

# Inverse retinotopy: Inferring the visual content of images from brain activation patterns

Bertrand Thirion,<sup>a,\*</sup> Edouard Duchesnay,<sup>b</sup> Edward Hubbard,<sup>d</sup> Jessica Dubois,<sup>c</sup>  
Jean-Baptiste Poline,<sup>c</sup> Denis LeBihan,<sup>c</sup> and Stanislas Dehaene<sup>d</sup>

<sup>a</sup>INRIA Futurs, Service Hospitalier Frédéric Joliot, 4, Place du Général Leclerc 91401 Orsay Cedex, France

<sup>b</sup>Unité INSERM ERM 0205, Service Hospitalier Frédéric Joliot, 4, Place du Général Leclerc, 91401 Orsay Cedex, France

<sup>c</sup>CEA, DSV, DRM, SHFJ 4, Place du Général Leclerc, 91401 Orsay Cedex, France

<sup>d</sup>Unité INSERM 562 “Neuroimagerie Cognitive” Service Hospitalier Frédéric Joliot, 4 Place du Général Leclerc, 91401 Orsay Cedex, France

Received 12 January 2006; revised 26 June 2006; accepted 28 June 2006

Available online 9 October 2006

Traditional inference in neuroimaging consists in describing brain activations elicited and modulated by different kinds of stimuli. Recently, however, paradigms have been studied in which the converse operation is performed, thus inferring behavioral or mental states associated with activation images. Here, we use the well-known retinotopy of the visual cortex to infer the visual content of real or imaginary scenes from the brain activation patterns that they elicit. We present two decoding algorithms: an explicit technique, based on the current knowledge of the retinotopic structure of the visual areas, and an implicit technique, based on supervised classifiers. Both algorithms predicted the stimulus identity with significant accuracy. Furthermore, we extend this principle to mental imagery data: in five data sets, our algorithms could reconstruct and predict with significant accuracy a pattern imagined by the subjects.

© 2006 Elsevier Inc. All rights reserved.

---

## Introduction

### *The neuroimaging inverse problem*

Validation of anatomo-functional knowledge produced from neuroimaging data is a difficult task. While statistical significance, reproducibility and multi-modal coherence are well-accepted proofs of consistency, neuroscientists lack a gold standard to assess the significance of their findings. A possible way to solve this issue is to reason as follows: understanding a cognitive subsystem of the brain means that the stimulus-to-activation chain has been identified. More precisely, although the detailed mechanisms of neural and hemodynamic activation are not fully understood, we can expect that a controlled stimulus (e.g. a flashing checkerboard) will produce a known pattern of activation.

When this holds, the processing chain can be inverted, leading to activation-to-stimulus inference. When possible, this inverse inference allows good performance characterization, since the results are expressed in terms of predicted versus true stimulus, in the well-known (and controlled) stimulus space.

This point of view has already been investigated in the case of motor experiments (Dehaene et al., 1998), mental imagery (O’Craven and Kanwisher, 2000), counting/subitizing (Piazza et al., 2003), the notion of object categories (Haxby et al., 2001; Carlson et al., 2003; Cox and Savoy, 2003), the orientation of visual stimuli (Haynes and Rees, 2005; Kamitani and Tong, 2005) and lie detection (Davatzikos et al., 2005). It has been popularized under the concept of brain reading. This novel approach in neuroimaging has been facilitated by the use of data classification techniques such as Linear Discriminant Analysis (LDA) (Carlson et al., 2003) and more recently, Support Vector Machines (SVM) (Cox and Savoy, 2003; LaConte et al., 2005) that can take functional images as input and classify them into categories (supervised classification). But in that case the activation-to-stimulus function remains implicit, i.e. it is embedded in a set of learning samples, each one being associated with a known stimulus. An important question is whether this binding may be made explicit.

There is at least one system in which the stimulus-to-activation coding is known explicitly: this is the case of retinotopy, where the spatial layout of an image is in the visual field also spatially encoded in the primary visual cortex (Sereni et al., 1995). The inverse problem consists in predicting the spatial layout of an activation pattern (stimulus) given a functional activation image. We address this problem with two kinds of analysis tools: supervised classification (based on SVMs) and an explicit inversion of the stimulus-to-activation function (inverse retinotopy). In this paper, we study two different situations: a visual stimulation experiment, in which the subject passively views a sequence of stimuli chosen among a discrete set and a mental imagery experiment in which the subject is asked to imagine a self-selected pattern chosen among the presented stimuli.

---

\* Corresponding author.

E-mail address: bertrand.thirion@inria.fr (B. Thirion).

Available online on ScienceDirect (www.sciencedirect.com).

### Retinotopy of the human visual cortex

It is well known that the human visual cortex is retinotopically organized, at least in early areas. Retinotopic mapping, based on a travelling wave paradigm, is a standard procedure in the fMRI literature (see e.g. Sereno et al., 1995; DeYoe et al., 1996; Tootell et al., 1996, 2003; Engel et al., 1997; Warnking et al., 2002; Dougherty et al., 2003; Wotawa et al., 2005). It is frequently performed in order to delineate the early visual areas (V1, V2, V3, V3a, VP), which can be characterized by a visual field sign (VFS) (Sereno et al., 1995). By contrast, we interpret here the retinotopic information as a forward mapping from the visual field to the visual cortex: we assume that there exists a transfer function that maps visual stimulation patterns to the primary visual cortex. In this work, the retinotopic data are used to estimate the transfer function. This kind of model has been suggested for V1 (Tootell et al., 1998b), and is supported by recent experiments (Hansen et al., 2004). In our setting, we take into account the receptive field structure (Smith et al., 2001) that characterizes the responsiveness of cortical neurons to retinal stimulation. Let us note, however, that such a model ignores some parts of the response (non-linear and/or negative components) (Shmuel et al., 2002, 2006).

In a recent paper (Vanni et al., 2005), a direct estimation of the transfer function has been proposed based on randomized visual stimulation in an event-related design. However, this procedure is not as generic as the phase-encoded retinotopic experiments and it can only delineate predefined regions of the visual field.

### Inverse reconstruction and classification

Assume that we are given a set of brain activation images  $\phi^1, \dots, \phi^n$  associated with a set of stimuli  $\sigma^1, \dots, \sigma^n$  chosen within a finite set  $S$ . Supervised classification and inverse reconstruction perform two kinds of characterization on these data:

- In the inverse reconstruction framework, we assume that a forward operator  $T$  that models the stimulus-to-activation process has been defined. The inverse reconstruction consists in estimating the stimulus pattern  $\hat{\rho}^i = T^{-1}\phi^i, i=1..n$  in order to identify the label  $\hat{\sigma}^i, i=1..n$  of the reconstructed pattern. The performance of the procedure can thus also be expressed in terms of correct prediction rate. While this procedure requires the prior knowledge of the operator  $T$ , it applies to any activation image, and understanding the failures of the system might be easier, since the intermediate results  $\hat{\rho}^i, i=1..n$  are available.
- In the supervised classification framework, a subset of the images  $\phi^1, \dots, \phi^r, r < n$  associated with stimulus labels  $\sigma^1, \dots, \sigma^r$  are used as a learning set. The learning algorithm (SVM typically) learns how to predict the stimulus label given the functional image. The test set, that consists of the remaining images  $\phi^{r+1}, \dots, \phi^n$  is used to predict labels  $\hat{\sigma}^{r+1}, \dots, \hat{\sigma}^n$ . The performance of the classifier is given by the rate of correct predictions. The advantage of such an approach is that it works efficiently, without requiring prior knowledge on the precise activation mechanisms (functional architecture, connectivity, hemodynamic phenomena). In that sense it is universal. The disadvantage is that it is hard to diagnose a failure in the system. Moreover, the interpretation is not straightforward (see e.g. Hanson et al., 2004). Last the ability to discriminate between activation patterns and to associate correct labels is restricted to the data set used in the learning procedure.

Our main experiment consists thus in the identification of visual patterns presented to the subjects, separately in the left and right hemifields. The inverse reconstruction is based on retinotopic information obtained in a traveling wave paradigm, while SVM classification is performed directly on activation images masked by the retinotopic regions. We also asked the subjects to perform a mental imagery experiment. Involvement of low-level visual areas has been reported during imagery tasks (Tootell et al., 1998a; Kosslyn et al., 1999), and hints of a retinotopic organization of mental images has been seen with fMRI (Klein et al., 2004; Slotnick et al., 2005). We propose to use this limit case as an additional benchmark to test how well the classification/inverse reconstruction techniques can decode subjective brain states.

### Materials and methods

#### Data acquisition and pre-processing

#### Subjects

Nine subjects participated to the study. One data set was discarded due to poor fixation during the experiment (see below). This provided us with a total of 16 data sets, each hemisphere being analyzed independently. The subjects gave written informed consent and the protocol was approved by the local ethics committee.

