Investigating Location-specific Contextual Modulations of Object Processing

Mariska Peeters
S1063625
07-10-2022
Cognitive Neuroscience Master, 2020-2022
Radboud University, Nijmegen
Master Thesis
Supervisors: Dr. Marius Peelen and Dr. Surya Gayet
Second reader: Prof. Dr. Rob van Lier
Correspondence: Mariska.peeters@donders.ru.nl

Investigating Location-specific Contextual Modulations of Object Processing.

In our everyday surroundings, objects never appear in isolation, but they are integrated within a scene context. This context is able to facilitate object recognition. However, it is unknown what specific contextual cues are contributing to this facilitation effect. Here, we investigate the influence of an object's estimated size, inferred from the viewing distance. In an fMRI experiment, participants viewed scenes in which the object's location and size was manipulated, such that they were either congruently (e.g., near and large) or incongruently (e.g., far and large) presented. Decoding analyses showed that congruently presented objects were better represented in object-selective areas than incongruently presented objects. This congruency effect was not found in early visual cortex. This indicates that an object's estimated size, which can be inferred from the viewing distance, contributes to the neural representation of that object in the object-selective cortex.

Keywords: Object perception, contextual feedback, fMRI, MVPA

Humans are highly efficient at recognizing objects within a cluttered environment. Without much effort, you can find your coffee mug on a messy table, or the actual table where you have left the mug. To this day, visual search in natural scenes remains a task where humans consistently outperform computers (Borji & Itti, 2014). Only when expectations about visual information are violated, computers are able to surpass human performance (Eckstein et al., 2017). Nevertheless, a lifelong amount of experience prevents humans from making mistakes that are common in artificial object recognition. For instance, hackers are able to make a self-driving car slam the brakes by flashing an image of a stop sign on a billboard (Nassi et al., 2020). Humans, who are perfectly capable of discriminating between an actual stop sign and a stop sign presented on a billboard, would not make this error. What critical cues of the object or its surroundings are contributing to this difference between human and artificial agents? Understanding which cues are crucial in developing and acting upon expectations about visual information, may aid research in optimizing artificial object recognition. Perhaps of more importance, investigating the influence of certain contextual cues on the neural representation of objects in the brain contributes to our understanding of object recognition in humans.

Expectations elicited by structural regularities

One of the main assumptions in the field of visual object recognition is that it is supported by the generation of expectations (Summerfield & De Lange, 2014). These expectations are derived from the fact that visual information contains structural regularities: visual compositions that have appeared consistently in the past, will have a higher probability to appear in the future. Visual compositions are comprised of a scene and objects acting within that scene. Several regularities can be observed in the relationship between scenes and objects.

First, a scene can elicit expectations about the identity of an object. A famous study demonstrates that recognizing a loaf of bread in a kitchen requires less effort than recognizing a drum or mailbox in this same context (Palmer, 1975). This is due to the fact that the loaf of bread matches the expectations that are elicited by the kitchen

context, while this is not the case for the drum and mailbox. Especially when an object is ambiguous (e.g., presented for a short amount of time, or degraded from obvious physical characteristics), contextual information derived from a scene can facilitate object recognition (Bar, 2004; Oliva & Torralba, 2007).

Second, scenes can elicit expectations about the appearance of an object. For example, a source of light in a scene predicts the color, brightness and shadow of an object, while a specific angle of view predicts the shape of an object. Furthermore, the size of objects can be estimated by their position in the scene and by their position relative to other objects (Biederman et al., 1982): a coffee mug that is placed on a table nearby will have a larger retinal size (i.e., it will comprise a larger portion of the visual field) than a coffee mug placed on a table in the back. In reality of course, the mug's real-world size is the same. If these expectations about the relation between the scene and object are violated, observers are less likely to identify a target object. This was demonstrated in an experiment where human observers were more likely to miss a target object when its size was not compatible with the rest of the scene, even though the object was displayed as having a larger size than usual (Eckstein et al., 2017).

Neural correlates of contextual facilitation

Recent neuroimaging studies have expanded on the behavioral theories of contextual facilitation above, in order to identify the neural correlates of this effect. Using functional Magnetic Resonance Imaging (fMRI), Brandman & Peelen (2017) investigated the interaction between the neural mechanisms of object and scene processing. Here, multivariate pattern analysis (MVPA) (see Box 1) was employed to compare participant's brain activity patterns during three conditions: 1) observing degraded (i.e., pixelated) objects in isolation, 2) observing scenes in isolation, and 3) observing degraded objects within scenes.

Box 1 | Multivariate pattern analysis – support vector machine

A consequence of having access to powerful neuroimaging tools, is the by-product of an overwhelming amount of data. In fMRI, a common method to extract information and patterns from these large amounts of data is multivariate pattern analysis (MVPA). Where univariate approaches mainly provide information on whether brain areas show differences in mean activation across conditions, MVPA takes into account the systematic differences in multi-voxel activity patterns across conditions. A technique frequently used in combination with MVPA is the linear support vector machine (SVM). This is a supervised machine learning algorithm that can be trained to distinguish between patterns of brain activity that are evoked by two sets of stimuli (e.g., clear images of animate or inanimate objects). These patterns can be derived from all brain voxels (whole-brain analysis) or from specific subsets of voxels (region-of-interest [ROI] analysis). Next, the algorithm is tested on a new set of patterns of brain activity that are evoked by a new stimuli set. This test returns an accuracy score that informs the user on how accurate the algorithm can classify the testing data according to the labels that were supplied by the training data. A higher accuracy score implies that the algorithm (or classifier) can more reliably decode certain patterns of activity from new data, on which the algorithm was not trained. This then reflects to what extend certain information is reflected in the brain activity within a given brain area.

Critically, the degraded objects were ambiguous when presented in isolation, but could be categorized without difficulty when presented within their scene. The results showed that object classification improved significantly when the object was presented in a scene, compared to the classification of

objects in presented in isolation or scenes in isolation. This contextual facilitation effect was observed in object-selective areas, but not in scene-selective areas. This implies that object-selective areas encompass a shaper object representation, elicited by the scene in which the object is presented.

