Explaining the hierarchy of visual representational geometries by remixing of features from many computational vision models

Seyed-Mahdi Khaligh-Razavi[1*], Linda Henriksson[1, 2], Kendrick Kay[3] and Nikolaus Kriegeskorte[1]

[1] MRC Cognition and Brain Sciences Unit, Cambridge, UK
[2] Brain Research Unit, O.V. Lounasmaa Laboratory, Aalto University, Espoo, Finland
[3] Department of Psychology, Washington University in St. Louis, St. Louis, MO, USA

*Corresponding author:
Seyed-Mahdi Khaligh-Razavi, Seyed.KalighRazavi@mrc-cbu.cam.ac.uk

Abstract

Visual processing in cortex happens through a hierarchy of increasingly sophisticated representations. Here we explore a very wide range of model representations (29 models), testing their categorization performance (animate/inanimate) and their ability to account for the representational geometry of brain regions along the visual hierarchy (V1, V2, V3, V4, and LO). We also created new model instantiations (85 model instantiations in total) by reweighting and remixing of the model features. Reweighting and remixing was based on brain responses to an independent training set of 1750 images. We assessed the models with representational similarity analysis (RSA), which characterizes the geometry of a representation by a representational dissimilarity matrix (RDM). In this study, the RDM is either computed on the basis of the model features or on the basis of predicted voxel responses. Voxel responses are predicted by linear combinations of the model features. The model features are linearly remixed so as to best explain the voxel responses (as in voxel/population receptive-field modelling). This new approach of combining RSA with voxel receptive field modelling may help bridge the gap between the two methods. We found that early visual areas are best accounted for by a Gabor wavelet pyramid (GWP) model. The GWP implementations we used performed similarly with and without remixing, suggesting that the original features already approximate the representational space, obviating the need for remixing or reweighting. The lateral occipital region (LO), a higher visual representation, was best explained by the higher layers of a deep convolutional network (Krizhevsky et al., 2012). However, this model could explain the LO representation only after appropriate remixing of its feature set. Remixed RSA takes a step in an important direction, where each computational model representation is explored more broadly by considering not only its representational geometry, but the set of all geometries within reach of a linear transform. The exploration of many models and many brain areas may lead to a better understanding of the processing stages in the visual hierarchy, from low-level image representations in V1 to visuo-semantic representations in higher-level visual areas.

Introduction

Primates can easily categorize objects in different variations. This is thought to rely on a high-level representation in higher stages of the visual hierarchy (i.e. inferior temporal cortex). This brain region has been intensely studied in primates (Bell et al., 2009; Hung et al., 2005; Kriegeskorte et al., 2008a). The representation in this higher visual area is the result of computations performed in stages across the hierarchy of the visual system. Although none of the stages in this cortical hierarchy is yet fully understood, there has been good progress in understanding and modeling early visual areas (Eichhorn et al., 2009; Kay et al., 2013; Güçlü and van Gerven, 2014). However, intermediate (e.g. V4) and higher visual areas (e.g. LO and IT) have been more difficult to understand.

The geometry of a representation can be usefully characterized by a representational dissimilarity matrix (RDM) computed by comparing the patterns of brain activity elicited by a set of visual stimuli. In this representational similarity analysis (RSA) framework, representations in two brain regions are compared by computing the correlation between their RDMs (Kriegeskorte, 2009; Nili et al., 2014). Each RDM contains a representational dissimilarity for each pair of stimulus-related response patterns (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008b). We use the RSA framework here to compare processing stages in computational models with the stages of processing in the hierarchy of ventral visual pathway. We use a previously published dataset (Kay et al., 2008) consisting functional magnetic resonance imaging (fMRI) responses to a training set of 1,750 natural images and an independent test set of 120 natural images (all grayscale).

In previous studies, we have compared the RDMs of a wide range of models (Khaligh-Razavi, 2014) with human IT (Khaligh-Razavi and Kriegeskorte, in-press; Kriegeskorte et al., 2008; Kriegeskorte, 2009). In this study we build on the results from Henriksson et al. (2014) and Khaligh-Razavi and Kriegeskorte (in-press), and compare the representational geometry of a wide range of object-vision models with several brain regions, from early to higher visual areas, using a dataset that has a larger set of stimuli than the one used in Khaligh-Razavi and Kriegeskorte (in-press). The focus of the previous study (Khaligh-Razavi and Kriegeskorte, in-press) was mainly on understanding and modelling the representation of visual information in IT. Here we investigate more brain regions (including early and intermediate visual areas).

In addition, we study the transformation of representational geometry across stages of the visual hierarchy by comparing representations between different brain regions. We have recently shown (Henriksson et al., 2014) that coherent fluctuations of overall activation between two brain regions can make the representational geometries of the two regions appear more similar than their purely stimulus-driven similarity. This can confound comparisons among different brain representations. To overcome this issue, we use cross-trial RDM comparison, where the stimulus-driven component is shared between regions, but not the intrinsic fluctuations of activity.

For model evaluation, in addition to RSA, we also take advantage of voxel/cell-population receptive-field (RF) modeling (Huth et al., 2012; Kay et al., 2008; Dumoulin and Wandell, 2008). Voxel RF modeling fits a linear transformation of the features of a computational model to predict a given voxel's response. The linear transform is fitted using a training data set of responses to an

independent sample of stimuli. We bring these two complementary approaches together by constructing RDMs based on voxel response patterns predicted by voxel-RF models. This constitutes a new method of model evaluation, *remixed RSA*, in which model features are first mapped to the brain space (as in voxel-RF modelling) and the predicted and measured RDMs are then statistically compared (as in RSA). The voxel-RF fitting stage is a way of *remixing* the model features so as to better predict brain responses. By remixing model features we can investigate the possibility that all essential nonlinearities are present in a model, and they just need to be appropriately linearly recombined to approximate the representational geometry of a given cortical area. Remixing provides quite a general transformation (also known as affine recoding), which includes feature reweighting as a special case. In practice, stable estimation of the remixing matrix requires regularisation. Here we used an L2 penalty (ridge regression) to fit the remixing matrix.

In interpreting results for fitted models, it is important to keep in mind that the prior implicit to any regularisation is part of the model. The L2 penalty implies a prior favouring small remixing weights. In addition to the general linear remixing, we also tried a heuristic approach for reweighting of model features. Remixing and reweighting was always performed on the basis of the independent training set of 1750 images and the 120-image test set is used for model comparisons.

Compared to the stimulus set used in Khaligh-Razavi and Kriegeskorte (in press), the Kay et al. (2008) dataset contains more stimuli in the training set, enabling the remixing approach, and the stimuli are natural grayscale images as opposed to colour images of isolated objects on a gray background. We will see to what extent results are consistent across these two different sets of images.

Results

In (Henriksson et al., 2014) we showed that the patterns of visual information are affected by non-stimulus driven effects (e.g. trial-to-trial variability of visually evoked responses and the coherent response fluctuations across visual cortex in the absence of stimuli). Therefore it is important to take into account the effects of intrinsic cortical dynamics when comparing models with several brain representations, otherwise the results may be misleading. For example, RDMs from two visual areas are more similar when the response patterns are estimated on the basis of the same trials (sharing intrinsic cortical dynamics) than when they are estimated on the basis of separate trials (sharing only the stimulus-driven component). In this study we build on the findings from (Henriksson et al., 2014), and in the context of representational similarity analysis (RSA) framework, we run cross-trial comparisons of brain ROIs with each other and with computational models of object-vision. We also extend our assessment of object-vision models (Khaligh-Razavi and Kriegeskorte, in-press) by comparing many model instantiations to several brain areas, form early to higher visual areas.

The transformation of visual information along the stages of visual stream: from shape information to category information

The cross-trial comparison of RDMs for several ROIs across the stages of visual stream is shown in Figure 1. We had early visual areas (i.e. V1, V2), intermediate level visual areas (V3, V3A, V3B, and V4), and LO as one of the higher visual areas. The RDMs for each ROI were calculated based on 120 test stimuli presented to the subject – there were 10 runs consisted of 12 distinct images presented 12 times each. For more information about the data set, and images see materials and methods or refer to (Kay et al., 2008; Henriksson et al., 2014). The representational dissimilarity patterns of all ROIs in the three subjects are compared in a second-order RDM (Figure 1A). One of the consistent patterns observed in this RDM is the high correlation between V1, V2, and V3 within a subject, and also across subjects. V4 is also correlated with these three ROIs; however the correlation is lower and therefore less prominent in the RDM. On the other hand, LO does not show a high correlation with other ROIs, particularly the early visual areas, showing the gradual transformation of visual information from early visual areas to higher visual areas.

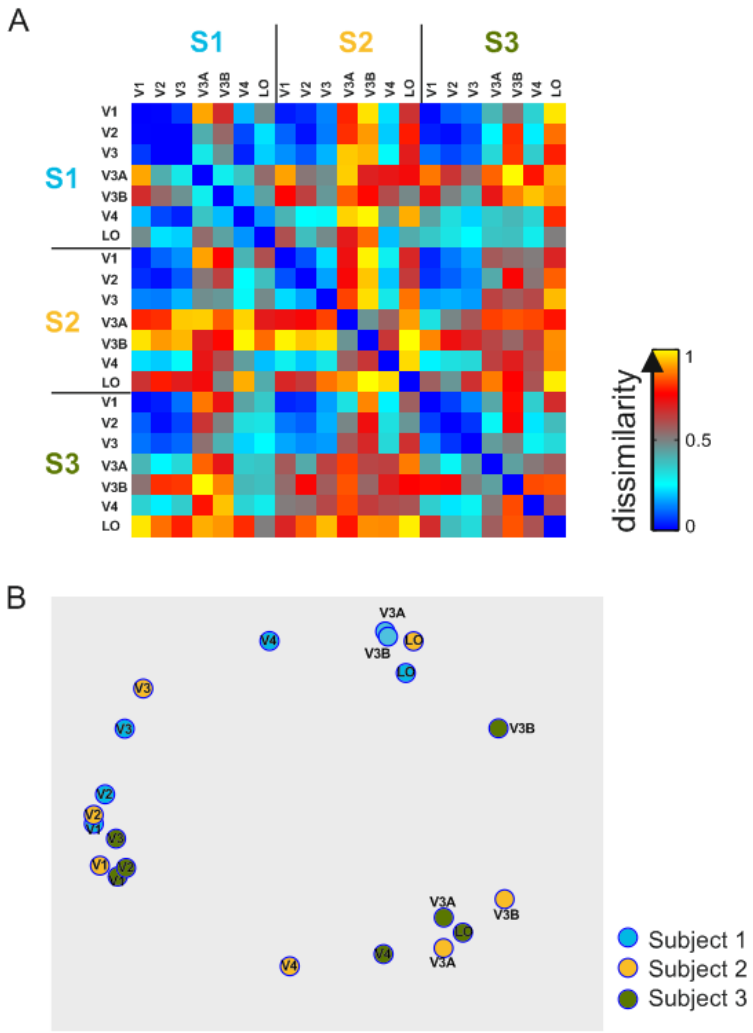

Figure 1. Cross-trial comparison of RDMs within and across subjects. A) A second-order representational dissimilarity matrix of RDMs of visual areas (V1, V2, V3, V3A, V3B, V4, LO) in all three subjects (S1, S2, S3) is shown. The results are based on 120 stimuli presented to the three subjects – there were 10 runs consisted of 12 distinct images presented 12 times each. The effects of coherent trial-to-trial fluctuations and stimulus presentation order were removed by comparing RDMs based on different trials, as well as by averaging across trials. Odd trials were averaged and compared to the average of even trials. B) A multidimensional-scaling arrangement of the second-order RDM similarities.

What are the possible underlying visual features that each ROI represents? To be able to answer this question we compared the ROIs with a wide variety of object-vision models (The RDM in figure 2A). Comparing the transformation of visual information with a wide variety of object-vision models can give us cues about the type of computations that happens in different stages of the visual hierarchy, and the type of features or visual information that is represented in each stage. The Gabor wavelet pyramid (GWP) model seems to be the most correlated model with early visual areas, confirming previous observations (Jones and Palmer, 1987) about the Gabor-like features in early visual areas. On the other side of the spectrum, LO seems to be best accounted for by the animacy model (this is best shown in subject 1). The animacy model is a simple model RDM that shows the animate-inanimate distinction. Interestingly, for intermediate visual areas, the intermediate layers of the HMAX model (HMAX C2, and C3) seem to have the highest correlation. This conforms with the visual information being gradually transformed from Gabor-like features in early vision to a high-level semantic representation of categories in higher visual areas.

In Figure 2A we saw that most of the models, apart from the animacy model, were not highly correlated with LO representation. One explanation is that all the object-vision models shown in this figure are unsupervised, therefore they do not explicitly receive semantic information to distinguish animates vs. inanimates. This may explain why the visual features extracted by these models are not good at discriminating semantic categories such as animates versus inanimates, which is what LO seems to care about. To show this we have evaluated the categorization performance of the models in an animate/inanimate categorization task (Figure 3). A linear SVM classifier was trained with the model features extracted from 1750 training images, and tested on the 120 test images. Only a few of models have performances significantly above chance, and the highest performance is below 70%.

