbled scenes could be discriminated well, both for the upright and inverted conditions. Critically, during later processing (in OPA/PPA and from 255ms) inversion effects (i.e., better decoding for upright than inverted scenes) revealed genuine sensitivity to spatial scene structure (b/c). To reveal sensitivity to categorical scene structure, we decoded between scenes with categorically intact and categorically jumbled parts (d). In this

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

analysis, no pronounced decoding and no inversion effects were found, neither across space (e) nor time (f). Error margins reflect standard errors of the difference. Significance markers denote inversion effects (p_{corr}<.05).

Sensitivity to categorical scene structure

To uncover where and when cortical processing is sensitive to categorical structure, we decoded between scenes whose categorical structure was intact or jumbled (Figure 2a).

For the fMRI (Figure 2e), the upright scenes' categorical structure could be decoded only from V1, $t(19)=3.11$, $p_{corr}=.017$, but not the scene-selective ROIs, both $t(19)<2.15$, $p_{corr}>.13$. Similarly, for the inverted scenes, significant decoding was only observed in V1, $t(19)=4.58$, $p_{corr}<0.001$, but not in the scene-selective ROIs, both $t(19)<2.29$, $p_{corr}>.10$. No inversion effects were observed, all $t(16)<0.60$, $p_{corr}>1$.

For the EEG (Figure 2f), we found only weak decoding between the categorically intact and jumbled scenes. In the upright condition, decoding was significant between 165ms and 175ms and between 215ms and 265ms, peak $z=2.32$, $p_{corr}=.02$, and in the inverted condition at 120ms, peak $z=1.97$, $p_{corr}=.049$. No inversion effects were observed, peak $z=1.64$, $p_{corr}=.10$.

Together, these results reveal no sensitivity to the categorical structure of a scene, at least when none of the scenes are fully coherent and when they are not relevant for behavior. This is in marked contrast with sensitivity for spatial scene structure, which is observed in the absence of behavioral relevance and is disrupted by stimulus inversion. Similar results were obtained in univariate analyses (see Supplementary Information).

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Discussion

Our findings provide the first spatiotemporal characterization of cortical sensitivity to natural scene structure. As the key result, we observed sensitivity to spatial (but not categorical) scene structure, which emerged in scene-selective cortex and from 255ms of vision. By showing that this effect is stronger for upright than for inverted scenes, we provide strong evidence for genuine sensitivity to spatial structure, rather than low-level properties.

Sensitivity to spatial structure may index mechanisms enabling efficient scene understanding. Previous work on object processing shows that in order to efficiently parse the many objects contained in natural scenes, the visual system exploits regularities in the environment, such as regularities in individual objects' positions (Kaiser & Cichy, 2018; Kaiser, Moeskops, & Cichy, 2018), relationships between objects (Kim & Biederman, 2011; Kaiser & Peelen, 2018; Kaiser, Stein, & Peelen, 2014; Roberts & Humphreys, 2010), and relationships between objects and scenes (Brandman & Peelen, 2017; Faivre, Dubois, Schwartz, & Mudrik, 2019). The current results suggest that also cortical scene analysis uses spatial regularities to efficiently handle complex visual information, in line with the view that real-world structure facilitates processing in the visual system across diverse naturalistic contents.

Our results also shine new light on the temporal processing cascade during scene perception. Sensitivity to spatial structure emerged after 255ms of processing, which is only after scene-selective peaks in ERPs (Harel et al., 2016; Sato et al.,

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

1999)[5] and after basic scene attributes are computed (Cichy, Khosla, Pantazis, & Oliva, 2017). Interestingly, after 250ms brain responses not only become sensitive to scene structure, but also to object-scene consistencies (Draschkow, Heikel, Fiebach, Võ, & Sassenhagen, 2018; Ganis & Kutas, 2003; Mudrik, Lamy, & Deouell, 2010; Võ & Wolfe, 2013). Together, these results suggest a dedicated processing stage for the structural analysis of objects, scenes, and their relationships, which is different from basic perceptual processing. However, whether these different findings indeed reflect a common underlying mechanism requires further investigation[6].

Perhaps surprisingly, our findings suggest more pronounced sensitivity to spatial structure than to categorical structure. This is in line with studies showing that scene-selective responses are mainly driven by spatial layout, rather than scene content (Dillon, Persichetti, Spelke, & Dilks, 2018; Harel, Kravitz, & Baker, 2013; Lowe, Rajsic, Gallivan, Ferber, & Cant, 2017; Kravitz, Peng, & Baker, 2011). However, the brain may be less sensitive to categorical structure when, as in our study, all scenes are jumbled to some extent and not behaviorally relevant.

