Transformation from independent to integrative coding of multi-object arrangements in human visual cortex

Daniel Kaiser[1,2,*], Marius V. Peelen[1]

[1]*Center for Mind/Brain Sciences, University of Trento, 38068 Rovereto (TN), Italy*

[2]*Institute of Psychology, Freie Universität Berlin, 14195 Berlin-Dahlem, Germany*

[*]Correspondence to:

Daniel Kaiser

Institute of Psychology, Freie Universität Berlin

Habelschwerdter Allee 45, 14195 Berlin-Dahlem, Germany

danielkaiser.net@gmail.com

Abstract

The human visual system has adapted to process cluttered scenes containing dozens of regularly arranged objects. The regularity among objects critically contributes to the efficiency of naturalistic vision. Recent studies investigating multiple object perception have demonstrated that visual cortex responses to multi-object displays can be accurately modeled by a linear combination of responses to individual objects, revealing independent processing of simultaneously presented objects. Here we use fMRI to show that this independence partly breaks down when objects are positioned according to frequently experienced configurations. Participants viewed pairs of objects that formed minimalistic two-object scenes (e.g., a "living room" consisting of a sofa and television) presented in their regularly experienced spatial arrangement or in an irregular arrangement (with interchanged positions). Additionally, single objects were presented centrally and in isolation. Multi-voxel activity patterns evoked by the object pairs were modeled as the average of the response patterns evoked by the two single objects forming the pair. In two experiments, this approximation in object-selective cortex (OSC) was significantly less accurate for the regularly than the irregularly positioned pairs, indicating integration of individual object representations. More detailed analysis revealed a transition from independent to integrative coding along the posterior-anterior axis of the visual cortex, with the independent component (but not the integrative component) being almost perfectly predicted by object selectivity across the visual hierarchy. These results reveal a transitional stage between individual object and multi-object coding in visual cortex, providing a possible neural correlate of efficient processing of regularly positioned objects in natural scenes.

Introduction

Our everyday environments are cluttered, consisting of a large number of separable objects. It is therefore of fundamental importance to understand how the visual system processes the multitude of objects contained in real-world scenes in an efficient way.

However, much of our knowledge about the neural mechanisms of object perception comes from neuroimaging studies using single objects as stimuli. These studies have linked object perception to activity in object-selective cortex (OSC), a region which preferentially responds to objects compared to scrambled objects (Grill-Spector, 2003; Malach et al., 1995) and exhibits reliably discriminable response patterns for different types of objects (Eger et al., 2008; Haxby et al., 2001; Kriegeskorte et al., 2008). Further neuroimaging work has demonstrated that OSC responses show some degree of size- and location-invariance (Cichy et al., 2011; Vuilleumier et al., 2002) and party reflect perceived rather than physical object properties (Haushofer et al., 2008; Kourtzi and Kanwisher, 2001), suggesting that OSC activation can be linked to behavioral performance in object recognition (Grill-Spector et al., 2000; Williams et al., 2007).

Recently, studies have moved towards more naturalistic conditions, measuring OSC responses when multiple objects are presented simultaneously. These studies have demonstrated that OSC codes multiple objects in an independent, linearly additive way: Responses to pairs of objects could be modeled as a linear combination (e.g., the average) of the responses to their constituent single objects (Agam et al., 2010; Baeck et al., 2013; Kaiser et al., 2014b; Kubilius et al., 2015; MacEvoy and Epstein, 2009; Reddy et al., 2009; Zoccolan et al., 2005), even when the objects were embedded in complex natural scenes (MacEvoy and Epstein, 2011).

63 Importantly, however, individual objects in natural scenes usually adhere to
64 typical and recurring positional structures: constrained by their function, their
65 relationship with other objects, and physical properties of the world, objects appear in
66 predictable locations relative to each other (e.g., sofas facing TV sets, or cars stopping in
67 front of traffic lights). Previous research has demonstrated that such inter-object
68 regularities facilitate behavioral performance in capacity-limited tasks (Bar, 2004; Chun,
69 2000; Kaiser et al., 2014a; Oliva and Torralba, 2007; Wolfe et al., 2011), suggesting that
70 regularly positioned object arrangements are more efficiently processed at a neural level.
71 In the current study, we provide evidence that the visual system integrates
72 representations of regularly positioned objects, revealing a transformation from
73 independent to integrative coding of multiple objects along the posterior-anterior axis of
74 the visual processing hierarchy.
75 In two fMRI experiments, participants viewed pairs of objects that formed
76 minimalistic versions of scenes (e.g., a "living room", consisting of a sofa and a TV; Fig. 1,
77 **A**). These pairs could be arranged in a regular, typically encountered way (e.g., the sofa
78 facing the TV) or in an irregular way (with the two objects exchanged). Additionally, each
79 single object was presented centrally and in isolation. The multi-voxel response pattern
80 evoked by each pair was then modeled by the average of the patterns evoked by its
81 constituent objects. Replicating previous observations (Kaiser et al., 2014b; Kubilius et al.,
82 2015; MacEvoy and Epstein, 2009, 2011), we find that OSC response patterns evoked by a
83 pair can be accurately predicted by the average response patterns evoked by its
84 constituent objects. Crucially, however, this linear approximation of the pair response
85 depended on the relative position of the objects: Pair approximation in OSC was less
86 accurate for regularly positioned pairs than for irregularly positioned pairs, showing that

the independence of multiple object processing breaks down when objects are positioned regularly. More detailed analyses revealed a transformation from independent to integrative representations of regularly positioned object pairs along the posterior-anterior axis of the visual cortex. The integrative component of multi-object processing reflects an emerging tuning to positional structures in real-world scenes.

Materials and Methods

Participants

Fifteen healthy adults (6 male; mean age 24.6 years, SD=3.3) took part in Experiment 1, and 16 healthy adults took part in Experiment 2 (8 male; mean age 24.7 years, SD=3.0). All participants had normal or corrected-to-normal visual acuity. All procedures were carried out in accordance with the Declaration of Helsinki and were approved by the ethical committee of the University of Trento. Due to excessive head movement, one participant was excluded from Experiment 1. Two participants were excluded from Experiment 2, one due to excessive head movement, and one due to bad functional localizer data, not allowing for a reliable definition of object-selective voxels.

Stimuli

The stimulus set comprised four pairs of objects, which each depicted a minimalistic version of a scene (Fig. 1, **A**). The four scenes were bathroom (toilet and sink), living room (sofa and television), street crossing (car and traffic lights), and playground (seesaw and slide). The stimuli were created by cutting the stimuli directly from a complete scene images, leaving the size and position the same as in the original scene ("regular" condition). For each scene, a second version was created where the object positions were swapped ("irregular" condition). These irregular versions contained the same objects, but the arrangement did not follow the typical arrangement normally experienced in scenes. For each pair, a total of eleven different stimuli were used. In addition to the pair stimuli we also showed the eight single objects centrally, and in isolation. For Experiment 2, additionally a set of 22 scenes (11 indoor rooms, 11 outdoor street scenes) was used; none of these scenes contained any of the objects that

constituted the pairs. All stimuli were surrounded by a black frame, which subtended a visual angle of 10° X 7.5°. Stimulus presentation was controlled using MATLAB and Psychtoolbox (Brainard, 1997). The stimulation was back-projected onto a translucent screen placed at the end of the scanner bore. Participants viewed the screen through a tilted mirror mounted to the head coil.