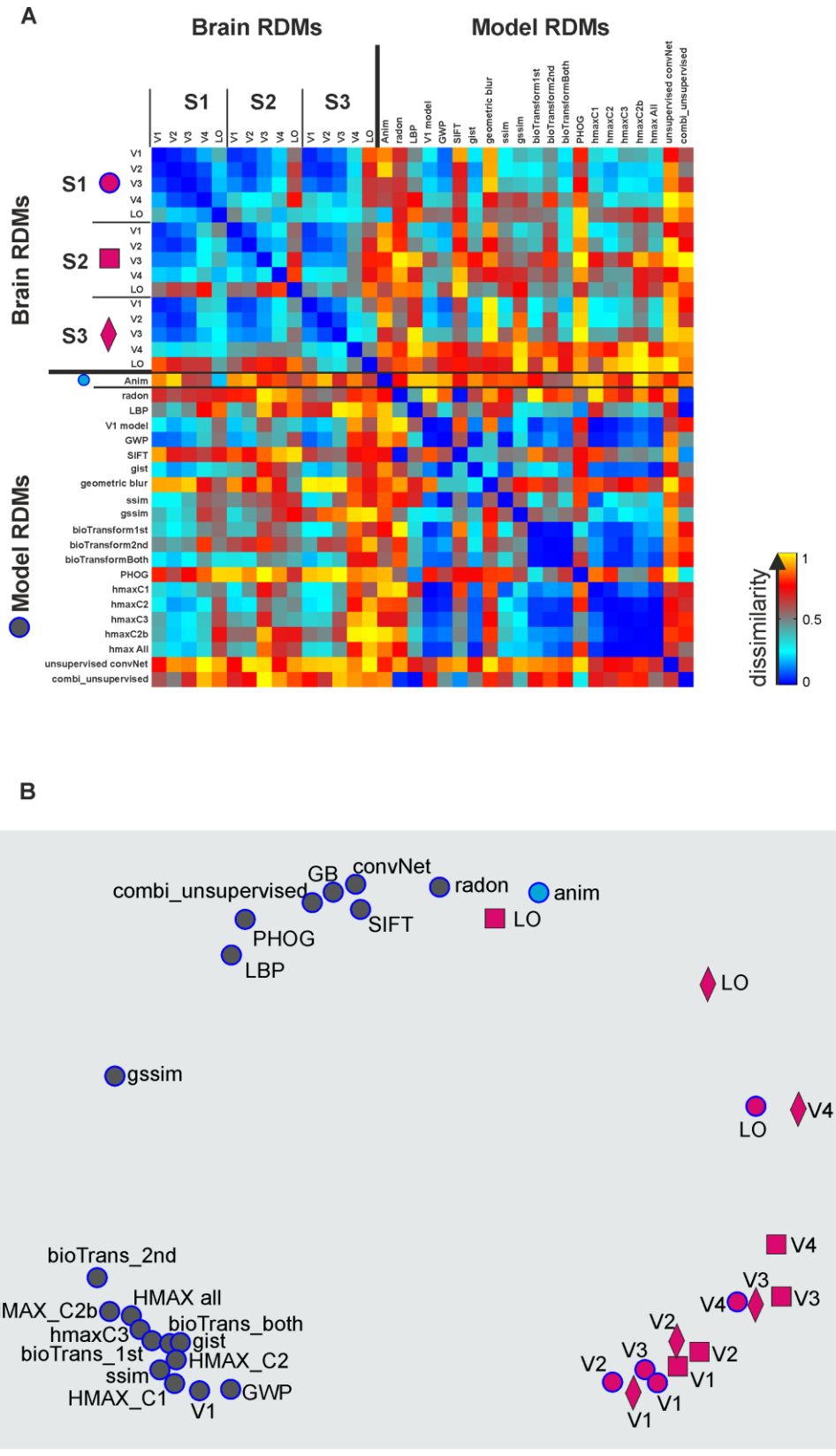

Figure 2. Cross-trial comparison of RDMs within and across subjects and Models. A) A second-order representational dissimilarity matrix of RDMs of visual areas (V1, V2, V3, V4, LO) in all three subjects (S1, S2, S3) and 20 object-vision model representations, and one category-model RDM (animate vs. inanimate) is shown. The results are based on 120 stimuli presented to the three subjects – there were 10 runs consisted of 12 distinct images presented 12 times each. The effects of coherent trial-to-trial fluctuations and stimulus presentation order were removed by comparing RDMs based on different trials, as well as by averaging across trials. Odd trials were averaged and compared to the average of even trials. B) A multidimensional-scaling arrangement of the second-order RDM similarities. Corr(diss,distance) = 0.68 (Pearson's correlation).

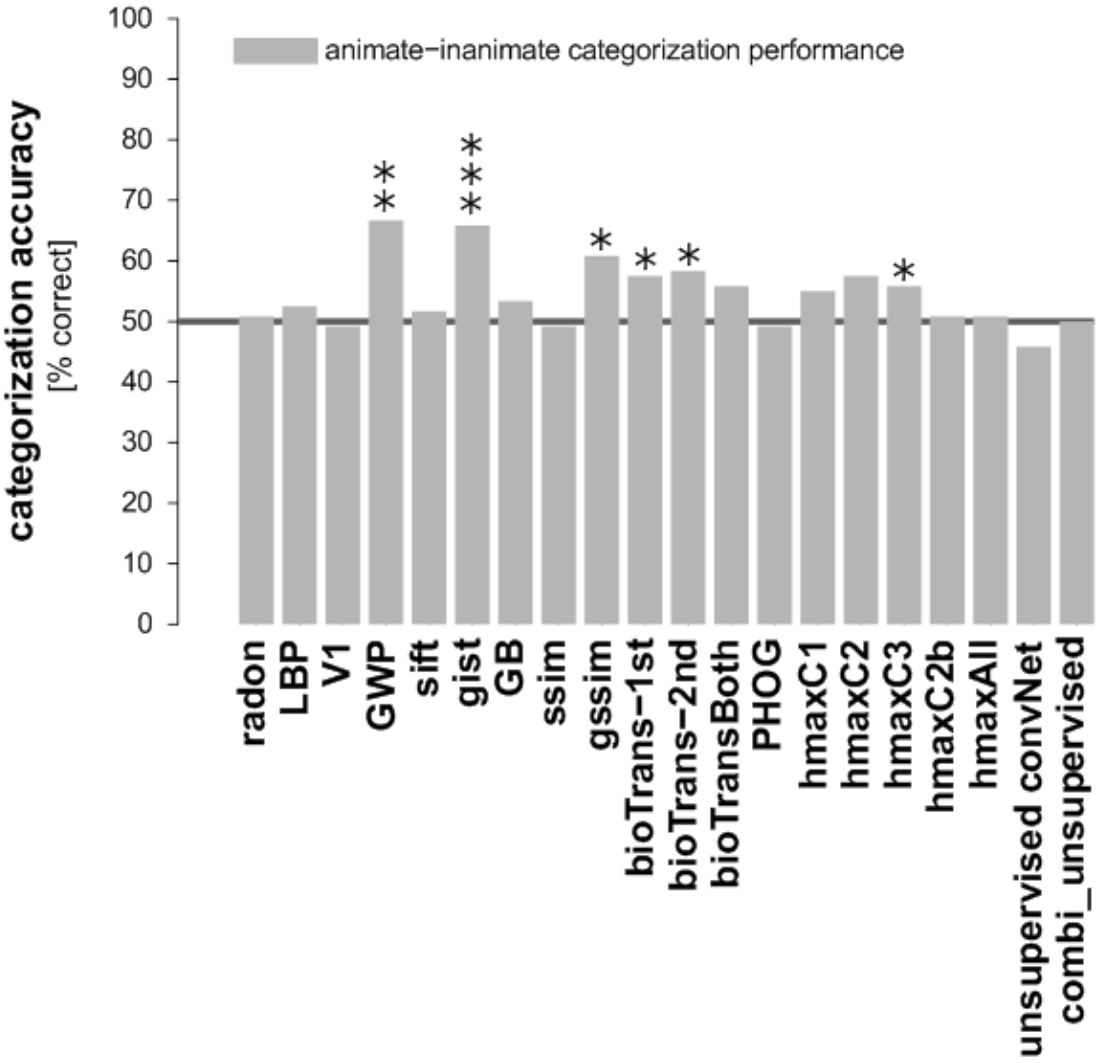

Figure 3. Models' animate-inanimate categorization performance. Bars show animate vs. inanimate categorization performance for each of the models shown on the X-axis. A linear SVM classifier was trained using 1750 training images and tested by 120 test images. P values that are shown by asterisks show whether the categorization performances significantly differ from chance [$p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***] . P values were obtained by random permutation of the labels (number of permutations = 10,000).

In summary, we may conclude that the physical similarity among objects, as instantiated in the unsupervised models that we tested, is more represented in early visual areas, and as we go higher in the visual hierarchy, the higher areas represent the perceived similarity (e.g. category information) rather than the physical similarity. Therefore the visual information across the hierarchy is transformed from shape information to category information. This suggests that the representations in higher visual areas may provide a substrate for perceptual and conceptual mental spaces.

A 2D arrangement of brain ROIs – using multidimensional scaling – is shown in Figure 1B, and a 2D arrangement for brain and model RDMs together is shown in Figure 2B.

Exploring model space through remixing and reweighting of model features

The rich training data, allowed for a more comprehensive assessment of the models. In other words, we could better explore the space of possible models (which we refer to as model space) by remixing and reweighting model features using the training data. We could use the training data (1750 stimuli) to fit the model parameters, and then test them with the testing set (120 stimuli). In the previous study (Khaligh-Razavi and Kriegeskorte, in-press) subjects were presented only with 96 images, and therefore we did not have a separate large enough training set to search the model space with.

For each of the model representations, in addition to the original model response (sometimes is referred to as an unfitted model), two other instantiations were created: **1) Remixed features:** remixing is done by voxel receptive field (RF) modelling so it is also called voxel receptive field-fitted (voxel RF-fitted) model instantiation. For each model the voxel RF-fitted features are made by predicting voxel responses from model features. Using the training data (voxel responses for 1750 stimuli) we learn a mapping from model features to brain voxels (Figure 4). The mapping is then used to map the model features to voxel responses (i.e. predicting responses of brain voxels) for the test stimuli (120 images). Figure 4 shows the procedure of voxel receptive field modeling (see Materials and Methods). We then construct an RDM from the predicted voxel responses and compare it to the reference brain RDM. **2) Reweighted features**: given a model, extract model features for each image, weight the features using a set of weights, and then compute the RDM. The weights were the average of the weights that were learned after building a receptive field model for each voxel. This is a heuristic weighting; the idea being that after we build a receptive-field model for each voxel using the set of model features, we find that some of the model features are very little used (see Figure S4, or S5). These are not informative features in predicting voxel responses; therefore we want to give these features a lower weight in constructing RDM matrices based on the model. In a sense, how different features are weighted can be thought of as just another way in which a model can vary, thus extending our model exploration.

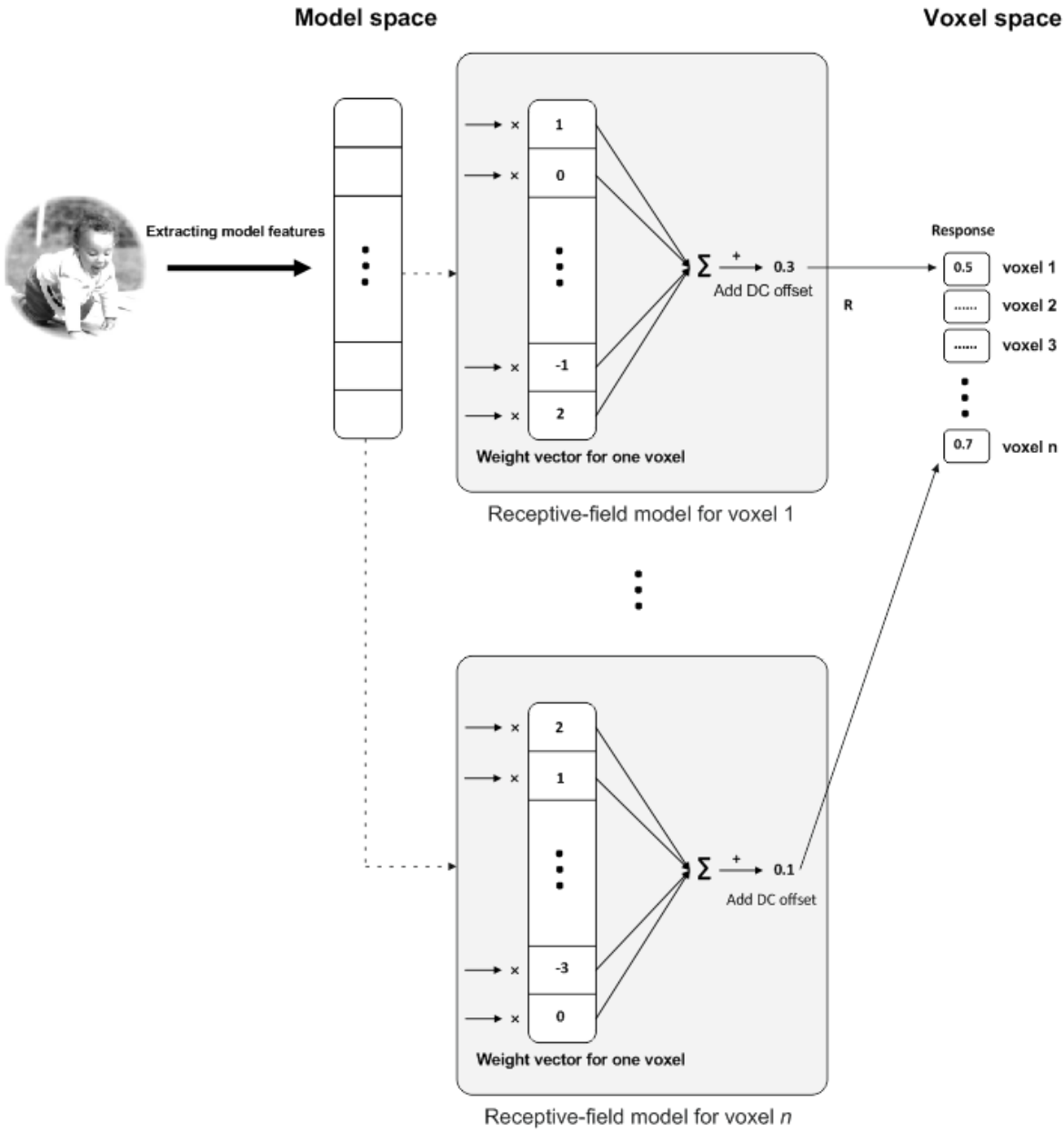

Figure 4. Mapping from model space to voxel space. The figure shows the process of predicting voxel responses using model features extracted from an image. There is one receptive field model for each voxel. In each receptive field model the weight vector and the DC offset are learnt in a training phase using 1750 training images for which we had model features and voxel responses. The weights are determined by gradient descent with early stopping. In the test phase, we used 120 test images (not included in the training images). For each image, model features were extracted and voxel responses for each ROI were predicted using the above procedure. The figure shows the process for a sample model; we did the same for all models.