On the contrary, it is worth stressing that sensitivity to spatial scene structure emerged in the absence of behavioral relevance, suggesting that spatial structure is analyzed automatically during perceptual processing. As in real-world situations we cannot explicitly engage with all aspects of a scene concurrently, this automatic analysis of spatial structure may be crucial for rapid scene understanding.

---

[5] In our study, ERP responses in posterior-lateral electrodes peaked at 235ms.

[6] One open question concerns whether these effects primarily reflect enhanced processing of consistent structure or responses to inconsistencies.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

References

Bar, M. (2004). Visual objects in context. *Nature Neuroscience, 5*, 617-629.

Bernstein, M., Oron, J., Sadeh, B., & Yovel G. (2014). An integrated face-body representation in the fusiform gyrus but not the lateral occipital cortex. *Journal of Cognitive Neuroscience, 26*, 2469-2478.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*, 77-80.

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology, 97*, 22-27.

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology, 103*, 597-600.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433-436.

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *Journal of Neuroscience, 37*, 7700-7710.

Brandman, T., & Yovel, G. (2016). Bodies are represented as wholes rather than their sum of parts in the occipital-temporal cortex. *Cerebral Cortex, 26*, 530-543.

Chan, A. W., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience, 13*, 417-418.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Cichy, R. M., 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.

Dillon, M. R., Persichetti, A. S., Spelke, E. S., & Dilks, D. D. (2018). Places in the brain: bridging layout and object geometry in scene-selective cortex. *Cerebral Cortex, 28*, 2365-2374.

Draschkow, D., Heikel, E., Võ, M. L.-H., Fiebach, C. J., Sassenhagen, J. (2018). No evidence for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia, 120*, 9-17.

Epstein, R. A. (2014). Neural systems for visual scene recognition. In M. Bar & K. Keveraga (Eds.), *Scene Vision* (pp. 105-134). Cambridge, MIT Press.

Faivre, N., Dubois, J., Schwartz, N., & Mudrik, L. (2019). Imaging object-scene relations processing in visible and invisible natural scenes. *Scientific Reports, 9*, 4567.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Brain Research Cognitive Brain Research, 16*, 123-144.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *eNeuro, 3*, ENEURO.0139-16.2016.

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: gradients of object and spatial layout information. *Cerebral Cortex, 23*, 947-957.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage, 60*, 2357-2364.

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology*, *120*, 848-853.

Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018) Typical retinotopic locations impact the time course of object coding. *NeuroImage, 176*, 372-379.

Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences USA, 111*, 11217-11222.

Kaiser, D., & Peelen, M. V. (2018) Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *NeuroImage 169*, 334-341.

Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? *Cerebral Cortex, 21*, 1738-1746.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychological Science, 21*, 1551-1556.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience, 31*, 7322-7333.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Lowe, M. X., Rajsic, J., Gallivan, J. P., Ferber, S., & Cant, J. S. (2017). Neural representation of geometry and surface properties in object and scene perception. *NeuroImage, 157*, 586-597.

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

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*, 520-527.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011,* 156869.

Oosterhof, N. N., Connolly, A. C., & Haxby, J V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics, 10,* 20.

Potter, M. C. (1975). Meaning in visual search. *Science, 187*, 965-966.

Potter, M. C. (2012). Recognition and memory for briefly presented scenes. *Frontiers in Psychology, 3*, 32.

Roberts, K. L., & Humphreys, G. W. (2010). Action relationships concatenate representations of separate objects in the ventral visual cortex. *NeuroImage 52*, 1541-1548.

Sato, N., Nakamura, K., Nakamura, A., Sugiura, M., Iko, K., Fukuda, H., & Kawashima, R. (1999). Different time course between scene processing and face processing: a MEG study. *Neuroreport, 10*, 3633-3637.

bioRxiv preprint doi: https://doi.org/10.1101/613885; this version posted April 18, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage, 44,* 83-98.

Sorisa Bauser, D., & Suchan, B. (2015). Is the whole the sum of its parts? Configuration processing of headless bodies in the right fusiform gyrus. *Behavioral Brain Research, 281*, 102-110.

Thorpe, S., Fize, D., & Marlot, D. (1996). Speed of processing in the human visual system. *Nature, 381*, 520-522.

Varakin, D. A., & Levin, D. T. (2008). Scene structure enhances change detection. *The Quarterly Journal of Experimental Psychology, 61*, 543-551.

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, https://doi.org/10.1016/j.copsyc.2019.03.009

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science, 24*, 1816-1823.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex, 25,* 3911-3931.

Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research, 50*, 2062-2068.