Experiment 1 procedure

The experiment consisted of multiple runs of 5 minutes each. Half of the runs were object pair runs, and half of the runs were single object runs. During object pair runs, the four different pair types (bathroom, living room, street crossing, playground) were shown in each of the two conditions (regular and irregular). During single object runs, all constituent single objects were shown. Each run consisted of 18 blocks. Within each block, twelve stimuli were presented, with each stimulus being an exemplar of the same condition (e.g., a bathroom/irregular block would contain only bathroom pairs in irregular configuration). Each block contained all eleven pairs (i.e., exemplars) of a specific condition, and one-back image-level repetition of a specific pair (i.e., exemplar). Each stimulus was shown for 400ms, with an inter-stimulus-interval of 900ms (Fig. 1, **B**); each block thus lasted 15.6s. The first nine blocks of each run consisted of eight stimulation blocks and one block of fixation baseline (in randomized order). The second nine blocks of a run were the same blocks in mirror-reversed order. The experiment comprised ten runs (two participants only completed eight runs), always starting with a pair run and then alternating between single object and pair runs. Participants had to report one-back image repetitions by pressing a button.

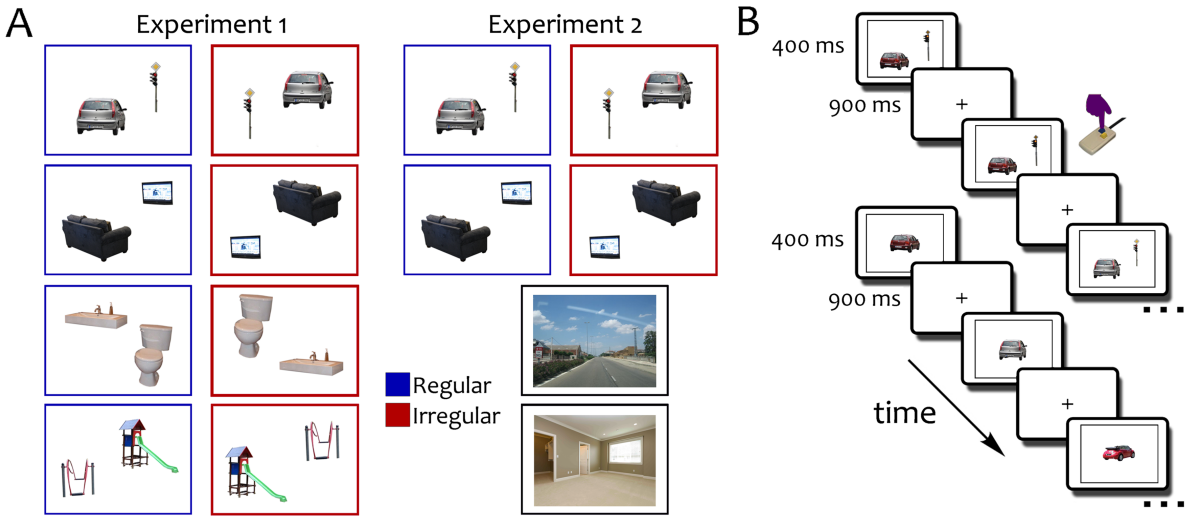

Fig. 1. Stimuli and paradigm. **A**, For Experiment 1, four pairs of objects were collected by directly cropping them from scenes of four types (bathroom, living room, street crossing, playground). Object pairs could be positioned in a regular way (as they appeared in the scenes) or in an irregular way (with their positions exchanged). In Experiment 2, only the living room and street crossing pairs were used and, in separate blocks, participants saw images of natural scene images (indoor room versus outdoor street scenes). **B**, Stimuli were presented for 400ms, separated by inter-stimulus intervals of 900ms. In each stimulation block, only pairs from one category and one regularity condition appeared (e.g., regularly positioned cars and traffic lights). Participants responded to one-back image-level repetitions. Similarly, the single objects belonging to the pairs were presented in separate runs, using the same block design and task.

Experiment 2 procedure

The procedure in Experiment 2 was identical to experiment 2 unless otherwise noted. In Experiment 2, only two of the object pairs were used (living room and street crossing). Additionally, in separate runs, the natural scene stimuli were presented (Fig. 1, **A**). These runs had an identical structure to the pair and single object runs. In alternating

order, only indoor rooms or outdoor street scenes appeared in a single block. In total, Experiment 2 consisted of 9 runs (3 pair runs, 3 single object runs, 3 scene runs), always starting with a pair run, followed by a single object run and by a scene run (this order was repeated three times).

Functional localizer

In both experiments, participants additionally completed two functional localizer runs of five minutes each. Participants performed a one-back task while viewing images of faces, houses, everyday objects, and phase-scrambled versions of these objects. Each stimulus category included 36 individual exemplars. Within each run, there were four blocks of each stimulus category and four blocks of fixation baseline, with all blocks lasting 16s. Block order was randomized for the first ten blocks and then mirror-reversed for the remaining ten blocks. Each non-fixation block included two one-back stimulus repetitions.

fMRI data acquisition and preprocessing

MR imaging was conducted using a Bruker BioSpin MedSpec 4T head scanner (Bruker BioSpin, Rheinstetten, Germany), equipped with an eight-channel head coil. During the experimental runs, T2*-weighted gradient-echo echo-planar images (EPI) were collected (repetition time TR=2.0s, echo time TE=33ms, 73° flip angle, 3 x 3 x 3mm voxel size, 1mm gap, 34 slices, 192mm field of view, 64×64 matrix size). Additionally, a T1-weighted image (MPRAGE; 1 x 1 x 1mm voxel size) was obtained as a high-resolution anatomical reference. All resulting data were preprocessed using MATLAB and SPM8. EPI volumes were realigned and coregistered to the structural image. Functional volumes