Early visual areas are best accounted for by the Gabor wavelet pyramid (GWP) and the gist model

The Gabor wavelet pyramid model was used in Kay et al. (Kay et al., 2008) to predict responses of voxels in early visual areas in humans. Gabor wavelets are directly related to Gabor filters, since they can be designed for different scales and rotations. The aim of GWP has been to model early stages of visual information processing, and it has been shown that 2D Gabor filters can provide a

good fit to the receptive field weight functions found in simple cells of cat's striate cortex (Jones and Palmer, 1987). Interestingly, one of the instantiations of the GWP model (reweighted GWP), had the highest RDM correlation with both V1, and V2 (Figure 5). The model comes very close to the noise ceiling of these two early visual areas (V1, and V2), although it does not reach the noise ceiling. Indeed the noise ceiling for these two areas is much higher than for the other areas. The highest correlation obtained between a model and a brain ROI is for the GWP model and the early visual areas V1 and V2. This suggests that early vision is better modelled or better understood, compared to other brain ROIs. Newer Gabor-based models of early visual areas (Kay et al., 2013) may explain early visual areas even better.

The next best model in explaining early visual areas was the weighted gist model. For V2, in addition to the GWP, and the gist model, the HMAX-C2 features (not fitted) also showed a high RDM correlation. All these models that better explained V1 and V2 are built based on Gabor-like features.

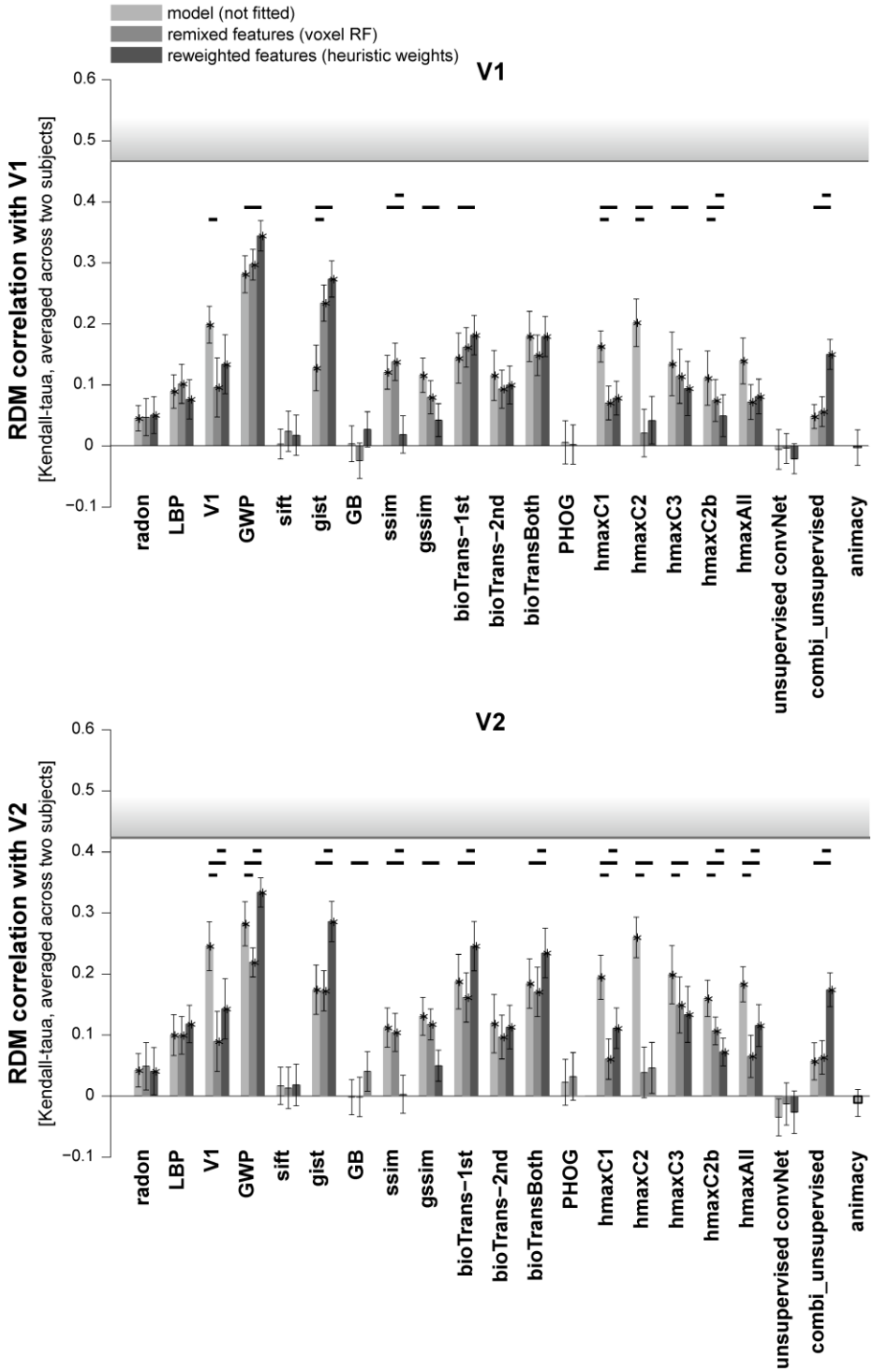

Figure 5. Kendall tau-a RDM correlation of models with early visual areas. Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) with V1, and V2 brain RDMs. There are three bars for each model. The first bar, 'model (not fitted)', shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels. The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI, using 1750 training images (refer to Figure 4 to see how the fitting is done). We used gradient descent to find the weight vectors (the mapping between a model feature vector and the voxels of a brain ROI). The weight vectors were then used to weight the model features, without mapping them to the brain space; the third bar (reweighted features) for each model shows the RDM correlation of the weighted model features with the reference brain ROI. Stars above each bar show statistical significance obtained by signrank test at 5% significance level. Black horizontal bars show that the difference between the bars for a model is statistically significant (signrank test, 5% significance level). The results are the average over the first two subjects. The shaded horizontal bar for each ROI indicates the lower bound of the noise ceiling. The lower bound is defined as the average Kendall-taua correlation of the ten 12x12 RDMs (120 test stimuli) between the two subjects. The animacy model is categorical, consisting of a single binary variable, so remixing has no effect on the predicted RDM rank order. We therefore only show the unfitted animacy model.

Intermediate and higher visual areas are best explained by models that are fitted/weighted to predict voxel responses in these areas

Several models show high correlations with V3, and V4, and some of them go within the noise ceiling for V4 –however, notice that the noise ceiling is lower in V4 compared to V1, V2, and V3 (Figure 6).

Similar to early visual areas, the weighted GWP and the weighted gist model, have high RDM correlations with V3, and V4, and the voxel RF-fitted instantiation of the two models have high correlations with LO (Figure 7), which comes close to the noise ceiling. In V4, the weighted gist model goes above the lower bound of the noise ceiling. This suggests that the gist features form the basis for explaining intermediate visual areas, and only by an appropriate reweighting of the features we could explain the V4 data. Overall from these results we may conclude that the gist model and the GWP model make a good basis for predicting voxel responses in the brain from early visual areas to intermediate levels, and to some degree in higher visual areas; and they just have to be reweighted appropriately.

More generally, intermediate visual areas are best accounted for by the reweighted features of the following models: GWP, gist, and bioTransform-1st stage. The unfitted instantiation of the intermediate layers of the HMAX model (HMAX C2, and C3) were also good in explaining the intermediate visual areas.

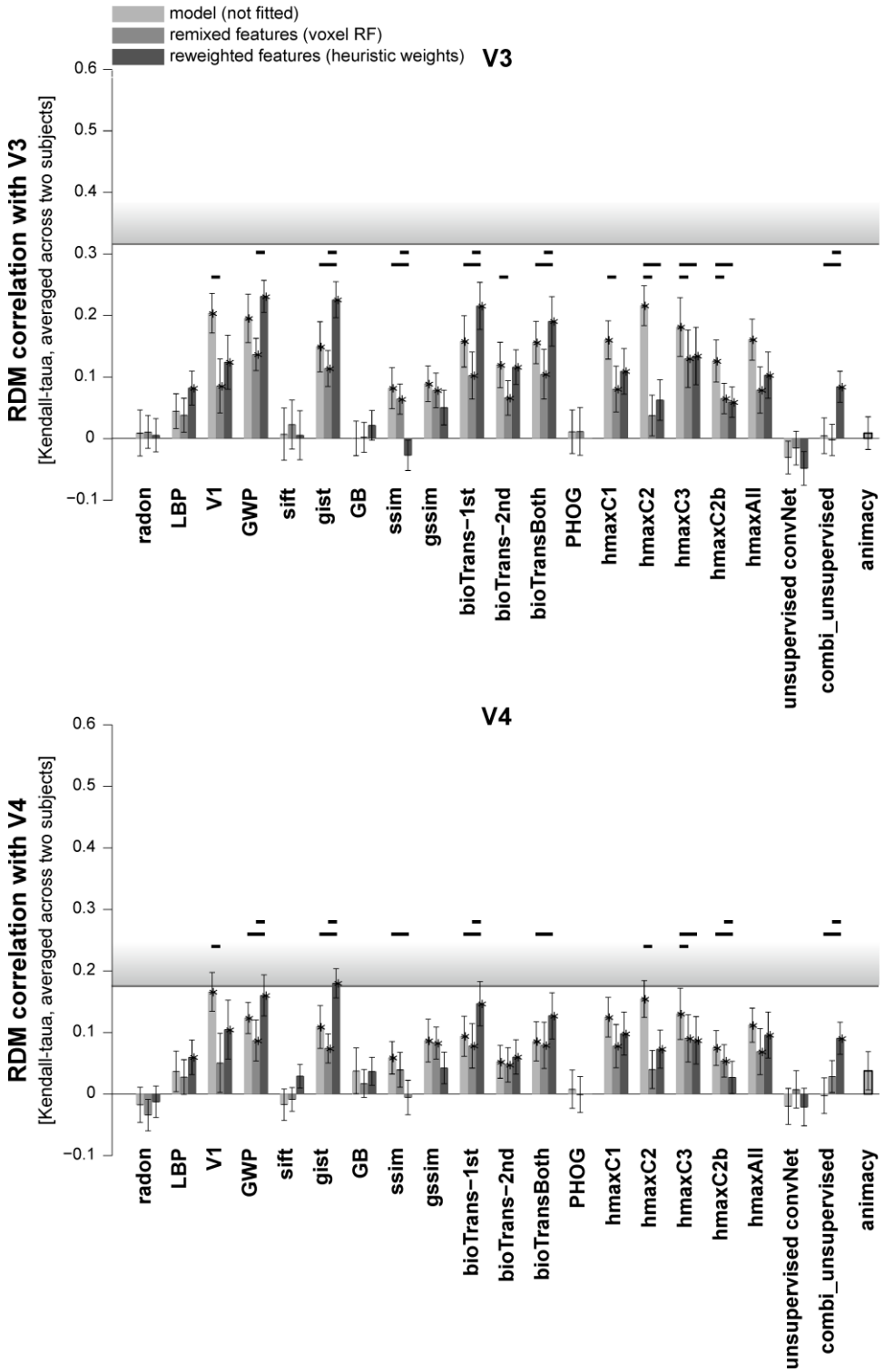

Figure 6. Kendall tau-a RDM correlation of models with the intermediate-level visual areas. Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) with V3, and V4 brain RDMs. There are three bars for each model. The first bar, 'model (not fitted)', shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels. The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI, using 1750 training images (refer to Figure 4 to see how the fitting is done). The third bar (reweighted features) shows the RDM correlation of the weighted model features with a reference brain ROI. The shaded horizontal bar in each panel indicates the lower bound of the noise ceiling. The statistical analyses and conventions here are analogous to Figure 5.

For the higher visual area, LO, the animacy model has the highest correlation, and comes close to the noise ceiling. The RF-fitted instantiation of the GWP and the gist model also come close to the LO noise ceiling.

