Published in final edited form as:
Neuroimage. 2015 January 15; 105: 215–228. doi:10.1016/j.neuroimage.2014.10.018.

A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes

Thomas Naselaris[1], **Cheryl A. Olman**[2,3], **Dustin E. Stansbury**[4], **Kamil Ugurbil**[3], and **Jack L. Gallant**[4,5,6]

[1]Department of Neurosciences, Medical University of South Carolina, SC, USA

[2]Department of Psychology, University of Minnesota, MN, USA

[3]Center for Magnetic Resonance Research, University of Minnesota, MN, USA

[4]Vision Science Group, University of California, Berkeley, CA, USA

[5]Helen Wills Neuroscience Institute, University of California, Berkeley, CA, USA

[6]Department of Psychology, University of California, Berkeley, CA, USA

Abstract

Recent multi-voxel pattern classification (MVPC) studies have shown that in early visual cortex patterns of brain activity generated during mental imagery are similar to patterns of activity generated during perception. This finding implies that low-level visual features (e.g., space, spatial frequency, and orientation) are encoded during mental imagery. However, the specific hypothesis that low-level visual features are encoded during mental imagery is difficult to directly test using MVPC. The difficulty is especially acute when considering the representation of complex, multi-object scenes that can evoke multiple sources of variation that are distinct from low-level visual features. Therefore, we used a voxel-wise modeling and decoding approach to directly test the hypothesis that low-level visual features are encoded in activity generated during mental imagery of complex scenes. Using fMRI measurements of cortical activity evoked by viewing photographs, we constructed voxel-wise encoding models of tuning to low-level visual features. We also measured activity as subjects imagined previously memorized works of art. We then used the encoding models to determine if putative low-level visual features encoded in this activity could pick out the imagined artwork from among thousands of other randomly selected images. We show that mental images can be accurately identified in this way; moreover, mental image identification accuracy depends upon the degree of tuning to low-level visual features in the voxels selected for decoding. These results directly confirm the hypothesis that low-level visual

Correspondence should be addressed to: Thomas Naselaris Medical University of South Carolina 96 Jonathan Lucas St. CSB 325H Charleston, SC 29425 843-792-6263 tnaselar@musc.edu.

Author Contributions TN designed the experiment. TN and CAO collected the data. TN analyzed the data. TN wrote the paper. CAO, DS, JLG and KU edited the paper and advised on all aspects of the work.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

features are encoded during mental imagery of complex scenes. Our work also points to novel forms of brain-machine interaction: we provide a proof-of-concept demonstration of an internet image search guided by mental imagery.

Keywords

mental imagery; voxel-wise encoding models; decoding; fMRI; vision; perception

1. Introduction

Spend a few moments examining “Betty” (Figure 1A, second image from left), a famous portrait by the artist Gerhard Richter. With eyes closed generate a mental image of the painting and maintain it for a few seconds. With eyes again open re-examine the painting. Which of its basic features were conserved in your mental image? The position of Betty’s head within the center of the frame? The vertical orientation of her torso? The spatial frequencies induced by the strands of her hair, the folds of her sweatshirt, the floral print along her sleeve?

Low-level visual features such as position, orientation, and spatial frequency are among the fundamental building blocks of visual perception. During perception of an external image these features are encoded in the activity of early visual cortical areas (i.e., V1 and V2), and provide an efficient basis for representing complex natural scenes (Olshausen and Field, 1996). An important and long-standing question in mental imagery research is whether these same low-level visual features contribute to the representation of complex mental images (Pylyshyn, 2002; Kosslyn et al., 2009).

Most of the fMRI research on mental imagery has addressed a closely related but importantly different question, namely: are patterns of activity in early visual cortex generated during mental imagery *similar* to patterns of activity generated during perception? Between 1993 and 2010 at least twenty studies addressed this question by estimating the amplitude of BOLD activity in early visual areas in subjects engaged in mental imagery. At least eight studies reported no significant activity above baseline in early visual cortex during mental imagery (D’Esposito et al., 1997; Ishai et al., 2000; Knauff et al., 2000; Trojano et al., 2000; Wheeler et al., 2000; Formisano et al., 2002; Sack et al., 2002; Daselaar et al., 2010), while at least twelve reported attenuated but significant activity (Le Bihan et al., 1993; Sabbah et al., 1995; Goebel et al., 1998; Chen et al., 1998; Klein et al., 2000; O’Craven and Kanwisher, 2000; Ishai et al., 2002; Lambert et al., 2002; Ganis et al., 2004; Handy et al., 2004; Amedi et al., 2003; Cui et al., 2004). Recent evidence suggests that the discrepancy can be explained by differences in experimental factors (Kosslyn and Thompson, 2003) and variation in the vividness of mental imagery across individuals (Cui et al., 2007). Thus, it is safe to conclude that primary visual cortex is weakly but significantly activated by mental imagery.

In recent years at least three studies have used multivoxel pattern classification (MVPC) to measure the similarity between patterns of activity during imagery and perception in early visual cortex (Cichy et al., 2011; Lee et al., 2012; Albers et al., 2013). MVPC is a useful tool for studying mental imagery because it is sensitive to information that is encoded in multi-voxel patterns of activity even when the amplitude of the activity is attenuated. The recent MVPC studies have shown that patterns of activity generated during mental imagery in V1 and V2 are discriminable; specifically, pattern classifiers that accurately discriminate patterns of activity generated during perception of simple external stimuli can also discriminate patterns of activity generated during mental imagery of the same stimuli. MVPC studies that targeted high-order visual areas have shown similarity between activity patterns in those visual areas as well (Stokes et al., 2009; Reddy et al., 2010, Johnson and Johnson, 2014). Results from MVPC studies investigating visual working memory (Harrison and Tong, 2009; Xing et al., 2013), and dreaming (Horikawa et al., 2013) also support the notion that patterns of activity generated during mental imagery and perception are similar in some way.

The finding that patterns of activity in early visual cortex during imagery are similar to patterns of activity during perception implies--but does not directly demonstrate--that low-level visual features are represented in both imagery and perception. In fact, this specific hypothesis is difficult to test using MVPC (or activation analysis) because MVPC does not provide an explicit model of the many sources of variation that can contribute to activity patterns during imagery and perception. It is well-established that low-level visual features are *not the only* source of variation in activity in early visual areas. Additional sources of variation include attention (Kamitani and Tong, 2005), reward expectation (Serences, 2008), the perception of coherent shape (Murray et al., 2002), global context (Joo et al., 2012), and even auditory stimulation (Vetter et al., 2014). Any one of these distinct sources of variation could induce similarity between activity patterns generated during imagery and perception. To directly test hypotheses about low-level visual features it is therefore essential to isolate the specific component of variation in activity that is due to low-level visual features. This can be done experimentally using reduced stimuli that depict *only* low-level visual features (e.g., an oriented grating); however, perception and imagery of multi-object, complex scenes will inevitably tap sources of variation that cannot be experimentally controlled. Thus, when considering mental imagery of complex scenes (the class of stimuli that is clearly most relevant to mental imagery as it occurs in everyday cognition), it is important to adopt an approach that provides analytical control over the multiple sources of variation in activity patterns.

Here we use a voxel-wise modeling and decoding approach (Naselaris et al., 2011) to demonstrate directly that low-level visual features are encoded in activity generated during mental imagery of complex scenes recalled from memory. The voxel-wise encoding model characterizes tuning of activity in each voxel to retinotopic location, spatial frequency and orientation. Unlike activation and MVPC analyses, the encoding model approach isolates the specific component of variation in activity that is due to low-level visual features. This component is then used to identify the mental image associated with a measured pattern of activity. Consequently, decoding is directly linked to the representation of low-level visual features, and can only succeed if the features are encoded in the activity of the voxels selected to perform decoding. Indeed, our key result is that the accuracy of mental image decoding is directly dependent upon how well-tuned the underlying voxel activity is to low-level visual features during perception.

An additional advantage of our encoding model-based approach is that, unlike MVPC analysis, it is not constrained by a fixed set of categories or restricted to any specific stimulus class (e.g., gratings). Our approach thus opens opportunities for mental imagery to mediate interactions between brain and machine that are not conceivable with MVPC-based decoding. Here we show that our approach can be used to sort the images returned by an internet search query according to their low-level similarity to a specific mental image.

2. Results

We used fMRI to measure blood oxygenation-level dependent (BOLD) activity in the visual cortex of three subjects. Scanning was conducted at 7-Tesla in order to exploit the sensitivity and specificity gains provided by ultrahigh fields (Yacoub et al., 2001; Olman and Yacoub, 2011). Each scanning session consisted of interleaved *model-fitting, model-testing, perception,* and *imagery* runs (Figure 1B). During the model-fitting and model-testing runs subjects fixated a small square at the center of the visual field while passively viewing a sequence of briefly presented color photographs. During the imagery runs subjects were cued to generate mental images of five previously memorized works of art (Figure 1A) while fixating a small square at the center of a blank gray screen. A perception run preceded each imagery run and was identical to it except that subjects viewed the works of art instead of generating mental images of them.

Brain activity measured during the model-fitting runs was used to construct a voxel-wise encoding model for each voxel in the acquired functional volumes (Figure 1C, left). The encoding model is based upon a Gabor wavelet decomposition of input images and characterizes tuning to low-level visual features (Figure 1D). Previous studies (Kay et al., 2008; Naselaris et al., 2009; Nishimoto et al., 2011) have shown that Gabor-wavelet encoding models are able to predict activity in individual voxels evoked by arbitrary visual stimuli. As in these previous publications, we identified the voxels that were accurately characterized by the encoding model by calculating the correlation between model predictions and activity measured during the model-testing runs (images in the model-fitting and model-testing runs did not overlap and did not include the artwork in the perception and imagery runs). The distribution of correlation coefficients (referred to as *model prediction accuracy*) across all voxels has a single mode at 0 and a heavy tail (Figure 2A). The mode indicates that the model does not accurately predict activity for a large fraction of voxels, while the tail indicates that there is a subset of voxels for which model predictions are quite accurate (Naselaris et al., 2009). As expected, this subset of *well-tuned voxels* includes the voxels in early visual areas V1 and V2 (Figure 2A, inset) and is restricted to voxels that occupy cortical grey matter (pink voxels in Figure 2B). Voxels for which model predictions are extremely poor are scattered throughout white matter from the posterior to anterior boundaries of the slice prescription (blue voxels in Figure 2B).

