Visual image reconstruction from human brain activity: A modular decoding approach

To cite this article: Yoichi Miyawaki et al 2009 J. Phys.: Conf. Ser. 197 012021

View the article online for updates and enhancements.

Related content
- Infrared Imaging: Thermographic monitoring of blood flow in arterio-venous fistulas
  F Ring, A Jung and J uber
- Uniform discretizations: a quantization procedure for totally constrained systems including gravity
  Miguel Campiglia, Cayetano Di Bartolo, Rodolfo Gambini et al.
- EIT of human brain activity
  A T Tidswell, A Gibson, R H Bayford et al.

Recent citations
- Spreading Codes Enables the Blind Estimation of the Hemodynamic Response with Short-Events Sequences
  M. A. Lopez-Gordo et al.
Visual image reconstruction from human brain activity: A modular decoding approach

Yoichi Miyawaki\textsuperscript{1,2}, Hajime Uchida\textsuperscript{3,2}, Okito Yamashita\textsuperscript{2}, Masa-aki Sato\textsuperscript{2}, Yusuke Morito\textsuperscript{4,5}, Hiroki C Tanabe\textsuperscript{5,4}, Norihiro Sadato\textsuperscript{5,4} and Yukiyasu Kamitani\textsuperscript{2,3}

\textsuperscript{1}National Institute of Information and Communications Technology, 2-2-2, Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan.
\textsuperscript{2}ATR Computational Neuroscience Laboratories, 2-2-2, Hikaridai, Seika, Soraku, Kyoto 619-0288, Kyoto, Japan.
\textsuperscript{3}Nara Institute of Science and Technology, 8916-5, Takayamacho, Ikoma, Nara 630-0192, Japan.
\textsuperscript{4}The Graduate University for Advanced Studies, Hayamacho, Miura, Kanagawa 240-0193, Japan.
\textsuperscript{5}National Institute for Physiological Sciences, 38, Nishigonaka, Myodaiji, Okazaki, Aichi 444-8585, Japan

E-mail: kmtn@atr.jp

Abstract. Brain activity represents our perceptual experience. But the potential for reading out perceptual contents from human brain activity has not been fully explored. In this study, we demonstrate constraint-free reconstruction of visual images perceived by a subject, from the brain activity pattern. We reconstructed visual images by combining local image bases with multiple scales, whose contrasts were independently decoded from fMRI activity by automatically selecting relevant voxels and exploiting their correlated patterns. Binary-contrast, 10 x 10-patch images (2\textsuperscript{100} possible states), were accurately reconstructed without any image prior by measuring brain activity only for several hundred random images. The results suggest that our approach provides an effective means to read out complex perceptual states from brain activity while discovering information representation in multi-voxel patterns.

1. Introduction
Prediction of perceptual contents from brain activity represents a major challenge in human neuroimaging. Previous fMRI studies have shown that visual features, such as orientation and motion direction [1,2], and visual object categories [3,4] perceived by a subject can be predicted from fMRI activity patterns by a statistical ‘decoder,’ which learns the relationship between brain activity patterns and pre-specified stimulus categories. However, such a simple classification approach is insufficient to capture the complexity of perceptual experience, since our perception consists of numerous possible states, and it is impractical to measure brain activity for all possible states. A recent study [5] has demonstrated that a presented image can be identified among a large number of candidate images using a receptive field model that predicted fMRI activity for visual images (see also [6] for a related approach). But the image identification was still constrained by the candidate image set. Even more challenging is visual image reconstruction (see the earlier work [7] using cat LGN spikes for...
reconstruction), which translates brain activity into an image, free from the pre-specified stimulus categories.

A possible approach is to utilize the retinotopy in the early visual cortex. Retinotopy associates a specific visual field location to an active cortical location, or voxel, providing a mapping from the visual field to the cortical voxels. Thus, one may predict local contrast information by monitoring the fMRI signals corresponding to the retinotopy map of the target visual field location. The retinotopy can be further elaborated using a voxel receptive-field model. By inverting the receptive-field model, a presented image can be inferred given the brain activity consistent with the retinotopic map [9].