182 collected during the functional localizer runs were additionally smoothed with a 6mm
183 full-width-half-maximum Gaussian kernel.
184
185 *Region on interest definition*
186 The BOLD-signal of each voxel in each participant in the localizer runs was
187 modeled using one regressor for each stimulus category, and six regressors for the
188 movement parameters obtained from the realignment procedure. Object-selective
189 cortex (OSC) (Malach et al., 1995) was localized using an object>scrambled contrast.
190 Regions for both hemispheres were concatenated to form a bilateral ROI (Experiment 1:
191 mean ROI size 1000 voxels, SE=35; Experiment 2: mean ROI size 1036 voxels, SE=45).
192 Additionally, an early visual cortex (EVC) ROI was defined by inverse-normalizing an
193 anatomical mask into individual-subject space (Experiment 1: mean ROI size 996 voxels,
194 SE=8; Experiment 2: mean ROI size 969 voxels, SE=9). For both ROIs, an additional voxel
195 selection criterion was adopted: as we expected that optimally pair-selective voxels
196 allow for better approximation of the pair patterns by single-object patterns (Baeck et
197 al., 2013; MacEvoy and Epstein, 2009), the OSC and EVC ROIs were intersected with
198 voxels that could optimally discriminate between the different pairs. These voxels were
199 selected from a whole-brain searchlight classification analysis (see below). A-priori, a
200 voxel count of the 100 most pair-selective voxels was set for all analyses; to rule out that
201 the results were specific to this particular number of selected voxels, the analysis was
202 repeated with different voxel counts ranging from 20 to 500 (see Supplementary
203 Information). For results obtained with a different selection method and for the results in
204 scene-selective regions see the Supplementary Information.
205

Univariate Analysis

The BOLD-signal of each voxel in each participant was modeled using one regressor for each stimulus type, and six regressors for the movement parameters obtained from the realignment procedure. Activation differences were computed by averaging beta-weight estimates across voxels for each ROI. Significant differences between the regularly and irregularly positioned pairs were assessed using t-tests.

Multi-voxel pattern analysis

Multivariate pattern analysis (MVPA) was carried out on a TR-based level using the CoSMoMVPA toolbox (Oosterhof et al., 2016). For each voxel belonging to a specific ROI, TRs corresponding to the conditions of interest were selected by shifting the voxel-wise time-course of activation by three TRs (to account for the hemodynamic delay). Subsequently, for each run separately, activation values were extracted from the EPI-volumes for each TR coinciding with the onset of a specific condition; the activation values were then normalized by z-scoring the values for each voxel. For the decoding analyses, linear discriminant analysis (LDA) classifiers were trained on all but one runs, and tested on the remaining run; this procedure was repeated so that every run was left out once. Pair classification was also computed in a searchlight decoding analysis (which was used for guiding ROI definition): for this analysis, a spherical neighborhood of 100 voxels was used to identify voxels in visual cortex that could optimally discriminate the different object pairs, irrespective of the regularity condition. Significant decoding was assessed by comparing classifier accuracy to chance performance, using t-tests. For the pair combination analysis, response patterns were averaged across all TRs belonging to a specific condition, resulting in a single response pattern for every condition. Single-object

response patterns for the two objects belonging to a specific pair (e.g., car and traffic light or sofa and TV) were averaged. Thus, for each pair three response patterns were available: (1) the response pattern evoked by the pair presented in the regular arrangement, (2) the response pattern evoked by the pair presented in the irregular arrangement, and (3) the average response pattern evoked by the two constituent objects. To assess pair approximation quality in the regular and irregular conditions separately, the average single object patterns were correlated to their corresponding pair response patterns (within-correlation) and to the pair patterns evoked by the non-corresponding pairs (between-correlation) (Fig. 2). For the scene approximation analysis in Experiment 2, pair response patterns were correlated with the corresponding scene response patterns (e.g., sofa/TV and indoor room scene; within-correlation) and the non-corresponding scene patterns (e.g., sofa/TV and outdoor street scene; between-correlation), separately for regular and irregular pairs. The difference between these within- and between-correlations was used as a measure of pair approximation quality. Approximation quality was then compared for the regular and irregular pairs by using paired t-tests. All correlations were Fisher-transformed prior to statistical analysis.

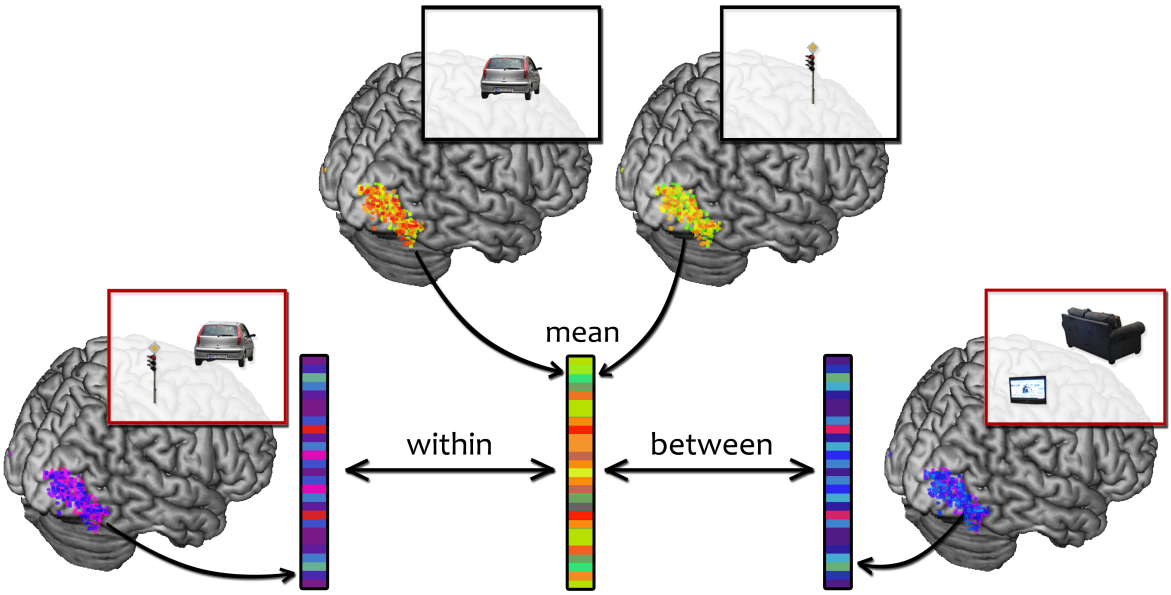

Fig. 2. Schematic illustration of the pair approximation analysis. For each ROI, the response pattern across hemispheres was extracted both for the single object and pair conditions. The single object patterns corresponding to each pair were then averaged and correlated with the pair-evoked pattern. The difference between the correlations with corresponding pairs (within-correlation) and non-corresponding pairs (between-correlation) was taken as a measure of approximation quality. This analysis was done separately for the regular and irregular pairs.

Combined searchlight analysis

To characterize the emergence of the regularity effect along the visual processing hierarchy, we performed a searchlight analysis, where we systematically moved a spatial voxel neighborhood along the anterior-posterior axis. For this analysis, the data from both experiments was combined. First, the occipitotemporal cortex (defined based on an anatomical mask) was divided into 40 slices. Each slice spanned 10mm in the anterior-posterior (Y) direction, with an inter-slice distance of 2mm (thus, individual slices were overlapping). X and Z coordinates of the occipitotemporal region were fully covered,

with the exception that the medial parts of occipital cortex were masked out to avoid sampling of EVC voxels. The 40 slices were then inverse-normalized to individual-participant space. Separately for each participant, within each slice the 100 voxels that allowed the most accurate pair decoding in a searchlight analysis were selected (similarly to the OSC analyses, see above). Subsequently, the pair approximation analysis (see above) was performed for these 100 voxels within each slice. To assess the general object-selectivity of the voxels in each slice, we computed the difference between the GLM beta weights for intact and scrambled object in the functional localizer runs (see above). To identify slices yielding significant effects, we used a threshold-free cluster-enhancement procedure (Smith and Nichols, 2009) with default parameters, using multiple-comparison correction 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 Z-values were thresholded at $Z>1.96$ (i.e., $p<0.05$) to reveal significant decoding performance.