The encoding models estimated for each voxel were used to perform external and mental image identification using the activity from the perception and imagery runs, respectively

(Figure 1C). Image identification is a decoding method that enables direct detection of putative low-level visual features encoded in voxel activity. With image identification, putative low-level features are detected by correlating measured brain activity against activity predicted by the voxel-wise encoding models. If the low-level visual features of a target image are encoded in measured brain activity, then measured brain activity should be more correlated with model predictions in response to the target image than to other randomly selected photographs. Via this logic, a single target image is *identified* by picking it out from a gallery of hundreds, thousands, or millions of other images (Kay et al., 2008) (Note that this method is radically different from MVPC, in which a single object category is discriminated from a small number of other categories using any aspect of the underlying activity that makes it discriminable). Of course, this method of decoding depends entirely upon the accuracy of the encoding models. If the encoding models do a poor job of predicting the activity of the underlying voxels, correlations between measured and predicted activity will be random and decoding will fail. Image identification thus provides a direct test of the hypothesis that the low-level visual features of remembered scenes are encoded in activity generated during mental imagery. If the hypothesis is true, remembered scenes should be accurately identified whenever the encoding models are accurate; identification accuracy should otherwise be poor.

To test the hypothesis, we segregated (Figure 2C) well-tuned voxels (whose activity was accurately predicted by the encoding model) from poorly-tuned voxels (whose activity was poorly predicted by the model). Specifically, we rank-ordered the voxels based on model prediction accuracy and then grouped them into populations of 1000 voxels based on the ranking. The ~1000 distinct *populations* constructed by this procedure varied smoothly from populations whose voxels were poorly tuned (e.g., voxels in white matter or in non-responsive cortex) to populations whose voxels were well-tuned (e.g., early visual cortex; Figure 2D). External and mental image identification was then performed independently for each population. The measure of identification accuracy for each population was a simple tally of *hits*: whenever *predicted* activity in response to the artwork was more correlated with *measured* activity than the predicted activity in response to a set of randomly selected photos, a *hit* was tallied. A thousand such comparisons were made using one thousand sets of randomly selected photos; perfect image identification would therefore correspond to one thousand hits, while chance would correspond to 500 (see Figure S1 for an illustration of the analysis).

Our key results are shown in Figure 3 (see also Figure S2). As expected, external images could be accurately identified using activity from the perception runs, and identification accuracy depended directly upon the accuracy of the underlying encoding models. Median image identification accuracy varied monotonically with the prediction accuracy of the voxel-wise encoding models (Figure 3, blue lines), indicating that activity sampled from populations of poorly-tuned voxels produced poor image identification, while activity sampled from populations of well-tuned voxels produced excellent image identification accuracy.

The same pattern of results was observed for mental images identified using activity from the imagery runs. Median mental image identification accuracy varied monotonically with the prediction accuracy of the voxel-wise encoding models (Figure 3, orange lines). For populations of well-tuned voxels mental image identification accuracy was well above chance (p < .01, permutation test) and in subjects 1 and 3 approached the accuracy of external image identification. Special voxel populations consisting exclusively of well-tuned voxels from V1 (Figure 3, triangles) or V2 (squares) were consistent with these findings. This result directly demonstrates that low-level visual features of remembered complex scenes are encoded in activity during mental imagery.

The fact that an encoding model for early visual areas can be used to accurately identify mental images suggests that current understanding of visual processing is already sufficient to exploit mental imagery for use in a brain-machine interface. As a proof-of-concept, we considered an image-search task in which a specific artwork must be selected from among a gallery of internet images associated with an artist's name. In this scenario, a user has observed an artwork by a known artist, but cannot remember the title of the work. In principle, the brain activity associated with a mental image of the artwork could be used to pick it out from a digital gallery of images returned by an internet query on the artist's name.

We performed a Google Images query on the names of each artist whose work was sampled in our experiment. For each artist 100 of the images returned by the query were downloaded and saved. The majority of images returned by the query were artwork by the artists, although images of book jackets, gallery photos, and unrelated miscellany were also returned. We then used the voxel-wise encoding models to predict the brain activity that would be evoked by viewing each of the 100 downloaded images. We rank-ordered the images according to how closely their predicted activity resembled the activity generated during mental imagery. This procedure consistently assigned high ranks to the imagined artwork (Figure 4B and 5B). For example, 90% of the time the imagined artwork was ranked in the top 50; 20% of the time it was ranked in the top 10 (Figure 4B). These rankings are lower than the rankings for the perceived artwork (Figure 4A and 5A) but are significantly higher than expected by chance (p < .0001; permutation test). A breakdown of results by individual image (Figure 4C and 4D) and subject (Figure 4E and 4F) show that accurate decoding is not driven by any single image or subject (although the pattern of variation in accuracy across artwork and subjects may reveal important information; see section 3.8). Inspection of the sorted galleries (Figure 5) suggests that obviously irrelevant images (Figure 5A and 5B, "worst" five images in the first 3 rows) were assigned low rankings. Images assigned high rankings typically had one or two structural features in common with the perceived/imagined artwork (e.g., the contour of the stained-glass window in the Richter block of 5A; the high-frequency textures in the El Greco blocks of 5A and 5B).

These results demonstrate the power of the encoding modeling approach for decoding mental imagery, and establish the feasibility of using brain activity driven by mental imagery to perform useful computational tasks.

3. Discussion

We have addressed the long-standing issue of whether basic elements of visual perception--what we have called low-level visual features--are conserved during mental imagery. We have directly demonstrated that they are conserved: low-level visual features in complex scenes (here, works of art) are encoded in brain activity generated during mental imagery. We were motivated to address this issue because of its historical obstinacy and its relevance to currently influential theories of perception. Three key innovations were deployed: a voxel-wise encoding model, the use of complex scenes as targets for mental imagery, and a voxel-selection procedure that depends upon tuning to low-level visual features instead of predetermined regions of interest. Below we discuss these motivations and innovations, address some potential confounds, and speculate on the future of brain-machine interfaces driven by mental imagery.

3.1 Mental imagery research in the past century

For most, mental imagery is a salient and obviously critical component of mental life that is experienced as an imprecise approximation to seeing. However, since the turn of the century influential philosophers and experimental psychologists have objected to this intuitive characterization of mental imagery, arguing that mental imagery is not a critical component of mental life, that it is unrelated to the phenomenon of seeing, and that it cannot be usefully described in visual terms. From early to mid-century, influential theories of mental imagery emphasized its “cognitive unimportance” (Thomas, 2014) and lack of clear functional role in reasoning or language (Thorndike, 1907; Sartre, 1936; Ryle, 1949; Wittgenstein, 1953). During the Behaviorist era of psychology mental imagery was largely ignored as a topic of research and was in fact argued to be non-existent (Watson, 1913). When interest in mental imagery resumed in the 1970s, experimental data suggesting that mental images, like external images, are representations of visual features (e.g., objects, edges, textures, and so on) distributed across visual space (Podgorny and Shepard, 1978; Kosslyn, 1978) were countered by “non-depictive” or “propositional” accounts of mental imagery (Pylyshyn, 1973). Such accounts held that the *apparent* similarity between perception and mental imagery and the subjective experience of visually inspecting mental images do not indicate a shared *depictive* format for visual perception and mental imagery. Debates about the depictiveness of mental imagery have dominated mental imagery research for the past three decades (see Pylyshyn, 2002 and Kosslyn et al., 2009 for reviews of the arguments).

The availability of fMRI, beginning in the 1990’s, has not by itself been sufficient to resolve any of the debates about mental imagery. As we argue below, the analytical techniques required to directly test hypotheses about the visual features putatively encoded in brain activity during mental imagery of complex scenes have only recently been developed. From this historical perspective, the significance of our work is that it provides the most direct demonstration to date that activity generated during visual perception and mental imagery of complex scenes encodes the same low-level visual features. The low-level, Gabor-like features encoded during mental imagery and perception are clearly depictive since they characterize visual features (e.g., spatial frequency) at specific regions of visual space and accurately characterize the visual mechanisms that operate during stimulus-driven perception (Kay et al., 2008; Naselaris et al., 2009). Our result thus provides a critical and until now missing piece of evidence in support of depictive theories and--more generally--of the intuitive characterization of mental imagery.

3.2 Mental imagery and predictive coding theory

The feedforward pathway into the early visual areas (EVA) is one of the most well-understood aspects of visual processing. Knowledge of the functional role of the *feedback* pathway into EVA is much less well-developed. Anatomical studies have revealed that EVA receives feedback projections from higher-order visual areas (Rockland and Pandya, 1979; Felleman and Van Essen, 1991; Markov et al., 2013a). The feedback projections are believed to be critical for visual processing (Angelluci et al., 2002; Bullier, 2006; Markov et al., 2013b), but their precise functional role is currently unknown. What, if anything, is represented by EVA when it is activated without retinal input?

One compelling answer has been advanced by the predictive coding theory of vision (Gregory, 1980; Rao and Ballard, 1999; Lee and Mumford, 2003; Yuille and Kersten, 2006; Bastos et al., 2012). According to this theory feedback corresponds to the outputs of an internal, hierarchical model of the visual environment (see Berkes et al., 2011 for a recent experimental test of this idea). The top nodes of the model hierarchy are equated with higher-order visual areas (Lee and Mumford, 2003). Feedback projections from these areas send representations of the objects in predicted, imagined or remembered scenes to EVA. When driven by this internal feedback, activity at any one location in EVA indicates that the visual features depicted by local receptive fields would *probably* be detected *if* the objects in the scene were presented to the retina (Albright, 2012). Thus a basic prediction of predictive coding theory is that, given an accurate model of the receptive fields in EVA, it should be possible to accurately decode the visual features associated with remembered scenes from activity that is driven entirely by internal feedback.