However, it may not be optimal to use retinotopy or the inverse of the receptive field model to predict local contrast in an image. These methods are based on the model of individual voxel responses given a visual stimulus, and multi-voxel patterns are not taken into account for the prediction of the visual stimulus. Recent studies have demonstrated the importance of the activity pattern, in particular the correlation among neurons or cortical locations in the decoding of a stimulus [10, 11]. Since even a localized small visual stimulus elicits spatially-spread activity over multiple cortical voxels, multi-voxel patterns may contain information useful for predicting the presented stimulus.

In addition, a visual image is thought to be represented at multiple spatial scales in the visual cortex, which may serve to retain the visual sensitivity to fine-to-coarse patterns at a single visual field location [12, 13]. The conventional retinotopy, by contrast, does not imply such multi-scale representation, as it simply posits a location-to-location mapping. It may be possible to extract multi-scale information from fMRI signals and use it to achieve better reconstruction.

---

**Figure 1.** Reconstruction procedure. fMRI activity is measured while a contrast-defined 10 x 10 patch image is presented. ‘Local decoders’ use linearly-weighted multi-voxel fMRI signals (voxel weights, w_i, w_j, …) to predict the contrasts (contrast values, C_i, C_j, …) of ‘local image bases (or elements)’ of multiple scales (1 x 1, 1 x 2, 2 x 1, and 2 x 2 patch areas, defined by colored rectangles). Local image bases are multiplied by the predicted contrasts, and linearly combined using ‘combination coefficients’ (λ_i, λ_j, …) to reconstruct the image. Contrast patterns of the reconstructed images are depicted by a gray scale. Image bases of the same scale (except the 1 x 1 scale) partially overlapped with each other, though the figure displays only non-overlapping bases for the purpose of illustration. Modified from [8].
In this paper, we summarize our recent work on visual image reconstruction (see [8] for full details). We present an approach to visual image reconstruction using multi-voxel patterns of fMRI signals and multi-scale visual representation (figure 1). We assume that an image is represented by a linear combination of local image elements of multiple scales, whose contrasts are independently predicted from multi-voxel patterns. As each local element has fewer possible states than the entire image, the training of local decoders requires only a small number of training samples. Hence, each local decoder serves as a ‘module’ for a simple image component, and the combination of the modular decoders allows us to represent numerous variations of complex images. The decoder is not explicitly informed about the retinotopy mapping, but automatically selects relevant voxels from the visual areas. We show that arbitrary visual images defined by 10 x 10 binary contrast (2100 variations), including geometric and alphabet letter shapes, can be reconstructed accurately, using only several hundreds of images for training the model. Quantitative analyses reveal that the multi-voxel pattern decoder, which exploits voxel correlations especially in V1, and the multi-scale reconstruction model both significantly contribute to the high quality of reconstruction.

2. Methods

2.1. Subjects
We first screened four subjects for head motion in preliminary scans, and two of them (male adults with normal or corrected-to-normal visual acuity) who showed the least head motion underwent the full experimental procedure. The subjects gave written informed consent. The study was approved by the Ethics Committee of ATR and National Institute for Physiological Sciences.

2.2. Visual stimulus and experimental design
We had three types of experimental sessions to measure the fMRI responses of the visual cortex: 1) the random image session, 2) the figure image session, and 3) the conventional retinotopy mapping session.

In the random image session, subjects observed images consisting of 12 x 12 small square patches (1.15 x 1.15 deg each) presented on a gray background. Each patch was either a flickering checkerboard (spatial frequency, 1.74 cycles/deg; temporal frequency, 6 Hz) or a homogeneous gray area, with equal probability. Each image presentation lasted for 6 s, followed by a 6-s rest period. To avoid the effects of the stimulus frame, the central 10 x 10 area was used for analysis. A total of 440 different random patterns were presented to each subject.

In the figure image session, stimulus images consisted of flickering checkerboard patches as in the random image session, but formed geometric shapes (‘square,’ ‘small frame,’ ‘large frame,’ ‘plus,’ and ‘X’) or alphabet letters (‘N,’ ‘E,’ ‘U,’ ‘R,’ and ‘O’). Each image was presented for 12 s followed by a 12-s rest period. A total of 6 -- 8 trials per stimulus were performed.