#### Stimuli

The experimental protocol consisted in three parts: (i) a retinotopic mapping of the subjects, (ii) a passive viewing experiment, in which the subjects were viewing so-called *domino* stimuli, (iii) an imagery experiment, in which the subjects had to *imagine* one of the domino stimuli when prompted to. Next we describe the stimuli used in these three parts.

- (i) The retinotopic experiment consisted in rotating wedges and expanding/contracting rings that flickered at a rate of 7.5 Hz. The checkerboard pattern was superimposed on a uniform grey field. The stimuli were projected onto a rear-projection viewing screen mounted within the scanner. Subjects were supine and viewed the display by means of a mirror placed above their eyes and housed in a custom-designed head piece. The duration of a complete stimulus movement was 32 s, and it was repeated eight times for either condition. The wedge stimuli had one single lobe, with a maximal eccentricity of 10.5° and an angular width of 40° (see Fig. 1(a)). The ring had an eccentricity between 0.8° and 10.5°. The size of the display, which matched the red circle in Fig. 1(b), was 21° diameter. The subjects were instructed to fixate a central cross, and fixation was controlled using an eye-tracker system.
- (ii) In the *domino* experiment, two grids, situated on the left and right parts of the visual field, and a central fixation cross were presented to the subjects. The grid was surrounded by a disk of 9.5° diameter. Every 8 s, a flickering pattern appeared in several sectors of the grid. These patterns belonged to a set of 6 possible shapes (see Fig. 1(c)). The patterns were presented simultaneously in the left and right visual field for a total of 36 combinations which were all presented once per fMRI run, in a randomized order. Each subject performed four sessions of this domino experiment.
- (iii) Then the subjects were asked to choose one of the six patterns. During the last session, the subjects viewed the

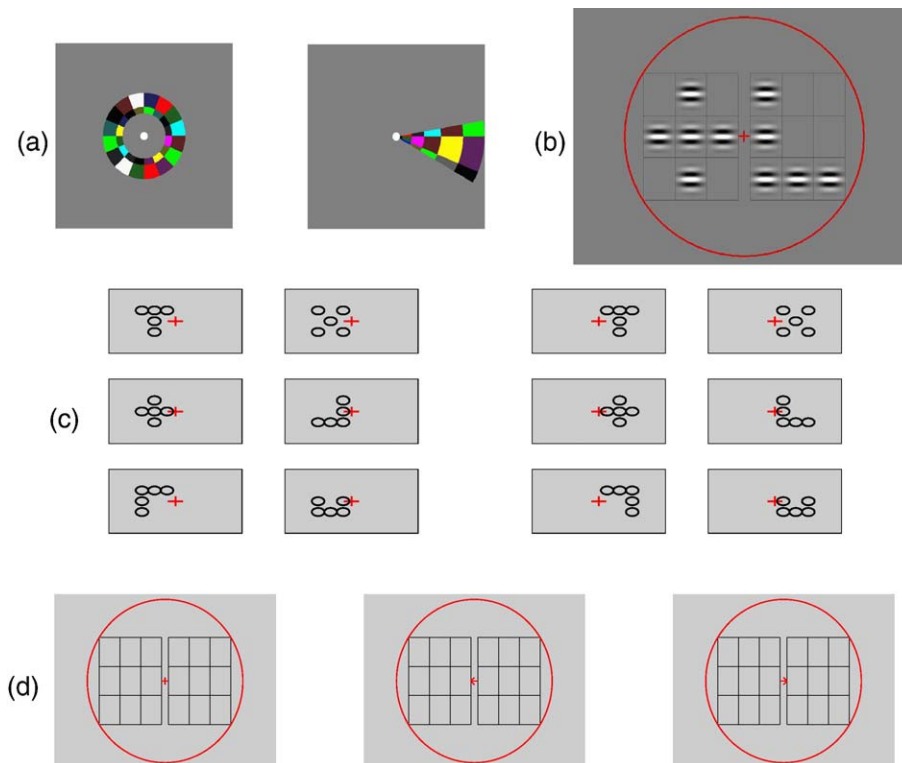


Fig. 1. Visual stimuli used in our experiments. (a) First the subject was involved in a classical retinotopic mapping experiment, in which he viewed flickering rotating wedges and expanding/contracting rings. (b) In the domino experiment, the subject viewed groups of quickly rotating Gabor filters in an event-related design. These disks appeared simultaneously on the left and right side of the visual field, superimposed on a low-contrast grid and a fixation cross. (c) There were 6 different patterns in each hemifield. (d) In a last session, the subject was presented with the same grid. When the central fixation cross (left) became a right arrow (middle) or a left arrow (right), the subject had to imagine one of the six patterns presented previously, either in the left or right hemifield.

same grid, but without any pattern presentation. The central cross was changed to a small left/right arrow of the same size ( $0.8^\circ$ ), prompting the subject to imagine the selected pattern on the left or right side (see Fig. 1(d)). The arrow occurred 4 s every 10 s interval and appeared a total of 36 times: 18 times on the left side, 18 times on the right side. Once the experiment was finished, the subject reported which pattern he or she had chosen for the imagery experiment.

During all the scanning sessions, the subjects were instructed to fixate the center of the screen. Eye movement were registered with an ISCAN eye-tracker system, in order to ensure that fixation was maintained. One subject did not fixate adequately, and the data set was eliminated from the analysis.

#### Acquisition parameters, pre-processing

Functional images were acquired on a 3 T Bruker scanner using an EPI sequence (TR=2000 ms, TE=40 ms, matrix size= $64 \times 64$ , FOV= $19.2 \text{ cm} \times 19.2 \text{ cm}$ ). Each volume consisted of 35 3 mm-thick axial slices without gap. The first four functional scans were discarded in order to allow the MR signal to reach steady state. Anatomical T1 images were acquired on the same scanner, with a spatial resolution of  $1 \times 1 \times 1.2 \text{ mm}^3$ .

Motion estimation was performed on each data set using SPM2 software (see e.g. Ashburner et al., 2004). The anatomical images were then normalized to the MNI template of the SPM2 software, and resampled. The interpolation of the functional data took into account motion estimates, so that the normalized images were also

realigned. Resolution after interpolation was  $2 \times 2 \times 2 \text{ mm}^3$ . No other pre-processing was performed.

#### First-level analysis of the data

All data sets were analyzed using the General Linear Model (GLM) implemented in the SPM2 software: retinotopic sessions were analyzed using sinusoidal regressors at the stimulus frequency; the other sessions were analyzed by convolving the activation onset vectors with a standard hemodynamic response; standard high pass filtering ( $hf_{\text{cut}}=80 \text{ s}$ ) and AR(1) noise whitening were used. Activation maps were produced for each experiment. In the retinotopic mapping experiment, these maps show regions with significant activity at the stimulus frequency, hence retinotopic regions. By contrast, the statistical images resulting from the analyses of the domino and imagery experiments were associated with occurrences of the stimuli. They could thus be readily interpreted as stimulus-induced activation patterns.

The parameter maps of the retinotopic experiments were further processed as indicated in (Sereni et al., 1995) in order to yield polar and eccentricity maps (see also Appendix A.2). False positives were discarded by retaining only the main connected component of supra-threshold voxels, after thresholding at  $P < 10^{-3}$  uncorrected. This systematically corresponded to a symmetric occipital cluster. This yielded  $V$  (~10000–15000) voxels, according to the subject. The retinotopic regions were divided into left and right hemispheres using the segmentation of the anatomical image by the Brainvisa analysis pipeline (Rivière et al., 2000).

The domino experiment was analyzed on a trial-by-trial basis, yielding trial-specific (*ts*) activation maps. Condition-specific (*cs*) contrasts and activation maps were also estimated. For further processing, both *cs* and *ts* maps were masked by the retinotopic regions. Similarly, the imagery experiment was analyzed in order to yield trial- and condition-specific images. All maps were masked as the images of the domino experiment.

*Explicit solution of the inverse problem*

The explicit reconstruction of images in the visual field requires the solution of a forward problem (definition of a mapping from retina to cortical activity) and then the solution of an inverse problem (visual image associated with a given activation image). The global setting is described in Fig. 2.

*Solution of the forward problem*

We define a visual image as a function  $\rho$  that associates an activity value  $\rho(p)$  with any point  $p$  on the retina  $R$ . In practice the retina will be discretized on a grid of size  $P$ . In our setting,  $P$  is  $100 \times 100$  to balance the competing demands of computational efficiency and resolution. An activation image is a function  $\phi$  that associates an activation value  $\phi(v)$  with any voxel  $v$  of the brain volume. In practice the brain volume is restricted to a set of  $V$  voxels that have retinotopically specific responses.

We use the following generative model: the visual stimulation  $\rho$  is mapped to a functional image  $\phi$  through a transfer operator  $T$ , i.e.  $\phi = T(\rho)$ . The forward problem consists in estimating  $T$ . Since the travelling wave paradigm used in the retinotopic mapping experiment performs a complete sweep of the visual field, we use the corresponding data to estimate  $T$ : let  $\rho^1, \dots, \rho^n$  be the visual images of the retinotopic mapping paradigm, and  $\phi^1, \dots, \phi^n$  the associated functional images, we search  $T$  such that

$$\phi^i = T(\rho^i) + \epsilon^i, \forall i \in [1 \dots n] \quad (1)$$

where  $\epsilon^i$  is an additive (measurement) noise that models possible mismatch. This noise will be assumed to be independently identically distributed Gaussian and centered.