We have tested this basic prediction by using a voxel-wise encoding model for early visual areas to decode mental images of remembered scenes. Our results support predictive coding theory by confirming the existence of signals in early visual areas that are not driven by retinal input and encode visual features of objects that need not be present. Other critical components of predictive coding theory--such as the existence of signals that encode prediction errors--were not tested here. Nonetheless, our study suggests that mental imagery can be effectively exploited to further test predictive coding theory. Mental imagery is by definition internally generated and therefore provides an ideal point of leverage for experimentally manipulating the visual features encoded in internal feedback.

3.3 Advantages of the encoding model approach for studying mental imagery

Our study used an encoding model and image identification approach to study mental imagery. Previously this approach had only been applied to visual perception (Kay et al., 2008). In what follows we will refer to this approach as the *voxel-wise modeling and decoding* method, or VM. VM is the only method we are aware of that is capable of directly answering the specific question posed here. It is therefore important to discuss how VM differs from the more frequently-used method of MVPC.

MVPC trades on the (dis)similarity between patterns of activity measured during differing sensory or cognitive states. Under MVPC, activity patterns are classified by measuring their (distance from) similarity to a classification boundary. In the case of support vector classification the classification boundary is itself a measured activity pattern. In the case of linear discriminant analysis the classification boundary is a linear combination of measured activity patterns.

Important questions about mental imagery have been answered by using MVPC to effectively compare the similarity of activity patterns generated during imagery and other sensory or cognitive states. Two recent excellent studies provide cases in point. In Lee et al., 2012 MVPC was used to reveal a gradient of similarity between perception and mental imagery along the cortical visual processing hierarchy. Their results indicate that the relationship between perception and imagery varies along different stages of visual processing. In Albers et al., 2013 MVPC was used to reveal a fundamental similarity between activity generated during visual working memory and mental imagery. Their results serve as a critical starting point for consolidating the extensive and largely parallel literatures on working memory and mental imagery (Harrison and Tong, 2009; Tong, 2013).

Yet despite its obvious utility, MVPC remains a fundamentally inadequate tool for addressing the specific question posed here, namely: is the representation of low-level visual features conserved during mental imagery? MVPC is inadequate because it provides no means of decomposing patterns of brain activity into their distinct sources of variation. Even in V1, patterns of brain activity are composed of multiple distinct sources of variation. For example, activity in V1 in response to simple line elements can vary significantly depending upon whether the stimuli is perceived as a coherent object or a collection of independent lines (Murray et al., 2002). Because MVPC simply compares activity patterns to one another, it cannot discriminate between the specific contributions that top-down factors (such as the perception of a coherent object) and low-level visual features (such as the orientation of lines) make in determining patterns of activity. Thus, MVPC can reveal *that* two patterns of activity are similar, but reveals very little about *why*. This is the fundamental limitation of MVPC.

An extreme example of this limitation can be found in a recent study that elegantly exploited it to solve a difficult engineering problem. Sorger et al. (2012) designed an fMRI-based speller that can accurately and rapidly read-out letters from patterns of brain activity. The authors constructed an ingenious combinatorial code for letters by arbitrarily associating each letter in the alphabet (plus a blank space) with one of three cognitive tasks, onset delays, and task durations (for a total of 27 different states). They then used a pattern classifier to identify letters by discriminating activation patterns associated with the particular combination of task parameters assigned to each letter. Letter read-out was remarkably accurate and the study is unquestionably an important advance in the development of brain-driven spellers. Most germane to this discussion, however, is the fact that accurate decoding of letters was achieved in Sorger et al. by applying MVPC to patterns of activity that had nothing to do with the native representation of letters in the brain. For someone unaware of the combinatorial code intentionally embedded in the experimental design--in other words, someone unaware of the hidden sources of variation--it would be natural to incorrectly infer that the accurate decoding reported in this study revealed something about how the brain represents letters. The Sorger et al. study provides an excellent demonstration that pattern classifiers can produce accurate decoding while revealing nothing about how the decoded stimuli are represented in the brain.

The fundamental limitation of MVPC can be partially circumvented via experiments that reduce or tightly control the multiple sources of variation that contribute to patterns of brain activity. Most studies on mental imagery have in fact adopted this approach, using extremely reduced visual stimuli such as blobs (Thirion et al., 2006), wedges (Slotnick, 2005), or gratings (Albers et al., 2013) for their experiments. However, the use of reduced stimuli necessarily entails a reduction in the generality of the experimental results. In using MVPC one is therefore forced to accept a trade-off between complexity and generality. Since the goal of this study was to understand mental images generated under natural conditions, we were compelled to use another approach.

The VM method was used in our study because it avoids the fundamental limitation of MVPC without sacrificing generality. When decoding is successful using VM, it is very clear why. This is because VM trades on the similarity between measured patterns of activity and the *predictions of an encoding model*. The predictions of the encoding model represent a single, explicit source of variation due entirely to the visual features embedded in the model itself. In our case, these features were Gabor wavelets that were each specified by a retinotopic location, spatial frequency, and orientation. If these features are not encoded in the measured activity, model predictions will be meaningless and decoding will fail. Conversely, if decoding succeeds, it can only be because the features embedded in the model are encoded in the activity. VM thus provides a method for directly testing if a specific set of features is encoded in activity. Because encoding models are designed to be applied to stimuli of arbitrary complexity (e.g., natural scenes), this boost in inferential power (relative to MVPC) comes at no cost to generality.

The importance of using an encoding model to perform decoding was previously articulated in Thirion et al., 2006. In their landmark study, Thirion et al. designed a voxel-wise model of tuning to retinotopic location (i.e., a receptive field model) and then used it to decode mental images of high-contrast blobs in various domino configurations. Prior to the current work it was unclear whether the results of the Thirion et al. study would generalize to complex mental images generated by remembering a natural scene. As we argue below, the ability to investigate the representation of complex, naturalistic scenes is critically important for understanding mental imagery.

3.4 The importance of complex natural scenes for studying mental imagery

The perception of complex scenes engages multiple levels of visual processing. Mental imagery of complex scenes need not. One’s mental image of “Betty” *could* consist of only those features most essential to the painting’s appeal (e.g., the source and softness of its illumination, or the eyes of Betty’s unseen face), and omit (or sample very sparsely) the low-level features we have decoded in the current work. In fact, many distinct mental images of “Betty” could in principle be generated by randomly sampling *any* subset of features in the painting. Many of the mental images generated in this way would not be consistent with the hypothesis that low-level features are encoded during mental imagery. Thus, the use of complex scenes in our experiment provided a robust and rigorous test of our hypothesis.

Mental imagery of complex scenes is subjectively very similar to mental imagery that accompanies recall of long-term memories. Thus, the use of complex scenes as targets for mental imagery mimics the kind of mental imagery that occurs during everyday mental life. Since we have shown that low-level visual features play a role in mental imagery of complex scenes, we can infer that these features are likely to support memory recall as it occurs in the course of everyday mental life. It is unclear if mental imagery of reduced stimuli is comparable to mental imagery that occurs naturally. Thus, it would be difficult to infer if low-level visual features encoded during mental imagery of reduced stimuli play any broader role in cognition.

3.5 Selection by tuning, not area

All fMRI studies that deploy a decoding analysis must include some rational, objective procedure for selecting the voxels that are used to perform decoding. In this study we selected voxel populations based upon encoding model prediction accuracy. This procedure was a natural one in our case because it facilitated a straightforward test of our hypothesis. A more conventional choice would have been to simply select voxels from early visual areas (V1 and V2), but this was less appropriate for two reasons. First, our hypothesis concerned the encoding of specific visual *features*, not the engagement of specific visual *areas*. By ranking voxels according to their degree of tuning we were able to test if features were encoded in activity independent of the visual area from which activity was sampled. Second, our procedure for voxel selection afforded a more robust test of our hypothesis than selection based upon region-of-interest. This is because our procedure sampled image identification accuracy in multiple voxel populations. The importance of this resampling can be seen in Figure 3. For voxel populations that are *not* tuned to low-level visual features there is a broad distribution of image identification accuracy. The distribution is exactly what would be expected if decoding accuracy were sampled randomly from a uniform distribution, leading to the correct conclusion that activity in voxels that are not tuned encodes no information about low-level visual features. If only one voxel population representative of well- and poorly-tuned voxels, respectively, had been sampled the results could have easily led to spurious conclusions.

Although our decision to select by tuning instead of area was appropriate and well-motivated, the question of the relative contributions of distinct visual areas to the representation of mental imagery is extremely important. For example, the results of Lee et al. (2012) suggest that mental imagery and perception converge toward parity with ascension of the visual hierarchy. Using our analysis, the quality of tuning to low-level visual features is a stronger determinant of decoding accuracy than visual area. Therefore, to examine this issue with the VM approach, encoding models that capture the basic representation in intermediate visual areas will be needed. The Gabor-wavelet encoding model used here is appropriate for V1 and V2, but it is increasingly less appropriate for V3, V4, etc. Although predictive encoding models for intermediate visual areas are under development, to our knowledge none are currently mature enough to use for studying mental images of complex scenes.

3.6 Potential confounds

Of potential concern is the possibility that our encoding model inadvertently captured tuning to object category via the correlations between object category and low-level features that may occur in natural scenes. However, it has been shown repeatedly in previous studies (Naselaris et al., 2009; Cukur et al., 2013) that the encoding model used here correlates poorly with object categories in natural scenes; therefore, we can safely rule out the possibility that identification of mental images depends upon object-category tuning that is correlated with tuning to low-level features.

Another potential concern is the possibility of a circumstantial correlation between the three-letter cues and the artworks we used in the study. Specifically, the concern is that accurate mental image identification could be the result of an unexpected correlation between the purely visual signals evoked by viewing the three-letter cues and the predicted activity of the encoding model in response to the artwork. A control analysis presented in the Supplementary Materials (Figure S3) provides explicit evidence that this is not the case: the three-letter cues presented during the imagery runs could not be accurately identified using our encoding model. Thus, we can safely discount circumstantial correlation between the cues and the artwork as a potential confound.