The retinotopy mapping session followed the conventional procedure [14] using a rotating wedge and an expanding ring of flickering checkerboard. The data were used to delineate the borders between visual cortical areas and to identify the retinotopy map on the flattened cortical surfaces. We used fMRI signals from areas V1 and V2 for analyses (unless otherwise stated).

Throughout these experiments, subjects were requested to view the stimulus sequence while maintaining fixation on a spot presented at the center of the screen.

2.3. MRI acquisition and data preprocessing
Preliminary experiments were performed using 3.0-Tesla Siemens MAGNETOM Allegra located at National Institute for Physiological Sciences. MRI data for the presented results were all obtained using a 3.0-Tesla Siemens MAGNETOM Trio A Tim scanner located at the ATR Brain Activity Imaging Center. We measured 30 slices of functional images around the visual cortex (TR, 2 s; voxel size, 3 cubic mm).
The acquired fMRI data were preprocessed and coregistered to the high-resolution anatomical image with standard procedures using SPM2. After voxels of extremely low signal amplitudes were removed, the fMRI data underwent linear trend removal, amplitude normalization, baseline correction, hemodynamic delay compensation (4-s data shift), and within-stimulus-block averaging. See [8] for details.

2.4. Labelling of fMRI data
Each fMRI data sample was labeled by the mean contrast values of local image elements in the corresponding stimulus image. Local image elements were 1 x 1, 1 x 2, 2 x 1, and 2 x 2 patch areas covering the entire 10 x 10 patch area with overlaps (a total of 361 elements; 1 x 1, 100; 1 x 2, 90; 2 x 1, 90; 2 x 2, 81). The mean contrast value of each local image element was defined as the number of flickering patches divided by the total number of patches (1 x 1, [0 or 1]; 1 x 2 and 2 x 1, [0, 0.5, or 1]; 2 x 2, [0, 0.25, 0.5, 0.75, or 1]).

2.5. Training of local decoders
Local decoders were defined to predict the mean contrast of each local image element. They were individually trained with fMRI data and the corresponding class labels representing the mean contrast values. Each local decoder consisted of a multi-class classifier, which classified fMRI data samples into the classes defined by the mean contrast values.

Our classification model is based on multinomial logistic regression, in which each contrast class has a linear discriminant function that calculates the weighted sum of the inputs (voxel values). The discriminant function for contrast class \( k \) in a local decoder is expressed as,

\[
y_{w_k}(r) = \sum_{d} w^d_k r^d + w_0^k,
\]

where \( w^d_k \) is a weight parameter for voxel \( d \) and contrast class \( k \), \( r^d \) is the fMRI signal of voxel \( d \), \( w_0^k \) is the bias, and \( D \) is the number of voxels. The probability that an fMRI signal pattern \( r = [r^1, r^2, \ldots, r^D] \) belongs to the contrast class \( k \) is defined using the softmax function,

\[
p_v(k \mid r) = \frac{\exp[y_{w_k}(r)]}{\sum_j \exp[y_{w_j}(r)]},
\]

where \( K \) is the number of the contrast classes. The predicted contrast class for \( m \)-th local image element, \( C_m(r) \), is chosen as the contrast class with the highest probability.

In conventional multinomial logistic regression, the weight parameters are determined by finding the values that maximize the likelihood function of the weight parameters given a training data set,

\[
p_v(S \mid w_1, \ldots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p_v(k \mid r_n)^{s_{nk}},
\]

where \( S \) represents a class label matrix whose element \( s_{nk} \) is 1 if the trial \( n \) corresponds to the contrast class \( k \) otherwise 0, \( w_k \) is the weight vector for contrast class \( k \) including the bias term ((\( D+1 \))x1 vector), and \( N \) is the number of trials.

In this study, we adopted a full-Bayesian approach to the estimation of weight parameters (‘sparse logistic regression’ [15]). The above likelihood function was combined with a prior distribution for each weight to obtain the posterior distribution. Weight parameters were estimated by taking the expectation of the posterior distribution for each weight.