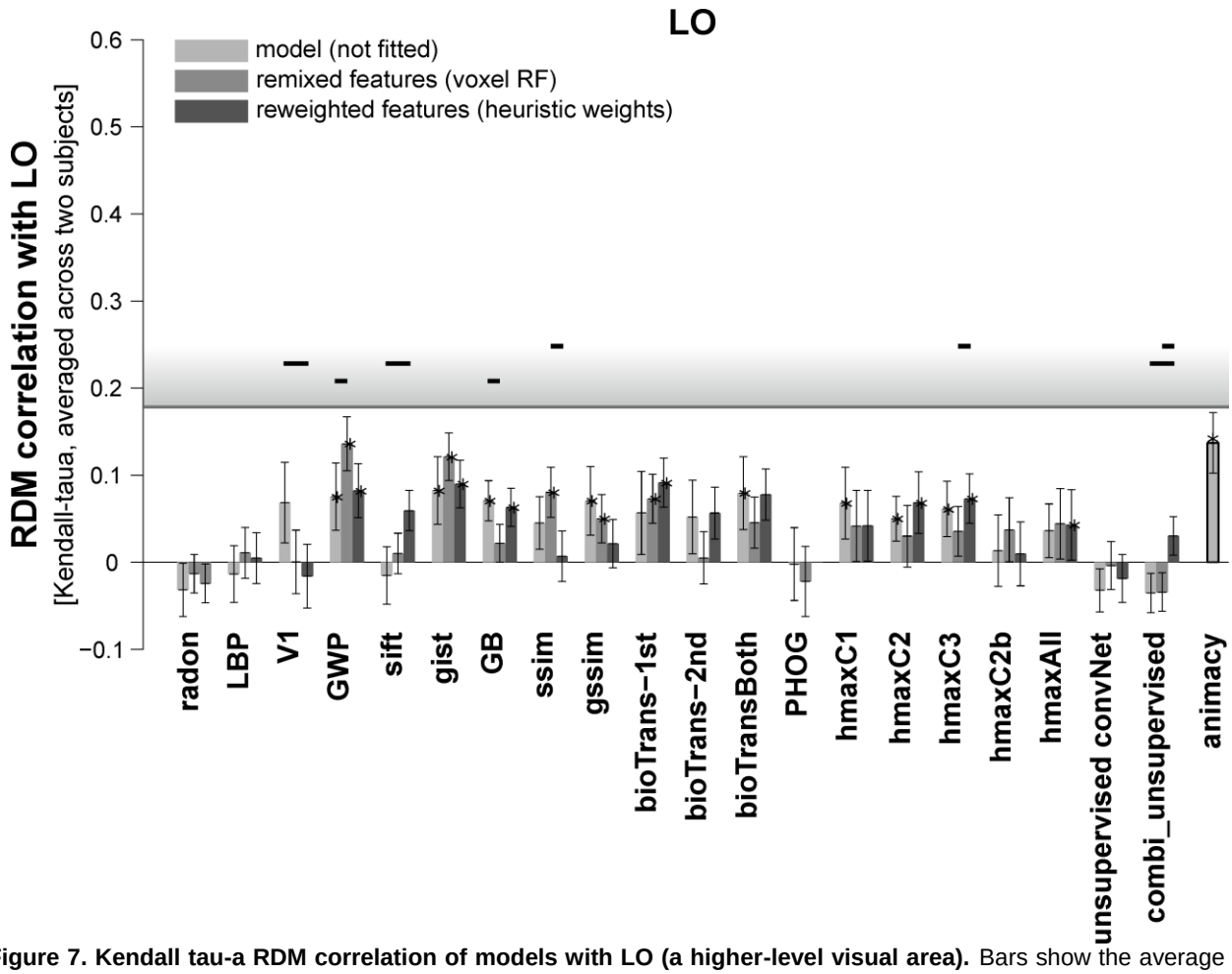

Figure 7. Kendall tau-a RDM correlation of models with LO (a higher-level visual area). Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) with the LO RDM. There are three bars for each model. The first bar, 'model (not fitted)', shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels. The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI, using 1750 training images (refer to Figure 4 to see how the fitting is done). The third bar (reweighted features) shows the RDM correlation of the weighted model features with a reference brain ROI. The shaded horizontal bar indicates the lower bound of the noise ceiling. The statistical analyses and conventions here are analogous to Figure 5.

From these results (remixing/reweighting of the unsupervised models) and also the results from remixing layers of the deep convNet (Figure 8), we can see that the reweighted/remixed models in general are doing a better job in explaining intermediate and higher visual areas than the unfitted models. This suggests that an appropriate linear recombination of model features (i.e. performing general affine transformations) can improve the performance of models in explaining intermediate and higher visual areas. Specially in the case of remixing, this can be better seen for the deep supervised convolutional network (Figure 8), where the voxel RF-fitted instantiation of the model layers (i.e. remixed deep convNet) explains intermediate and higher visual areas significantly better than the unfitted model instantiation.

Higher visual areas are best explained by the animacy model and higher layers of the remixed deep convolutional network

In recent years, a deep supervised convolutional neural network that is trained with 1.2 million labelled images from imageNet (Deng et al., 2009) (1000 category labels) has been proved to be very successful in object recognition tasks, and has achieved top-1 and top-5 error rates on the ImageNet data that is significantly better than previous state-of-the art results on this dataset (Krizhevsky et al., 2012). The network has 8 layers: 5 convolutional layers, followed by 3 fully connected layers. We tested this model, and compared the representation of the different layers of the model with the representation of visual areas along the hierarchy (Figure 8). Interestingly, the early layers of the model are more correlated with early visual areas and less correlated with intermediate or higher visual areas. As for the intermediate visual areas, layer 4 of the model has the highest correlation with V4, and reaches the noise ceiling. And the higher layers of the model have higher correlations with LO, compared to other regions.
Overall among all models, the one that best explains LO is the voxel RF-fitted version of layer 6 of the deep convolutional network, referred to as 'fc6'. This layer also has the highest animate/inanimate categorization accuracy (Figure 9). Apart from layer 6 of the deep convolutional network, the animacy model also gives a good account of the LO. The remixed version of layer 6 reaches the noise ceiling for LO and the animacy model comes very close to the noise ceiling. The remixed version of some other layers of the deep convolutional network also come very close to the noise ceiling (i.e. layer 3, layer 4, layer 5, layer 7, and layer 8), however none of the unfitted instantiations come close to the noise ceiling.

The voxel RF-fitted instantiations of the layers of the deep convolutional network work much better than the unfitted features, in terms of explaining the brain data, particularly for higher visual areas. This suggests that the features from the deep convolutional network form a good basis for predicting voxel responses, and they just have to be appropriately recombined (by linear remixing). This is consistent with our previous study (Khaligh-Razavi and Kriegeskorte, in-press), in which we also showed that by remixing and reweighting the features from the deep supervised convolutional network we could explain the data that we had from IT for a set of 96 stimuli. Having said that, whether the features of the deep convolutional network rely on computational mechanisms similar to the human visual system is yet to be determined.

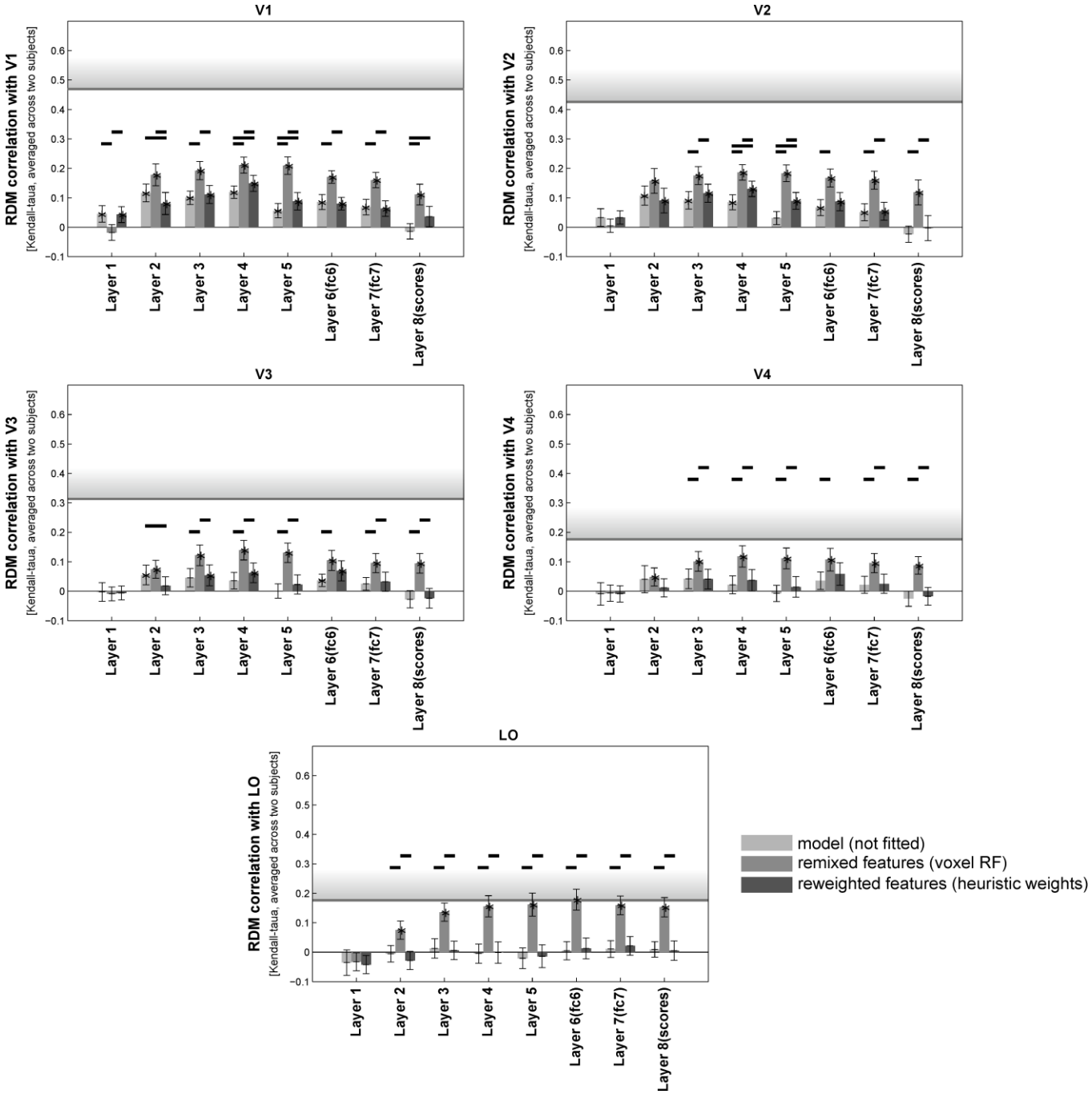

Figure 8. Kendall tau-a RDM correlation of the deep convNet layers across the hierarchy of visual areas. Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) between different layers of the deep convolutional network with each of the brain ROIs. There are three bars for each layer of the model: model (not fitted), voxel RF-fitted, and weighted features. They are defined in a similar way as explained in Figure 5. The shaded horizontal bar in each panel indicates the lower bound of the noise ceiling. The statistical analyses and conventions here are analogous to Figure 5.

Comparing the animate/inanimate categorization accuracy of the layers of the deep convolutional network (Figure 9) with other models (Figure 3) shows that the deep convolutional network is generally better at this task, particularly the higher layers of the model. Given that the animacy model explains a significant non-noise variance of LO, this may explain why the higher layers of this model that are good at animate/inanimate discrimination also better explain LO. The deep convolutional network is trained with so many labelled images as opposed to other models in Figure 3 that are all unsupervised models. This could explain why the deep convolutional network is better in animate/inanimate categorization task.

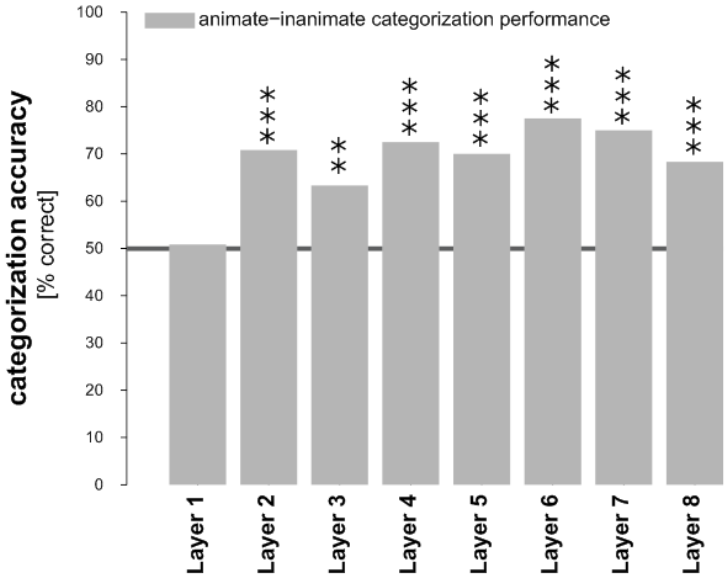

Figure 9. Animate-inanimate categorization performance based on different layers of the deep convNet. Bars show animate vs. inanimate categorization performance for each of the layers of the deep convolutional network. A linear SVM classifier was trained using 1750 training images and tested by 120 test images. P values that are shown by asterisks show whether the categorization performances significantly differ from chance [$p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***] . P values were obtained by random permutation of the labels (number of permutations = 10,000).