Results

Experiment 1

To investigate whether positional regularities impact neural responses in object-selective visual cortex (OSC), we presented participants with pairs of objects that commonly appear together in real world scenes, such as in a bathroom (toilet and sink), in a living room (sofa and TV), on a street crossing (car and traffic lights), or at a playground (seesaw and slide). These objects were directly cut from natural scene images and presented on a blank background ("regular" condition; Fig. 1, **A**). To investigate whether the way in which these object typically co-occur in the real world influences neural responses, we added a condition where we reversed the object locations ("irregular" condition, Fig. 1, **A**), keeping everything else identical. Additionally, the single objects belonging to the pairs were presented centrally (Fig. 1, **B**).

Functional MRI analyses primarily focused on two regions of interest: functionally defined object-selective cortex (OSC) and anatomically defined early-visual cortex (EVC) (see below for data from scene-selective ROIs). While we expected an effect of regularity on the representation of object pairs in OSC, no effect was expected in EVC, where more basic visual analysis takes place. Based on previous work (Baeck et al., 2013; MacEvoy & Epstein, 2009), we expected stronger effects in areas of OSC that preferentially represent information about the object pairs. To define voxels that contribute most to the coding of the object pairs, a whole-brain searchlight analysis was performed (see Materials and Methods), where linear classifiers were used to discriminate the four object pairs, regardless of their positioning (i.e., collapsed across the regular and irregular pairs). Based on this searchlight analysis, within OSC and EVC the 100 voxels

301 (see Supplementary Information for results with different voxel counts) that contributed
302 most to the classification of the four pairs were selected (Fig. 3, **A**).
303 First, we tested whether the four object pairs were reliably discriminable in both
304 OSC and EVC. Separately for regular and irregular pairs, a multivariate classification
305 analysis of response patterns across voxels was performed. In a leave-one-run-out
306 fashion, linear discriminant analysis (LDA) classifiers were used to classify response
307 patterns to the four different pairs (see Materials and Methods). In both OSC and EVC,
308 these classifiers performed highly above chance level (OSC: 42%; EVC: 40%). For both
309 regions, no reliable difference in decoding accuracy was found when comparing the
310 regular and irregular conditions directly (OSC: $t[13]=0.8$, $p=0.41$; EVC: $t[13]=1.2$, $p=0.25$).
311 Similarly, no difference in univariate activation values between the two conditions was
312 observed (OSC: $t[13]=0.7$, $p=0.48$; EVC: $t[13]=1.1$, $p=0.29$). This pattern of results argues
313 against overall differences in signal quality, for example due to differences in attentional
314 engagement for the two conditions: such an effect would be expected to result in a
315 univariate and/or classification difference between the regular and irregular conditions.
316 Next, we tested whether the average of the response patterns evoked by the
317 single objects could more accurately approximate the pattern evoked by the
318 corresponding pairs than the pattern evoked by the non-corresponding pairs: for
319 example, is the average of the response patterns for cars and traffic lights more similar
320 to the response pattern for the car and traffic light pair than to the other pairs (e.g., to a
321 TV and a sofa)? This analysis was performed separately for regular and irregular pairs (Fig.
322 2). Correlations between the response patterns evoked by object pairs and the average
323 patterns coming from the corresponding single objects (within-correlations) or from non-
324 corresponding single objects (between-correlations) were computed (see Materials and

Methods). For both regular and irregular pairs separately, the difference between within- and between-correlations was taken as a measure of pair approximation quality.

In OSC, both regular and irregular pairs were discriminable using the average single object patterns (regular: $t[13]=6.33$, $p<0.001$; irregular: $t[13]=11.2$, $p<0.001$; Fig. 3, **B**). Despite the position change of the objects, also EVC patterns allowed for reliably discriminating between the four pairs (regular: $t[13]=6.1$, $p<0.001$; irregular: $t[13]=4.7$, $p<0.001$). Next, to investigate whether regularly positioned pairs were processed differently from irregularly positioned pairs, discriminability was compared directly for regular and irregular pairs. Crucially, in OSC the average single object patterns more closely resembled the response patterns evoked by irregular pairs than by regular pairs ($t[13]=2.5$, $p=0.02$). By contrast, no such difference was found in EVC ($t[13]=1.0$, $p=0.32$; interaction with region: $t[13]=2.9$, $p=0.01$). These findings suggest that object processing in OSC is sensitive to typical real-world positioning.

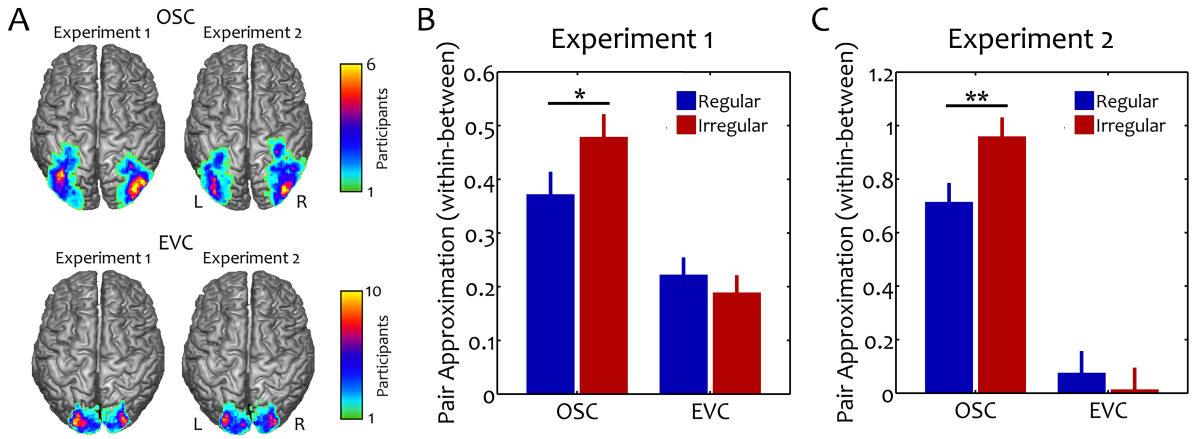

Fig. 3. Pair approximation results. **A**, Object selective cortex (OSC) and early visual cortex (EVC) maps depicting voxel-wise overlap between participants: Regions were defined on a single-subject level by selecting the 100 voxels that best discriminated the four pairs across the regularity conditions (see Materials and Methods) within either a functional