A final concern is related to the fact that the analysis of mental image identification accuracy in Figure 3 employed a set of 1000 randomly selected photographs to compare against the artwork used in the perception and imagery runs. It is possible the activity measured during the imagery runs encodes just enough information about the imagined artwork to enable a coarse distinction between art and randomly selected photographs. However, the mental image identification analysis presented in Figures 4 and 5 employed a set of images consisting primarily of work by the very artists whose paintings and photographs were used in our experiments. Thus, it appears that early visual areas encode more about the details of specific mental images than a coarse distinction between art and random photographs.

3.7 Using mental images to drive machines

We have demonstrated that, in principle, activity generated during mental imagery could be used to sort the results of an internet image search. This demonstration establishes that current knowledge of visual perception (embodied in the encoding model) could be used to exploit mental imagery for brain-machine interfacing. Mental images are a rich source of information about internal subjective states; however, unlike like internal speech, mental images are not easily communicated to others using language. A brain-machine interface (BMI) that taps mental imagery could therefore enable a very useful new mode of communication. Development along these lines should include an investigation of the role that long-term memory plays in enabling accurate mental imagery. In this study, subjects were exposed to the imagined artwork during perception runs that occurred just minutes before the onset of imagery runs. It will be important to measure how the duration of the interval between presentation and imagination affects mental imagery decoding. Furthermore, although we consider fMRI in its current stage of development to be an excellent tool for establishing the basic science and algorithmic foundations for a BMI driven by mental imagery, the fast, portable brain imaging technology that will be needed to produce a practical BMI is a goal for future work.

3.8 Differences between mental imagery and perception

Our results show that decoding accuracy for mental imagery was more variable than decoding accuracy for perception (compare Figure 4C to 4D; compare Figure 4E to 4F). For mental images, decoding accuracy varied considerably across individual artworks (Figure 4D) and across individual subjects (Figure 4F). The relative increase in variation of decoding accuracy for mental imagery could be due to two potential sources. The first potential source is *noise*. During mental imagery there is no driving input from a reliable stimulus (a static image in our case) so it is not surprising that activity evoked by imagining a scene is less reliable than activity evoked by seeing it. The second potential source of difference between imagery and perception is *bias*. The memories upon which much of mental imagery is based are imperfect. It is possible that even a noise-free mental image of an artwork could differ from its perceptual counterpart if the mental image was in some way distorted--a missing object, a blurred texture, etc.

Our data suggest that noise does indeed contribute to variation in mental imagery decoding accuracy. As seen in Figure 4A-B, decoding of mental images is much more accurate for populations of 1000 well-tuned voxels than for populations of 100, while for perceived images there is little difference in decoding accuracy over this ten-fold increase in population size. This finding suggests that the features needed to identify most of the artworks are present in most populations of well-tuned voxels, but that during mental imagery the noise in the signal makes pooling over 10 times more voxels much more beneficial than during perception. Ultrahigh field fMRI--which accommodates high resolution scanning (and therefore larger numbers of voxels) without catastrophic loss of signal-to-noise--may therefore be invaluable for studying mental imagery.

Our data provide some evidence that bias is also a source of variation in mental image decoding accuracy. Figures 4C and 4D show that there is more variation in decoding accuracy across artworks in mental imagery than during perception. Although noise may explain this variation, there is some indication that mental images of some works of art may be more distorted (relative to the original) than others.

Supporting this interpretation is the fact that El Greco’s *View of Toledo* is significantly *less* likely to be ranked into the top 20% of images than would be expected by chance. This suggests that in our subject pool mental images of El Greco’s painting were consistently distorted with respect to the painting itself.

Figures 4E and 4F also show that there is more variation in decoding accuracy across subjects for mental imagery than for perception. This observation is consistent with previous evidence that subjects vary in the amount of V1 activation elicited by mental imagery (Cui et al., 2007). Again, this variation could be due entirely to noise; however, it may also be due to variation in the amount of distortion in the subjects’ mental images. An observation consistent with this interpretation is that for subject 2 the imagined artworks are ranked in the top 25% significantly *less* often than would be expected by chance. This suggests that one or more of this subject’s mental images are systematically distorted in ways that the other subjects’ are not. Additional experiments will be needed to confirm these intriguing observations.

4. Materials and Methods

4.1 Ethics statement

Experimental protocols were approved by the Institutional Review Board at the University of Minnesota. Subjects gave written informed consent prior to participation.

4.2 Subjects

Three healthy adult subjects with normal or corrected-to-normal vision participated in the experiment. Subject 2 was an author of the study. Participants gave written informed consent before taking part in the experiment. Prior to scanning sessions subjects were required to thoroughly inspect and commit to memory five works of art (Figure 1A). Subjects were also required to memorize a three-letter cue associated with each artwork.

4.3 Experimental design and stimuli

The experiment included four distinct types of *runs* (i.e., a set of contiguous trials during which BOLD activity is measured): model-fitting, model-testing, perception, and imagery. Data from the model-fitting (Figure 1B, left) and model-testing runs were used to estimate the parameters of voxel-wise encoding models (Figure 1C, left and Figure 1D). Data from model-testing runs were used to calculate the prediction accuracy of the encoding models. Data from the perception (Figure 1B, middle) and imagery (Figure 1B, right) runs were used for the image identification analyses.

During model-fitting runs 14° X 14° (400 X 400 pixels) color natural photographs were presented (for an example see Figure 1B, left). Subjects were instructed to fixate on a colored square (0.14°, 4 pixels) at the center of each photograph. The color of the fixation spot changed three times per second to ensure that it was visible regardless of the content of the photographs. Photographs were presented in successive 2 s trials (Figure 1B). During each trial a “dummy” cue (the string “000” centered on a grey screen and filling a .62° × 0.42° rectangle) was displayed for 600 ms, followed by a photograph presented for 1.4 s. The dummy cue was included to ensure that trials during the model-fitting runs would have the same temporal sequence as trials during the perception and imagery runs. During each 1.4 s presentation the photograph was flashed ON-OFF-ON-OFF-ON-OFF-ON where ON corresponds to presentation of the photograph for 0.2 s and OFF corresponds to presentation of a grey screen for 0.2 s. Each trial was followed by an inter-trial interval during which a grey screen of the same brightness as the cue screen was presented. Inter-trial interval duration was specified in units of TR (=2 s) and varied randomly from 1 to 1+*j* TR’s, where *j* was sampled from a Poisson distribution with a mean parameter of 0.7. Each run began with a 24 s presentation of grey screen, and ended with a 16 s presentation of grey screen. For each subject a total of 20 model-fitting runs was completed. Each of the model-fitting runs consisted of 72 photographs presented two times each. Eight model-testing runs were also completed. Model-testing runs had an identical temporal design, but consisted of 12 photographs presented 12 times each. The photographs presented during each model-fitting and model-testing run were randomly selected and mutually exclusive; thus, a total of 1536 unique photographs were presented during the 28 model-fitting/testing runs.

During perception runs 5 color works of art were presented (image and fixation square dimensions as above). The works of art were "*View of Toledo*" (c. 1600) by Doménikos Theotokópoulos (El Greco); "*Night Sleeper*" (1979) by Andrew Wyeth; "*Betty*" (1988) by Gerhard Richter; "*Ruhrtal*" (1989) by Andreas Gursky; "*Horse Bath*" (2004) by Odd Nerdrum. Each trial consisted of a cue displayed for 0.6 s, followed by a 1.4 s stimulus epoch. The cue for each artwork was a 3-letter abbreviation of the artists' name ("elg", "gur", "rch", "ner", and "wyt"; same dimensions and location as dummy cue above). A single artwork was presented during each perception epoch using the same ON-OFF sequence as above. Intertrial intervals were fixed at 4 s. Each artwork was presented 12 times during each perception run.

Imagery runs were identical to perception runs in every way except that during the stimulus epoch a slightly brightened grey screen (1.4% brighter than the baseline illumination level of the cue screen) was presented instead of the cued artwork. Subjects were instructed to mentally project the cued artwork onto the slightly brightened grey screen and to cease imagining it when the screen returned to the baseline illumination level. Perception runs were always immediately followed by imagery runs.

For subject 1 and subject 2, three perception and three imagery runs were collected. For subject 3, six perception and six imagery runs were collected.

Data from each subject were collected across 5 to 6 separate scanning sessions spanning approximately 2 months. For subject 1 and subject 2 the first two sessions consisted entirely of model-fitting / testing runs. For subject 3 model-fitting/testing and at least one perception and imagery run were collected during each session.

4.4 MRI parameters

MRI data were collected at the Center for Magnetic Resonance Research at the University of Minnesota (Minneapols, MN) using a 7-Tesla Siemens MR scanner and an open-faced 4-channel loop transmit / 9-channel receive array coil specifically designed for high resolution fMRI at 7T (Adriany et al., 2011). Data were acquired from 39 coronal slices that covered occipital cortex: slice thickness 1.5 mm, field-of-view 167 mm X 167 mm. For functional data, a T2*-weighted, single-shot, slice-interleaved, gradient-echo EPI pulse sequence was used: matrix size 108 X 108, TR 2 s, TE= 0.021 s, flip angle 65°, GRAPPA acceleration factor = 2. The nominal spatial resolution of the functional data was 1.5 mm X 1.5 mm X 1.5 mm.

4.5 Data Preprocessing

Functional volumes from each run were motion-corrected using the FSL MCFLIRT routine. A brain-mask was extracted from the first run of the first session using the FSL BET routine. All subsequent analysis were performed on masked volumes. Runs across all 5-6 sessions were aligned to the first run of the first session by step-wise application of the FSL FLIRT and FNIRT routines. For the FLIRT registration step the normalized correlation cost-function was used (--cost = ‘normcorr’). For the FNIRT registration step a warp resolution of 5 mm (--warpres = 5,5,5) and a 4-stage sub-sampling procedure (--subsamp = 4,2,1,1) were used. Inspection of the processed volumes revealed that nonlinear registration (i.e., FNIRT) using these parameters noticeably improved the quality of cross-session registration relative to application of linear registration (i.e., FLIRT) alone. Image transformation and deformation parameters obtained from the MCFLIRT, FLIRT and FNIRT applications were concatenated and applied to each volume using the FSL applyWarp routine. This motion correction and registration pipeline was implemented using the *nipype* Python library (Gorgolewski et al., 2011).