Early layers of the deep convolutional network are inferior to GWP and gist model in explaining the early visual areas

Although the higher layers of the deep convolutional network successfully work as the best model in explaining higher visual areas, the early layers of the model are not very successfully in explaining the early visual areas. The early visual areas (V1 and V2) are best explained by GWP model, and then the gist model. For example, the RDM correlations of the first two layers of the deep convolutional network with V1 are 0.04 (layer 1–not fitted) and 0.18 (layer 2–RF-fitted), respectively. However, the V1 correlation with GWP model (weighted features), which is 0.34, is significantly higher ($p < 0.001$, signrank test). Therefore, GWP provides a much better account of the early visual system than the early layers of the deep convolutional network. It may well be the case that improving the features in early layers of the deep convolutional network, in a way that makes it more similar to early visual areas, could improve the model performance in higher layers of the model.

Materials and methods

Presented stimuli, and fMRI data

In this study we used the experimental stimuli and fMRI data from (Kay et al., 2008; Naselaris et al., 2009). The data have been previously described and analysed to address different questions. The stimuli were gray-scale natural images that were masked with a 20°-diameter circle. There were 1750 training stimuli that were presented to subjects in 5 scanning sessions with 5 runs in each session (overall 25 experimental runs). Each run consisted of 70 distinct images presented two times each. The testing stimuli were 120 gray-scale natural images. The data for testing stimuli were collected in 2 scanning sessions with 5 runs in each session (overall 10 experimental runs). Each run consisted of 12 distinct images presented 13 times each.

Data from three subjects were analysed (S1–S3). The regions-of-interests (i.e. V1, V2, V3, V4, LO, V3A and V3B) were identified using a retinotopic mapping procedure. The data for retinotopic mapping was collected in separate scan sessions. See (Kay et al., 2008; Naselaris et al., 2009) for further experimental details.

The data were pre-processed using an updated protocol which included slice-timing correction, motion correction, upsampling to (1.5 mm)^3 resolution and improved co-registration between the functional data sets. The data were modelled with a variant of the general linear model including discrete cosine basis set for the hemodynamic response function (HRF) estimation. The beta weights characterizing the amplitude of the BOLD response to each stimulus were transformed to Z scores. Our analysis was restricted to voxels with signal-to-noise ratio greater than 1.5 (median value observed across all images).

Representational similarity analysis (RSA)

RSA enables us to relate representations obtained from different modalities (e.g. computational models and fMRI patterns) by comparing the dissimilarity patterns of the representations. In this framework representational dissimilarity matrix (RDM) is a square symmetric matrix in which the diagonal entries reflect comparisons between identical stimuli and are 0, by definition. Each off-diagonal value indicates the dissimilarity between the activity patterns associated with two different stimuli. RDM summarizes the information carried by a given representation from an area in the brain or a computational model. In this study, the fMRI response patterns evoked by the different natural images were compared to each other using representational dissimilarity matrices (RDMs). The measure for dissimilarity was correlation distance (1- Pearson linear correlation) between the response patterns.
There were 120 testing stimuli. For each brain ROI, we had ten 12x12 RDMs. That is one RDM for each experimental run (10 runs consisted of 12 distinct images each = 120 distinct images overall). Each image was presented 13 times. The trials were divided to two independent data sets (odd and even trials, the 13th trial excluded from the analysis). To remove the effects of coherent trial-to-trial fluctuations and stimulus presentation order, odd trials were averaged and compared to the average of even trials (cross-trial comparison). All comparisons among brain ROIs were done in this way.

To calculate the correlation between model and brain RDMs, within each experimental run, all trials were averaged, which gives us one 12x12 RDM for each run. This gives us ten 12x12 RDMs, each of which were correlated with the model RDM obtained from the same 12 images. The reported correlations are the average of these ten correlations.

Voxel receptive field modelling

Voxel-receptive-field modelling (Kay et al., 2008, 2013) aims to construct a computational model for each fMRI voxel and to predict the voxel responses for new stimuli. Voxel-receptive field mapping therefore requires a linear model (predicting the measured responses from the model representation) to be fitted with one data set and tested with a separate data set (different stimuli). Both RSA framework and voxel receptive field modelling has been used separately to evaluate computational models. RSA compares the response-pattern dissimilarities between models and brain ROIs; whereas, voxel receptive-field modelling uses computational-model representations to predict the measured response patterns of brain voxels.

Combining voxel receptive field modeling with RSA: In this study we have taken advantage of both, and we bridge the gap between voxel receptive field modeling and RSA: using voxel receptive field modelling, and based on training images we first learn a mapping between model representations and each of the brain voxels. Then we predict the response-pattern of brain voxels for the test stimuli using the internal representation of the models. Finally, we use RSA to compare the pattern-dissimilarities between the predicted voxel responses and the brain voxels. This gives us the additional benefit of finding visual features that better predict brain responses, and might enable us to further understand the nature of features that brain uses in each level of the hierarchy of the ventral visual pathway. Mapping model responses to voxel responses through voxel receptive field modeling is a way of remixing of model features. By linear **remixing of features** (affine recoding), we go beyond stretching and squeezing the representational space along its original axes (feature reweighting) and attempt to create new features as linear combinations of the original features. This provides a more general transformation, affine recoding, which includes feature reweighting as a special case.

Figure 4 shows how the internal representation of each of the object-vision models is used to learn and then to predict responses of brain voxels to presented stimuli. During the learning process, for each of the brain voxels we learn a weight vector and a DC offset value that maps the internal representation of an object-vision model to the responses of brain voxels. We only use the 1750 training images and the voxel responses to these stimuli. The weights are determined by gradient descent with early stopping [see (Kay et al., 2008) for further details]. Early stopping is a form of regularization (Skouras et al., 1994) where the magnitude of model parameter estimates is shrunk in order to prevent overfitting. A new mapping is learnt for each of the object-vision models. Finally in the testing phase, we use the learned mapping to predict voxel responses to the 120 test stimuli. For a given model and a presented image, we use the extracted model features and calculate the inner product of the feature vector with each of the weight vectors that were learnt in the training phase for each voxel. We then add the learnt DC offset value to the results of the inner product for each voxel, which gives us the predicted response value for that voxel.

Weighting model features

Heuristic reweighting: For each of the object-vision models we made a weighted set of model features for each ROI (dimensionality of the weight vector: number of model features * number of voxels in the ROI). The weights for each ROI that were learnt to map model responses to voxel responses –using voxel receptive field modeling (Figure 4) – were averaged across all voxels; the averaged weight vector was then used to weight the model features. In other words, for an object-vision model, the weighted model features for an ROI are simply the result of multiplying the model features with the mean of the weight vectors learnt for each voxel in the ROI. This approach emphasizes on the model features that are more predictive of brain voxels, by giving higher weights to these features, and lower weights to less relevant features. In Figures 5, 6, 7, and 8 the weighted models are shown by *'reweighted features'*.

Kendall τ_A (tau-a) correlation and noise ceiling

To judge the ability of a model RDM in explaining a brain RDM, we used Kendall's rank correlation coefficient τ_A (which is the proportion of pairs of values that are consistently ordered in both variables). When comparing models that predict tied ranks (e.g. category model RDMs) to models that make more detailed predictions (e.g. brain RDMs, object-vision model RDMs) Kendall's τ_A correlation is recommended. In these occasions τ_A correlation is more likely than the Pearson and Spearman correlation coefficients to prefer the true model over a simplified model that predicts tied ranks for a subset of pairs of dissimilarities. For more information in this regard please refer to the RSA Toolbox paper (Nili et al., 2014).

The noise in the brain activity data has imposed limitations on the amount of dissimilarity variance that a model RDM can explain. Therefore an estimation of noise-ceiling was needed to indicate how much variance of a brain RDM –given the noise level– was expected to be explained by an ideal model RDM (i.e. a model RDM that is able to perfectly capture the true dissimilarity structure of the brain RDM).

The lower bound of the noise-ceiling for each ROI is defined as the average Kendall τ_A correlation of the ten 12x12 RDMs (120 test stimuli) between the first two subjects (Figures 5, 6, 7, and 8). This is consistent with the definition of the noise ceiling in (Nili et al., 2014). We did not estimate the upper bound of the noise ceiling as it did not serve any purpose for our analysis in this study. In relating different model instantiations with brain ROIs, we left out the third subject, given that the data from the third subject was noisier and less consistent with the other two. To see the results for each subject individually see supplementary Figures S1, S2, and S3).

Object-vision models

We used a wide range of computational models (Khaligh-Razavi, 2014) to explore many different ways for extracting visual features. We selected some of the well-known bio-inspired object recognition models as well as several models and feature extractors from computer vision. Furthermore, to search the model space more comprehensively, in addition to the model

representation itself, we make two other instantiations from each model representation (therefore for each model representation we have three model instantiations). The instantiations are, the voxel-RF fitted models (i.e. voxel responses that are predicted from model representation), and the weighted features (i.e. model features that are weighted using the weights obtained in the voxel receptive field modeling step).
The visual area RDMs were compared to all these models. In a model-RDM, each cell reflects the dissimilarity of an image pair predicted by the computational model. The comparison between a brain-RDM and a model-RDM was based on Kendall's tau-a rank correlation distance of the values in the upper triangles of the RDMs.
Below is a description for all models used in this study.

Gabor wavelet pyramid: The Gabor wavelet pyramid model was adopted from Kay et al. (2008). Each image was represented by a set of Gabor wavelets of six spatial frequencies, eight orientations and two phases (quadrature pair) at a regular grid of positions over the image. To control gain differences across wavelets at different spatial scales, the gain of each wavelet was scaled such that the response of that wavelet to an optimal full-contrast sinusoidal grating is equal to 1. The response of each quadrature pair of wavelets was combined to reflect the contrast energy of that wavelet pair. The outputs of all wavelet pairs were concatenated to have a representational vector for each image.

Gist: The spatial envelope or gist model aims to characterize the global similarity of natural scenes (Oliva and Torralba, 2001). The gist descriptor is obtained by dividing the input image into 16 bins, and applying oriented Gabor filters in 8 orientations over different scales in each bin, and finally calculating the average filter energy in each bin[1].

Animate–inanimate distinction: The natural images were labelled as animate if they contained one or several humans or animals, bodies of humans or animals, or human or animal faces. In the animate–inanimate model-RDM, the dissimilarities are either 0 (identical responses) if both images are of the same category (animate or inanimate) or 1 (different responses) if one image is animate and the other is inanimate. Because the animacy model is essentially one-dimensional, remixing it will not change the representational space beyond a scaling factor. Therefore we did not include the remixed version of the animacy model.

Radon: The Radon transform of an image is a matrix, in which each column corresponds to a set of integrals of the image intensities along parallel lines of a given angle. The Matlab function Radon was used to compute the Radon transform for each luminance image.

Unsupervised convolutional network: A hierarchical architecture of two stages of feature extraction, each of which is formed by random convolutional filters and subsampling layers (Jarrett et al., 2009). Convolutional layers scan the input image inside their receptive field. Receptive Fields (RFs) of convolutional layers get their input from various places on the input image, and RFs with identical weights make a unit. The outputs of each unit make a feature map. Convolutional layers are then followed by subsampling layers that perform a local averaging and subsampling, which

[1] http://people.csail.mit.edu/torralba/code/spatialenvelope/

make the feature maps invariant to small shifts (Bengio et al., 1995). The convolutional network which we used[2] had two stages of unsupervised random filters, that is shown by RR in table 1 in Jarret et al. (2009) (Jarrett et al., 2009). The obtained result for each image was then vectorized. The parameters were exactly the same as used in (Jarrett et al., 2009) .

Deep supervised convolutional neural network: The deep supervised convolutional network by Krizhevsky et al. (Donahue et al., 2013; Krizhevsky, A., Sutskever, I. and Hinton, G. E., 2012) is trained with 1.2 million labelled images from ImageNet (1000 category labels), and has 8 layers: 5 convolutional layers, followed by 3 fully connected layers. The output of the last layer is a distribution over the 1000 class labels. This is the result of applying a 1000-way softmax on the output of the last fully connected layer. The model has 60 million parameters and 650,000 neurons. The parameters are learnt with stochastic gradient descent.[3]

Biological Transform (BT): BT is a hierarchical transform based on local spatial frequency analysis of oriented segments. This transform has two stages, each of which has an edge detector followed by an interval detector (Sountsov et al., 2011). The edge detector consists of a bar edge filter and a box filter. For a given interval I and angle θ, the interval detector finds edges that have angle θ and are separated by an interval I. In the first stage, for any given θ and I, all pixels of the filtered image were summed and then normalized by the squared sum of the input. They were then rectified by the Heaviside function. The second stage was the same as the first stage, except that in the first stage θ was changing between 0-180 ° and I between 100-700 pixels and the input to the first stage had not a periodic boundary condition on the θ axis (repeating the right-hand side of the image to the left of the image and vice versa); but in the second stage the input, which is the output of the first stage, was given a periodic boundary condition on the θ axis, and I was changing between 15-85 pixels.

Geometric Blur (GB): 289 uniformly distributed points were selected on each image, then the Geometric Blur descriptors(Belongie et al., 2002; Berg et al., 2005; Zhang et al., 2006) were calculated by applying spatially varying blur around the feature points. We used GB features that were part of multiple kernels for image classification described in (Vedaldi et al., 2009)[4]. The blur parameters were set to $\alpha=0.5$ and $\beta=1$; the number of descriptors was set to 300.