344 OSC mask or an anatomical EVC mask. For illustration purposes, the regions were
345 normalized into standard space and overlaid on a brain template using MRIcron (Rorden
346 and Brett, 2000). **B**, The pair approximation analysis for Experiment 1 revealed a more
347 accurate approximation of the irregularly positioned pairs than of the regularly
348 positioned pairs in OSC, while no effect was found in EVC. **C**, In Experiment 2, this pattern
349 of results was replicated, demonstrating sensitivity for object regularity structure in OSC,
350 but not EVC. All error bars reflect standard errors of pairwise differences. (*p<0.05,
351 **p<0.01)
352
353 *Experiment 2*
354 In a second fMRI experiment we aimed to replicate the pattern of results. The
355 structure of the experiment was largely identical to Experiment 1 (see Materials and
356 Methods) and participants viewed images from two of the pairs also used in Experiment
357 1 (sofa and TV; car and traffic lights).
358 First, as in Experiment 1, discriminability of the response patterns for the regular
359 and irregular pairs was assessed using a two-way linear classification analysis in OSC and
360 EVC. In both regions, classifiers performed highly above chance level (OSC: 69%; EVC:
361 64%) and no difference in decoding accuracy was found when comparing the regular and
362 irregular conditions directly (OSC: t[13]=0.02, p=0.99; EVC: t[13]=0.7, p=0.51). Additionally,
363 no difference was observed in univariate activation values for the two conditions (OSC:
364 t[13]=0.01, p=0.99; EVC: t[13]=1.0, p=0.32).
365 Second, separately for regular and irregular pairs, constituent single-object
366 response patterns were averaged for each pair to test how well these average response
367 patterns could approximate the two pairs. As in Experiment 1, in OSC the average of the

368 response patterns evoked by the single objects was more similar to the pattern evoked
369 by the corresponding than by the non-corresponding pairs, both for the regular pairs
370 ($t[13]=11.1$, $p<0.001$) and the irregular pairs ($t[13]=11.7$, $p<0.001$; Fig. 3, **C**). By contrast (and
371 unlike Experiment 1), the average single object patterns did not accurately approximate
372 the pair patterns in EVC (regular pairs: $t[13]=0.8$, $p=0.44$; irregular pairs: $t[13]=0.2$,
373 $p=0.87$). Most importantly, Experiment 2 replicated the key finding of Experiment 1: a
374 difference between the approximation of regular and irregular pairs in OSC ($t[13]=3.4$,
375 $p=0.004$), with no such difference in EVC ($t[13]=0.77$, $p=0.46$; interaction across regions:
376 $t[13]=3.3$, $p=0.006$). These results closely resemble the pattern of results obtained in
377 Experiment 1, providing further evidence for an influence of positional regularities on
378 multi-object coding in OSC.
379 Experiment 2 was additionally designed to test a possible explanation for the
380 regularity effect: perhaps regularly positioned object pairs more strongly induced a
381 mental representation of the corresponding scene, thereby distorting the representation
382 of the individual objects. To test this account, participants viewed images of room scenes
383 and street scenes during separate blocks of Experiment 2 (Fig. 1, **A**). None of these
384 scenes contained any of the objects belonging to the pairs. OSC response patterns to
385 these scenes were reliably discriminable in a two-way classification analysis (mean
386 accuracy: 64%; $t[13]=9.1$, $p<0.001$). To determine whether the relatively inaccurate pair
387 approximation for regularly positioned pairs in OSC reflected the activation of the
388 corresponding scene representations (e.g., living room for sofa-TV pairs), we correlated
389 the activation patterns evoked by the pair stimuli to the activation patterns of the scene
390 stimuli. Separately for the regular and irregular conditions, this correlation was
391 computed within scene type (i.e., sofa/TV and indoor room scene, and car/lights and

392 outdoor street scene) and between scene type (i.e., sofa/TV and outdoor street scene,
393 and car/lights and indoor room scene). The difference of these within- and between-
394 correlations was taken as a measure of scene approximation quality. In OSC, this scene
395 approximation was not significantly different from chance, both for regularly ($t[13]=0.4$,
396 $p=0.69$) and irregularly ($t[13]=0.2$, $p=0.87$) positioned pairs and no difference was found
397 when directly comparing discriminability for regular and irregular pairs ($t[13]=0.7$,
398 $p=0.50$). Similar results were obtained in scene-selective regions of visual cortex (see
399 Supplementary Information). These results suggest that the differential coding of regular
400 and irregular pairs in OSC does not reflect a difference in the degree to which they
401 activate scene representations.
402
403 *Transformation from independent to integrative coding*
404 The results from Experiments 1 and 2 demonstrate that a linear combination of
405 single object response patterns accurately predicts the response patterns evoked by
406 both regularly and irregularly positioned pairs. However, a relative disruption of this
407 approximation was observed in the regular condition. This pattern of results suggests
408 that responses to regularly positioned pairs reflect both an independent processing
409 component and an integrative component that disrupts this independent processing to
410 some degree. To reveal the transformation from independent to integrative processing
411 at finer anatomical resolution we tracked the overall approximation quality (reflecting
412 the independent component of multi-object representations) and the regularity effect
413 (reflecting the relative disruption of independent coding in the regular condition) along
414 the visual hierarchy. Combining the data from both experiments, we performed a
415 searchlight analysis along the posterior-anterior axis of the occipitotemporal cortex (see

416 Materials and Methods). In brief, anatomically defined occipitotemporal cortex was
417 divided into slices (width: 10mm) varying along the y-coordinate (in MNI space) in steps
418 of 2mm (see Fig. 4, **A**). For each of these slices, comparably to the OSC analyses for
419 Experiments 1 and 2, the 100 best pair-representing voxels were selected based on
420 individual-participant searchlight maps. Then, for each slice separately, the pair
421 approximation analysis was repeated, leading to a pair approximation value (within-
422 correlation vs. between-correlation) and a regularity effect for each slice (Fig. 4, **B**).
423 Generally, approximation quality across the two regularity conditions (i.e., the mean of
424 the two conditions) was very robust and significantly above chance for all slices centered
425 between Y=-100 and Y=-34 (Fig 4, **C**). To detect regularity effects, the difference of the
426 linear approximation quality in the regular and irregular conditions was compared to
427 zero. This analysis revealed an effect that emerged between Y=-80 and Y=-84, and later
428 between Y=-64 and Y=-38 (Fig 4, **D**). This pattern of results shows that highly accurate
429 linear approximation can be observed at earlier processing stages than the relative
430 breakdown of this approximation in the regular condition, suggesting that the
431 integrative multi-object coding for regularly positioned objects is preceded by an
432 independent representational stage.
433 To assess how the overall approximation quality and the regularity effect relate to
434 object selectivity, the difference of the intact and scrambled object conditions in the
435 functional localizer was computed for each slice (see Materials and Methods). All slices
436 between Y=-98 and Y=-30 exhibited significant object-selectivity (Fig. 4, **C**/**D**).
437 Interestingly, across slices object-selectivity almost perfectly predicted overall
438 approximation quality ($r=0.97$, $p<0.001$). By contrast, object-selectivity did not
439 significantly predict the regularity effect across slices ($r=-0.05$, $p=0.75$). These results

suggest that the independent component of multi-object processing is tightly linked to the degree of object-selectivity, while the encoding of object regularity emerges later in the visual hierarchy. In sum, this analysis shows that overall linear approximation (reflecting the independent component of multi-object processing) and its relative breakdown (reflecting an effect of positional structure) can be anatomically dissociated, revealing a transformation from independent to integrative coding of multi-object displays along the visual processing hierarchy.