A priori  $T$  is a-possibly nonlinear-operator from  $R^P$  to  $R^V$ . Given that the sizes  $P$  and  $V$  are well above  $10^3$ , the direct estimation of  $T$

from Eq. (1) is impossible. Thus, we first assume that  $T$  is linear, which is equivalent to a spatial superposition principle of visual activations; this hypothesis is supported by recent experiments (Hansen et al., 2004). If  $T$  is linear, it is fully specified by its behavior on spatial Dirac functions ( $\delta_p, p \in [1 \dots P]$ ) in the input space. At this point, we use physiological prior knowledge to estimate  $T$ . For each voxel  $v \in [1 \dots V]$ , we assume that there exist a point  $p_v$  of the retina, a positive real number (radius)  $\lambda_v$ , and a real number  $\gamma_v$  (gain) so that

$$[T\delta_p](v) = \gamma_v \exp\left(-\frac{\|p - p_v\|^2}{2\lambda_v^2}\right) + \epsilon(v) \quad (2)$$

where  $\delta_p$  is a Dirac function on  $p$ , and  $[T\delta_p](v)$  is the associated functional image evaluated at voxel  $v$ . This simply means that voxel  $v$  is associated with a receptive field, i.e. a Gaussian kernel centered on  $p_v$ , with width  $\lambda_v$ , and that the gain of the filter is  $\gamma_v$ . The receptive fields are assumed to be isotropic. This model is illustrated in Fig. 3.

Given model (2), the estimation of  $T$  boils down to the estimation of the parameters ( $p_v, \lambda_v, \gamma_v$ ). Given Eq. (1), this amounts to solving the following equations

$$p_v, \lambda_v, \gamma_v = \operatorname{argmin}_{p, \lambda, \gamma} \sum_{i=1}^n \left\| \phi^i(v) - \gamma \int_R \rho^i(r) \exp\left(-\frac{\|r - p\|^2}{2\lambda^2}\right) dr \right\|^2 \quad (3)$$

However, due to the non-linear nature of the estimation problem, we find an approximate solution by estimating (i)  $p_v$  first, then (ii)  $\lambda_v$ , then (iii)  $\gamma_v$ .

- (i) The estimation of  $p_v$  is standard in retinotopic mapping experiments (Serenio et al., 1995). For completeness, we detail it in the Appendix A.2.
- (ii) The size  $\lambda_v$  of the receptive field could be determined from the retinotopic data (Smith et al., 2001; Duncan and Boynton, 2003); however, here we prefer to rely on two models:

$$(M_1) \lambda_v = l_0 \quad (4)$$

$$(M_2) \lambda_v = l_1 \|p_v\| \quad (5)$$

In the first model the width of the receptive field is constant. In the second model, the width is proportional to the eccentricity of its center. These two models are two possible simplifications of the current physiological knowledge about receptive field size, which corresponds to an increasing affine function whose characteristics depend on the visual area considered (Smith et al., 2001). Model ( $M_1$ ) might be a more robust choice on real data, given the strong non-linear dependence of the model (3) on  $\lambda_v$ . An illustration of the results of the inverse problem using ( $M_1$ ) or ( $M_2$ ) is given in Fig. 4. Thereafter we retain the model ( $M_1$ ), where the constant  $l_0$  is  $0.75^\circ$  in the visual field.

- (iii) Last, the estimation of  $\gamma_v$  from Eq. (3) is now straightforward and is performed by linear regression.

Our estimation procedure thus yields

$$[\hat{T}\rho](v) = \hat{\gamma}_v \int_R \rho(r) \exp\left(-\frac{\|r - \hat{p}_v\|^2}{2\hat{\lambda}_v^2}\right) dr \quad (6)$$

*Solution of the inverse problem*

Once the operator  $T$  has been estimated, it can be used to infer the visual image  $\rho$  associated with any activation map  $\phi$ . In our

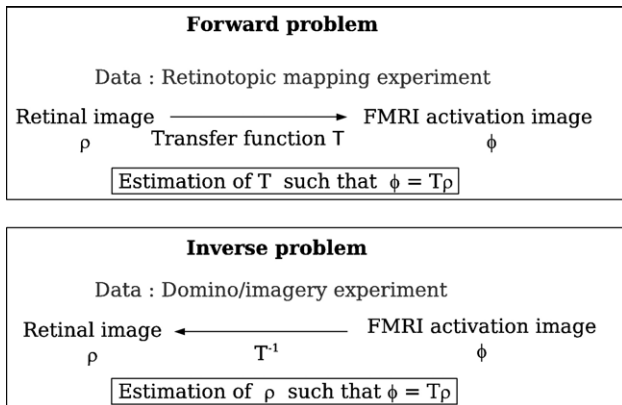


Fig. 2. Illustration of the forward and inverse problem in an inverse retinotopy framework. The forward problem consists in estimating explicitly a transfer operator that maps a stimulus into an activation image. The inverse problem consists in predicting the stimulus associated with an activation image, given the transfer operator.

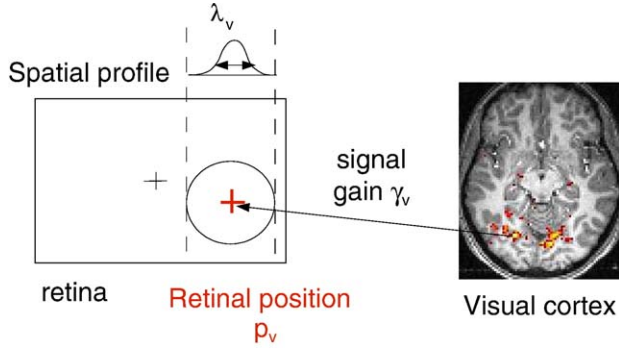


Fig. 3. Receptive field model that is implemented in the forward model. Any voxel  $v$  of the visual cortex is associated with a kernel centered on a retinal point  $p_v$ , with a width  $\lambda_v$ . The gain that maps the magnitude of visual activity to BOLD activity is modeled by a parameter  $\gamma_v$ .

setting, the images  $\phi$  are now those obtained from the domino or imagery experiment.

If  $\hat{T}$  was invertible, the straightforward estimate of  $\rho$  would be

$$\hat{\rho} = \hat{T}^{-1} \phi \quad (7)$$

However,  $\hat{T}$  might not be invertible—neither in theory nor in practice. The estimation of  $\rho$  must be regularized. This can be simply cast in a Bayesian framework

$$P(\rho|\phi) \propto P(\rho)P(\phi|\rho) \quad (8)$$

Given our Gaussian noise model hypothesis, the likelihood writes

$$P(\phi|\rho) \propto \exp\left(-\frac{1}{2}(\phi - \hat{T}\rho)' \Delta^{-1}(\phi - \hat{T}\rho)\right) \quad (9)$$

where  $\Delta$  models the uncertainty about the measurement  $\phi$ . Note that, assuming that  $\Delta$  is diagonal,<sup>1</sup> an estimate of this uncertainty is provided by the GLM analysis when  $\phi$  is a parametric image.

Given the model (9) for the likelihood, it is natural to choose the conjugate, hence normal prior

$$P(\rho) = \exp\left(-\frac{1}{2}\rho'K^{-1}\rho\right) \quad (10)$$

This means that visual activations are expected to be zero, with a spatial correlation structure provided by  $K$ . The prior can be a simple shrinkage prior ( $K = \mu^{-1}I_{\mathcal{P}}$ ),  $I_{\mathcal{P}}$  being the  $\mathcal{P} \times \mathcal{P}$  identity matrix and  $\mu$  a positive constant, or it may involve some spatial modeling (e.g.  $K_{ij} = k(p(i) - p(j))$ ), where  $k$  is some decreasing function of the distance  $|p(i) - p(j)|$ .