Dense SIFT: For each grayscale image, SIFT descriptors (Lowe, 2004) of 16x16 pixel patches were sampled uniformly on a regular grid. Then, all the descriptors were concatenated in a vector as the SIFT representation of that image. We used the dense SIFT descriptors that were used in (Lazebnik et al., 2006) to extract PHOW features, described below.

Pyramid Histogram of Gradients (PHOG): The canny edge detector was applied on grayscale images, and then a spatial pyramid was created with four levels (Bosch et al., 2007). The histogram of orientation gradients was calculated for all bins in each level. All histograms were

[2] http://koray.kavukcuoglu.org/code.html

[3] http://caffe.berkeleyvision.org/

[4] http://www.robots.ox.ac.uk/~vgg/software/MKL/#download

then concatenated to create PHOG representation of the input image. We used Matlab implementation that was freely available online[5]. Number of quantization bins was set to forty, number of pyramid levels to four and the angular range to 360^0.

Local Self-Similarity descriptor (ssim): This is a descriptor that is not directly based on the image appearance; instead, it is based on the correlation surface of local self-similarities. For computing local self-similarity features at a specific point on the image, say *p*, a local internal correlation surface can be created around *p* by correlating the image patch centred at *p* to its immediate neighbours (Chatfield et al., 2009; Shechtman and Irani, 2007)**.** We used the code available for ssim features that were part of multiple kernels for image classification described in (Vedaldi et al., 2009)[6]. The ssim descriptors were computed uniformly at every five pixels in both X and Y directions.

Global Self-Similarity descriptor (gssim): This descriptor is an extension of the local self-similarity descriptor mentioned above. Gssim uses self-similarity globally to capture the spatial arrangements of self-similarity and long range similarities within the entire image (Deselaers and Ferrari, 2010). We used gssim Matlab implementation available online[7]. Number of clusters for the patch prototype codebook was set to 400, with 20000 patches to be clustered. D1 and D2 for the self-similarity hypercube were both set to 10.

Local Binary Patterns (LBP): Local binary patterns are usually used in texture categorization. The underlying idea of LBP is that a 2-dimensional surface can be described by two complementary measures: local spatial patterns and gray scale contrast. For a given pixel, LBP descriptor gives binary labels to surrounding pixels by thresholding the difference between the intensity value of the pixel in the center and the surrounding pixels (Ojala et al., 2001, 2002; PietikÃ¤inen, 2010). We used LBP Matlab implementation freely available online[8]. Number of sampling points was fixed to eight.

V1 model: A population of simple and complex cells were modelled and were fed by the luminance images as inputs. Gabor filters of 4 different orientations (0^o, 90^o, -45^o, and 45^o) and 12 sizes (7-29 pixels) were used as simple cell receptive fields. Then, the receptive field of complex cells were modelled by performing the MAX operation on the neighboring simple cells with similar orientations. The outputs of all simple and complex cells were concatenated in a vector as the V1 representational pattern of each image.

HMAX: The HMAX model developed by Serre et al.(Serre et al., 2007) has a hierarchical architecture inspired by the well-known simple to complex cells model of Huble & Wiesel (Hubel and Wiesel, 1968; HUBEL and WIESEL, 1962). The HMAX model that is used here adds three

[5] http://www.robots.ox.ac.uk/~vgg/research/caltech/phog.html

[6] http://www.robots.ox.ac.uk/~vgg/software/SelfSimilarity/

[7] http://www.vision.ee.ethz.ch/~calvin/software.html

[8] http://www.cse.oulu.fi/CMV/Downloads/LBPMatlab

more layers –ends at S4- on the top of the complex cell outputs of the V1 model described above. The model has alternating S and C layers. S layers perform a Gaussian-like operation on their inputs, and C layers perform a max-like operation, which makes the output invariant to small shifts in scale and position. We used the freely available version of the HMAX model[9]. All simple and complex layers were included until the S4 layer. We used the pre-trained version of the HMAX model (i.e. trained with large number of natural images).

Combi_Unsupervised: This is the concatenation of features extracted by the 19 unsupervised model representations. Given an input stimulus, features from all of the above-mentioned models were extracted. Because the dimension for extracted features differs across models, we used principle component analysis (PCA) to reduce the dimension of all of them to a unique number. We used the first 119 PCs from each of the models and concatenated them along a vector (119 was the largest possible number of PCs that we were able to use, because we had 120 testing images; so the covariance matrix has only 119 non-zero eigenvalues).

Discussion

Higher visual areas present a difficult explanatory challenge and can be better studied by considering the transformation of representations across the stages of the visual hierarchy from lower- to higher-level visual areas. Here we investigated the progress of visual information through the hierarchy of visual cortex by comparing the representational geometry of several brain regions with a wide range of object-vision models. The models tended to have higher correlations with early visual areas, than with higher visual areas. In comparing models with brain regions, we also considered the effect of coherent response-pattern fluctuations between visual areas. To compare the stimulus-driven component of the brain representations in two brain regions, we used separate set of trials, which have been presented in independent random orders, to estimate the response patterns for each area (Henriksson et al., 2014). By doing so we minimize the confounding effect of correlated intrinsic fluctuations in dissimilarity patterns. Importantly, we presented a new methodological framework for testing models, remixed RSA, which bridges the gap between RSA and voxel-RF modeling, both of which have been used separately but not in combination in previous studies (Kriegeskorte et al., 2008a; Nili et al., 2014; Khaligh-Razavi and Kriegeskorte, in-press; Kay et al., 2008, 2013). Using remixed RSA, we evaluated the performance of many models and several brain areas.

Performance of different models across the visual hierarchy

The models that we tested here were all feedforward models of vision, from simple feature extractors (e.g. SIFT), to a complex convolutional neural network model that has recently been shown to be very successful at visual tasks. We first compared the representational geometry of this wide range of models with the stages of visual hierarchy. We then employed voxel-receptive-field modeling to construct new computational models based on the features extracted by each of

[9] http://cbcl.mit.edu/software-datasets/pnas07/index.html

the object-vision models and to predict voxel responses for the test stimuli (nonoverlapping set of natural images). We explored a wide range of models (i.e. three model instantiations for each of the 28 model representations + the animacy model = 85 model representations in total) and extended previous findings on this model set (Khaligh-Razavi and Kriegeskorte, in-press; Khaligh-Razavi and Kriegeskorte 2013) by showing which models best explain each brain representation across the hierarchy of visual areas.

Unsupervised models partially explain the lower-level representations without remixing

The unsupervised models explained substantial variance components of the early visual representations, although they did not reach the noise ceiling. These model may reflect low-level image similarity in a similar way to the representation in lower visual areas.

The remixed versions of the unsupervised models often performed significantly worse than the original versions of those models. For lower visual areas, the original models already approximate the representational space quite well. The space of remixed models, of course, contains the original model as a special case, but the remixing might be hampered to some extent by overfitting. In addition, the L2 regularisation implies a prior favouring a remixing matrix with small distributed weights. However, to recover the original model, the remixing matrix would have to be a permuted version of the identity matrix, which would incur a large L2 penalty. This illustrates the fact that the regulariser in voxel-RF modelling implies a prior that is part of the model and can hurt model performance when it is inappropriate.

For higher visual areas, the unsupervised models were not successful either with or without remixing. None of the unsupervised models came close to the LO noise ceiling. One explanation for this is that these models are missing the visuo-semantic nonlinear features needed to explain these representations, consistent with findings from a different data set in Khaligh-Razavi and Kriegeskorte (in press). In that previous study, the not-strongly-supervised models similarly failed to explain the higher ventral-stream representations with or without remixing. (The remixing in that study was based on weights set to optimise categorisation performance, because the data did not include training set of fMRI data for a rich separate set of stimuli.)

The deep supervised network explains the higher-level representations with remixing

The deep supervised network performed worse than the unsupervised models at explaining the early visual representations. However, it performed better at explaining the higher-level LO representation. Interestingly, in LO the remixed layers of the deep supervised network are often better than the animacy model.

Whereas the remixed versions of the unsupervised models performed worse than the original versions of those models, the remixed versions of the layers of the deep supervised convolutional network, tended to perform significantly better than the original versions of these representations. This again is broadly consistent with our previous results (Khaligh-Razavi and Kriegeskorte, in press) showing that remixing of the deep supervised network's features (by optimising categorisation performance) improves the explanation of the IT representational geometry. Remixing appears to be essential for the deep supervised model to account for the semantic

categorical clusters observed in higher visual areas (Connolly et al., 2012; Kiani et al., 2007; Kriegeskorte et al., 2008a; Mur et al., 2012; Naselaris et al., 2012).

Remixed RSA: bridging the gap between RSA and voxel-RF modelling

Remixing creates new features as linear combinations of the original features (thus performing general affine transformations). Remixed RSA combines RSA (Kriegeskorte, 2009; Kriegeskorte et al., 2008b; Nili et al., 2014) and voxel-receptive-field modeling (Kay et al., 2008; 2013; Dumoulin and Wandell, 2008), two complementary approaches to testing computational models with brain-activity data. Using training data acquired with a separate set of stimuli enabled us to fit a remixing matrix that predicts voxel responses from model features. We could then compare the representational geometry of the predicted voxel responses with that of actual voxel responses using the RSA framework.

Remixing enables us to investigate whether a linear recombination of model features can provide a better explanation of the brain representational geometry. This helps address the question of whether the model features (a) provide a good basis for explaining a brain region and just need to be appropriately linearly recombined or (b) the model features do not provide a good basis of the brain representation. However, remixing requires a combination of (a) substantial additional training data for a separate set of stimuli and (b) prior assumptions (e.g. implicit to the regularising penalty) about the remixing weights. The former is costly and the latter affects the interpretation of the results, because the prior is part of the model. The lower performance of the remixed unsupervised models (compared to their original unremixed versions) illustrates that remixing should not in general be interpreted as testing the best of all remixed models.

A special case of remixing is reweighting, which stretches and squeezes the representational space along its original axes. For the unsupervised models, this approach (although implemented in a rather ad-hoc heuristic fashion) often significantly improved model performance over the original and also over the remixed versions of a model. Since reweighting is a special case of remixing, the only explanation for its superior performance is that the remixing (even with L2-regularisation) is hampered by overfitting.

The reweighting was heuristic in the present analyses. We did not perform a search to find the optimal weights for which the RDM of the weighted model features has the highest correlation with the brain RDM. Instead we averaged the weights obtained from the voxel-RF modeling. These weights emphasize important model features, features that are more predictive of voxel responses, but are not necessarily the optimal weights. In Khaligh-Razavi and Kriegeskorte (in press), we used non-negative least-square fitting (which can equivalently be applied to squared Euclidean distance matrices rather than the original features) to find optimal weighted combinations of model features. Future studies should explore using nonnegative least-squares for weighting model representations (as in Khaligh-Razavi and Kriegeskorte, in press) or individual features of a model. Reweighting spans a smaller space of model variants that is easier to interpret and can be more stably estimated than general remixing. Reweighting of individual features requires estimation of only one parameter per model feature, much fewer than remixing, which requires estimation of one parameter per model feature for every voxel or recorded neuron. Reweighted RSA therefore

provides an interesting stepping stone between traditional and remixed RSA and might usefully complement these methods for elucidating the computational mechanisms of the brain.

References

Bell, A.H., Hadj-Bouziane, F., Frihauf, J.B., Tootell, R.B.H., and Ungerleider, L.G. (2009). Object Representations in the Temporal Cortex of Monkeys and Humans as Revealed by Functional Magnetic Resonance Imaging. J Neurophysiol 101, 688–700.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence 509–522.

Bengio, Y., Lecun, Y. (1995). Convolutional Networks for Images, Speech, and Time-Series.

Berg, A.C., Berg, T.L., and Malik, J. (2005). Shape matching and object recognition using low distortion correspondences. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, pp. 26–33.

Bosch, A., Zisserman, A., and Munoz, X. (2007). Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, (New York, NY, USA: ACM), pp. 401–408.

Chatfield, K., Philbin, J., and Zisserman, A. (2009). Efficient retrieval of deformable shape classes using local self-similarities. (IEEE), pp. 264–271.

Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.-C., Abdi, H., and Haxby, J.V. (2012). The Representation of Biological Classes in the Human Brain. J. Neurosci. 32, 2608–2618.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009, pp. 248–255.

Deselaers, T., and Ferrari, V. (2010). Global and efficient self-similarity for object classification and detection. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 1633–1640.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2013). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. arXiv:1310.1531 [cs].

Dumoulin, S. O., Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. Neuroimage 39, 647–660.

Eichhorn, J., Sinz, F., and Bethge, M. (2009). Natural Image Coding in V1: How Much Use Is Orientation Selectivity? PLoS Comput Biol 5, e1000336.

Güçlü, U., and van Gerven, M.A. (2014). Unsupervised Feature Learning Improves Prediction of Human Brain Activity in Response to Natural Images. PLoS Computational Biology 10, e1003724.

Henriksson, L., Khaligh-Razavi, S.-M., Kay, K., and Kriegeskorte, N. (2014). Intrinsic cortical dynamics dominate population responses to natural images across human visual cortex. bioRxiv 008961.

Hubel, D.H., and Wiesel, T.N. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology 195, 215.

Hubel, D., and Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology 160, 106–154.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science 310, 863.

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. Neuron 76, 1210–1224.

Jarrett, K., Kavukcuoglu, K., Ranzato, M.A., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In Computer Vision, 2009 IEEE 12th International Conference on, pp. 2146–2153.

Jones, J.P., and Palmer, L.A. (1987). An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. J. Neurophysiol. 58, 1233–1258.

Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature 452, 352–355.

Kay, K.N., Winawer, J., Rokem, A., Mezer, A., and Wandell, B.A. (2013). A Two-Stage Cascade Model of BOLD Responses in Human Visual Cortex. PLoS Comput Biol 9, e1003079.

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (in-press). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol.

Khaligh-Razavi, S.-M., and Kriegeskorte, N. (2013). Object-vision models that better explain IT also categorize better, but all models fail at both. Cosyne Abstracts, Salt Lake City USA.

Khaligh-Razavi, S.-M. (2014). What you need to know about the state-of-the-art computational models of object-vision: A tour through the models. arXiv:1407.2776 [cs, Q-Bio].

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object Category Structure in Response Patterns of Neuronal Population in Monkey Inferior Temporal Cortex. J Neurophysiol 97, 4296–4309.

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. Frontiers in Neuroscience 3, 363–373.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends in Cognitive Sciences 17, 401–412.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008a). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. Neuron 60, 1126–1141.

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

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In NIPS, (Lake Tahoe, Nevada),.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, pp. 2169–2178.

Lowe, D.G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision 60, 91–110.

Mur, M., Ruff, D.A., Bodurka, J., De Weerd, P., Bandettini, P.A., and Kriegeskorte, N. (2012). Categorical, Yet Graded – Single-Image Activation Profiles of Human Category-Selective Cortical Regions. J. Neurosci. 32, 8649–8662.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Bayesian Reconstruction of Natural Images from Human Brain Activity. Neuron 63, 902–915.

Naselaris, T., Stansbury, D.E., and Gallant, J.L. (2012). Cortical representation of animate and inanimate objects in complex natural scenes. Journal of Physiology-Paris 106, 239–249.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. PLoS Comput Biol 10, e1003553.

Ojala, T., Pietikäinen, M., and Mäenpää, T. (2001). A generalized local binary pattern operator for multiresolution gray scale and rotation invariant texture classification. Advances in Pattern Recognition—ICAPR 2001 399–408.

Ojala, T., Pietikainen, M., and Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on 24, 971–987.

PietikÃ¤inen, M. (2010). Local Binary Patterns. Scholarpedia 5, 9775.

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences 104, 6424–6429.

Shechtman, E., and Irani, M. (2007). Matching Local Self-Similarities across Images and Videos. In IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07, pp. 1–8.

Skouras, K., Goutis, C., and Bramson, M.J. (1994). Estimation in linear models using gradient descent with early stopping. Statistics and Computing 4, 271–278.

Sountsov, P., Santucci, D.M., and Lisman, J.E. (2011). A biologically plausible transform for visual recognition that is invariant to translation, scale, and rotation. Frontiers in Computational Neuroscience 5.

Vedaldi, A., Gulshan, V., Varma, M., and Zisserman, A. (2009). Multiple kernels for object detection. In Computer Vision, 2009 IEEE 12th International Conference on, pp. 606–613.

Zhang, H., Berg, A.C., Maire, M., and Malik, J. (2006). SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, pp. 2126–2136.

Funding: This work was supported by Cambridge Overseas Trust and Yousef Jameel Scholarship to SK; an Aalto University Fellowship Grant to LH and a European Research Council Starting Grant (261352) to NK. The authors declare no competing financial interests.

Supporting information

Supporting Text 1: The weights that are given to two well-performing object vision models (gist, GWP) by the voxel receptive field fitting are shown in Figure S4. It shows that for the GWP model that has so many features (43,680) only a small number of features are highly weighted and the rest are zero. This suggests that a sparse representation of features should be enough for predicting voxel responses. Interestingly the same features are highly weighted for other brain ROIs (Figure S5). The pattern of weights given to the features of GWP model by voxels from different brain ROIs look very similar, suggesting that the weighted features are commonly important features for all these ROIs regardless of where they are in the hierarchy of vision.

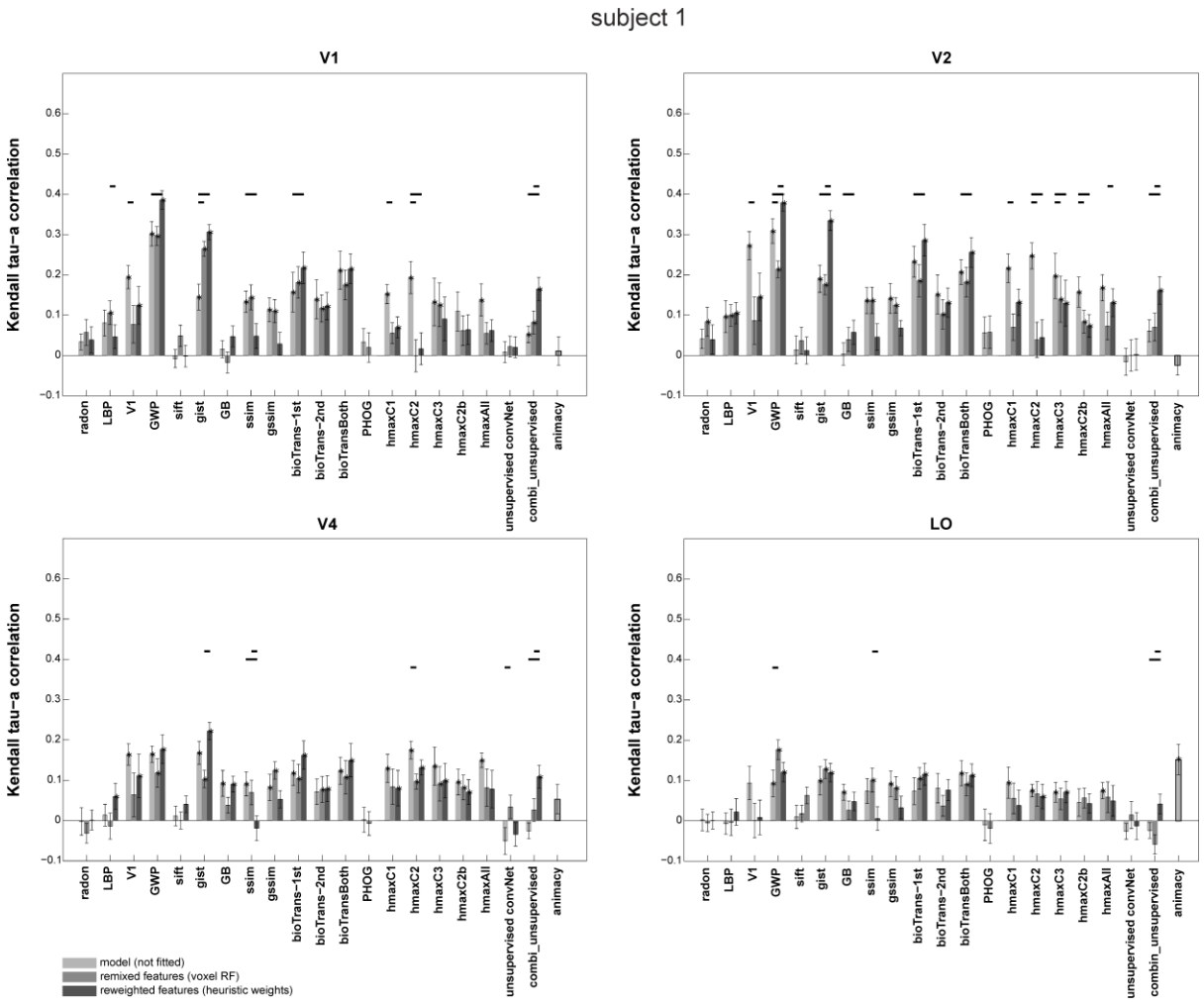

Figure S1. Kendall tau-a RDM correlation of the models with the brain ROIs of subject 1 across the hierarchy of visual areas. Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) with V1, V2, V4, and LO brain RDMs for subject 1. There are three bars for each model. The first bar, 'model (not fitted)', shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels. The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI, using 1750 training images (refer to Figure 4 to see how the fitting is done). We used gradient descent to find the weight vectors (the mapping between a model feature vector and the voxels of a brain ROI). The weight vectors were then used to weight the model features, without mapping them to the brain space; the third bar (reweighted features) for each model shows the RDM correlation of the weighted model features with the reference brain ROI. Stars above each bar show statistical significance obtained by signrank test at 5% significance level. Black horizontal bars show that the difference between the bars for a model is statistically significant (signrank test, 5% significance level).

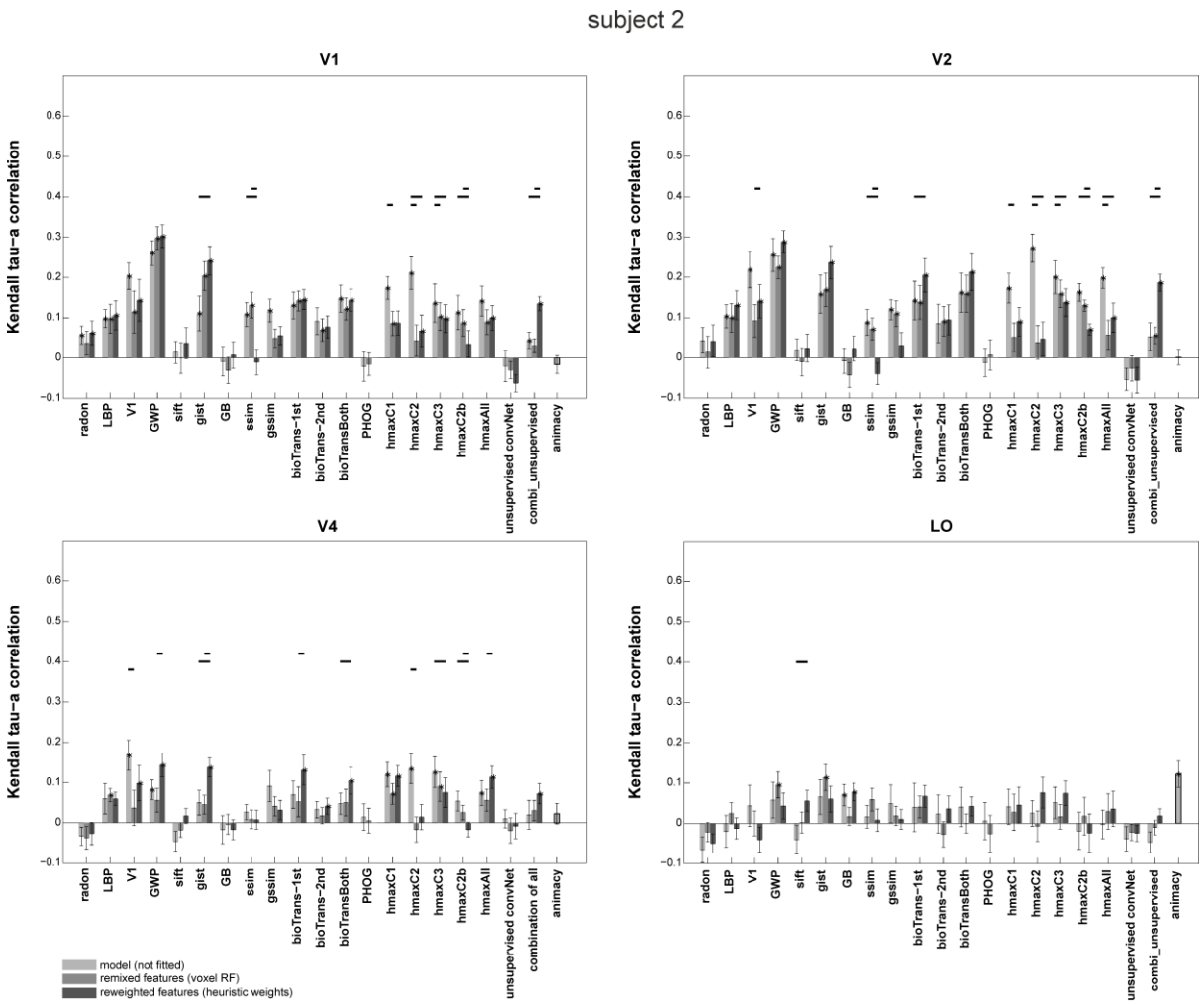

Figure S2. Kendall tau-a RDM correlation of the models with the brain ROIs of subject 2 across the hierarchy of visual areas. Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) with V1, V2, V4, and LO brain RDMs for subject 2. There are three bars for each model. The first bar, 'model (not fitted)', shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels. The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI, using 1750 training images (refer to Figure 4 to see how the fitting is done). We used gradient descent to find the weight vectors (the mapping between a model feature vector and the voxels of a brain ROI). The weight vectors were then used to weight the model features, without mapping them to the brain space; the third bar (reweighted features) for each model shows the RDM correlation of the weighted model features with the reference brain ROI. Stars above each bar show statistical significance obtained by signrank test at 5% significance level. Black horizontal bars show that the difference between the bars for a model is statistically significant (signrank test, 5% significance level).