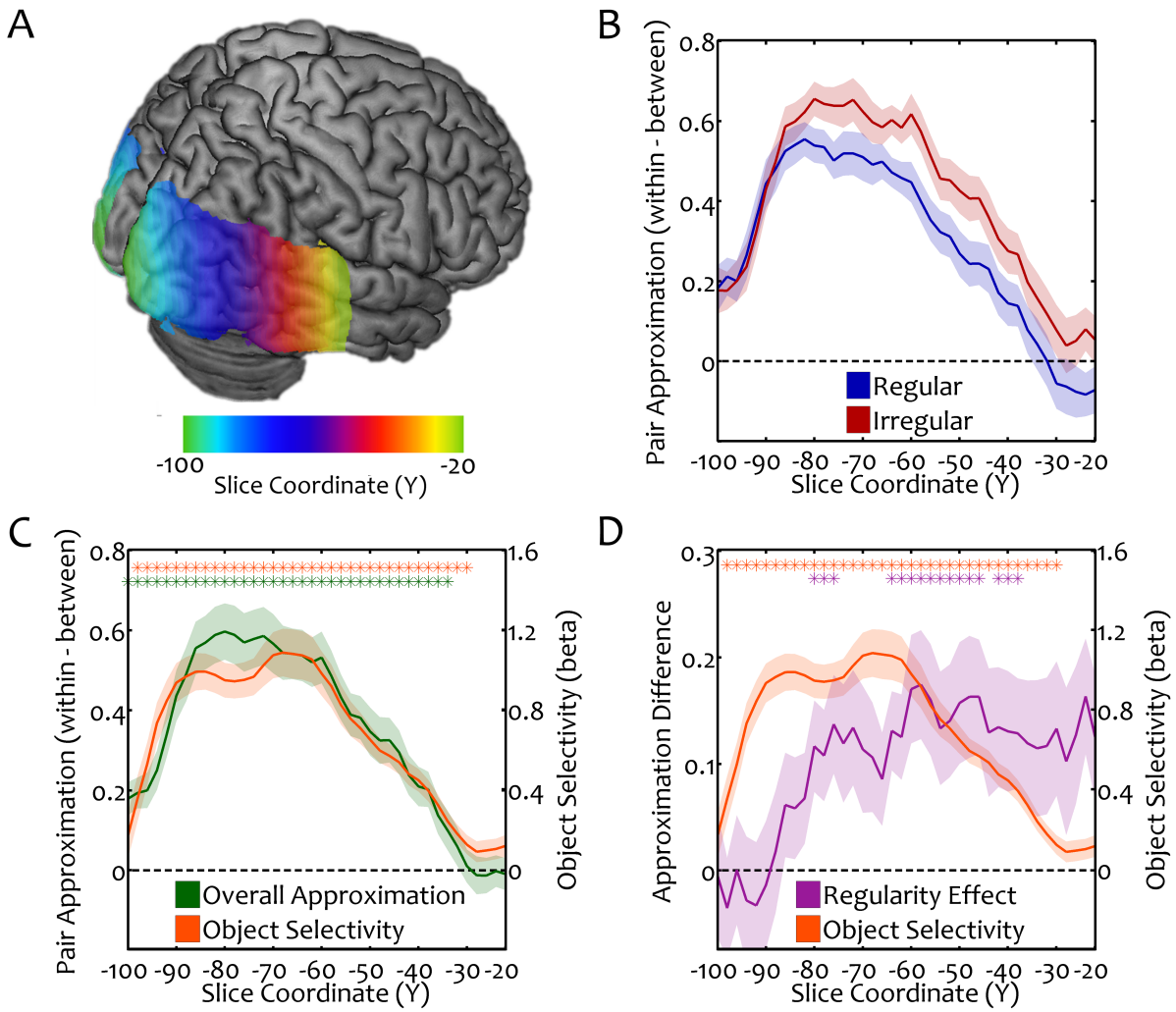

Fig. 4. Combined searchlight analysis. **A,** Searchlight analysis was performed for slices of occipitotemporal cortex (10mm width) with varying y-coordinate, in steps of 2mm. Slices

were defined in standard space (based on MNI coordinates) and back-projected into individual subject space. Subsequently, within each slice, the 100 best pair-representing voxels were selected on an individual-participant level. **B**, Pair approximation (expressed as the difference between within- and between correlations) for regular and irregular pairs in slices along the y-coordinate. **C**, Overall approximation (i.e., the mean of the regular and irregular conditions) was generally very accurate throughout the visual hierarchy, and showed a striking correspondence with the degree of object-selectivity. **D**, Differences between the regular and irregular conditions emerged starting from Y=-80. In contrast to the overall approximation, the regularity effects along the hierarchy did not correspond well with the degree of object-selectivity. (*$p<0.05$, corrected for multiple comparisons)

463 **Discussion**

464 In two fMRI experiments, we demonstrate non-independence in the visual cortex
465 representation of multi-object displays, revealing a transformation from independent to
466 integrative representations along the visual processing hierarchy. Response patterns in
467 OSC evoked by regularly positioned object pairs were less accurately modeled by the
468 response patterns evoked by their constituent objects. The relative breakdown of this
469 linear response approximation reflects a disruption of independent multi-object coding
470 when objects adhere to their typical real-world positions. The regularity effect was
471 significant in functionally localized OSC (but not EVC and scene-selective regions; see
472 Supplementary Information). However, further analyses revealed that object selectivity
473 was closely related, along the posterior-anterior axis, to independent object coding
474 rather than integrative object coding, which emerged later in the visual hierarchy.

475 The searchlight analysis along the posterior-anterior axis of the occipitotemporal
476 cortex suggests that the coding of multiple, regularly positioned objects is reflecting two
477 dissociable components: an independent processing component and an integrative
478 component. Independent processing can be observed starting at early stages of the
479 visual object processing hierarchy, and is tightly linked to object selectivity. The striking
480 correspondence between object-selectivity and pair approximation suggests that the
481 same neural mechanisms support single-object coding and (independent) multi-object
482 coding. By contrast, at later stages of the processing hierarchy, pair approximation
483 quality diverges for regular and irregular pairs, reflecting additional, integrative
484 processing for regularly positioned pairs that (partly) disrupts independent coding.
485 Interestingly, this relative disruption cannot be explained by differences in object-
486 selectivity along the visual hierarchy. We speculate that the independent and integrative

487 components of multi-object coding are reflecting activity in different sub-regions of
488 object-selective cortex, with a more posterior sub-region – possibly reflecting LO1
489 (Larsson and Heeger, 2006) (Y=-90, SE=5) – primarily reflecting independent coding and a
490 more anterior sub-region – possibly reflecting LO2 (Larsson and Heeger, 2006) (Y=-83,
491 SE=6) – containing also an integrative component. This assumption is consistent with a
492 previously reported difference in spatial summation along the posterior-anterior axis,
493 with a higher rate of information compression in the more anterior LO2 (Kay et al., 2013).
494 However, further studies employing retinotopic mapping techniques are needed to test
495 whether the effect can indeed be pinpointed to a specific sub-region of object-selective
496 occipitotemporal cortex.