Later, this claim was causally supported by a study using Transcranial Magnetic Stimulation (TMS) (Wischnewski & Peelen, 2021). In a similar setup, participants observed objects in isolation and objects presented in scenes. For each stimulus, participants had to determine as fast as possible to which of eight categories the object belonged. After the stimulus was presented, two TMS pulses were delivered with various latencies after stimulus onset: 60-100ms (early), 160-200ms (middle), and 260-300ms (late). The targeted regions were the object-selective lateral occipital cortex (LOC), the scene-selective occipital place area (OPA), and the early visual cortex (EVC). The main results showed that object recognition in scenes was significantly impaired after stimulation of the OPA at the middle time window, and after stimulation of the LOC at the late time window. This demonstrates that the object-selective LOC receives feedback from the scene-selective OPA, suggesting that scene-context provides important information that is needed to disambiguate the identity of an object.

The role of object size and inferred viewing distance

Although the neural evidence for scene-contexts being able to facilitate object recognition is accumulating, it is unclear what specific contextual cues contribute to such facilitation effects. Previous work has demonstrated the important role of the expected object size, in which the distance between the object and observer acted as a contextual cue (Gayet & Peelen, 2022). Using an fMRI approach, a cue instructed participants to search for either a melon or a box, while simultaneously instructing participants on whether the object would appear nearby (in the bottom plane of the screen) or far away (in the top plane of the screen). Accordingly, this cue allowed participants to make a prediction about the upcoming object's retinal size. In half of the trials, the cue was followed by a scene with the melon or box appearing at the location of the cue, while in the other half of the trials the cue was followed by a scene without objects. This allowed the researchers to measure neural responses evoked by search preparation. A classifier that was trained on brain activity patterns of participants observing intact melons and boxes was able to classify these objects above chance when tested on the brain activity patterns evoked by search preparation. This was the case for the object-selective LOC, but not for EVC. Critically, a second analysis revealed that classification accuracies were higher when the classifier was trained on object sizes that corresponded with the viewing distance: when the classifier was trained on brain activity patterns evoked by large objects, classification accuracies were higher when the classifier was tested on brain activity patterns evoked by preparatory search activity for objects nearby, than for objects far away. This reveals that observers make predictions about the object size of a search target, based on the distance between the observer and the object.

The study by Gayet and Peelen (2022) demonstrates how the expected object location, and thus inferred distance from which the object is observed, can elicit predictions that are reflected in the neural representations about the upcoming object's size: objects appearing in the top plane of the

screen will appear to be viewed from a larger distance, resulting in a smaller estimated object size, while objects appearing in the bottom plane of the screen will appear to be viewed from a smaller distance, resulting in a larger estimated object size. This contributes to the existing literature that indicates that object size is an important contextual cue in visual search (Eckstein et al., 2017).

The point at issue in this study concerns the role of such size-distance cues in object recognition. Is the expected object size one of the important contextual cues that can facilitate the object's neural representation in the visual cortex? Considering the findings that 1) contextual cues are able to facilitate object recognition (Palmer, 1975; Eckstein et al., 2017) and object representation in the object-selective cortex (Brandman & Peelen, 2017; Gayet & Peelen, 2022), and 2) observers make predictions about object size based on viewing distance (Gayet & Peelen, 2022), we will investigate the influence of facilitating contextual cues, specifically the expected object size, derived from the inferred viewing distance, on the representation of objects in the object-selective cortex. Specifically, we ask: are visual object representations in object-selective cortex modulated by inferred viewing distance?

Here, we use fMRI to test whether the object's representation is enhanced when the size and location of that object matches the expectations of the observer compared to when the object's size and location are manipulated, such that they do not match the expectations of the observer. Additionally, we test whether these representations are enhanced in specific areas of the visual cortex compared to other areas. It is expected that objects that match the expectations of the observers are better represented in the visual cortex than objects that do not match the expectations. The analyses will be performed in the object-selective LOC and in the EVC, which are known to be involved in object size perception (Murray et al., 2006) and object recognition (Grill-Spector et al., 2001). Additionally, the analyses will be performed in the intraparietal sulcus (IPS), which is involved in object-scene scale consistency.

Following the conclusions of Gayet & Peelen (2022) on the one hand, it is expected that the congruency effect will be more profound in the LOC compared to the EVC. For example, it might be the case that although the EVC is involved in object size perception, these perceived differences do not contribute to the ability to recognize or classify an object. Instead, these tasks might depend to a greater extent on higher-level visual areas (such as the LOC). On the other hand, a recent study that investigated the influence of TMS on the Ponzo illusion effect elicits contrasting expectations (Zeng et al., 2020). In the Ponzo illusion, two horizontal lines are placed at different heights of two converging vertical lines (Ponzo, 1910). The upper line is estimated as longer, due to the sense of depth that is created by the converging lines. The authors revealed that, although TMS affected the strength of the Ponzo illusion effect after stimulation of both the EVC and LOC, strongest effects were found for stimulating the EVC at a late time window. They described a possible scenario in which the EVC forwards low-level stimulus information to the LOC, while the LOC receives contextual information from scenes, which is subsequently fed back to the EVC. Following this line of thought, we might expect to observe a congruency effect in both the LOC as well as in the EVC.

Methods

Participants

A total of 35 participants (21 female, mean age = 24.8, SD = 3.92) were recruited via the Radboud University participant pool (SONA systems). Upon participation, participants received a monetary reward. The study was conducted in accordance with the institutional guidelines of the local ethical committee (CMO region Arnhem/Nijmegen, the Netherlands, Protocol CMO2014/288).

One participant was removed from all analyses, since an initial analysis revealed that animacy information could not be reliably decoded from the participant's brain activity.

Apparatus

Stimuli were presented on a BOLDscreen 32, MR-proof monitor, synchronized lag-free to a 1920 x 1080 pixels @ 120 Hz video signal. Before the session, participants were guided into the MRI bore to adjust the mirror that was mounted on the head coil. When the stimulus screen was fully visible in this mirror, the experiment proceeded.