The prior distribution of a weight parameter is described by a zero-mean normal distribution with a variance, whose inverse is treated as a hyperparameter,
\[ p(w^d_k | \alpha^d_k) = N \left( 0, \frac{1}{\alpha^d_k} \right), \]

where \( N \) represents a normal distribution, and \( \alpha^d_k \) is the hyperparameter denoting the inverse of the variance, or precision, of the weight value for voxel \( d \) and contrast class \( k \). The hyperparameter \( \alpha^d_k \) is also treated as a random variable, whose distribution is defined by,

\[ p(\alpha^d_k) = \frac{1}{\alpha^d_k}. \]

These prior distributions are known to lead to ‘sparse estimation’ in which only a small number of parameters have non-zero values and the remaining parameters are estimated to be zero [16]. Thus, the prior distributions implement the assumption that only a small number of voxels are relevant for the decoding of each local image element. This sparseness assumption may be validated by the fact that a spatially localized visual stimulus gives rise to neural activity only in small regions of the early visual cortex. The sparse parameter estimation avoids overfitting to noisy training data by pruning irrelevant voxels, thereby helping to achieve high generalization performance [15].

Since the direct evaluation of the posterior distribution is analytically intractable, we used a variational Bayesian method to approximate the distribution. The algorithm for the parameter estimation is presented in [8, 15].

2.6. Combination of local decoders

The outputs of the local decoders were combined by a linear model of the corresponding local image elements,

\[ \hat{I}(x | r) = \sum_{m} \lambda_{m} C_{m}(r) \phi_{m}(x), \]

where \( \phi_{m}(x) \) represents a local image element, or a basis, \((\phi_{m}(x) = 1 \text{ if location } x \text{ is contained in the area of the local image element, otherwise } \phi_{m}(x) = 0, C_{m}(r) \) is the predicted contrast, and \( \lambda_{m} \) is the combination coefficient. Combination coefficients, \( \lambda_{m} \), were determined so as to minimize square errors of reconstructed images for the training data set.

2.7. Evaluation of performance

The trained reconstruction model was tested with independent samples. We performed two types of reconstruction tests. First, to obtain a quantitative and unbiased evaluation, we conducted cross-validation analysis using the samples in the random image session. Second, to illustrate the quality of reconstructed images, the model obtained from the random image session was used to reconstruct the images presented in the figure image session.

3. Results

3.1. Reconstructed visual images

Reconstructed images from all trials of the figure image session are illustrated in figure 2. Even though the geometric and alphabet shapes were not used for the training of the reconstruction model, the reconstructed images reveal essential features of the original shapes.

We also found that reconstruction was possible even from 2-s single-volume data without block averaging. The results show the temporal evolution of volume-by-volume reconstruction including the rest periods. All reconstruction sequences are shown as a movie data. The movie data can be found on [17].

In the following sections, we examine how multi-voxel patterns and multi-scale image representation contributed to the high reconstruction performance.
3.2. Weight distribution on the cortical surface
We first examined the distributions of voxel weights of local decoders in comparison with the conventional retinotopy. Cortical surface maps show the distributions of weight magnitudes for a foveal and a peripheral patch (figure 3). The largest weight values are found around the cortical locations consistent with the retinotopic representation of the patch locations, showing that local decoders mainly used voxels corresponding to the retinotopic locations for their target patches. The weight distribution tended to be blurred for peripheral patches, suggesting that peripheral decoders failed to select retinotopic voxels.

3.3. Advantage of multi-voxel pattern decoders
To examine whether our local decoders exploit multi-voxel patterns for the prediction of target contrast, we devised other types of local decoders that only used retinotopic voxels (‘retinotopic decoders’). By applying the standard general linear model to the data from the random image session, we identified a single voxel with the highest t-value, or a group of significantly responsive voxels (p < 0.05, false discovery rate (FDR) corrected for multiple comparisons) for each patch. We used 1) the most responsive voxel, and 2) the average of the significantly responsive voxels. The decoders consisted of the standard univariate logistic regression model. The performance of these decoders was compared with that of the multi-voxel pattern decoder.