497 The disruption of independent coding observed here is unlikely to reflect
498 differential semantic associations between objects. Previous electrophysiological studies
499 have demonstrated that learned associations between objects can alter neural tuning
500 properties in visual cortex: The associative pairing of two objects can shape neural
501 responses in a way that neurons become selective to both of the individual objects
502 (Messinger et al., 2001; Sakai et al., 1991). However, in our study the objects were equally
503 related in their semantic content in both the regular and irregular conditions – the crucial
504 difference was the positional structure. Thus, our results are more aptly explained by
505 visual group representations that are shaped by co-occurrence structure of objects in
506 real-world environments. Such group representations have been described in the cortical
507 representation of people, where information from the face and body is integrated into a
508 person representation, which is sensitive to the correct relative positioning of both parts
509 (Bernstein et al., 2014; Harry et al., 2016; Schmalzl et al., 2012; but see Kaiser et al., 2014b).
510 Similarly, regularly positioned objects may be grouped in visual cortex, giving rise to

511 different response patterns to the whole than the sum of the parts. Further studies are
512 required to probe the nature of these representations at a finer grained level.
513 A number of previous studies have explored relational coding of objects in the
514 context of action relationships. Some of these studies have reported univariate
515 activation differences between objects positioned correctly or incorrectly for performing
516 a specific action (Kim and Biederman, 2011; Kim et al., 2011; Roberts and Humphreys,
517 2010). Others have used similar methods as the current study to demonstrate that action
518 relationships alter combination rules in the processing of multiple objects (Baeck et al.,
519 2013; Baldassano et al., 2016). Notably, the regularity effects observed in these studies
520 have been attributed to processing asymmetries favoring the "active" object (e.g., a
521 hammer) over the "passive" object (e.g., a nail) in a pair (Baeck et al., 2013; Riddoch et
522 al., 2003). By contrast, our approach focused on real-world object structure more
523 generally, avoiding such direct functional interactions of the two objects. Furthermore, in
524 our study the positional structure of the object pairs was solely manipulated by
525 exchanging the position of the single objects, without introducing changes of low-level
526 visual characteristics (like physically connecting the objects differently; see Kim and
527 Biederman, 2011; Baldassano et al., 2016). Therefore, the relative disruption of
528 independent coding in the regular condition observed in our data cannot be explained by
529 low-level visual interactions between the two stimuli. Our results thus provide strong
530 evidence for sensitivity for positional structure in visual cortex, which is not attributable
531 to a shift of weights in favor of one of the objects and is independent of low-level visual
532 content.
533 How does this sensitivity for positional regularities arise from the typical relative
534 positioning of objects? In the present study, we interchanged the position of the object

535 pairs to disrupt positional structure while keeping the individual objects identical. In the
536 irregular displays, positional structure was disrupted on multiple levels: The objects did
537 not adhere to their typical relative viewpoints and sizes, the laws of physics were partly
538 violated, and the typical functional relationships between the objects were disrupted. All
539 these factors constitute properties of positional object structures in real-world scenes, so
540 that all of them potentially contribute to the effects observed here. Thus, while our study
541 provides clear evidence for a role of positional structure in multi-object representations,
542 further studies are needed to disentangle the relative contributions of the different
543 aspects our manipulation entailed.

544 The sensitivity for positional structure observed here may reflect the grouping (or
545 integration) of multiple, regularly positioned objects. Such grouping may provide an
546 efficient way to deal with the large number of objects contained in natural scenes (Kaiser
547 et al., 2014a). As a result of limitations in visual capacity, parallel processing of objects is
548 associated with competition for representation between individual stimuli: When
549 multiple objects need to be processed at the same time, neural processing becomes less
550 efficient and behavioral performance decreases (Desimone and Duncan, 1995; Franconeri
551 et al., 2013; Kastner and Ungerleider, 2001). Processing objects in meaningful chunks
552 might help to reduce the complexity of a scene by reducing the number of competing
553 objects. Behavioral evidence suggests that regularly positioned objects can be more
554 efficiently processed in a number of visual tasks (Gronau and Shachar, 2014, 2015; Kaiser
555 et al., 2014a, 2015; Stein et al., 2015). These behavioral benefits have been linked to a
556 reduction in competitive interactions in visual cortex when objects follow positional
557 regularities (Kaiser et al., 2014a). Our current results provide an explanation for this
558 reduction of neural competition: Regularly positioned arrangements of objects are not

recruiting independent and competing object representations, but are to some degree integrated into group representations at higher stages of visual cortex. Such group representations may constitute an integral step in the visual processing hierarchy, bridging the gap between individual object coding and the representation of complex visual scenes.

Acknowledgements

The research was supported by the Autonomous Province of Trento, Call "Grandi Progetti 2012", project "Characterizing and improving brain mechanisms of attention – ATTEND".

570 **References**

571 Agam Y, Liu H, Papanastassiou A, Buia C, Golby AJ, Madsen JR, Kreiman G (2010) Robust
572 selectivity to two-object images in human visual cortex. Curr Biol 20:1-8.
573 Baeck A, Wagemans J, Op de Beeck HP (2013) The distributed representation of random
574 and meaningful object pairs in human occipitotemporal cortex: the weighted
575 average as a general rule. Neuroimage 70:37-47.
576 Baldassano C, Beck DM, Fei-Fei L (2016) Human-object interactions are more than the
577 sum of their parts. Cereb Cortex, DOI: 10.1093/cercor/bhw077.
578 Bar M (2004) Visual objects in context. Nat Rev Neurosci 5:617-629.
579 Bar M, Aminoff EM (2003) Cortical analysis of visual context. Neuron 38:347-358.
580 Bernstein M, Oron J, Sadeh B, Yovel G (2014) An integrated face-body representation in
581 the fusiform gyrus but not the lateral occipital cortex. J Cogn Neurosci 26:2469-
582 2478.
583 Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433-436.
584 Chun MM (2000) Contextual cueing of visual attention. Trends Cogn Sci 4:170-178.
585 Cichy RM, Chen Y, Haynes JD (2011) Encoding the identity and location of objects in
586 human LOC. Neuroimage 54:2297-2307.
587 Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev
588 Neurosci 18:193-222.
589 Eger E, Ashburner J, Haynes JD, Dolan RJ, Rees G (2008) fMRI activity patterns in human
590 LOC carry information about object exemplars within category. J Cogn Neurosci
591 20:356-370.
592 Epstein RA, Kanwisher N (1998) A cortical representation of the local visual environment.
593 Nature 392:598-601.