The solution of the inverse problem consists in minimizing the following functional

$$\Psi(\rho) = (\phi - \hat{T}\rho)' \Delta^{-1}(\phi - \hat{T}\rho) + \rho'K^{-1}\rho \quad (11)$$

Note that the covariance of the estimator  $\hat{\rho}$  can be estimated as

$$\hat{A}_{\rho} = (K^{-1} + \hat{T}' \Delta^{-1} \hat{T})^{-1} \quad (12)$$

<sup>1</sup> This amounts to assuming that the errors in the forward model are uncorrelated. This oversimplification is necessary for computational efficiency. This allows to estimate the likelihood of an activation at a given

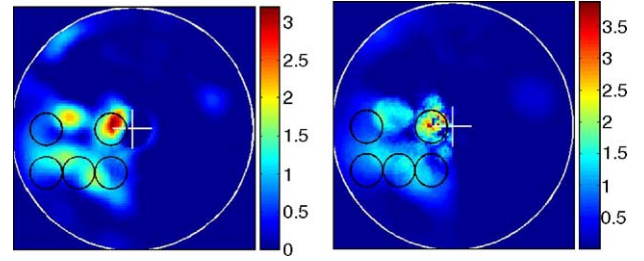


Fig. 4. Comparison of receptive fields models  $M_1$  and  $M_2$  when applied in the solution of inverse problem described in Solution of the forward problem. Left: the model  $M_1$  is used, resulting in a spatially stationary smoothness of the reconstructed image. Right: using model  $M_2$ , the reconstructed images are rougher in the foveal region and smoother at the periphery. The true stimulus is represented by five circles in both cases. Note that these differences have little or no impact on pattern identification.

point  $r$  of the retina through the statistic

$$\tau(r) = \frac{\hat{\rho}(r)}{\sqrt{\hat{A}_{\rho}(r,r)}} \quad (13)$$

This neglects the covariance between neighboring points, but allows for an easy interpretation, since it yields the probability that  $\rho(r)$  is indeed positive given our observation.

In practice, we initialize  $\rho$  to 0, and iterate the update rule

$$\rho^{(i+1)} = \rho^{(i)} - \eta \nabla \Psi(\rho^{(i)}) \quad (14)$$

$\eta$  being small enough to ensure convergence. We have tried two possible alternatives for  $K^{-1}$ , namely  $K^{-1} = \mu I_{\mathcal{P}}$  and

$$K_{ij}^{-1} = \begin{cases} 1 & \text{if } i = j \\ -0.25 & \text{if } p(i) \text{ and } p(j) \text{ are four-neighbors} \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

The final difference was not very important but model (15) performed slightly better and was used in our experiments. The factor  $\mu > 0$  – which characterizes the amount of regularization – has to be set a priori. It was chosen to be proportional to the norm of  $\hat{T}$ . We noticed that halving it had little impact on the resulting image.

The set of parameters used in the forward/inverse problem is summarized in Table 1. Last, we have approximated  $A_{\rho}$  (see Eq. (12)) through the inverse of the diagonal part of  $K^{-1} + \hat{T}' \Delta^{-1} \hat{T}$ .

Table 1  
Values of the parameters used in the forward/inverse retinotopy model

Parameter	Value
$p_v$	Polar coordinates estimated from the retinotopy data (voxel-based)
$\lambda_v$	0.75°
$\gamma_v$	Estimated by linear regression (voxel-based)
$\mu$	0.0001 $\sum_{ij}  T_{ij} $
$\eta$	0.01

Table 2

Correct classification rate of the trial-specific functional patterns across subjects and hemispheres after explicit reconstruction of the stimuli in the domino experiment

Subject	Correct classification rate	
	Left hemisphere (%)	Right hemisphere (%)
bru2773	41	65
bru2774	71	51
bru2782	69	53
bru2783	60	47
bru2784	42	43
bru3070	63	61
bru3071	69	56
bru3072	50	56

The chance level is 1/6, and a score of 27% is above chance level with a  $P$ -value of  $10^{-3}$ .

*Evaluation of the inverse reconstruction*

We have computed the correlation of the reconstructed pattern with the different candidate patterns  $\text{corr}(i, s) = \langle \tau^i | s \rangle \forall s \in S$ ; the predicted stimulus is then the best correlating one  $s^*(i) = \text{argmax}_s \text{corr}(i, s)$ . The performance of the stimulus decoding can then be assessed as the correct prediction rate  $P(s^*(i) = \sigma^i)$ .

*Classification of the trial-specific images*

Let  $\phi^1, \dots, \phi^n$  be a set of brain activation images associated with a set of known stimuli  $\sigma^1, \dots, \sigma^n$ . These images are those of the domino experiment. Any classifier proceeds by learning to discriminate between the images associated with a given label  $\Phi_s = \{\phi^i | \sigma^i = s\}$  and the other images. Each image is defined by its values on a number  $V$  of retinotopically specific voxels.

We describe in the Appendix A.1, how classification techniques can be used to learn the association between the set of stimuli and functional data. Let us simply mention that it relies on (a) the selection of a subsample of the voxels whose activity is used to discriminate between stimuli, (b) the construction of a decision function that gives generalizable results and (c) a validation procedure that yields unbiased prediction rates for the classification-based identification. It is important to note that, in this procedure, the learning of the association is performed on the domino data; this procedure does not use the information collected in the retinotopic mapping experiment, except for the definition of retinotopically specific voxels. We have nevertheless studied in which visual areas the most discriminative voxels of the classifier could be found across subjects and hemispheres.

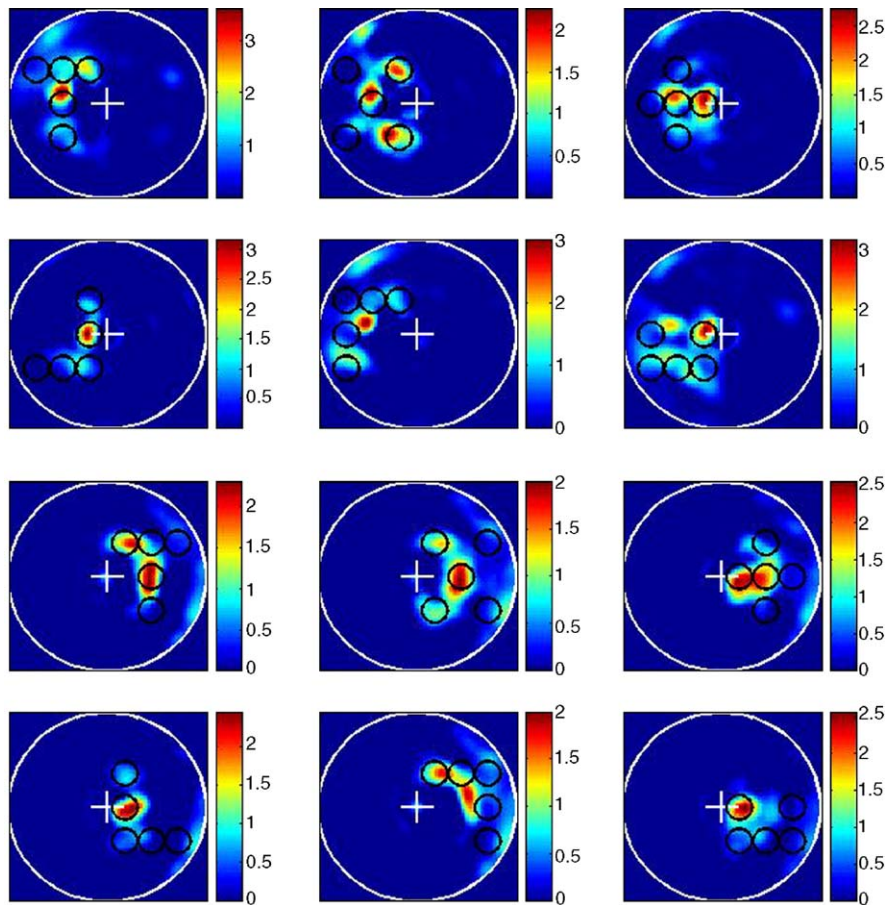


Fig. 5. Explicit reconstruction of the condition-specific visual patterns  $\tau$  obtained for subject bru3070: (top) left part of the visual field; (bottom): right half of the visual field. The true stimulus is defined by five disks whose contour is superimposed on the reconstructed visual image. In spite of the imperfections of the reconstruction, all patterns can be identified with the true stimulus for both hemispheres by correlation analysis.

## Results

### Explicit reconstruction of the visual stimuli

The reconstructed visual images  $\tau(r)$  (see Explicit solution of the inverse problem) were correlated with the true stimuli, so that the prediction was the best correlated input image. Note that these are *ts* images, i.e. one for each trial. The rate of correct responses is given in Table 2 for each subject and hemisphere. This rate varies between 41% and 71%, hence is significantly ( $P < 10^{-11}$ ) above chance level (1/6) in all cases, in spite of significant between subject variability.

An example of the condition-specific reconstructed maps  $\tau(r)$  is given for the two hemispheres of subject bru3070 in Fig. 5. The correlation of the reconstructed visual patterns with the different candidate shapes is given in Table 3 for this subject. Although the most lateral part of the stimulus was imperfectly inferred, this reconstruction allows for an unambiguous recognition of the true stimulus in both hemispheres.

For comparison, the reconstructed images and correlation with the candidate patterns are given as Supplementary Fig. 1 and Table 1 for subject bru3071. In general, the images reconstructed from the other data sets have similar quality and correlation scores. The average of the correlation scores across all subjects is given in Supplementary Table 2.

We have also performed the reconstruction of the stimulus using only voxels from area V1, which has been delineated from the retinotopy experiment. This gives quite similar results as the reconstruction from all retinotopic voxels, in terms of visual appearance and in terms of correlation. Reconstructed images and a correlation table with the candidate patterns are provided as Supplementary material Fig. 2 and Table 3 respectively for subject bru3070.