Participants used a handheld button box to carry out a 1-back task. The button box was held in the right hand, and operated with the index finger.

General experimental procedure

Participants registered for a time slot via SONA systems. Before entering the scanner facilities, participants received a short, verbal explanation of the tasks they were about to perform inside the scanner. Participants were instructed to pay close attention to the stimuli, because they had to answer a question concerning these stimuli at the end of the experiment. Inside the scanner, participants were able to practice the task during the five-minute anatomical scan. After every run, the experimenter checked if the participant was doing okay, or whether they needed a break.

The scanner session started with an anatomical scan, followed by 13 experimental runs: four so-called training runs, six so-called testing runs, and three functional localizer runs. The order of the runs was balanced, such that consecutive runs were not the same. This was the same for every participant. Every run lasted approximately five minutes. Apart from the IPS localizer run, participants performed a 1-back task during every run. After the scanner session, the participants completed a separate behavioral task on a laptop, measuring the participant's susceptibility to the Ponzo size illusion (Ponzo, 1910).

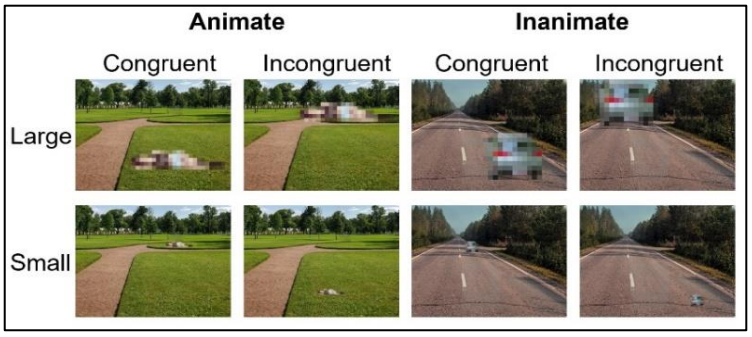
Experimental design & stimuli

Testing runs. The testing runs consisted of degraded objects presented within a scene. The objects were either animate (containing the categories: humans, dogs, kangaroos, boars, bears, apes and cows) or inanimate (containing the categories: signs, tables, chairs, cars, motorcycles, suitcases, and wheelbarrows). The degradation of the objects was done by using the filter 'Pixelate' → 'Mosaic' function of Photopea (an online, free version of Photoshop). A cell size of 5 or 6 was chosen to make sure that the object would not be recognizable in isolation, but would be recognizable within a scene. The size of the object was either large or small, while the position of the object was either nearby or far away. This way, an object was either congruently located (e.g., a small object far away), or incongruently located (e.g., a small object nearby) within a scene. Stimuli were created using Photopea. Photographs of 64 unique scenes were selected from www.unsplash.com. The scenes elicited a clear sense of depth, did not contain other objects, and contained some

texture, such that the contrast between scene and object did not lead to an immediate saliency effect of the object. Every scene was able to contain all object categories with approximately equal probabilities, implying that the type of scene was not predictive of the type of object that was located within the scene. Photographs of 64 unique objects were selected from vhv.rs. The objects varied in orientation, color and posture, such that each category was broadly represented. Inanimate objects with bright colors such as red and green were avoided, since these colors are not common among animate objects, and are therefore more likely to be categorized as inanimate. Across all categories, animate and inanimate objects were selected to be comparable in real-world size and global shape, to increase the potential to be confused with another category. For examples of the testing run stimuli, see Figure 1.

Figure 1

Examples of testing run stimuli



Note. Objects were either animate or inanimate, and the different combinations of sizes and locations of the object determined whether the object was congruently or incongruently located within its scene.

The stimuli were selected from a previously conducted online study (unpublished), in which participants observed the objects for only 150ms, after which they had to categorize the object within the scene (either located in a congruent or incongruent location) as living or non-living. The 64 out of 84 stimuli that showed the largest congruency effect (better performance for congruently located objects than incongruently located objects) were selected for the fMRI study.

Participants had to press a button when they saw the exact same stimulus twice in a row. The six training runs each contained four mini blocks, in which a block of animate objects (16 stimuli) was followed by a block of inanimate objects (16 stimuli). Within a mini block, this pattern was repeated twice, such that every mini block consisted of 16 x 4 = 64 consecutive image presentations, resulting in 64 x 4 = 256 image presentations in total for each run.

The runs were preceded by a 16s fixation cross. Every stimulus appeared on the screen for 150ms, with an 850ms interval between every stimulus. The training runs consisted of 4 mini blocks, with a longer fixation (16s) after every mini block.

The testing runs were counterbalanced such that every participant was exposed to every scene, but only one variation of the object's location and size (near-large, far-large, near-small or far-small) within that scene. This prevented participants from recognizing objects from an earlier trial with the same scene but a different condition. Across the whole experiment, all 16 categories x 2 congruency conditions x 2 distance conditions were counterbalanced. Additionally, the mini blocks were counterbalanced within-subjects, such that each mini block contained every condition (animate-congruent, inanimate-congruent, animate-incongruent and inanimate-incongruent), and all eight categories within an animacy condition.

Training runs. The training runs consisted of isolated (not including a background scene), large objects of high resolution, that were presented at fixation. These objects were novel exemplars from the same categories

that were used in the testing runs. The training runs allowed for the extraction of optimal brain activation patterns in response to animate and inanimate objects.

The task, timings, number of mini blocks and number of trials were similar to the testing runs.

Functional localizer runs. Two localizer runs were used to define regions of interest (ROIs) at the individual subject level. The specifics on individual ROI selection are described later.

The stimuli consisted of intact objects, scrambled objects, scenes, and faces. Participants had to press a button when they saw the exact same stimulus twice in a row. The stimulus appeared on the screen for 300ms, followed by a fixation cross for 450ms. Every localizer run contained 4 mini blocks, with the mini block consisting of separate blocks containing solely intact objects, scrambled objects, faces or scenes. Each block contained 20 trials, resulting in 20 x 4 = 80 consecutive image presentations per mini block, and 80 x 4 = 320 image presentations in total for each run.