Cross-validation analysis using the random image trials revealed that the multi-voxel pattern decoder achieved significantly higher correct rates than either of the two retinotopic decoders (two-way ANOVA, Bonferroni-corrected p < 0.05 for multiple comparisons), while the difference gradually diminished at the periphery approaching the chance level (figure 4a). Although the figure illustrates the performance only for the 1 x 1 scale, the decoders of other scales showed similar results. The number of the significantly responsive voxels was larger than the number of the voxels selected by the multi-voxel pattern decoder for the foveal to middle eccentricity. Since in this range of eccentricity the multi-voxel pattern decoder largely outperformed the retinotopic decoders, the higher performance of
The multi-voxel pattern decoder is not merely due to noise reduction by pooling multi-voxel signals. These results indicate that our local decoders did not simply depend on the mapping between a cortical location and a stimulus location, but that they effectively exploited multi-voxel patterns.

One of the key features of multi-voxel patterns is the correlation between voxels. To examine how voxel correlations contribute to decoding accuracy, we trained the decoder with fMRI data in which voxel correlations were removed, and compared the performance with that of the original decoder. The data were created by shuffling the order of the trials with the same stimulus label in each voxel [10]. This shuffling procedure removes voxel correlations that are independent of the stimulus label. Note that since the stimuli were random images, the voxel correlations observed in the original training data do not reflect the correlations between stimulus patches. The trained decoder was tested with independent non-shuffled data.

The performance with shuffled data was significantly lower than that with the original data (two-way ANOVA, p < 0.05), particularly at the middle range of eccentricity (figure 4b). The results suggest that the multi-voxel pattern decoder makes effective use of voxel correlation to achieve high decoding performance.

3.4. Reconstruction using individual visual areas

We next compared the reconstruction between individual visual areas by using the voxels in each of V1, V2, and V3 as the input. As illustrated in figure 5a, reconstruction quality progressively deteriorated along the visual cortical hierarchy. Quantitative comparison was performed by calculating the reconstruction errors for the images from the random image session (squared difference between the presented and the reconstructed contrast in each patch averaged over each entire image). Higher visual areas showed significantly larger errors than V1 (figure 5b; ANOVA, Bonferroni-corrected p < 0.05 for multiple comparisons), indicating that V1 contains most reliable information for reconstructing visual images.

Figure 3. Distributions of voxel weights on the flattened cortex for a foveal and a peripheral decoder. Voxel weights are shown on the right visual cortical surface of subject S1. The location of each patch (1 x 1) is indicated in the inset of the top-right corner. The white lines are the V1 and V2 borders. Modified from [8].
The lower reconstruction accuracy at higher visual areas could be accounted for largely by the less ordered retinotopic organization. However, we also found differences in the contribution of voxel correlations across visual areas. The shuffling analyses showed that the removal of voxel correlations impaired reconstruction performance most severely in V1. (figure 5c; ANOVA, Bonferroni-corrected p < 0.05 for multiple comparisons), indicating the critical role of voxel correlations in V1. These findings suggest that the reliable information available in V1 is represented not only in the retinotopic organization, but also in the correlated voxel patterns.

3.5. Advantage of a multi-scale reconstruction model
We then tested the significance of the multi-scale representation by comparing the multi-scale model with single-scale models that consisted of optimally combined, single-scale image bases (1 x 1, 1 x 2, 2 x 1, or 2 x 2; V1 and V2 voxels used as the input). Representative examples of the reconstructed images obtained from the figure image session are presented in figure 6a. The reconstructed image from the 1 x 1 scale model shows fine edges but exhibited patchy noise. By contrast, the 2 x 2 scale model produced a spatially blurred image. The images from the 1 x 2 and 2 x 1 scale models contained horizontally and vertically elongated components. The reconstructed image from the multi-scale model appears to have balanced features of these individual scales. The reconstruction error of the multi-scale model, calculated with the images from the random image session, was significantly smaller than those of the single scale models (figure 6b; ANOVA, Bonferroni-corrected p < 0.05 for multiple comparisons).