For each run and each individual voxel in the brain-masked volumes, BOLD activities were normalized to mean 0 and standard deviation 1 and then detrended using a 3rd-order polynomial.

Visual areas V1 and V2 were identified in separate retinotopic mapping experiments. Borders of retinotopic areas were defined using standard methods (Engel et al., 1997).

Encoding model—An encoding model was estimated for each individual voxel in the brain-masked functional volume. Let v_{it} be the (normalized and detrended) signal from voxel i at time t. The encoding model for this voxel is:

$$v_{it}=\sum_{\tau=0}^{\tau=10} h_{\tau}^{T} f\left(S_{t-\tau}\right)+\epsilon$$

Here $s_{t-\tau}$ is an image presented at time $t-\tau$ $f(s)$ is the image transformation that implements a basic model of low-level visual processing, the h_{τ}’s are each a vector of model parameters that indicate the sensitivity to a particular feature at τ timesteps (each timestep is 2 s = 1 TR) after image presentation (the T superscript indicates transposition), and $\in$is zero-mean Gaussian additive noise. The transformation $f(s)$ is a Gabor wavelet transform of the image followed by a compressive nonlinearity:

$$f\left(s\right)=log\left(\left|W^{T}s\right|+1\right)$$

where f is an F × 1 vector (F = 570, the number of wavelets used for the model), W denotes a matrix of complex Gabor wavelets, and || denotes an absolute value operation that removes phase sensitivity. W has as many rows as there are pixels in s, and each column contains a different Gabor wavelet; thus, W has dimensions 64^2 × 570 (images were presented at a resolution of 400 × 400 pixels but were downsampled to 64 × 64 pixels for this analysis). The features are the *log* of the magnitudes obtained after filtering the image by each wavelet. The *log* is applied because we have found that a compressive nonlinearity improves prediction accuracy.

The wavelets in W occur at five spatial frequencies: 2,4,8,16 and 32 cycles per field of view (FOV = 14°). Filters are positioned on a square grid that covers the FOV, with spacing determined separately for wavelets at each spatial frequency so that adjacent wavelets are separated by 3.5 standard deviations of the spatial Gaussian envelope. At each grid position, wavelets occur at orientations of 0° and 90°.

The model parameters $H = (h_{\tau=0},..., h_{\tau=10})$ were fit using ridge regression. Each component h_{τ} is an vector of weights so the total number of model parameters is 5,700. The regularization parameter was determined by testing 7 log-spaced values from 10 to 10,000,000. For each value of the regularization parameter the model parameters H were estimated for each voxel and then prediction accuracy was measured using a single held-out model-fitting run. For each voxel the model parameters H that yielded the highest prediction accuracy on this held-out run were retained for subsequent analysis. Model prediction accuracy is the Pearson's correlation between predicted activity and the measured activity on the model-testing runs.

4.6 Image identification analysis

Decoding of external and mental images was performed using the method of image identification (Kay et al., 2008) (see Figure S1A). In image identification, the activity of a population of voxels is used to pick out an observed or imagined image from a gallery of images that were not observed or imagined. Let V be a $T \times N$ matrix in which the columns are time-series of T BOLD activity measurements, and the rows, denotedV_t, indicate the activity measured in the N voxels at time-point t (these row-vectors are illustrated as dashed rectangles in Figure S1B). Let $S = (s_0,..., s_T)$ be a sequence of images, and $\hat{V}\left(S\right)$ be the $T \times N$ matrix of activities predicted by the encoding model for each voxel in response to the sequence S. We define the image identification *score* for the sequence S given the population V in a manner similar to (Kay et al., 2008):

$$\text{score}\left(S|\,V\right)=\sum_{t=0}^{t=T}\left\langle V_t,\,\hat{V}_t\left(S\right)\right\rangle$$

where $\langle\cdot\rangle$ denotes Pearson's correlation (see Figure S1 for an illustration). Thus, at each time-point, the vector of responses across all N voxels is correlated against the pattern of predicted responses across all N voxels. These correlations are summed over time to generate the score for a particular image sequence.

The image identification accuracy for a particular population V is calculated by comparing the score for the sequence of perceived/imagined artwork to 1000 randomly selected sequences. Let S^{art} be the sequence of artwork perceived/imagined during perception/imagery runs. In this sequence s_t is set to a blank gray screen during inter-stimulus intervals. The cue frames and slightly brighter gray frames presented during the imagery runs are ignored. Let s_m^{rand} be the m^{th} of 1000 sequences of images constructed by substituting each of the 5 works of art utilized in the perception/imagery runs with 5 randomly selected images (blank gray screens presented during inter-stimulus intervals are left in place). We quantified image identification accuracy for the voxel population V as the number of *hits*:

$$\text{hits}\left(V\right)=\sum_{m=1}^{m=1000}u\left(m\right)$$

where

$$u\left(m\right)=\begin{cases}1, & \text{score}\left(S^{\text{art}}|V\right)>\text{score}\left(S_m^{\text{rand}}|V\right)\\ 0, & \text{otherwise}\end{cases}$$

For each subject we measured image identification accuracy for ~1000 separate voxel populations (dots, squares and triangles in Figures 3 and S2). Voxel populations were constructed according to the following procedure: (1) All voxels within the brain-masked functional volume of a single subject were rank-ordered by their model prediction accuracy (as calculated using the model-testing runs). The lowest ranked voxel had the lowest prediction accuracy while highest ranked voxel had the highest prediction accuracy. (2) The rank-ordered voxels were binned into overlapping groups of 3000. Each group of 3000 contained the 1500 highest-ranked voxels of the group below it and the 1500 lowest-ranked voxels of the group above it. (3) For each group of 3000 voxels, *R* populations were constructed. Each population consisted of 1000 voxels sampled randomly (without replacement) from the group. For each subject the value of *R* was chosen so that the total number of populations for the subject was ~1000. For subject 1 *R*=9 and the total number of populations was 1116. For subject 2 *R*=9 and the total number of populations was 1017. For subject 3 *R*=12 and the total number of populations was 1008. For each subject, two special populations consisting of the 1000 voxels in V1 with the highest model prediction accuracy (Figure 3 and Figure S2, triangles) and the 1000 voxels in V2 with highest model prediction accuracy (Figure 3 and Figure S2, squares) were also constructed.

In Figure 3 and Figure S2 we investigate the relationship between the lower bound on model prediction accuracy and image identification accuracy. The lower bound on prediction accuracy for a specific voxel population is the model prediction accuracy for the lowest-ranked voxel within the population. Let $\{V\}_x$ be the set of voxel populations with lower-bound ≥x (indicated on the x-axis of Figure 3). The median curves in Figure 3 and Figure S2 show the median image identification accuracy (i.e., median number of hits) of *all* populations in $\{V\}_x$ as x varies in linearly spaced increments of 0.02. Thus, the leftmost point on the median curves shows the median identification accuracy across each of the ~1000 populations; the rightmost point shows the median accuracy for only the subset of populations with the highest model prediction accuracy. Because the median is taken across fewer populations as the lower-bound increases, the threshold for chance performance ($p <$ 0.01; gray shading in Figure 3 and S2) increases monotonically. Note that because Figure S2 pools voxel populations from all three subjects lower bounds were divided by the maximum lower bound for each subject.

4.7 Hypothesis testing

We tested the hypothesis that median image identification accuracy for a particular lower bound is greater than chance. Chance, in this case, corresponds to drawing *n* numbers from a uniform distribution over the integers between 1 and 1000, and then taking the median. Here, *n* is the number of voxel populations with a given lower-bound on prediction accuracy (as specified by the *x*-axis in Figure 3 and Figure S2). In the analysis of Figure 3, *n* decreases as we move from left to right on the *x*-axis, and this must obviously be taken into account when computing significance. Thus, the hypothesis testing procedure was implemented as follows: Let $\mathrm{randscores}{=}\left(\mathrm{score}\left(S_{\mathit{1}}^{\mathrm{rand}}|\,V_k\right),\ldots,\mathrm{score}\left(S_{\mathit{1000}}^{\mathrm{rand}}|\,V_k\right)\right)$ be the set of scores assigned to 1000 random image sequences for population $V_k \in \{V\}_x$. For each population V_k in $score(S_{art}|V_k)$ is swapped with a randomly selected score from *randscores*. Then $hits(V_k)$ is recalculated for each population V_k and the median image identification accuracy across all populations in $\{V\}_k$ is determined. This procedure was performed 10,000 times for each $\{V\}_k$ resulting in a distribution of median hits for each value of the lower bound *x*. The 99*th* percentile of this distribution (corresponding to p <0 .01) is indicated by gray shading in Figure 3 and S2. This value grows as the lower bound increases because the number of populations in $\{V\}_k$ shrinks as the lower bound *x* increases.

4.8 Sorting of art galleries

We consider a scenario in which brain activity generated by imagining a specific artwork is used to sort the images returned by an internet query on the artist's name. The procedure for sorting the images was similar to the image identification analysis described in section 4.6. In this case the imagined artwork for each artist was substituted with a different artwork by the same artist (instead of a randomly selected image). The details of the sorting procedure were as follows. A *Google Images* query was performed on the name of the artist of each of the five perceived/imagined works of art used in our experiment. The first 100 images returned from each query were downloaded and saved. The majority of images returned by the queries were works by the artist, although book jackets, gallery photos, text, photographs of the artist, and miscellaneous images were sometimes returned. Let S_m^{gur} be a sequence of images formed by substituting into S^{art} the m^{th} image returned for the query "Andreas Gursky" (the other four works of art remain in place). For a specific voxel population *V*, $\mathrm{score}\left(s_m^{gur}|\,V\right)$ was calculated for each of the $m \in [1,100]$ substitutions. These scores, along with $socre(S^{art}8V)$, were sorted from lowest to highest and the resulting rank of $socre(S^{art}8V)$ was retained. This procedure was performed independently for each artist using 100 voxel populations of size 100, 1000, and 10,000 (for a total of 300 populations per subject). Voxel populations were constructed by random sampling from the group of 30,000 voxels with the highest model prediction accuracies. For each population size the ranks obtained for S^{art} were combined across all populations, artists, and subjects to generate the cumulative density plots in Figure 4A and 4B. Chance performance on this image-sorting task (grey curves in Figure 4A and 4B) was computed by generating 10,000 cumulative histograms obtained by random permutation of rankings for each artwork, voxel population, and subject. In Figure 4C and 4D, data for all subjects were pooled, the number of voxels per population was fixed at 1000, and rankings were calculated for each artwork independently. In Figure 4E and 4F, data for all artworks were pooled, the number of voxels per population was fixed at 1000, and rankings were calculated for each subject independently.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by P30 NS076408 and P41 EB015894 to KU, NEI R01-EY019684 to JLG, and NEI R01-EY023384 to TN.