A third localizer run was used to identify the IPS (adopted from Welbourne et al., 2021). Welbourne and colleagues (2021) provided evidence for the involvement of the IPS in object-scene scale consistency. This was reflected in stronger BOLD responses to objects that were congruently located within a scene (object-scene scale consistent), than to objects that were incongruently located within a scene (object-scene scale inconsistent). To localize this region, Welbourne and colleagues (2021) describe a localizer task in which participants need to fixate on a fixation dot during alternating blocks lasting 20s, where the fixation dot was either moving or not moving. In blocks where the fixation dot is moving, the dot jumped to randomized positions along the horizontal axis of the screen, such that participants had to make saccades. The fixation dot stayed there for 500ms. In blocks where the fixation dot was not moving, the fixation dot remained in the center of the screen for 20s. The run contained 11 blocks of fixation, and 10 block of saccades.

Considering the IPS's involvement in object-scene scale consistency, it is expected that a congruency effect will be observed in this area, implying enhanced object representation for congruently located objects, compared to incongruently located objects.

Ponzo illusion task. As described earlier, the Ponzo illusion (Ponzo, 1910) is the visual illusion where two converging, vertical lines create the illusion of depth. When two parallel, horizontal lines of equal length are placed on top of the vertical lines, it appears as if the upper horizontal line is longer than the lower horizontal line, while in reality the lines are equally wide. It is expected that individual differences that will be observed in the main task, will mirror the individual differences in the Ponzo illusion task: the congruency effect in the main task would correlate with the degree to which objects in the top plane are estimated as larger than reality, compared to objects located in the bottom plane. It is hypothesized that participants who are strongly influenced by the implied distance in the Ponzo illusion, would be hindered in object representation by incongruently located objects (or helped by congruently located objects) in the testing runs.

The (behavioral) task used here is similar as one of the tasks that was reported in Gayet & Peelen (2019). Participants first observed a fixation dot that was displayed for 1s, after which a scene with a depth percept

appeared for 1s. The scene contained an object that was either located nearby or far away. Next, the same object was displayed in isolation, but with a different, randomized size. Participants were to adjust the size of the object, until it matched the size of the object they just observed within the scene.

Participants completed two blocks of 32 trials each, in which the location of the object (near or far) was balanced.

Data acquisition and preprocessing

The data was collected with a 3T Magnetom Skyra MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany), using a 32-channel head coil. A T1-weighted anatomical scan was acquired, using an MPRAGE sequence (TR 2.3s, TE 3.03ms, flip angle: 8°, 1mm isotropic voxels, 192 sagittal slices, FOV 256mm). Functional images were acquired by a T2*-weighted gradient echo EPI sequence (TR 1s, TE 34ms, flip angel: 60°, 2mm isotropic voxels, 66 slices).

Preprocessing of fMRI data was carried out using SPM12. The data were field-map corrected, unwarped, realigned, co-registered with the participant's anatomical data, segmented, normalized into MNI152 space and smoothed with a 3mm full width half maximum Gaussian filter.

ROI selection

Object-selective cortex (OSC). Using a general linear model (GLM), the response evoked by participants observing intact objects vs scrambled objects during the functional localizer runs was modeled. The regressor of interest was derived from the individual mini blocks, of which a boxcar function was created. This was then convolved with the canonical hemodynamic response function (HRF). Six additional motion regressors and one run-based regressor were included as nuisance regressors. For each participant, voxels in pre-informed subregions were selected that significantly ($p_{uncorrected} < .05$) responded stronger to intact objects than to scrambled objects. The pre-informed subregions consisted a functional-defined mask of the LOC, retrieved from Julian et al. (2012). The LOC ROI size varied across participants and hemispheres (left hemisphere: *M* = 1364, *SD* = 657, min = 147, max = 2937; right hemisphere: *M* = 1143, *SD* = 615, min = 193, max = 2485).

Early visual cortex (EVC). Early visual cortex regions of interest were identified in a similar manner as the object-selective cortex. The ROIs were acquired for each participant, by selecting voxels in pre-informed subregions that significantly ($p_{uncorrected} < .05$) responded stronger to stimuli (intact objects + scrambled objects + scenes + faces) than to no stimuli. The pre-informed subregions consisted of an anatomical mask of Brodmann's Areas 17 and 18, corresponding to the primary and secondary visual cortex. The EVC ROI size varied across participants and hemispheres (left hemisphere: M = 3387, SD = 617, min = 2042, max = 4901; right hemisphere: M = 3788, SD = 600, min = 2658, max = 5313).

The ROI selection procedure for scene- and face-selective regions have been reported in Supplement 1.

Statistical analyses

General Linear Model estimation – animacy. A GLM was used to model the evoked response of each individual participant for all four training runs. In these training runs, the regressors of interest were based on animacy, resulting in the regressors 'object' and 'animal'. A boxcar function was created for each regressor, which was convolved with the canonical HRF. Six additional motion

regressors and one run-based regressor were included as nuisance regressors.

Similarly, a GLM was used to model the evoked response of each participant for all six testing runs. Here, the regressors of interest were based on animacy, as well as congruency, resulting in the regressors 'animal-congruent', 'object-congruent', 'animal-incongruent' and 'object-incongruent'. Six additional motion regressors and one run-based regressor were included as nuisance regressors.

Multivariate Pattern Analysis. The multivariate pattern analyses were conducted by making use of The Decoding Toolbox (Hebart et al., 2015), using MATLAB. First, to verify whether the ROIs indeed contained animacy information, from which animate and inanimate objects could be classified, a within-participants leave-one-out cross-validation approach was used. Here, the SVM was trained on the first training run, and tested on the remaining three training runs. This pattern was repeated for the other three training runs. In the same manner, a within-participants leave-one-out cross-validation approach was applied to the testing runs. Since the testing runs did not only contain information about animacy, but also on congruency, the SVM's classification accuracy for discriminating between animate and inanimate conditions was calculated separately for the two congruency conditions (by comparing congruent animate objects vs congruent inanimate objects, as well as incongruent animate objects vs incongruent inanimate objects). Provided that the SVM would be able to accurately decode animacy from the individual training and testing run data, this would imply that animacy information is indeed available in the tested ROIs. Under these terms, it is justifiable to continue with the main analyses.