Figure 4. Advantage of multi-voxel pattern decoders. (a) Performance of the multi-voxel pattern decoder and retinotopic decoders. The binary classification performance for 1 x 1 patches is plotted as a function of eccentricity. Classification was performed using 1) a multi-voxel pattern, 2) the most responsive voxel for each patch (with the highest t-value), and 3) the mean of significantly responsive voxels for each patch (p<0.05, FDR corrected for multiple comparisons). The performance was evaluated by cross-validation using data from the random image session. The average performance was calculated in each 0.5-deg eccentricity bin (two subjects pooled; error bars, s.d., dashed line, chance level). (b) Effect of voxel correlation in training data. Performance is compared between the multi-voxel pattern decoders trained with the original data and the same decoders trained with ‘shuffled’ data, in which voxel correlations were removed. The results for the multi-voxel pattern decoder are the same as those displayed in (a). Modified from [8].
We also calculated reconstruction errors at each eccentricity (figure 6c). For all scales, the reconstruction error increased with eccentricity, but the profiles were different. The error sharply increased with eccentricity for the 1 x 1 model, while the profile was rather flat for the 2 x 2 model. As a result, the errors for these models were reversed at the fovea and the periphery. The 1 x 2 and 2 x 1 models showed intermediate profiles. Statistical analysis revealed a significant interaction between scale and eccentricity (p < 0.05 for interaction between eccentricity and scale, two-way ANOVA). The multi-scale model exhibited an error profile matching the minimum envelope of those for the single scale models. Thus, the multi-scale model appears to optimally find reliable scales at each eccentricity.

Consistent with the above observation, analyses of the combination coefficients of the multi-scale model showed that the reconstruction of the foveal and peripheral regions relied on the fine- and coarse-scale decoders, respectively [8]. These results indicate that the optimization of combination coefficients indeed found reliable local decoders at each visual field location to achieve high reconstruction performance.

Figure 5. Reconstruction using individual visual areas. (a) Reconstructed images. Examples from the figure image session (S1, ‘small frame’) are shown. (b) Reconstruction performance with entire images. The bar graph shows reconstruction errors, averaged across all test images in the random image session (two subjects pooled; error bars, s.d.). The dashed line indicates the chance level (1/3), which is achieved when a contrast value for each patch is randomly picked from the uniform distribution of 0 to 1. (c) Effect of correlations on reconstructed images. The distribution of difference in reconstruction error (error with the shuffled) - error with the original, for each image) is plotted for each visual area (two subjects pooled). Modified from [8].
4. Discussion
We have shown that contrast-defined arbitrary visual images can be reconstructed from fMRI signals of the human visual cortex. By combining the outputs of local decoders that predicted local contrasts of multiple scales, we were able to reconstruct a large variety of images (out of $2^{100}$ possible images) using only several hundred random images to train the reconstruction model. Analyses revealed that both the multi-voxel and the multi-scale aspects of our method were essential to achieve the high accuracy. Our automatic method for identifying relevant neural signals uncovered information represented in correlated activity patterns, going beyond mere exploitation of known functional anatomy.

4.1. Decoding from multi-voxel patterns
A major difference of our approach from the previous studies is that we directly computed the decoding model, instead of elaborating or inverting an encoding model. In our decoding approach, the model is optimized so as to best predict individual stimulus parameters given a multi-voxel pattern while taking into account voxel correlations. In contrast, the encoding models in the previous studies were optimized so as to predict individual voxel responses given a stimulus without considering voxel correlations when estimating the model parameters [5, 9].
Recent imaging studies suggest that there is a better combination of population responses to decode a given visual stimulus than using a signal from the most responsive cortical location or an averaged signal over the responsive cortical locations [1, 11]. In particular, if signals from multiple locations are correlated, a successful decoder should optimally assign various weights, including negative ones, to each location depending on the correlation structure [10, 11].