References

Adriany G, Waks M, Tramm B, Schillak S, Yacoub E, de Martino F, Van de Moortele P, Naselaris T, Olman C, Vaughan T, Ugurbil K. An Open Faced 4 ch. Loop Transmit / 16 ch. Receive Array Coil for HiRes fMRI at 7 Tesla. International Society for Magnetic Resonance in Medicine. 2011

Albers AM, Kok P, Toni I, Dijkerman HC, de Lange FP. Shared representations for working memory and mental imagery in early visual cortex. Curr Biol. 2013; 23:1427–1431. [PubMed: 23871239]

Albright TD. On the perception of probable things: neural substrates of associative memory, imagery, and perception. Neuron. 2012; 74:227–245. [PubMed: 22542178]

Amedi A, Malach R, Pascual-Leone A. Negative BOLD differentiates visual imagery and perception. Neuron. 2005; 48:859–872. [PubMed: 16337922]

Angelucci A, Levitt JB, Walton EJ, Hupe JM, Bullier J, Lund JS. Circuits for local and global signal integration in primary visual cortex. J Neurosci. 2002; 22:8633–8646. [PubMed: 12351737]

Bastos AM, Usrey WM, Adams RA, Mangun GR, Fries P, Friston KJ. Canonical microcircuits for predictive coding. Neuron. 2012; 76:695–711. [PubMed: 23177956]

Berkes P, Orbán G, Lengyel M, Fiser J. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. Science. 2011; 331:83–87. [PubMed: 21212356]

Bullier, J. What is fed back?. In: Sejnowski, T.; Hemmen, J., editors. Twenty-three problems in systems neuroscience. Oxford University Press; 2006.

Chen W, Kato T, Zhu XH, Ogawa S, Tank DW, Ugurbil K. Human primary visual cortex and lateral geniculate nucleus activation during visual imagery. Neuroreport. 1998; 9:3669–3674. [PubMed: 9858377]

Cichy RM, Heinzle J, Haynes JD. Imagery and perception share cortical representations of content and location. Cereb Cortex. 2011; 22:372–380. [PubMed: 21666128]

Cui X, Jeter CB, Yang D, Montague PR, Eagleman DM. Vividness of mental imagery: individual variability can be measured objectively. Vision Res. 2007; 47:474–478. [PubMed: 17239915]

Çukur T, Huth AG, Nishimoto S, Gallant JL. Functional subdomains within human FFA. J Neurosci. 2013; 33:16748–16766. [PubMed: 24133276]

Daselaar SM, Porat Y, Huijbers W, Pennartz CMA. Modality-specific and modality-independent components of the human imagery system. NeuroImage. 2010; 52:677–685. [PubMed: 20420931]

D’Esposito M, Detre JA, Aguirre GK, Stallcup M, Alsop DC, Tippet LJ, Farah MJ. A functional MRI study of mental image generation. Neuropsychologia. 1997; 35:725–730. [PubMed: 9153035]

Engel SA, Glover GH, Wandell BA. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. Cereb Cortex. 1997:181–192. [PubMed: 9087826]

Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate. Cereb Cortex. 1991; 1:1–47. [PubMed: 1822724]

Formisano E, Linden DEJ, Di Salle F, Trojano L, Esposito F, Sack AT, et al. Tracking the mind’s image in the brain I: time-resolved fMRI during visuospatial mental imagery. Neuron. 2002; 35:185–194. [PubMed: 12123618]

Ganis G, Thompson WL, Kosslyn SM. Brain areas underlying visual mental imagery and visual perception: an fMRI study. Cognitive Brain Res. 2004; 20:226–241.

Goebel R, Khorram- Sefat D, Muckl L, Hacker H, Singer W. The constructive nature of vision: direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. Eur J Neurosci. 1998; 10:1563–1573. [PubMed: 9751129]

Gorgolewski K, Burns CD, Madison C, Clark D, Halchenko YO, Waskom ML, Ghosh SS. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Front. Neuroimform. 2011; 5:13. Available: http://www.frontiersin.org/Neuroinformatics/10.3389/fninf.2011.00013/full.

Gregory RL. Perceptions as hypotheses. Philos T R Soc B. 1980; 290:181–197.

Handy TC, Miller M, Schott B, Shroff N, Janata P, Van Horn J, Inati S, Grafton S, Gazzaniga M. Visual imagery and memory: do retrieval strategies affect what the mind's eye sees? Eur J Cogn Psychol. 2004; 16:631–652.

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. Nature. 2009; 458:632–635. [PubMed: 19225460]

Horikawa T, Tamaki M, Miyawaki Y, Kamitani Y. Neural decoding of visual imagery during sleep. Science. 2013; 340:639–642. [PubMed: 23558170]

Ishai A, Ungerleider LG, Haxby JV. Distributed neural systems for the generation of visual images. Neuron. 2000; 28:979–990. [PubMed: 11163281]

Ishai A, Haxby JV, Ungerleider LG. Visual imagery of famous faces: effects of memory and attention revealed by fMRI. NeuroImage. 2002; 17:1729–1741. [PubMed: 12498747]

Johnson MR, Johnson MK. Decoding individual natural scene representations during perception and imagery. Frontiers in Human Neuroscience. 2014:8. [PubMed: 24478674]

Joo SJ, Boynton GM, Murray SO. Long-range, pattern-dependent contextual effects in early human visual cortex. Curr Biol. 2012; 22:781–786. [PubMed: 22503498]

Kamitani Y, Tong F. Decoding the visual and subjective contents of the human brain. Nature Neurosci. 2005; 8:679–685. [PubMed: 15852014]

Kay KN, Naselaris T, Prenger RJ, Gallant JL. Identifying natural images from human brain activity. Nature. 2008; 452:352–355. [PubMed: 18322462]

Klein I, Paradis AL, Poline JB, Kosslyn SM, Le Bihan D. Transient activity in the human calcarine cortex during visual-mental imagery: an event-related fMRI study. J Cogn Neurosci. 2000; 12:15–23. [PubMed: 11506644]

Knauff M, Kassubek J, Mulack T, Greenlee MW. Cortical activation evoked by visual mental imagery as measured by fMRI. Neuroreport. 2000; 11:3957–3962. [PubMed: 11192609]

Kosslyn SM, Ball TM, Reiser BJ. Visual images preserve metric spatial information: evidence from studies of image scanning. Journal of experimental psychology. Human perception and performance. 1978; 4:47–60. [PubMed: 627850]

Kosslyn SM, Thompson WL. When is early visual cortex activated during visual mental imagery. Psychol Bull. 2003; 129:723–746. [PubMed: 12956541]

Kosslyn, SM.; Thompson, W.; Ganis, G. The case for mental imagery. Oxford University Press; Oxford: 2009. p. 248

Lambert S, Sampaio E, Scheiber C, Mauss Y. Neural substrates of animal mental imagery: calcarine sulcus and dorsal pathway involvement--an fMRI study. Brain Research. 2002; 924:176–183. [PubMed: 11750903]

Le Bihan D, Turner R, Zeffiro TA, Cuénod CA, Jezzard P, Bonnerot V. Activation of human primary visual cortex during visual recall: a magnetic resonance imaging study. PNAS. 1993; 90:11802–11805. [PubMed: 8265629]

Lee SH, Kravitz DJ, Baker CI. Disentangling visual imagery and perception of real-world objects. NeuroImage. 2012; 59:4064–4073. [PubMed: 22040738]

Lee TS, Mumford D. Hierarchical bayesian inference in the visual cortex. J Opt Soc AmA. 2003; 20:1434–1448.

Markov NT, Vezoli J, Chameau P, Falchier A, Quilodran R, Huissoud C, Lamy C, et al. The anatomy of hierarchy: feedforward and feedback pathways in macaque visual cortex. J Comp Neurol. 2013a; 522 In press.

Markov NT, Ercsey-Ravasz M, Van Essen DC, Knoblauch K, Toroczkai Z, Kennedy H. Cortical high-density counterstream architectures. Science. 2013b; 342 In press.

Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL. Shape perception reduces activity in human primary visual cortex. PNAS. 2002; 99:15164–15169. [PubMed: 12417754]

Naselaris T, Prenger RJ, Kay KN, Oliver M, Gallant JL. Bayesian reconstruction of natural images from human brain activity. Neuron. 2009; 63:902–915. [PubMed: 19778517]

Naselaris T, Kay KN, Nishimoto S, Gallant JL. Encoding and decoding in fMRI. NeuroImage. 2011; 56:400–410. [PubMed: 20691790]

Nishimoto S, Vu AT, Naselaris T, Benjamini Y, Yu B, Gallant JL. Reconstructing visual experiences from brain activity evoked by natural movies. Curr Biol. 2011; 21:1641–1646. [PubMed: 21945275]

O'Craven KM, Kanwisher N. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. Jour Cogn Neurosci. 2000; 12:1013–1023. [PubMed: 11177421]

Olman CA, Yacoub E. High-field FMRI for human applications: an overview of spatial resolution and signal specificity. Open Neuroimag J. 2011; 5:74–89. [PubMed: 22216080]

Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996; 381:607–609. [PubMed: 8637596]

Podgorny P, Shepard RN. Functional representations common to visual perception and imagination. Journal of experimental psychology. Human perception and performance. 1978; 4:21–35. [PubMed: 627848]

Pylyshyn ZW. What the mind's eye tells the mind's brain: A critique of mental imagery. Psychological Bulletin. 1973; 80:1–24.