Important to note is, that even though it is possible to calculate a congruency score from the testing run data above (by comparing animacy classification accuracies in congruent vs incongruent data), this does not disclose whether there is a disparity in the degree of object representation between congruent and incongruent stimuli. After all, the classifier was trained and tested on the same type of data. A potential congruency effect in testing run data does not provide insight on what specific type of information is underlying this congruency effect. To make a statement about the difference in the degree of object representation between congruent and incongruent stimuli, the SVM needs to be trained on the isolated, intact, large objects presented at fixation. This way, the training data only provides information on whether an object is animate or inanimate. How well the classifier can make a distinction between animate and inanimate objects can then be tested for both congruent and incongruent objects in a scene. A potential difference in the classification accuracies between animate and inanimate objects is then purely driven by congruency, rather than other cues.

This is precisely what was being investigated in the cross-classification analysis. Here, the SVM was trained on training run data and tested on testing run data. This way, the SVM learned to distinguish between patterns of brain activity based on animacy (animacy classifier), and was then tested on the ability to generalize this information to a different dataset that was until then unfamiliar.

The analyses above were carried out for the object-selective LOC, the IPS and the EVC. The degree of accuracy was reported using z-score averaged distance to bound, scaled as such that -1 equaled minimal performance, and 1 equaled maximum

performance. A score of zero corresponded to chance level performance.

Ponzo illusion task. The main result of the Ponzo illusion task comprised the difference in the estimated size for objects located nearby and objects located far away. This was accomplished by calculating the average overestimation in size, for both conditions (near and far), for each participant. To test whether these averages were different from zero, two one-sample t-tests were conducted. Next, the difference between these averages was computed, after which a one-sample t-test was used to test whether this difference was different from zero.

Thereafter, the correlation between the congruency effect on the one hand, and the performance on the Ponzo illusion task on the other hand was calculated. This analysis was carried out for the object-selective LOC, IPS and EVC, provided that a significant congruency effect was found here. Kendall's τ correlations are reported because of the violated assumption of normality.

Results

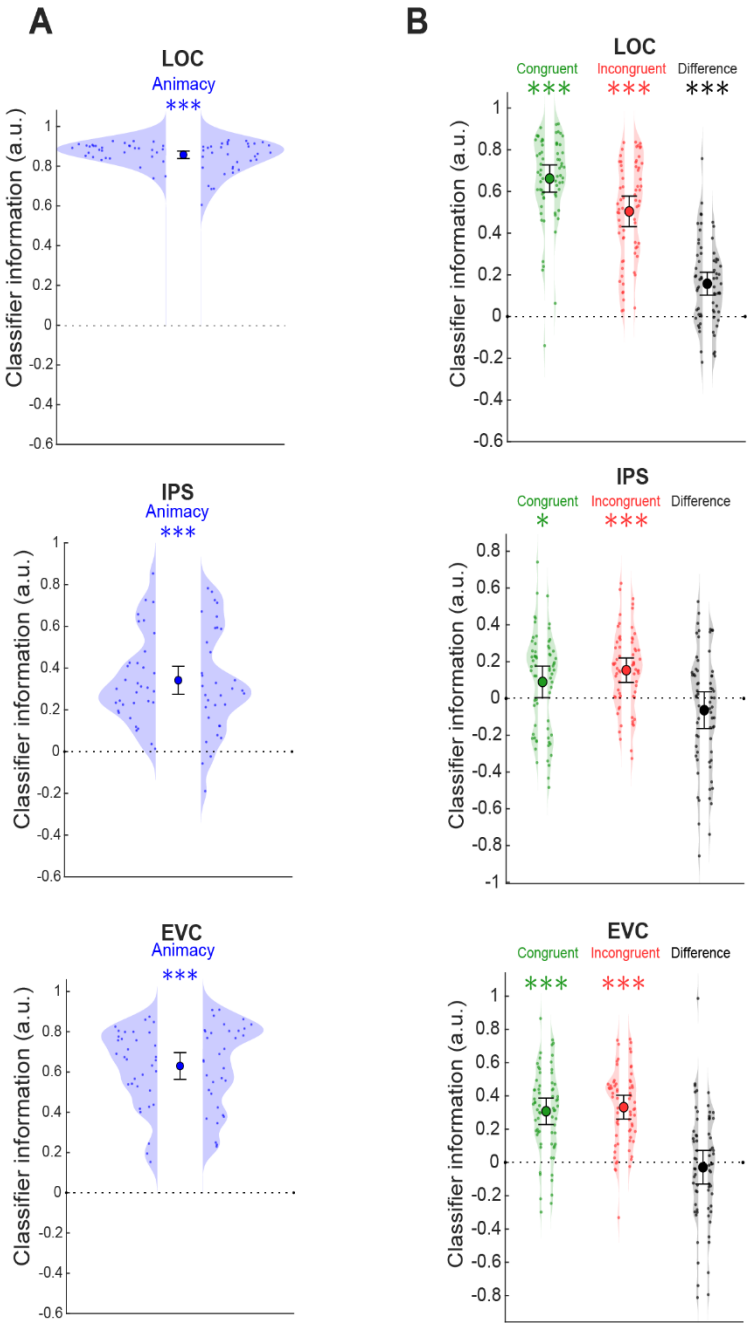
Within run type-classification – training. Before investigating whether animacy information can be decoded directly from the data, it is required to verify that the ROIs on which the SVMs are trained and tested actually contain animacy information. This was done by training an SVM according a leave-one-out cross-validation approach, in which the SVM was trained and tested on training data (consisting of isolated, intact, large objects presented at fixation). As can be seen in Figure 2A, results showed that animacy information could reliably be decoded in the LOC (M = .86, p = <.001, 95% CI = [.84, .88]), the IPS (M = .34, p = <.001, 95% CI = [.27, .41]), and the EVC (M = .63, p = <.001, 95% CI = [.56, .70]). This suggests that activity in these ROIs contains information that allows for distinguishing between animacy conditions.