594 Franconeri SL, Alvarez GA, Cavanagh P (2013) Flexible cognitive resources: competitive
595 content maps for attention and memory. Trends Cogn Sci 17:134-141.
596 Grill-Spector K (2003) The neural basis of object perception. Curr Opin Neurobiol 13:159-
597 166.
598 Grill-Spector K, Kushnir T, Hendler T, Malach R (2000) The dynamics of object-selective
599 activation correlate with recognition performance in humans. Nat Neurosci 3:837-
600 843.
601 Gronau N, Shachar M (2014) Contextual integration of visual objects necessitates
602 attention. Atten Percept Psychophys 76:695-714.
603 Gronau N, Shachar M (2015) Contextual consistency facilitates long-term memory of
604 perceptual detail in barely seen images. J Exp Psychol Hum Percept Perform
605 41:1095-1111.
606 Harry BB, Umla-Runge K, Lawrence AD, Graham KS, Downing PE (2016) Evidence for
607 integrated visual face and body representations in the anterior temporal lobes. J
608 Cogn Neurosci 28:1178-1193.
609 Hasson U, Levy I, Behrmann M, Hendler T, Malach R (2002) Eccentricity bias as an
610 organizing principle for human high-order object areas. Neuron 34:479-490.
611 Haushofer J, Livingstone MS, Kanwisher N (2008) Multivariate patterns in object-
612 selective cortex dissociate perceptual and physical shape similarity. PLoS Biol
613 6:e187.
614 Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and
615 overlapping representations of faces and objects in ventral temporal cortex.
616 Science 293:2425-2430.

617 Kaiser D, Stein T, Peelen MV (2014a) Object grouping based on real-world regularities
618 facilitates perception by reducing competitive interactions in visual cortex. Proc
619 Natl Acad Sci USA 111:11217-11222.
620 Kaiser D, Stein T, Peelen MV (2015) Real-world spatial regularities affect visual working
621 memory for objects. Psychon Bull Rev 22:1784-1790.
622 Kaiser D, Strnad L, Seidl KN, Kastner S, Peelen MV (2014b) Whole-person evoked fMRI
623 activity patterns in human fusiform gyrus are accurately modeled by a linear
624 combination of face- and body-evoked activity patterns. J Neurophysiol 111:82-90.
625 Kastner S, Ungerleider LG (2001) The neural basis of biased competition in the human
626 visual cortex. Neuropsychologia 39:1263-1276.
627 Kay KN, Winawer J, Mezer A, Wandell BA (2013) Compressive spatial summation in
628 human visual cortex. J Neurophysiol 110:481-494.
629 Kim JG, Biederman I (2011) Where do objects become scenes? Cereb Cortex 21:1738-1746.
630 Kim JG, Biederman I, Juan CH (2011) The benefit of object interactions arises in the lateral
631 occipital cortex independent of attentional modulation from the intraparietal
632 sulcus: A TMS study. J Neurosci 31:8320-8324.
633 Kourtzi Z, Kanwisher N (2001) Representation of perceived object shape by the human
634 lateral occipital complex. Science 293:1506-1509.
635 Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA
636 (2008) Matching categorical object representations in inferior temporal cortex of
637 man and monkey. Neuron 60:1126-1141.
638 Kubilius J, Baeck A, Wagemans J, Op de Beeck HP (2015) Brain-decoding fMRI reveals
639 how wholes relate to the sum of parts. Cortex 72:5-14.

640 Larsson J, Heeger DJ (2006) Two retinotopic visual areas in human lateral occipital
641 cortex. J Neurosci 26:13128-13142.
642 MacEvoy SP, Epstein RA (2009) Decoding the representation of multiple simultaneous
643 objects in human occipitotemporal cortex. Curr Biol 19:943-947.
644 MacEvoy SP, Epstein RA (2011) Constructing scenes from objects in human
645 occipitotemporal cortex. Nat Neurosci 14:1323-1329.
646 Malach R, Reppas JB, Benson RR, Kwong KK, Jiang H, Kennedy WA, Ledden JP, Brady TJ,
647 Rosen BR, Tootell RB (1995) Object-related activity revealed by functional
648 magnetic resonance imaging in human occipital cortex. Proc Natl Acad Sci USA
649 92:8135-8139.
650 Messinger A, Squire LR, Zola SM, Albright TD (2001) Neural representations of stimulus
651 associations develop in the temporal lobe during learning. Proc Natl Acad Sci USA
652 98:12239-12244.
653 Oliva A, Torralba A (2007) The role of context in object recognition. Trends Cogn Sci
654 11:520-527.
655 Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: multi-modal multivariate
656 pattern analysis of neuroimaging data in Matlab / GNU Octave. Front Neuroinform
657 10:20.
658 Reddy L, Kanwisher NG, VanRullen R (2009) Attention and biased competition in multi-
659 voxel object representations. Proc Natl Acad Sci USA 106:21447-21452.
660 Riddoch MJ, Humphreys GW, Edwards S, Baker T, Willson K (2003) Seeing the action:
661 neuropsychological evidence for action-based effects on object selection. Nat
662 Neurosci 6:82-89.

663 Roberts KL, Humphreys GW (2010) Action relationships concatenate representations of
664 separate objects in the ventral visual system. Neuroimage 52:1541–1548.
665 Rorden C, Brett M (2000) Stereotaxic display of brain lesions. Behav Neurol 12:191-200.
666 Sakai K, Miyashita Y (1991) Neural organization for the long-term memory of paired
667 associates. Nature 354:152-155.
668 Schmalzl L, Zopf R, Williams MA (2012) From head to toe: evidence for selective brain
669 activation reflecting visual perception of whole individuals. Front Hum Neurosci
670 6:108.
671 Smith SM, Nichols TE (2009) Threshold-free cluster enhancement: Addressing problems
672 of smoothing, threshold dependence and localisation in cluster inference.
673 Neuroimage 44:83-98.
674 Stein T, Kaiser D, Peelen MV (2015) Interobject grouping facilitates visual awareness. J Vis
675 15:10.
676 Vuilleumier P, Henson RN, Driver J, Dolan RJ (2002) Multiple levels of visual object
677 constancy revealed by event-related fMRI of repetition priming. Nat Neurosci
678 5:491-499.
679 Williams MA, Dang S, Kanwisher NG (2007) Only some spatial patterns of fMRI response
680 are read out in task performance. Nat Neurosci 10:685-686.
681 Wolfe JM, Võ ML-H, Evans KK, Greene MR (2011) Visual search in scenes involves selective
682 and nonselective pathways. Trends Cogn Sci 15:77-84.
683 Zoccolan D, Cox DD, DiCarlo JJ (2005) Multiple object normalization in monkey
684 inferotemporal cortex. J Neurosci 25:8150-8164.
685