### Classification of the trial-specific activation images

In the analysis of the domino experiment, we have selected the voxels based on their ANOVA score, keeping only voxels with an

Table 3  
Correlation of the reconstructed pattern with the different candidate patterns for subject 3070

Reconstructed pattern	Candidate pattern					
	T	⊗	+	⌋	⌈	⊘
<i>Left hemisphere</i>						
T	<b>0.6055</b>	0.3904	0.3442	0.2102	0.3941	0.0736
⊗	0.4259	<b>0.4411</b>	0.2489	0.3897	0.2417	0.2368
+	0.2722	0.1674	<b>0.5217</b>	0.3280	0.0116	0.3211
⌋	0.1227	-0.0766	0.0590	<b>0.3182</b>	-0.2499	0.1973
⌈	0.2379	0.1198	0.1296	0.0503	<b>0.3653</b>	0.0327
⊘	0.2075	0.3084	0.4839	0.5154	0.1406	<b>0.6000</b>
<i>Right hemisphere</i>						
T	<b>0.5634</b>	0.4418	0.2522	0.1395	0.2668	-0.0771
⊗	0.5289	<b>0.5803</b>	0.3946	0.3531	0.3231	0.2651
+	0.3385	0.2886	<b>0.5412</b>	0.2869	0.1399	0.2888
⌋	0.0620	0.1483	0.2369	<b>0.4620</b>	-0.0576	0.3053
⌈	0.4219	0.2483	0.2177	0.0700	<b>0.4618</b>	-0.0305
⊘	0.1135	0.2035	0.4869	0.4376	-0.0260	<b>0.4929</b>

Ideally, the diagonal value should be 1, and the off-diagonal values should be between 0 and 0.8, reflecting the correlation between the true stimuli. In the present case, the maximal values of each row, indicated in bold font, are actually in the diagonal, within the [0.3 0.8] interval.

Table 4

Correct classification rate of the trial-specific functional patterns across subjects and hemispheres in the visual stimulation experiment

Subject	Correct classification rate			
	4 folds cross-validation		LOO cross-validation	
	Left hemisphere	Right hemisphere	Left hemisphere	Right hemisphere
bru2773	81%	81%	70%	74%
bru2774	77%	73%	85%	70%
bru2782	78%	80%	85%	86%
bru2783	92%	96%	91%	94%
bru2784	83%	75%	85%	76%
bru3070	81%	88%	86%	90%
bru3071	93%	83%	96%	83%
bru3072	72%	87%	75%	88%
Means	82.5% ( $P < 10^{-15}$ )		83.4% ( $P < 10^{-15}$ )	

The selection of significant features ( $P < 0.1$ , FDR corrected), is followed by a linear SVM analysis. Two cross-validation methods are used, left: learn on 3 sessions, then test on the fourth; right learn on all samples except one then test on the left out sample. The chance level is 1/6, and a score of 27% is above chance level with a P-value of  $10^{-3}$ .

activity significantly modulated by the domino category ( $P < 0.1$ , FDR corrected), see Appendix A.1.2. The activity of the selected voxels is the input to a linear SVM classifier. Table 4 presents results with two different cross-validation schemes: on the left part of the table, three sessions (108 trials, see section Stimuli) are used as the training samples and the fourth session (36 trials) as the independent test set. This procedure is repeated four times and classification rate is averaged across the four runs. On the right part of the table, we performed a Leave-One-Out (LOO) procedure where all samples except one are used to train the discriminant model, which is then tested against the left-out sample. We obtained between 70% and 96% correct classification, according to the subject and the hemisphere. All 16 data sets were classified significantly ( $P < 10^{-11}$ ) above the chance level (1/6 or 16.7% correct responses).

Across subjects and hemispheres, we found that 50–60% of the most discriminative voxels were in V1, while only 20% were in V2 (ventral and dorsal). We did not try to study other visual areas, since their delineation was not reliable enough from our retinotopic maps.

### Mental imagery: explicit reconstruction of the patterns

The imagery activation images were also submitted to the inverse reconstruction procedure. We have tried to identify the pattern that was imagined by the subject using separately the data from the left and right hemisphere using condition-specific activation images, i.e. the images being averaged across trials, and trial-specific activation images. An example of condition-specific reconstructed pattern is given in Fig. 6, together with its correlation with the correlations with the candidate shapes.

In five out sixteen hemispheres (bru2774, right, bru2783, right, bru3070, right, bru3071, left, bru3072, left), we were able to predict the stimulus that the subject had imagined – or reported to imagine – by correlation of the reconstructed pattern with the candidate patterns.

On a trial-by trial basis, we were able (i) to identify the laterality (left or right hemifield) of the imagined pattern with significant

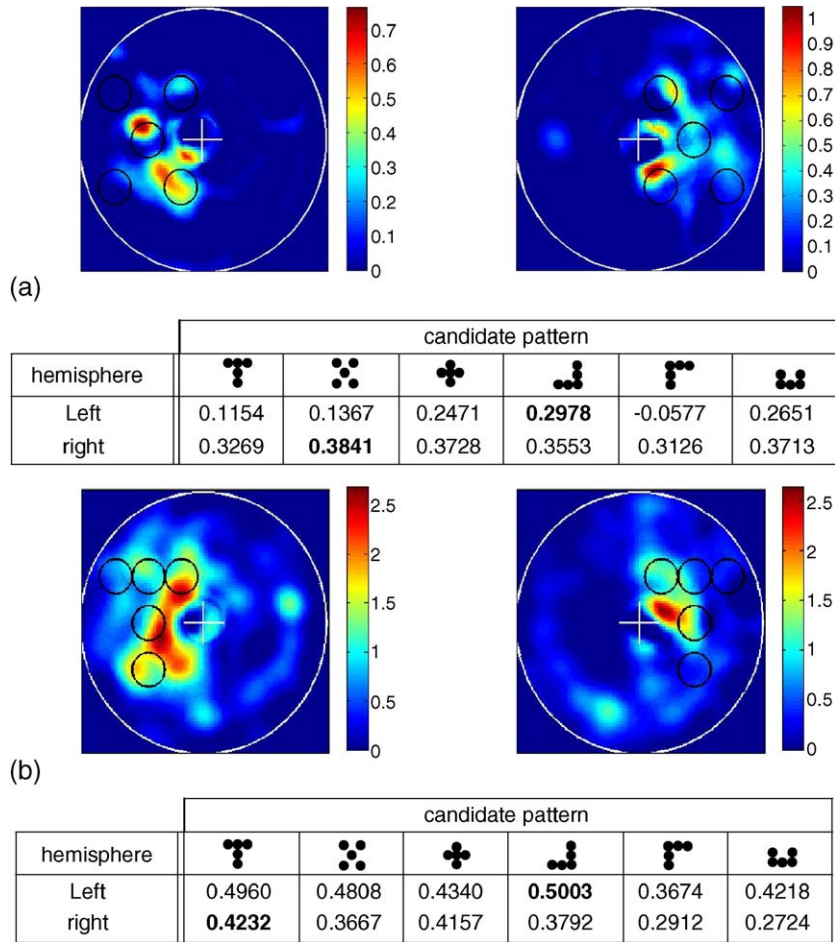


Fig. 6. (a) Explicit reconstruction of the condition-specific imagery pattern obtained for subject bru3070, in the left and right hemifields. The true pattern (as stated by the subject) is given by the set of circles superimposed on the figure. Note from the color scale that the activation magnitude is clearly smaller than in Fig. 5. (bottom) Correlation with the candidate shapes. The left hemifield is not identified correctly, while the right hemifield is. (b) The same data, for subject bru3072. In that case, the imagined stimulus is also identified correctly in the right hemifield.

accuracy, in five out of eight subjects and (ii) to identify the imagined stimulus with significant accuracy in three hemispheres: bru2783, right (78% accuracy,  $P < 10^{-8}$ ), bru2784, left (44% accuracy,  $P < 0.006$ ) and bru3071, left (39% accuracy,  $P < 0.03$ ).

*Mental imagery: identification of trial-specific patterns through classification*

In this experiment we tried to identify the pattern that was imagined by the subject using the data from the left and right hemispheres separately. The method is similar to the one used for the classification of the trial-specific domino activation maps (Classification of the trial-specific activation images), except for two points.

- (i) We compare a stringent feature selection scheme ( $P < 0.05$  after Bonferroni correction, see the left part of Table 5) with the initial scheme ( $P < 0.1$  FDR corrected, see the right part of Table 5).
- (ii) The second difference lies in the validation procedure: for each subject, we used the 144 visual trials of the domino experiment as training samples, and the 18 trials of the

imagery experiment as test samples. The correct classification rate thus indicates the correspondence between the predicted shape and the shape reported by the subjects.