Within run type-classification – testing. Similarly, an SVM was trained and tested on testing run data (consisting of objects in scenes). Figure 2B displays the results which reveal that animacy information could reliably be decoded in the LOC. This was the case for congruent objects (M = .66, p = <.001, 95% CI = [.60, .73]), as well as incongruent objects (M = .50, p = <.001, 95% CI = [.43, .58]). The difference in classification accuracy between congruent and incongruent objects was significant (M = .16, p = <.001, 95% CI = [.11, .21]), indicating that in the LOC, the classifier's ability to classify an object as animate or inanimate is improved when objects are congruently located within a scene, compared to when objects are incongruently located within a scene.

Different results were found for the IPS and EVC. Animacy information could reliably be decoded for congruent objects (IPS: M = .09, p = .041, 95% CI = [.00, .18]; EVC: M = .31, p = <.001, 95% CI = [.23, .39]) as well as incongruent objects (IPS: M = .15, p = <.001, 95% CI = [.09, .22]; EVC: M = .33, p = <.001, 95% CI = [.26, .40]). Critically, however, the classification accuracy between congruent and incongruent objects was not significant (IPS: M = -.06, p = .203, 95% CI = [.-.16, .04]; EVC: M = -.02, p = .624, 95% CI = [.-.13, .08]). This indicates that in the IPS and EVC, the classifier's ability to categorize an object as animate or inanimate is not affected by the object being congruently or incongruently located within the scene.

Figure 2

Within run-type validation for training and testing runs



Note. (A) Within run-type validation – training runs. An algorithm was trained on training runs, and tested on a different run from the same data type. (B) Within run-type validation – testing runs. An algorithm was trained on testing runs, and tested on a different run from the same data type. *p < .05, ***p < .0005. Error bars reflect 95% CI of the mean.

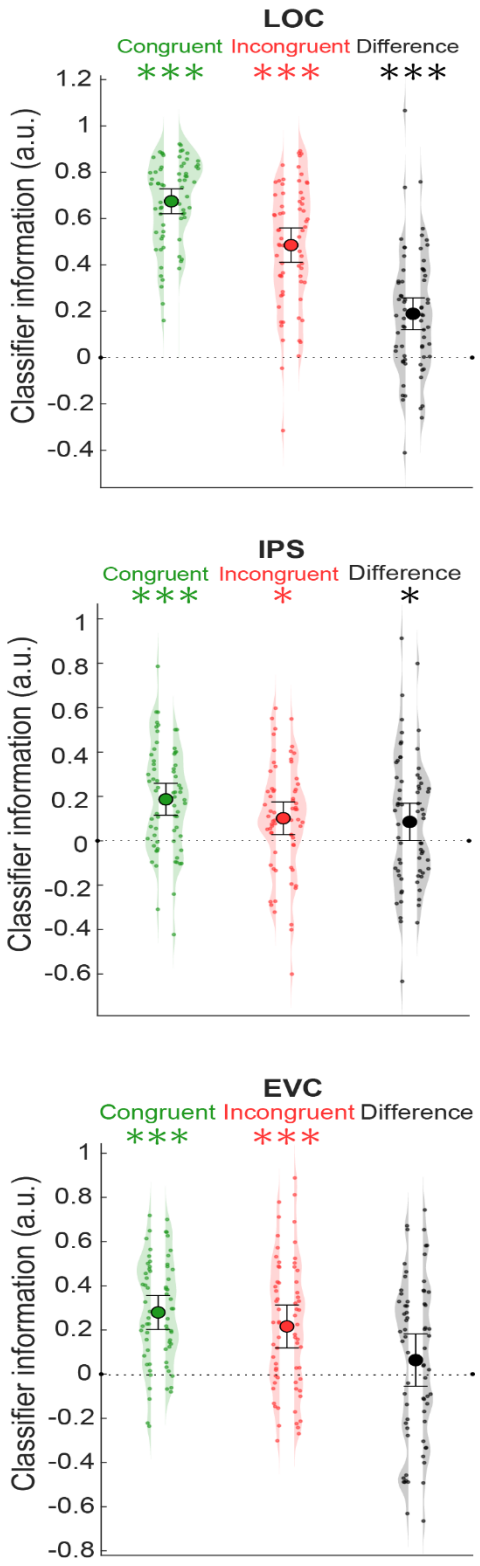
Cross-classification. For the main analysis in this study, an SVM was trained on training run data, and tested on testing run data. The results are displayed in Figure 3. Animacy information could reliably be decoded in the LOC and IPS for congruent objects (LOC: *M* = .67, *p* = <.001, 95% CI = [.62, .73]; IPS: *M* = .19, *p* = <.001, 95% CI = [.11, .26]), as well as incongruent objects (LOC: *M* = .49, *p* = <.001, 95% CI = [.41, .56]; IPS: *M* = .10, *p* = .009, 95% CI = [.03, .18]). The difference in classification accuracy between congruent and incongruent objects was significant (LOC: *M* = .19, *p* = <.001, 95% CI = [.12, .26]; IPS: *M* = .09, *p* = .048, 95% CI = [.00, .17]), indicating that in both the LOC and the IPS, the classifier's ability to classify an object as animate or inanimate is improved when objects are congruently located within a scene, compared to when objects are incongruently located within a scene.

In the EVC, the classifier could reliably make a distinction between animate and inanimate objects for congruent objects (*M* = .28, *p* = <.001, 95% CI = [.20, .36]) as well as for incongruent objects (*M* = .22, *p* = <.001, 95% CI = [.12, .31]). Again, the critical difference between the classification accuracies of congruent and incongruent objects did not reveal the congruency effect that was observed for the LOC and the IPS (*M* = .06, *p* = .284, 95% CI = [-.06, .18]). Ultimately, this reveals a gain in classification accuracy when objects are congruently located in a scene for the LOC and IPS, but not for the EVC.