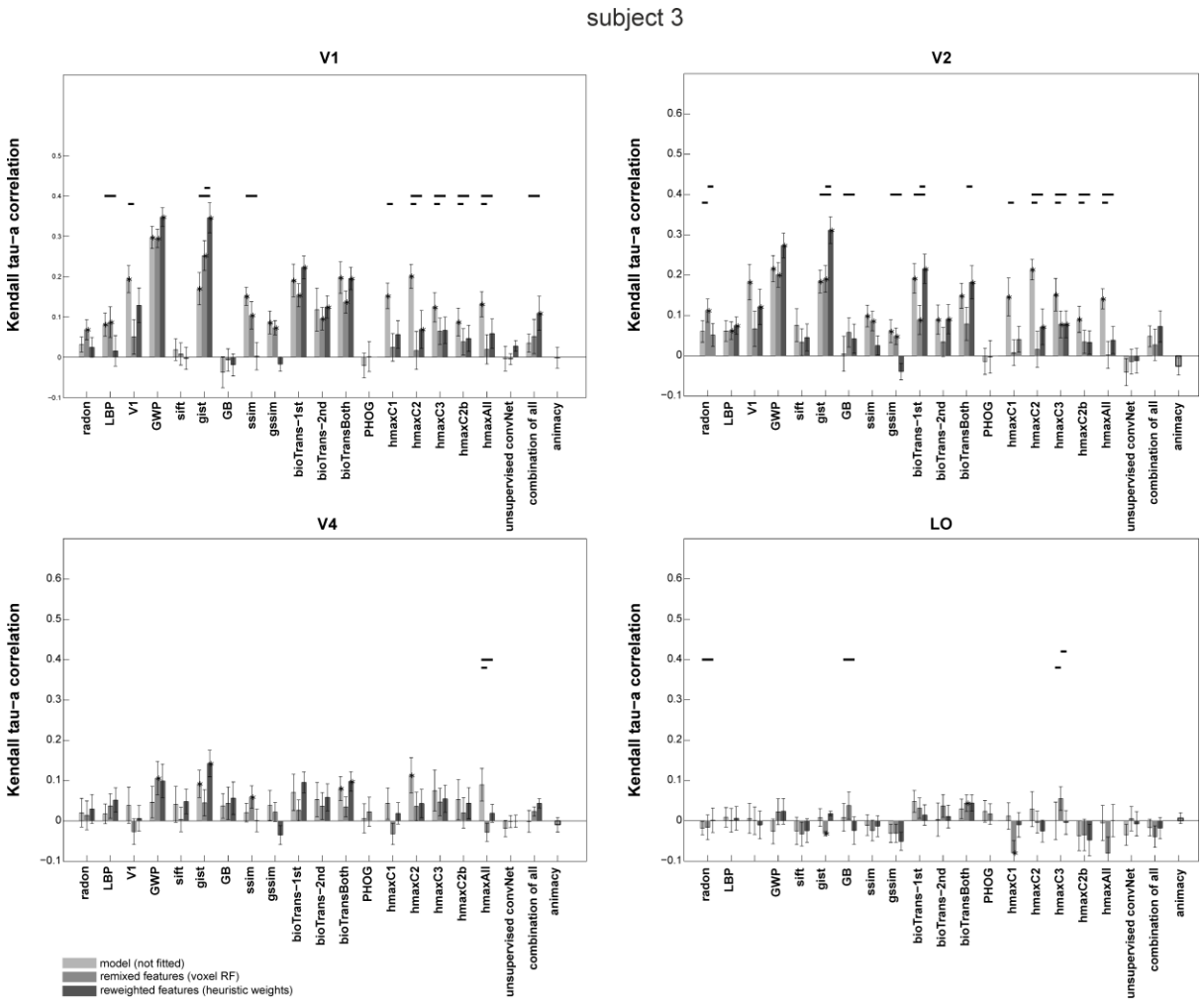

Figure S3. Kendall tau-a RDM correlation of the models with the brain ROIs of subject 3 across the hierarchy of visual areas. Bars show the average of ten 12x12 RDM correlations (120 test stimuli in total) with V1, V2, V4, and LO brain RDMs for subject 3. There are three bars for each model. The first bar, 'model (not fitted)', shows the RDM correlation of a model with a brain ROI without fitting the model responses to brain voxels. The second bar (voxel RF-fitted) shows the RDM correlations of a model that is fitted to the voxels of the reference brain ROI, using 1750 training images (refer to Figure 4 to see how the fitting is done). We used gradient descent to find the weight vectors (the mapping between a model feature vector and the voxels of a brain ROI). The weight vectors were then used to weight the model features, without mapping them to the brain space; the third bar (reweighted features) for each model shows the RDM correlation of the weighted model features with the reference brain ROI. Stars above each bar show statistical significance obtained by signrank test at 5% significance level. Black horizontal bars show that the difference between the bars for a model is statistically significant (signrank test, 5% significance level).

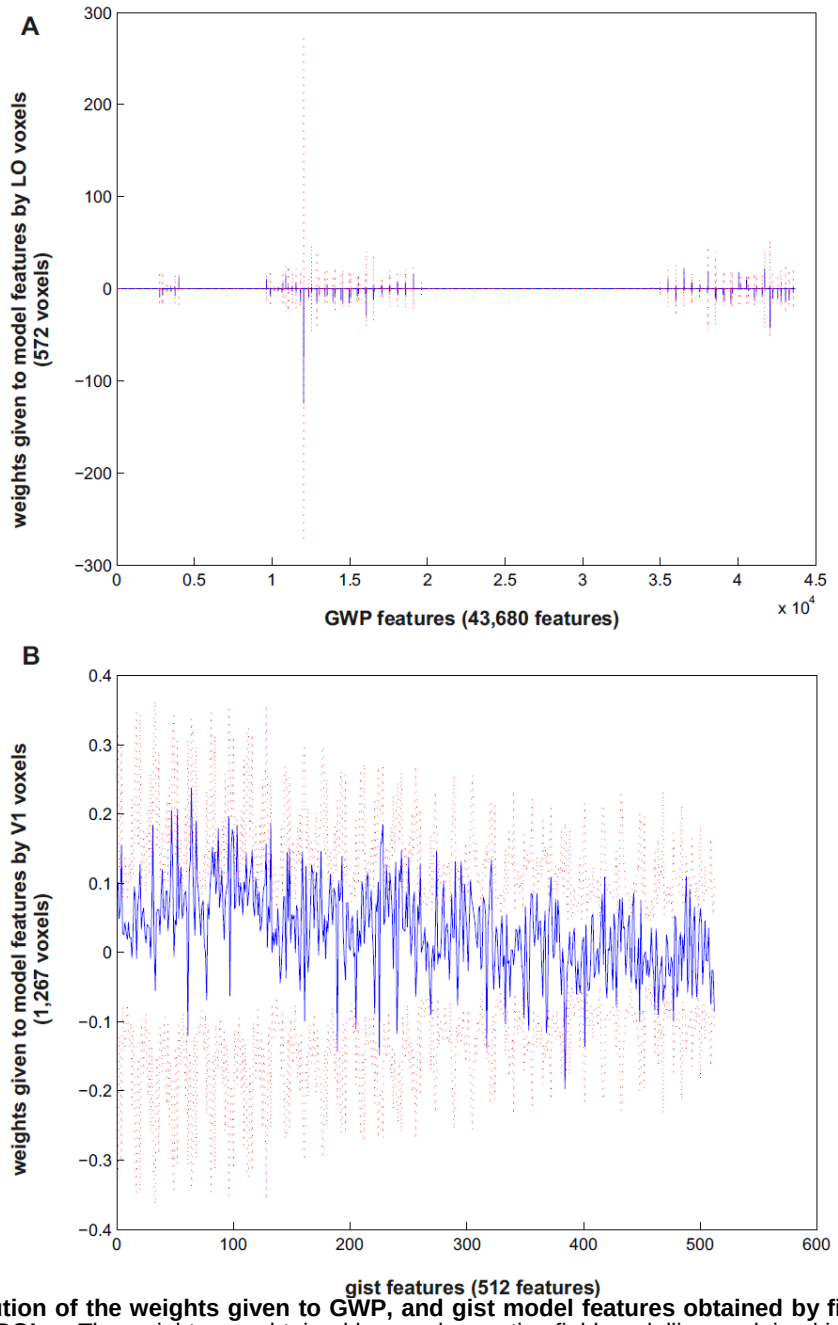

Figure S4. Distribution of the weights given to GWP, and gist model features obtained by fitting the features to voxels from brain ROIs. The weights are obtained by voxel receptive field modelling explained in Figure 4. Each voxel assigns a weight vector to GWP features; the weights map the model features to the voxel space. For each ROI, the plotted weight distribution is the average of weight vectors assigned to the voxels in that ROI. The GWP model features are fitted to LO voxels (A), and the gist features are fitted to V1 voxels (B). Figure 5 shows that only in these two occasions the fitted voxel receptive field model significantly outperforms the raw model representation (in terms of their RDM correlation with the reference RDM). (A) GWP model has a very high-dimensional features space (43,680 features), however the weight distribution for this model is sparse, suggesting that the brain prefers a sparse representation and only few number of GWP features are informative. On the hand, (B) the dimensionality for gist model is small (512 features), and the weight distribution is almost uniform, meaning that all features are equally important. The dotted red lines show the standard deviation of the mean.

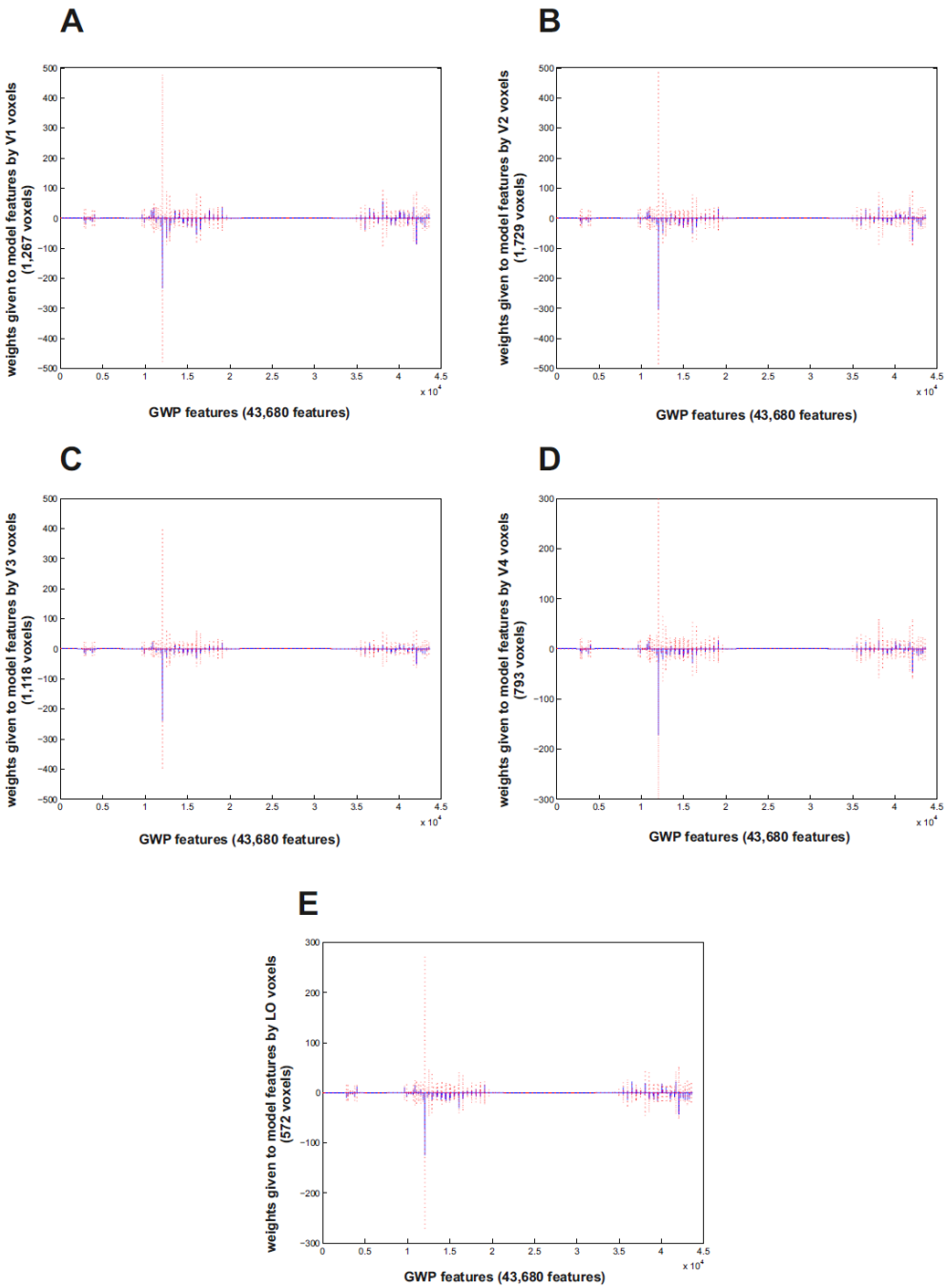

Figure S5. Distribution of the weights given to GWP model features obtained by fitting the features to voxels from the hierarchy of visual areas V1 (A), V2 (B), V3 (C), V4 (D), and LO (E). The weights are obtained by voxel receptive field modelling explained in Figure 4. Each voxel assigns a weight vector to GWP features; the weights map the model features to the voxel space. For each ROI, the plotted distribution of weights is the average of weight vectors assigned to the voxels in that ROI. The distribution is sparse, suggesting that only a small proportion of GWP features are informative. The sparse distribution of weights may also suggest a sparse representation for brain ROIs. The consistency of weight distributions across all ROIs, apparently means that all ROIs are emphasizing on the same set of features but with slightly different weights. These are the results for one of the subjects (subject 1), but the patterns are similar across all three subjects. The dotted red lines show the standard deviation of the mean.