In the first experiment (Bonferroni  $P < 0.05$ ), we obtain significantly above chance recognition rates in five of out of sixteen hemispheres. Using the less stringent feature selection method (FDR  $P < 0.1$ ), four cases remain above chance, but it should be noticed that it dramatically improves (up to 83%) the recognition rate in subject bru2773, right hemisphere. A possible reason is that lenient feature selection schemes yield high performance in some data sets with high signal levels, and poor performance in data sets with lower signal levels.

**Discussion**

*Inverse retinotopy*

We have presented a model-based scheme to decode the information carried by the occipital retinotopic cortex that was successful in identifying the presented stimuli, significantly for all subjects and hemispheres (see Table 2).



Table 5

Correct classification rate of the 18 trials of the imagery experiment, using a discriminant model built on the 144 trials of the domino experiment of the corresponding subject and hemisphere

Subject	Correct classification rate computed by LOO ( <i>P</i> -value)			
	Bonferroni <i>P</i> <0.05		FDR <i>P</i> <0.1	
	Left hemisphere	Right hemisphere	Left hemisphere	Right hemisphere
bru2773	11% (0.83)	<b>44% (0.005)</b>	11% (0.83)	<b>83% (1.04e-9)</b>
bru2774	28% (0.17)	0% (0.96)	17% (0.6)	0% (0.96)
bru2782	33% (0.07)	17% (0.6)	17% (0.6)	17% (0.6)
bru2783	0% (0.96)	6% (0.96)	0% (0.96)	11% (0.83)
bru3070	<b>67% (2.19e-5)</b>	<b>40% (0.03)</b>	<b>73% (1.94e-6)</b>	20% (0.47)
bru3071	0% (0.96)	11% (0.83)	0% (0.96)	6% (0.96)
bru3072	<b>50% (1.13e-3)</b>	<b>38% (0.02)</b>	<b>44% (5.3e-3)</b>	<b>39% (0.02)</b>

The left and right parts of the table differ on the feature selection method. Each rate is given with *P*-values computed relative to the null hypothesis that the classifier was operating at chance level. Significantly greater than chance results are emphasized with a bold font. Non-significant results are in grey. The chance level is 1/6, and a score of 33% is above chance level with a *P*-value of 0.05.

These results confirm that (i) retinotopic activations in the primary visual cortex are reproducible across trials and sessions, (ii) the retinotopic information obtained with the now classical traveling wave paradigm (Engel et al., 1997) can be used as a code, as suggested in (Tootell et al., 1998b; Vanni et al., 2005), and (iii) a linear filter model for V1 (Hansen et al., 2004), that we use in our forward model, holds as a first approximation.

An important new feature of our approach compared to recent contributions (Haynes and Rees, 2005; Kamitani and Tong, 2005) is that we compare classification techniques that model implicitly the stimulus/activation relationship with the explicit resolution of an inverse problem. Importantly, these techniques are based on different hypotheses:

- Supervised classification assumes that reproducible differences might be found between functional images, so that the associated stimulus can be inferred. The important issue is to identify and select the discriminating information and to assess its reliability. Cross-validation and heuristic arguments are used to solve the problem. The acquired knowledge is restricted to the categories presented in the learning set.
- Inverse reconstruction builds on a model of the activation process, with explicit simplifying assumptions. The main issue is to find a simplified model that remains consistent with the data. The inverse problem can then be solved in a rather systematic way, and with any kind of input data: the initial retinotopic mapping is assumed here to yield a generalizable model of any visual activation.

As we have noticed, SVM-based classification yields more accurate results than the inverse problem; the price to pay is that it is not as general. But a key point is that the high performance of classifiers indicates that sufficient discriminant information is indeed present in the data, even if it was not explicitly decoded: some identification failures in the inverse problem can thus be attributed to shortcomings of the model rather than insufficient

information in the data (which might e.g. be related to the performance or attention of the subjects).

### Mental imagery

Moreover, we were able to extrapolate our predictions from passive viewing experiments to mental imagery in some of the subjects, with particularly strong evidence when using classification tools. These latter findings are consistent with the hypothesis that mental imagery involves activation in the primary visual areas, and that the spatial structure of these activations is accounted for by standard retinotopic mapping (Kosslyn et al., 1999; Klein et al., 2004; Slotnick et al., 2005).

If inter-subject differences play a mild role in the performance of inverse retinotopy algorithms applied to actual visual stimuli, they might have more impact on the results of the imagery experiment. In particular, the different subjects reported more or less subjective difficulty in the task performance, as already noticed in the literature (Kosslyn et al., 1984). This is clear in Table 5, where the performance in the prediction of the laterality varies strongly across subjects. One can also notice that in O'Craven and Kanwisher (2000), mental imagery activations were decoded in three out of eight subjects. In view of this, the good performance achieved in five hemispheres is thus an important result (if responses were random, the probability of obtaining significant values in five hemispheres would be  $P < 0.00043$ ). The explicit identification of the imagined pattern is a challenging task, due to the weak signals that are obtained (see Fig. 6). Note that, besides the well-known weakness of retinotopic activations in imagery experiments (Klein et al., 2004), the subjects had to keep their eyes open and fixate the grid during this experiment, which might have reduced the level of activation in primary visual areas.

### Technical aspects

Care should be taken when evoking fMRI-based brain reading experiments. In particular, any method rests on a deconvolution of the hemodynamic responses on a voxel-by-voxel basis. We have performed this using a standard GLM procedure, which is reasonable given that the temporal linearity hypothesis might be fulfilled given our inter-stimulus intervals (6 s) (Boynton et al., 1996; Soltysik et al., 2004). One might think, however, that unmodeled spatio/temporal interactions may be present in the data.

Another set of simplifications were introduced in our formulation of the forward model: linearity of the transfer operator, isotropic Gaussian receptive field (RF) structure at the voxel level, constant RF size, linear gain. While this model might be partially supported by current knowledge about V1 (Tootell et al., 1998b; Hansen et al., 2004), it is obviously an over-simplification (Olshausen and Field, 2005), especially if one considers higher visual areas. However, it is important to keep in mind that fMRI signals represent in each voxel the average of the activity of thousands of neurons, so that some hypotheses, e.g. the spatial linearity or superposition principle used here, that are known to be violated at a microscopic level, may hold approximately at the much lower spatial resolution and/or using standard field strength. Clearly, our forward problem framework may be a good benchmark to test violations of different hypotheses, e.g. the spatial linearity of visual activity as seen in fMRI (see e.g. Shmuel et al., 2002, 2006). For instance, we have also implemented Mexican hat filters (Laplacian of the Gaussian), but did not find significant





improvements in the results. Our interpretation is that the main bottleneck in the forward/inverse problem is the correctness of the estimate of  $p_v$  in each voxel (see Eq. (2) and Appendix A.2). Assuming that this parameter is perfectly known, modeling non-linear effects or non-isotropic and non-Gaussian Receptive fields would become worthwhile.

We found that a majority (50 to 60%) of the most discriminant voxels used in the classifier were in V1, while a much smaller proportion (around 20%) were in V2. Since this delineation was not the primary goal of our analysis, we were not sure to obtain reliable boundaries for other visual areas and did make further identification of the discriminative information. One could indeed expect that most of the information on the spatial layout of the stimulus would be encoded in V1. In the Supplementary material, we show that performing the inverse reconstruction based only on V1 voxels yields a very good approximation of what is achieved when considering all the retinotopic voxels (see Supplementary Fig. 2 and Table 3).

The inverse problem is also limited by the possibility of evaluating correctly the precision of the results (see e.g. Eq. (12)). Approximations must be performed, so that the probabilistic interpretation of the visual patterns is not fully assessed. For this reason, we based our test procedure on the correlation of the reconstructed pattern with the possible candidates, rather than on a fully probabilistic interpretation of the reconstructed maps.

#### Power and limits of inverse retinotopy

First, it should be noticed that the retinotopic mapping was performed in less than 20 min in each subject, so that the limited accuracy of the retinotopic information may be the main limit in this experiment.

Not unexpectedly, many confusions occurred between spatial patterns that overlapped. For instance, the -shaped and -shaped patterns in Fig. 1(c) were often confounded. Moreover, the -shaped pattern was rarely identified in the framework of the inverse problem: it is interesting to note that this pattern is the least compact. This might be attributable to the low-pass filtering inherent to the inverse problem. This effect is particularly evident from Fig. 5 and Supplementary Fig. 1, and affects the results at the group level (see Supplementary Table 2). Interestingly, the -shaped stimulus was not confused with other patterns when using classification tools. Thus the problem described here might be an intrinsic shortcoming of the forward/inverse problem solution. Another important effect is that the portion of the stimulus closest to the center of the visual field is apparently much better reconstructed than the activity in the peripheral regions, which is often smoothed out in Fig. 5—and similarly for all the data sets studied. This weakness is apparently not related to the receptive field size model (see Fig. 4), and might be related to spatial/attentional modulations. For instance, the fact that left and right patterns were presented simultaneously facilitates fixation, but possibly increases foveal attention at the expense of the periphery. Note that this foveal emphasis effect is apparently also present in the imagery data (see Fig. 6).