Critically, when comparing the congruency effect between LOC and EVC, it is revealed that they differ significantly (*t*(33) = 2.31, *SD* = 0.32, *p* = .03, 95% CI = [.01, .24]). This, however, is not the case for IPS and EVC (*t*(33) = 0.32, *SD* = 0.40, *p* = .754, 95% CI = [-.12, .16]).

Figure 3

Cross-classification



Note. Cross classification. An algorithm was trained on training run data and tested on testing run data. *p < .05, ***p < .0005. Error bars reflect 95% CI of the mean.

Ponzo illusion task. Two one-sample t-tests were conducted to test whether the average estimation of object sizes in both conditions were different from zero. For objects located far away, participants were inclined to overestimate its size (M = 0.12, $t(33)$ = 5.55, SD = 0.12, p = < .001, 95% CI = [.08, .16]). For objects located nearby, participant were not inclined to over- or underestimate its size (M = 0.04, $t(33)$ = 1.80, SD = 0.12, p = .081, 95% CI = [< .01, .08]). The one-sample t-test confirmed that the difference in average estimated size between objects nearby and objects far away was different from zero ($t(33)$ = 2.83, SD = 0.17, p = .008, 95% CI = [.02, .14]. This indicates that participant are inclined to overestimate the object sizes of objects located far away, compared to the object sizes of objects located nearby. This is in line with the classic results from the Ponzo illusion, which refers to the tendency to overestimate the size of an object in the top plane of a depth-inducing scene, in comparison to the size of an object in the bottom plane of a scene.

No correlations were found, however, between the congruency effect and the magnitude of the Ponzo illusion across participants. This was the case for both for the LOC (τ = .05, p = .680) and the IPS (τ = -.05, p = .702). This indicates that there is no relation between the extent to which object congruency influences object representation in the brain, and size estimation of an object. The analysis was not performed in the EVC, since no significant congruency effect was found in this ROI.

Discussion

In the present study we investigated how the size and location of an object in a scene modulates its representation in different brain areas. We predicted that objects that were congruently located within a scene (i.e.,

when the inferred viewing distance matched the size of the object), evoked sharper representations in the visual cortex than objects that were incongruently located within a scene (i.e., when the inferred viewing distance did not match the size of the object). Specifically, we expected an enhanced representation in the LOC and IPS. For the EVC, separate scenarios were composed.

The results confirm a sharper object representation in LOC and IPS, but not in EVC. What is notable, is that animacy information, whether it is congruently or incongruently presented, is available in all ROIs: a classifier was able to reliably make a distinction between animate and inanimate objects, independent on their size and location within the scene. However, when we inspect the difference in animacy classification accuracies between congruent and incongruent objects, we notice a significant congruency effect in the LOC and IPS, but not in the EVC. This suggest that, although the EVC is reported to be involved in object size perception (Murray et al., 2006), these perceived differences in EVC do not play a role in object classification. Instead, this role is reserved for higher-level visual areas such as the LOC and IPS.

Furthermore, no significant relation was found between the magnitude of the congruency effect and the magnitude of the Ponzo illusion effect. In contrast to the hypothesis, participants who were strongly influenced by the implied distance in the Ponzo illusion do not seem to be hindered in object representation by incongruently located objects (or helped by congruently located objects) in the testing runs.

Based on the current data alone, we cannot be sure that the congruency effect resulted from the extraction of distance-related information from the scene. Alternatively, it might reflect that visual cortex is more sensitive to discriminating smaller objects in the upper visual hemifield and larger objects in the lower visual hemifield, as this is how objects typically appear in real-life. A previously conducted online study revealed that this is probably partially the case. In a first behavioral experiment, participants performed a classifying task (animate vs inanimate) on the stimuli from the testing runs. As expected, a clear congruency effect was observed. In a second experiment, a new set of participants performed the same task on the same stimuli, except for the fact that the objects were placed on a mean luminance background, rather than a depth-inducing scene. Although a congruency effect was observed from the data, the strength of the effect was half the size of the first experiment. This suggests that the size of an object as a function of its position is important, but adding a depth-inducing scene substantially contributes to the congruency effect.

A factor that might have hindered the aim of the study, was the fact that due to the unexpected locations of incongruent objects, it is no guarantee that all objects were observed. This could have been prevented by presenting the objects at fixation, causing the scene to shift from its central position. This might have strengthened the effects we found. Additionally, despite our best efforts, it might have been possible that the objects used in this experiment did not only differ in animacy, but also in shape, color, or the ratio between straight vs round edges. However, this does not affect the interpretation of the results: although the objects in the training runs were novel, and different from the objects in the testing runs, they were selected to be of the same categories as in the testing runs. Therefore, the training run stimuli can be generalized to the testing run stimuli. The classifier is thus tested on object features that were also included in the intact isolated

objects, causing the classifier's performance to be influenced only by the congruency of the objects.

The current results pave the way for an interesting follow-up study. According to previous and current literature, scene-context provides information that is important to disambiguate the identity of an object (Palmer, 1975; Bar, 2004; Oliva & Torralba, 2007). This was complemented by neuroimaging studies, revealing that contextual information enhances object representation in the object-selective cortex. More specifically, Wischnewski & Peelen (2021) provide evidence that the object-selective LOC receives this contextual feedback from the scene-selective OPA at an intermediate time window of 160-200ms. In this study, we extracted one type of contextual information, and showed that the expected object size is one of the important contextual cues that can facilitate object recognition. The next step is to investigate whether the congruency effect that was found here, is caused by projections from the scene-selective OPA to the object-selective LOC.

This can be explored by a TMS study following a similar design as the study by Wischnewski & Peelen (2021). However, rather than presenting participants with isolated, intact objects and degraded objects in scenes, we would use the congruent and incongruent images from this study's testing run stimuli. For congruent stimuli, it is then likely that the feedback which is projected from the OPA to the LOC contains useful information, which will help object recognition. However, for incongruent stimuli, the contextual information does not match prior expectations. Therefore, it is likely that the feedback that is projected from the OPA to the LOC will hinder object recognition. Considering that these feedback projections are occurring at an intermediate time window, it is expected that stimulation of the OPA during this time window will decrease participants' classification performance for congruent trials, therefore decreasing the congruency effect. Similarly, it is expected that when the feedback reaches the LOC, stimulation of the LOC at this later time window will decrease participants' classification performance on congruent trials.