Pylyshyn ZW. Mental imagery: in search of a theory. Behav Brain Sci. 2002; 25:157–237. [PubMed: 12744144]

Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999; 2:79–87. [PubMed: 10195184]

Reddy L, Tsuchiya N, Serre T. Reading the mind's eye: Decoding category information during mental imagery. NeuroImage. 2010; 50:818–825. [PubMed: 20004247]

Rockland KS, Pandya DN. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. Brain Res. 1979; 179:3–20. [PubMed: 116716]

Ryle, G. The Concept of Mind. Hutchinson; London: 1949.

Sabbah P, Simond G, Levrier O, Habib M, Trabaud V, Murayama N, et al. Functional magnetic resonance imaging at 1.5 T during sensorimotor and cognitive task. Eur Neurol. 1995; 35:131–136. [PubMed: 7628491]

Sack AT, Sperling JM, Prvulovic D, Formisano E, Goebel R, Di Salle F, et al. Tracking the mind's image in the brain II: transcranial magnetic stimulation reveals parietal asymmetry in visuospatial imagery. Neuron. 2002; 35:195–204. [PubMed: 12123619]

Sartre, JP. Imagination: A Psychological Critique. Williams, F., translator. University of Michigan Press; Ann Arbor: 1936.

Serences JT. Value-based modulations in human visual cortex. Neuron. 2008; 60:1169–1181. [PubMed: 19109919]

Slotnick SD, Thompson WL, Kosslyn SM. Visual mental imagery induces retinotopically organized activation of early visual areas. Cereb Cortex. 2005; 15:1570–1583. [PubMed: 15689519]

Sorger B, Reithler J, Dahmen B, Goebel R. A real-time fMRI-based spelling device immediately enabling robust motor-independent communication. Curr Biol. 2012; 22:1333–1338. [PubMed: 22748322]

Stokes M, Thompson R, Cusack R, Duncan J. Top-down activation of shape-specific population codes in visual cortex during mental imagery. J Neurosci. 2009; 29:1565–1572. [PubMed: 19193903]

Thirion B, Duchesnay E, Hubbard E, Dubois J, Poline JB, Le Bihan D, Dehaene S. Inverse retinotopy: Inferring the visual content of images from brain activation patterns. Neuroimage. 2006; 33:1104–1116. [PubMed: 17029988]

Thomas, NJT. Zalta, EN., editor. Mental imagery. The Stanford Encyclopedia of Philosophy. 2014 Edition2014 Spring. URL = <http://plato.stanford.edu/archives/spr2014/entries/mental-imagery/>

Thorndike EL. On the function of mental imagery. Journal of Philosophy, Psychology and Scientific Methods. 1907; 4:324–327.

Tong F. Imagery and visual working memory: one and the same? Trends in cognitive sciences. 2013; 17:489–490. [PubMed: 23958465]

Trojano L, Grossi D, Linden DEJ, Formisano E, Hacker H, Zanella FE, et al. Matching two imagined clocks: the functional anatomy of spatial analysis in the absence of visual stimulation. Cereb Cortex. 2000; 10:473–481. [PubMed: 10847597]

Vetter P, Smith FW, Muckli L. Decoding sound and imagery content in early visual cortex. Curr Biol. 2014; 11:1256–62. [PubMed: 24856208]

Watson JB. Psychology as the behaviorist views it. Psychological Review. 1913; 20:158–177.

Wheeler ME, Petersen SE, Buckner RL. Memory's echo: vivid remembering reactivates sensory-specific cortex. PNAS. 2000; 97(20):11125–11129. [PubMed: 11005879]

Wittgenstein, L. Philosophical Investigations. Anscombe, GEM., translator. Blackwell; Oxford: 1953.

Xing Y, Ledgeway T, McGraw PV, Schluppeck D. Decoding working memory of stimulus contrast in early visual cortex. J Neurosci. 2013; 33:10301–10311. [PubMed: 23785144]

Yacoub E, Shmuel A, Pfeuffer J, Van De Moortele PF, Adriany G, Andersen P, Vaughan JT, Merkle H, Ugurbil K, Hu X. Imaging brain function in humans at 7 Tesla. Magn Reson Med. 2001; 45:588–594. [PubMed: 11283986]

Yuille A, Kersten D. Vision as bayesian inference: analysis by synthesis? Trends Cogn Sci. 2006; 10:301–308. [PubMed: 16784882]

Highlights

A model of representation in early visual cortex decodes mental images of complex scenes.

Mental imagery depends directly upon the encoding of low-level visual features.

Low-level visual features of mental images are encoded by activity in early visual cortex.

Depictive theories of mental imagery are strongly supported by our results.

Brain activity evoked by mental imagery can be used to guide internet image search.

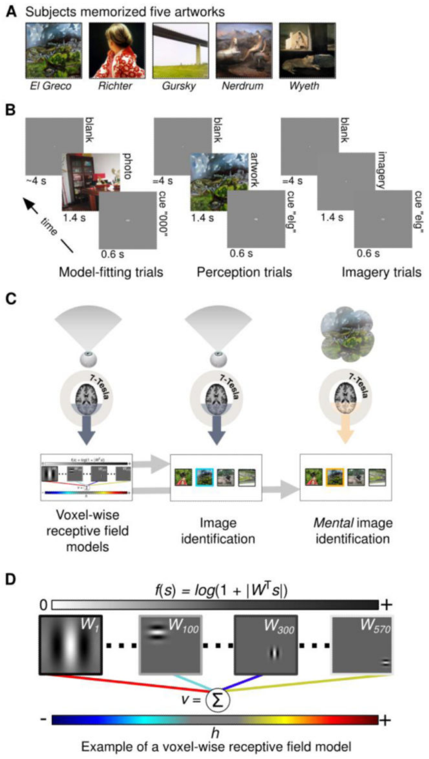

Figure 1. Experimental design

A) Prior to scanning subjects familiarized themselves with five works of art. **B**) Scans were organized into separate runs of contiguous trials. During each trial of the *model-fitting* (left) *and model-testing* (not shown) runs subjects fixated a central dot while viewing randomly selected photographs (duration of presentation = 1.4 s). Each photograph was preceded by a brief dummy cue (“000”; duration = 0.6 s) and followed by a blank gray screen (mean duration = 4 s). During each trial of the *perception* runs (middle) subjects viewed only the five works of art. Each artwork was preceded by a distinct 3-letter cue (an abbreviation of the artist’s name) and followed by a blank gray screen (duration = 4 s). *Imagery* runs (right) were identical to perception runs except that subjects imagined the five works of art while fixating at the center of a gray screen that was 1.4% brighter than the cue screen. **C**) High-field (7-Tesla) fMRI measurements of BOLD activity in the occipital lobe were obtained during each run. Activity measured during the model-fitting runs (left) was used to construct voxel-wise encoding models (bottom left). The voxel-wise encoding models and activity from the perception runs (middle) were used to perform image identification (bottom middle). The same voxel-wise encoding models and activity from the *imagery* runs (right) were used to perform *mental* image identification (bottom right). During image identification a target image (outlined in blue/orange for the perception/imagery runs) is picked out from among a set of randomly selected images. **D**) A simplified illustration of the voxel-wise encoding model. To produce predicted activity an observed or imagined scene ($\boldsymbol{s}$) is filtered through a bank of 570 complex Gabor wavelets (represented by the matrix W). Each wavelet is specified by a particular spatial frequency, spatial location, and orientation (four examples shown here). The filter outputs ($|W^{\mathrm{T}}\boldsymbol{s}|$, where | | denotes an absolute value operation that removes phase information) are passed through a compressive nonlinearity

($f(\mathbf{s}) = log(1+|W^{\mathrm{T}}\boldsymbol{s}|)$; outputs are represented by the grayscale bar at top) and then multiplied by a set of model parameters (colored lines; dark to light blue lines indicate negative parameters; yellow to red lines indicate positive parameters). The sum of the weighted filter outputs is the voxel's predicted activity (v) in response to the stimulus. The model parameters ($\boldsymbol{h}$) are learned from the training data only and characterize each voxel's tuning to spatial frequency, spatial location, and orientation.