#### Conclusion and future work

We have presented an inverse retinotopy framework that builds on two complementary points of view: an explicit inverse reconstruction approach that builds on the knowledge of a retinotopic experiment to decode any activation image projected

on early visual areas, and an SVM-based classification approach that best classifies images into a discrete categories. We could partly extrapolate these models from passive viewing to mental imagery experiments, confirming the retinotopic nature of imagery activations. Future work might concentrate on intermediate approaches where the forward/inverse problem could benefit from the support vector framework. Long-term research might address more realistic simulations of activation phenomena in the visual cortex, making the forward model more realistic.

#### Acknowledgment

We are very thankful to Jean Lorenceau, Laboratoire de Neurosciences Cognitives et Imagerie Cérébrale LENA-CNRS UPR 640, who helped us to generate the stimuli that we used in the Jeda environment that he has designed, as well as for his kind and insightful advice on retinotopic stimulation.

#### Appendix A

##### A.1. Classification of functional images with retinotopically-specific information

Let  $\phi^1, \dots, \phi^n$  be a set of brain activation images associated with a set of known stimuli  $\sigma^1, \dots, \sigma^n$ . Any classifier proceeds by learning to discriminate between the images associated with a given label  $\Phi_s = \{\phi^i | \sigma^i = s\}$  and the other images. Each image is defined by its values on a number  $V$  of retinotopically specific voxels.

##### A.1.1. Multivariate analysis and the risk of overfit

The discrimination function is efficient if it combines the information from many of these voxels: classifiers are thus inherently multivariate methods. Henceforth, we call *feature* a voxel-based information, *sample* an image of the learning or test set and *label* the indicator of the stimulus associated with an image. Many features, i.e. non-retinotopically specific voxels, do not carry any discriminant information, and thus they do not improve the classifier performance. When the proportion of such useless features increases, some of them are simply correlated with the associated label within the training set by chance, and their information cannot generalize to another set of samples (test samples). Those *fake positives* may dramatically harm the classifier performance. This problem, known as the *curse of dimensionality*, is illustrated in Fig. 7. The number  $N$  of features increases along the horizontal axis, so that the ratio  $\frac{n}{N}$  decreases: the training space becomes sparser. As shown by the blue line, the classifier rapidly reaches 100% of correct recognition on the training samples. In parallel, the performance of the classifier on an independent test set of images increases until it reaches a maximum of 86% recognition for  $N=150$  as shown by the red line. Then it starts to decrease as  $N$  further increases.

To overcome this problem, we first apply a feature selection (described in Appendix A.1.2); the selected features are then given as input to a linear Support Vector Machines (SVM) classifier, presented in Appendix A.1.3.

##### A.1.2. Classification step one: feature (voxels) selection

Feature selection is a crucial step of classification: it improves the generalization power of a classifier and it is also useful to select

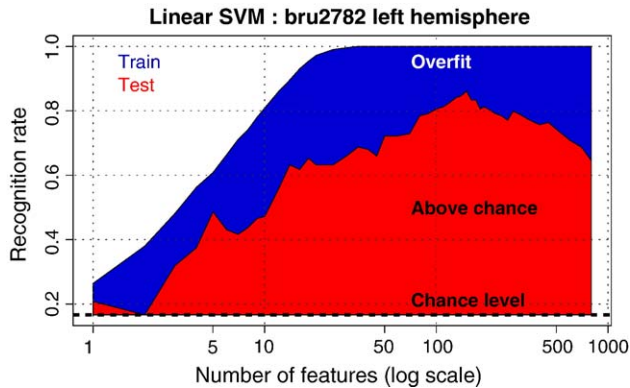


Fig. 7. Recognition rate (evaluated by a Leave-One-Out scheme) as a function of the number of input features (voxels) for the subject bru2782, left hemisphere. We measure the recognition rate of the classifier on the training set and on an independent test set when the number  $N$  of input features varies from 1 to 800.

a small subset of discriminant features, which is a requirement to interpret results in biological applications such as neuroscience or even genomics. Among feature selection methods (Guyon and Elisseeff, 2003), we choose a supervised univariate method based on the computation of an ANOVA  $F$  and  $P$ -values. This simple feature selection approach belongs to the family of methods called *filters*. Filters are supervised univariate methods that rank features independently of the context of others features, according to their ability to separate the populations. Such methods are computationally efficient, which makes them tractable even on thousands of features as in our case. Unlike PCA, filters select the features in the original feature space which eases the interpretation in terms of discriminant information; moreover, these methods are less prone to overfitting than multivariate selection methods in general.

We first perform an ANOVA to compute to which extent the features are label-related; this yields an  $F$  and a  $P$ -value. We select the features with two different methods for the control of false positives.

In the first case, we select significant features ( $P < 0.05$ ) after a Bonferroni correction. When doing so, we have a strong control of the type I error (the number of false positives) selecting few but reliable voxels which minimize the risk of overfit. We use this very stringent method in difficult problems like mental imagery (Mental imagery: identification of trial-specific patterns through classification).

In the case of visual stimulation (trial-specific images of the domino experiment, see the results in Classification of the trial-specific images), the risk of overfit is lower. Hence we want to reduce the type II error in order to grab more discriminant features as input of the classifier. Thus we select significant features ( $P < 0.1$ ) after a False Discovery Rate (FDR) (Benjamini and Hochberg, 1995) correction.

#### A.1.3. Classification step two: linear SVM

Support Vector Machines (SVMs) (Schölkopf and Smola, 2002) have recently been successfully used in fMRI applications (Cox and Savoy, 2003; LaConte et al., 2005; Kamitani and Tong, 2005). Briefly speaking SVMs build their discriminant model as a linear combination of critical training samples. Those samples called *Support Vectors* (SVs) are either samples that lie close to the boundary of the two classes or samples that cannot be

correctly classified. The success of SVMs on real data may be explained by their design which properly deals with few samples in high dimensional spaces: in a  $N$ -dimensional space with  $n$  samples, SVMs are fully parameterized with  $n+1$  parameters, while e.g. Linear Discriminant Analysis requires the estimation of  $N(N+3)/2$  parameters. This simple fact may explain the good behavior of SVMs in high dimensional spaces. Another argument is that the SVM model enhances the *parsimony* of the discriminant model: SVMs not only attempt to perform a good classification of the training samples, as a perceptron algorithm does, but also constrain the discriminant model to be as simple as possible, i.e. a model in which the number of SVs is minimal. The choice of a linear SVM instead of a radial SVM (Schölkopf and Smola, 2002) has been done after simple experiments conducted on one of the subjects (bru2782, left hemisphere), without any feature selection procedure. The linear SVM reaches 62% of correct classification while the radial SVM only reaches 22%. It is noticeable that the superiority of linear SVM has also been reported in Cox and Savoy (2003). The cost parameter (Schölkopf and Smola, 2002) of the linear SVM has been set to 1 and the implementation comes from LIBSVM (<http://www.csie.ntu.edu.tw/~cjlin/libsvm>).

#### A.1.4. Validation

Validation is a simple but crucial point that must be carefully conducted in order to assess the quality of a discriminant model without any methodological bias. The classical way is to perform an out-of-sample validation which consists of: (i) setting aside an independent set of subjects (the test set), (ii) learning on the remaining subjects (the training set) and (iii) testing the discriminant model on the test set. Cross-validation or bootstrap validation repeats the previous procedure and averages the errors on test sets. The limit case of cross-validation is the Leave-One-Out procedure where only one subject is set aside. It should be noticed that feature selection is the first step of the discriminant model, thus it must be performed within the cross-validation loop, only on the training samples, and not as a pre-processing on all samples before the cross-validation.

#### A.2. Analysis of the retinotopic maps: estimation of $p_v$

We detail here how the voxel-based time courses during the retinotopy experiment can be used to infer the center  $p_v$  of the receptive field associated with any voxel  $v$ . Note that this is the standard procedure used to analyze retinotopic mapping experiments (see e.g. Sereno et al., 1995).

Let  $\phi^i(v)$ ,  $i=1:n$  be the values of the fMR images at voxel  $v$  during one retinotopic mapping session (e.g. the clockwise rotating wedge). The stimulation is periodic with a period  $T_0=32$  s, i.e. a rotation speed  $\omega_0 = \frac{2\pi}{T_0}$ . Fourier analysis of  $\phi^i(v)$  yields

$$\phi^i(v) = A(v)\cos(\omega_0 i - \theta(v)) + w^i \quad (16)$$

where  $A(v) > 0$  is an estimate of the amount of activity at frequency  $\frac{\omega_0}{2\pi}$  in  $\phi^i(v)$ ,  $\theta_v$  the phase associated with voxel  $v$  and  $w$  the unmodeled signal. Note that the estimation of  $(A(v), \theta(v))$  can be cast in the framework of the general linear model since Eq. (16) is equivalent to

$$\phi^i(v) = A^1(v)\cos(\omega_0 i) + A^2(v)\sin(\omega_0 i) + w^i \quad (17)$$

where  $A^1(v) = A(v) \cos(\theta(v))$  and  $A^2(v) = A(v) \sin(\theta(v))$ . This enables us to use standard high-pass filtering and AR(1) residual whitening that are commonly used in fMRI data analysis (Ashburner et al., 2004). Note that Eqs. (16)–(17) rest on an approximation in which only the fundamental frequency of the stimulus is considered. Although we have also used more complete models of the retinotopic activity, there was little difference regarding the estimation of  $\theta(v)$ , which is ultimately the parameter of interest.