The current study contributes to the accumulating evidence on contextual facilitation in object recognition. Where previous work has showed that this facilitation of object recognition is reflected in the neural correlates of the visual cortex, here, we narrowed down the importance of specific contextual cues. We conclude that the estimated object size, which can be inferred from the viewing distance is able to modulate the object's representation in object-selective cortex. The findings fit in the framework that contextual facilitation is endorsed by expectations that are evoked by scene context (Bar, 2004, Summerfield & De Lange, 2014). Additionally, the results help explain why humans still prevail in object recognition tasks in natural scenes, compared to artificial agents (Borji & Itti, 2014).

Besides size-distance relations, there are more specific contextual cues that might contribute to enhanced object recognition or object representations in visual cortex. It is worth investigating whether a similar role is reserved for contextual cues such as the source of light in a scene, that can predict the color, brightness and shadow of an object, or the angle of view from which a scene is captured, that can predict the shape of an object.

Conclusion

In the current study we investigated whether visual object representations in object-selective cortex are modulated by inferred viewing distance. The results reveal that object representations are substantially enhanced for size-congruent objects, compared to size-incongruent objects. This was the case for object-selective cortex, but not early visual cortex. Concluding, scene-context can elicit expectations about the real-world size of an object. These expectations are able to modulate the object's representation in object-selective cortex.

References

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617–629. https://doi.org/10.1038/nrn1476

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-x

Borji, A., & Itti, L. (2014). Human vs. computer in scene and object recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 113-120).

Brandman, T., & Peelen, M. V. (2017). Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. *The Journal of Neuroscience, 37*(32), 7700–7710. https://doi.org/10.1523/jneurosci.0582-17.2017

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. *Current Biology, 27*(18), 2827–2832.e3. https://doi.org/10.1016/j.cub.2017.07.068

Gayet, S., & Peelen, M. V. (2019). Scenes Modulate Object Processing Before Interacting With Memory Templates. *Psychological Science, 30*(10), 1497–1509. https://doi.org/10.1177/0956797619869905

Gayet, S., & Peelen, M. V. (2022). Preparatory attention incorporates contextual expectations. *Current Biology, 32*(3), 687-692.e6. https://doi.org/10.1016/j.cub.2021.11.062

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision research, 41*(10-11), 1409–1422. https://doi.org/10.1016/s0042-6989(01)00073-6

Hebart, M. N., Görgen, K., & Haynes, J. D. (2015). The Decoding Toolbox (TDT): a versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00088

Julian, J., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage, 60*(4), 2357–2364. https://doi.org/10.1016/j.neuroimage.2012.02.055

Murray, S. O., Boyaci, H., & Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nature neuroscience, 9*(3), 429–434. https://doi.org/10.1038/nn1641

Nassi, B., Mirsky, Y., Nassi, D., Ben-Netanel, R., Drokin, O., & Elovici, Y. (2020). Phantom of the ADAS: Securing Advanced Driver-Assistance Systems from Split-Second Phantom Attacks. Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. Published. https://doi.org/10.1145/3372297.3423359

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Palmer, T. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3*(5), 519–526. https://doi.org/10.3758/bf03197524

Ponzo, M. (1910). lntorno ad alcune illusioni nel campo delle sensazioni tattili, sull'illusione di Aristotele e fenomeni analoghi [On some tactile illusions, Aristotle's illusion, and similar phenomena]. *Archive für die Gesamte Psychologie*, *16*, 307-345.

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience, 15*(11), 745–756. https://doi.org/10.1038/nrn3838

Welbourne, L. E., Jonnalagadda, A., Giesbrecht, B., & Eckstein, M. P. (2021). The transverse occipital sulcus and intraparietal sulcus show neural selectivity to object-scene size relationships. *Communications Biology, 4*(1). https://doi.org/10.1038/s42003-021-02294-9

Wischnewski, M., & Peelen, M. V. (2021). Causal neural mechanisms of context-based object recognition. *eLife*, *10*. https://doi.org/10.7554/elife.69736

Zeng, H., Fink, G. R., & Weidner, R. (2020). Visual Size Processing in Early Visual Cortex Follows Lateral Occipital Cortex Involvement. *The Journal of Neuroscience, 40*(22), 4410–4417. https://doi.org/10.1523/jneurosci.2437-19.2020

Supplement 1

ROI selection

Scene- and face-selective areas. Scene- and face-selective regions of interest were identified in a similar manner as the object-selective cortex. Scene-selective ROIs were acquired for each participant, by selecting voxels in pre-informed subregions that significantly ($p_{uncorrected} < .05$) responded stronger to scenes than to intact and scrambled objects. The pre-informed subregions consisted a functional-defined mask of the parahippocampal place area (PPA) (left hemisphere: *M* = 358, *SD* = 88, min = 114, max = 564; right hemisphere: *M* = 342, *SD* = 77, min = 109, max = 465), the retrosplenial cortex (RSC) (left hemisphere: M = 405, SD = 201, min = 36, max = 815; right hemisphere: M = 569, SD = 246, min = 73, max = 1238), the occipital place area (OPA) (left hemisphere: M = 92, SD = 34, min = 10, max = 130; right hemisphere: M = 184, SD = 49, min = 73, max = 250. Note that for two participants the alpha level was adjusted to .10 and .23) and the IPS (left hemisphere: M = 2533, SD = 1263, min = 163, max = 5271; right hemisphere: M = 2684, SD = 1506, min = 149, max = 6764).

Face-selective ROIs were acquired for each participant, by selecting voxels in pre-informed subregions that significantly ($p_{uncorrected} < .05$) responded stronger to faces than to objects and scenes. The pre-informed subregions consisted a functional-defined mask of the fusiform face area (FFA) left hemisphere: M = 155, SD = 86, min = 34, max = 383; right hemisphere: M = 364, SD = 128, min = 72, max = 606). All masks were retrieved from Julian et al. (2012).