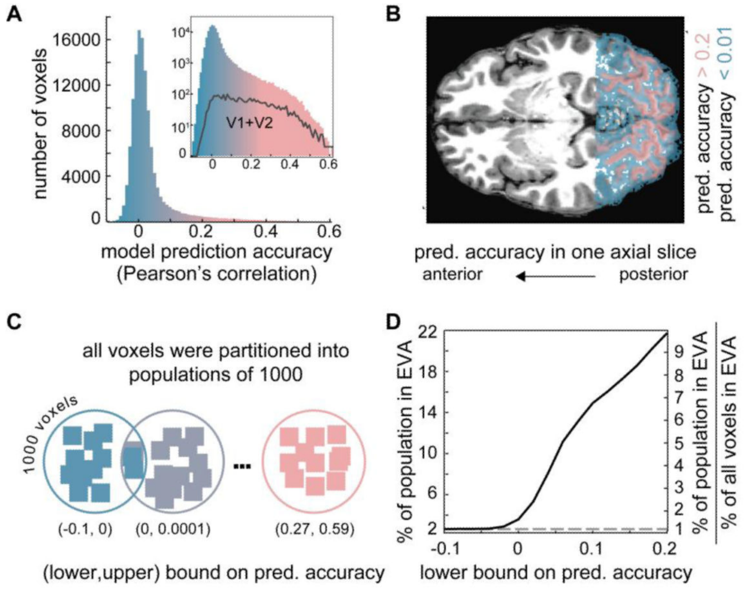

Figure 2. Voxel-wise encoding model performance and voxel-binning procedure

A) Histogram of voxel-wise encoding model prediction accuracy for all voxels in the functional volume acquired for Subject 1. For each voxel the model prediction accuracy is the correlation between the encoding model predictions and activity measured during the model-testing runs. The histogram has a mode near 0 and a heavy tail that is more easily appreciated on a log scale (inset). **B**) Overlay of model prediction accuracy onto a single axial slice (Subject 1). Voxels in which activity is poorly predicted (shown in blue) are scattered throughout white matter across the posterior-anterior extent of the scanned area. Voxels in which activity is accurately predicted (shown in pink) are confined to gray matter and thus track the convolutions of the cortical surface. **C**) To facilitate the image identification analyses voxels were rank-ordered by model prediction accuracy and then binned into populations each containing 1000 voxels. Populations are illustrated as circles surrounding schematized voxels (squares) whose color indicates model prediction accuracy. Low-rank populations (blue) contain poorly tuned voxels that had low prediction accuracy. High-rank populations (pink) contained well-tuned voxels that had high prediction accuracy. The lower (upper) bound on prediction accuracy for a population is determined by the voxel in the population with the lowest (highest) model prediction accuracy. **D**) The percent of early visual area (EVA; V1 and V2) voxels in each of the voxel populations used for image identification. The x-axis indicates the lower bound on model prediction accuracy for each population. The left y-axis indicates the percentage of the voxels in EVA across all populations with a lower-bound greater than or equal to the value on the x-axis. The right y-axis indicates the percentage of voxels in EVA in each population *relative* to the percentage of voxels in EVA in the functional volume (~2% for all subjects). The black curve shows how the percentage increases as the lower-bound on model accuracy increases. In the population with the highest lower bound > 20% of the voxels are in EVA. This is greater than 9 times the percent of voxels in EVA contained in the functional volume. The dashed line indicates the absolute (left y-axis) and relative (right y-axis) percentage of voxels in

EVA that would be obtained if populations were constructed by randomly sampling voxels from the functional volume.

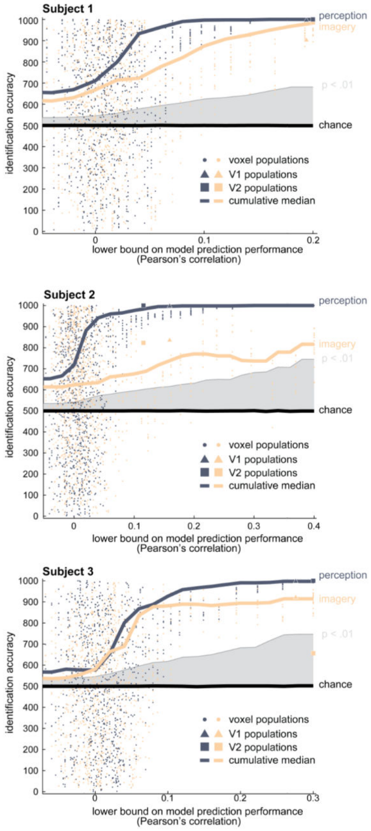

Figure 3. Accurate identification of mental imagery depends upon accuracy of encoding models

Each panel displays data for a single subject. Each dot corresponds to a single population of 1000 voxels. Triangles and squares indicate populations formed exclusively from voxels in V1 and V2, respectively. For each voxel population image identification accuracy was quantified by correlating the measured response to the five works of art presented during the perception runs (blue) or imagery runs (orange) against the encoding models' predicted response to the five works of art. The measured activity was also correlated against the predicted response to 1000 sets of five randomly selected images. Accuracy of image identification was quantified as the number of *hits*, which are cases where the predicted responses to the perceived/imagined works of art were more correlated with the measured response than the predicted responses to the randomly selected images. The position of each dot along the y-axis indicates the image identification accuracy in hits for the voxel population. Position along the x-axis indicates the lowest model prediction accuracy (i.e., the lower-bound) of all voxels in the population. Solid curves show the cumulative median image identification accuracy (y-axis; blue=perception, orange=imagery) of *all* voxel populations whose lower-bound is greater than or equal to the model prediction accuracy indicated on the x-axis (i.e., all voxel populations to the right of the lower bound indicated on the x-axis). The median identification performance for both perception and imagery increases monotonically as the lower-bound on model prediction accuracy increases. Chance performance is indicated by the black line near *hits* = 500; a statistical significance threshold of $p < .01$ (permutation test) is indicated by gray shading. These data establish that accurate identification of mental images is possible and depends upon the accuracy of the underlying encoding model.

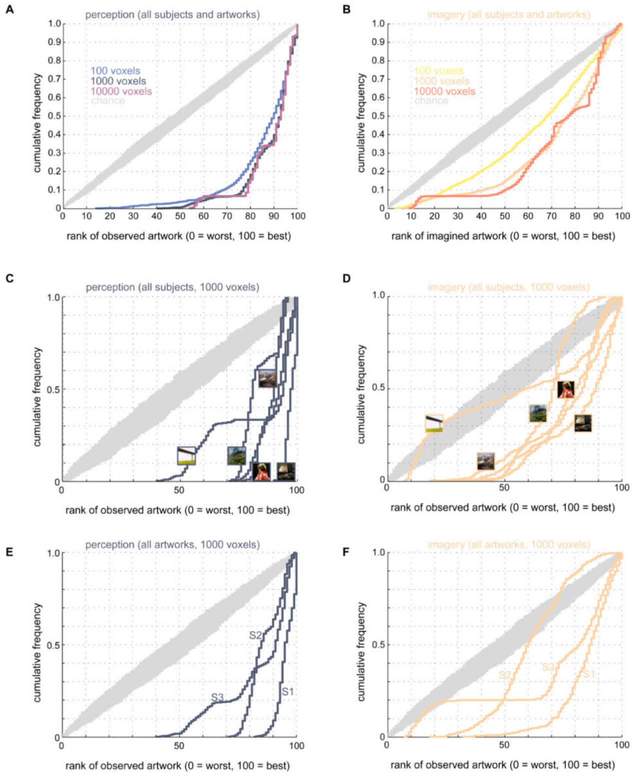

Figure 4. Artists' galleries sorted by mental imagery

A-B) Effect of population size on mental image identification. Measured activity in populations of 100, 1000, or 10000 voxels was used to sort digital galleries associated with each of the five artists whose artwork was used in our experiments. Voxel populations were created by random sampling of the 30,000 voxels with highest model prediction accuracy. Activity measured in these populations was used to rank-order the perceived or imagined artwork with respect to 100 images retrieved from a *Google Images* query on each artist's name. For each voxel population measured responses were correlated against the response predicted by the voxel-wise encoding models to all the images for a specific artist (including the artwork that was perceived/imagined during the experiment). Images were then ranked according to the strength of the correlation between the measured and predicted responses. The cumulative histograms show the cumulative probability (y-axis) of rankings (x-axis) for the artwork actually perceived (blue-violet curves in **A**) or imagined (yellow-orange curves in **B**) during the experiment. Rankings vary from 0 (worst) to 100 (best). Chance performance is indicated by grey curves (10,000 cumulative histograms obtained by random permutation of rankings). Rankings are significantly higher than expected by chance wherever the histograms for perception and imagery runs dip below the histograms for chance performance. In these panels data for all subjects and artworks are combined. **C-D**) Mental image identification accuracy for individual artworks. In these panels data for all subjects were pooled and the number of voxels per population was fixed at 1000. Thumbnail images indicate artwork corresponding to each curve. Although no single artwork can account for the ability to decode mental images using low-level features, there is more variation in decoding accuracy across artworks for mental images than perceived images. **E-**

F) Mental image identification accuracy for individual subjects. In these panels data for all artworks were pooled and the number of voxels per population was fixed at 1000. Each curve corresponds to a different subject (labeled S1, S2, and S3). Identification accuracy varies more across subjects for mental images than for perceived images.

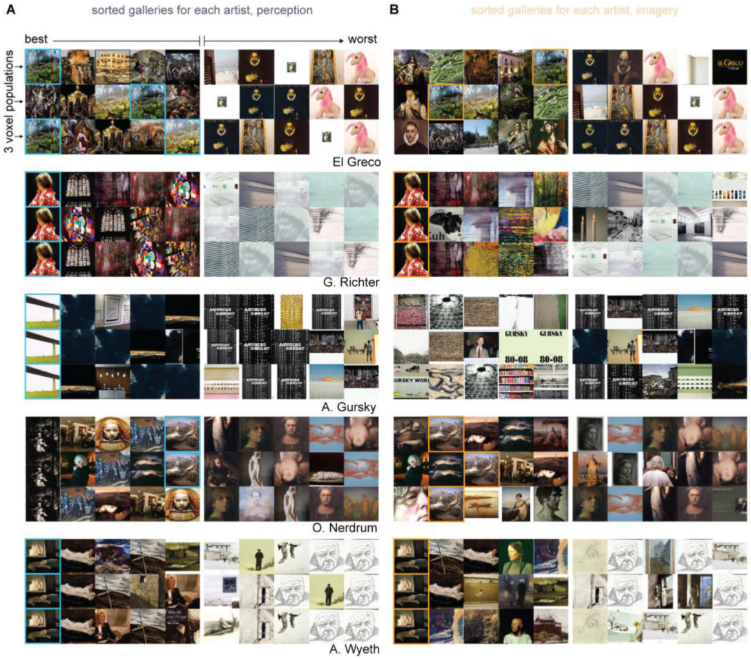
A
sorted galleries for each artist, perception
B
sorted galleries for each artist, imagery
best
worst
3 voxel populations
El Greco
G. Richter
A. Gursky
O. Nerdrum
A. Wyeth

Figure 5. Examples of sorted artist's galleries

Each 3 *x* 10 block of images shows how the images returned by the *Google Images* query for a single artist (named in the lower right corner of each block) were sorted by the decoding procedure. Rows within each block correspond to one population of 1000 voxels. For each block the three populations selected for display were the three "best" in the sense that activity evoked by the perceived or imagined artwork were more highly correlated with model predictions in these populations than in all other populations. Columns indicate the ranking of each image. The top 5 images are shown to the left of the small vertical division within each block. The bottom 5 images are shown to the right of the division. In **C** the perceived works of art are outlined in blue. In **D** the imagined works of art are outlined in orange. Data for all three subjects are pooled in this figure.