Before turning to the analysis of the phase information, let us first notice that the assessment through a standard Fisher statistic that  $\|A(v)\|^2 > 0$  yields retinotopically specific voxels.

Then  $\theta(v) \in [-\pi, \pi]$  measures the delay of  $\phi^v$  with respect to the activity in a reference region, which is equal to the polar angle between  $p_v$  and the reference direction, biased by a hemodynamic offset. This offset is nicely canceled out by averaging the value of  $\theta(v)$  across both directions of phase change. We end up with an estimation of the polar angle  $\alpha(v) \in [-\pi, \pi]$  of  $p_v$ . Similarly, the analysis of expanding/contracting ring stimulus provides us a phase eccentricity value  $\theta_v \in [-\pi, \pi]$ , which is not biased by the hemodynamic delay. The latter value is converted to a physical eccentricity  $\zeta(v) = \frac{\delta}{2\pi}(\pi + \theta(v))$ , where  $\delta = 9.5^\circ$  is the maximal visual eccentricity achieved by the stimulus.

The values define uniquely an estimate  $\hat{p}_v = \begin{pmatrix} \zeta(v)\cos(\alpha(v)) \\ \zeta(v)\sin(\alpha(v)) \end{pmatrix}$  of  $p_v$ .

## Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.neuroimage.2006.06.062](https://doi.org/10.1016/j.neuroimage.2006.06.062).

## References

- Ashburner, J., Friston, K., Penny, W., 2004. Human Brain Function, 2nd Ed. Academic Press.
- Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* 57 (1), 289–300.
- Boynton, G.M., Engel, S.A., Glover, G.H., Heeger, D.J., 1996. Linear systems analysis of functional magnetic resonance imaging in human V1. *J. Neurosci.* 16, 4207–4221.
- Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. *J. Cogn. Neurosci.* 15 (5), 704–717.
- Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) “brain reading”: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19 (2), 261–270.
- Davatzikos, C., Ruparel, K., Fan, Y., Shen, D.G., Acharyya, M., Loughead, J.W., Gur, R.C., Langleben, D.D., 2005. Classifying spatial patterns of brain activity with machine learning methods: application to lie detection. *NeuroImage* 28 (3), 663–668.
- Dehaene, S., Clec'H, G.L., Cohen, L., Poline, J.-B., de Moortele, P.-F.V., Bihan, D.L., 1998. Inferring behavior from functional images. *Nat. Neurosci.* 1 (7), 549–550.
- DeYoe, E.A., Carman, G.J., P.B., Bandettini, P., Glickman, S., Wieser, J., Cox, R., Miller, D., Neitz, J., 1996. Mapping striate and extrastriate visual areas in human cerebral cortex. *Proc. Natl. Acad. Sci. U S A* 93 (6), 2382–2386.
- Dougherty, R.F., Koch, V.M., Brewer, A.A., Fischer, B., Modersitzki, J., Wandell, B.A., 2003. Visual field representations and locations of visual areas V1/2/3 in human visual cortex. *J. Vis.* 3 (10), 586–598.
- Duncan, R.O., Boynton, G.M., 2003. Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron* 38 (4), 659–671.
- Engel, S.A., Glover, G.H., Wandell, B.A., 1997. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex* 7 (2), 181–192.
- Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.* 3, 1157–1182 (Special issue on variable and feature).
- Hansen, K.A., David, S.V., Gallant, J.L., 2004. Parametric reverse correlation reveals spatial linearity of retinotopic human V1 BOLD response. *NeuroImage* 23 (1), 233–241.
- Hanson, S.J., Matsuka, T., Haxby, J.V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a “face” area? *NeuroImage* 23 (1), 156–166.
- Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293 (5539), 2425–2430.
- Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.* 8, 686–691.
- Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8, 679–685.
- Klein, I., Dubois, J., Mangin, J.-F., Kherif, F., Flandin, G., Poline, J.-B., Denis, M., Kosslyn, S.M., Bihan, D.L., 2004. Retinotopic organization of visual mental images as revealed by functional magnetic resonance imaging. *Brain Res. Cogn. Brain Res.* 22 (1), 26–31.
- Kosslyn, S.M., Brunn, J., Cave, K.R., Wallach, R.W., 1984. Individual differences in mental imagery ability: a computational analysis. *Cognition* 18 (1–3), 195–243.
- Kosslyn, S.M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J.P., Thompson, W.L., Ganis, G., Sukel, K.E., Alpert, N.M., 1999. The role of area 17 in visual imagery: convergent evidence from PET and rTMS. *Science* 284 (5411), 167–170.
- LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. *NeuroImage* 26 (2), 317–329.
- O’Craven, K.M., Kanwisher, N., 2000. Mental imagery of faces and places activates corresponding stimulus-specific brain regions. *J. Cogn. Neurosci.* 12 (6), 1013–1023.
- Olshausen, B.A., Field, D.J., 2005. How close are we to understanding V1? *Neural Comput.* 17 (8), 1665–1699.
- Piazza, M., Giacomini, E., Le Bihan, D., Dehaene, S., 2003. Single-trial classification of parallel pre-attentive and serial processes using functional magnetic resonance imaging. *Phil. Trans. R. Soc. Lond. B.* 270, 1237–1245.
- Rivière, D., Papadopoulos-Orfanos, D., Poupon, C., Poupon, F., Coulon, O., Poline, J.-B., Frouin, V., Régis, J., Mangin, J.-F., 2000. A structural browser for human brain mapping. In *Proc. 6th HBM. NeuroImage* 11 (5), 912 (San Antonio, Texas).
- Schölkopf, B., Smola, A., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA.
- Sereno, M.I., Dale, A.M., Reppas, J.B., Kwong, K.K., Belliveau, J.W., Rosen, B.R., Tootell, R.B., 1995. Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science* 268, 889–893.
- Shmuel, A., Yacoub, E., Pfeuffer, J., de Moortele, P.F.V., Adriany, G., Hu, X., Ugurbil, K., 2002. Sustained negative BOLD, blood flow and oxygen consumption response and its coupling to the positive response in the human brain. *Neuron* 36 (6), 1195–1210.
- Shmuel, A., Augath, M., Oeltermann, A., Logothetis, N.K., 2006. Negative functional MRI response correlates with decreases in neuronal activity in monkey visual area V1. *Nat. Neurosci.* 9 (4), 569–577.
- Slotnick, S.D., Thompson, W.L., Kosslyn, S.M., 2005. Visual mental

- imagery induces retinotopically organized activation of early visual areas. *Cereb. Cortex* 15 (10), 1570–1583.
- Smith, A.T., Singh, K.D., Williams, A.L., Greenlee, M.W., 2001. Estimating receptive field size from fMRI data in human striate and extrastriate visual cortex. *Cereb. Cortex* 11 (12), 1182–1190.
- Soltysik, D.A., Peck, K.K., White, K.D., Crosson, B., Briggs, R.W., 2004. Comparison of hemodynamic response nonlinearity across primary cortical areas. *NeuroImage* 22 (3), 1117–1127.
- Tootell, R.B., Dale, A., Sereno, M., Malach, R., 1996. New images from human visual cortex. *Trends Neurosci.* 19, 481–489.
- Tootell, R.B.H., Hadjikhani, N., Mendola, J.D., Marett, S., Dale, A.M., 1998a. From retinotopy to recognition: fMRI in human visual cortex. *Trends Cogn. Sci.* 2, 174–183.
- Tootell, R.B., Hadjikhani, N.K., Vanduffel, W., Liu, A.K., Mendola, J.D., Sereno, M.I., Dale, A.M., 1998b. Functional analysis of primary visual cortex (V1) in humans. *Proc. Natl. Acad. Sci. U. S. A.* 95 (3), 811–817.
- Tootell, R.B., Tsao, D., Vanduffel, W., 2003. Neuroimaging weighs in: humans meet macaques in “primate” visual cortex. *J. Neurosci.* 23, 3981–3989.
- Vanni, S., Henriksson, L., James, A.C., 2005. Multifocal fMRI mapping of visual cortical areas. *NeuroImage* 27 (1), 95–105.
- Warnking, J., Dojat, M., Guerin-Dugue, A., Delon-Martin, C., Olympieff, S., Richard, N., Chéhikian, A., Segebarth, C., 2002. fMRI retinotopic mapping—step by step. *NeuroImage* 17, 1665–1683.
- Wotawa, N., Thirion, B., Castet, E., Anton, J.-L., Faugeras, O., 2005. Human retinotopic mapping using fMRI. Technical Report 5472, INRIA Sophia-Antipolis.