Consistent with this observation, our decoder using a multi-voxel pattern outperformed a decoder using only a single responsive voxel or an average of responsive voxels (figure 4a). The shuffling of training data, which removed voxel correlations, impaired the decoding performance, indicating the critical role of voxel correlation for constructing an optimal decoder (figure 4b). Careful inspection of the weight distributions in figure 3 indicates that a decoder trained with the original data uses both positive and negative weights, which are found at nearby locations, particularly at the middle to peripheral range of eccentricity. Further analyses showed that the magnitudes of negative weights decreased after shuffling the training data, suggesting that negative weights served to exploit voxel correlation [8].

Although the previous animal study [11] suggested that neural activity in V1 contains significant spatial correlations that can be useful for decoding a visual stimulus, it has been unclear whether such informative correlations are present in other areas of the early visual cortex. Our analysis (figure 5) showed that much of the information available in V1 was represented in voxel correlations, while other areas were less dependent on them.

There are many possible sources of voxel correlation. As the neural populations in nearby voxels are likely to be synapticly coupled, correlated fMRI signals could be spontaneously induced. Nearby voxels might also show correlations through their vascular coupling. Physiological status (e.g., cardiac and respiratory noise) and fMRI scanner conditions (e.g., gradient coil heating) might also cause slow fluctuations correlated among voxels. However, they are unlikely to be major sources of the voxel correlations because the decoder’s performance was not affected by filtering out slow components from the data [8]. In addition, head motions of a subject and spatial reinterpolation during preprocessing are also unlikely to be the source, since they cannot account for the area-specific effects of the voxel correlations (figure 5c).

4.2. Multiple scales of visual representation

Our multi-scale reconstruction model achieved higher reconstruction accuracy than single-scale models by combining reliable scales at each location. The reliable scales largely depended on eccentricity, which can be related to the receptive field size and the cortical magnification factor. The receptive field size of visual cortical neurons is known to increase with eccentricity [18] and, in parallel, the cortical magnification factor decreases with eccentricity [19, 20]. The receptive-field size for the human visual cortex was estimated at about 1 – 2 deg at 7 deg eccentricity, which is near the most peripheral patch in our stimulus image, while the cortical magnification factor at 7 deg is about 2 – 3 mm/deg. These estimates suggest that single voxels (3 cubic mm) for the peripheral representation carry retinotopic information about more than a single peripheral patch, and thus are not suitable for the decoding of fine-scale (1 x 1) patches, consistent with our reconstruction results (figure 6c). Such eccentricity-dependent changes in the scale of visual representation may partly account for the superior reconstruction by the multi-scale model.

However, it should also be noted that the reconstruction model did not exclusively select a single scale at each eccentricity. At every location except the most foveal region, all scales were effectively combined to improve the reconstruction accuracy (figure 6e) [8]. Previous studies have shown variability in receptive field size among neurons whose receptive fields overlap [13, 21]. Even though each fMRI voxel should contain numerous neurons with receptive fields of various sizes, it may be possible to extract scale-specific information by combining many voxels with a weak scale bias in each, analogous to the extraction of orientation information from coarse voxel sampling of cortical columns [1].
4.3. Modular decoding and its applications

Our approach provides a general procedure to deal with complex perceptual experience consisting of numerous possible states by using multiple decoders as modules. If a perceptual state can be expressed by a combination of elemental features, a modular decoder can be trained for each feature with a small number of data, but their combination could predict numerous states including those that have never been experienced.

Although we focused here on the reconstruction of contrast patterns, our approach could be extended to reconstruct visual images defined by other features, such as color, motion, texture, and disparity. Besides, the motor function may also be dealt with our approach. A large variety of motor actions could be described by a combination of putative modules [22]. Thus, the modular decoding approach may greatly improve the flexibility of prediction, which could also expand the capacity of neural prosthetics or brain-machine interfaces.

More interesting are attempts to reconstruct subjective states that are elicited without sensory stimulation, such as visual imagery, illusions, and dreams. Several studies have suggested that these subjective percepts occur in the early visual cortex [23], consistent with the retinotopy map [9]. One could address this issue by attempting to reconstruct a subjective state using a reconstruction model trained with physical stimuli. Reconstruction performance can also be compared among cortical areas and reconstruction models. Thus, our approach could provide valuable insights into the complexity of perceptual experience and its neural substrates.

Acknowledgements

The authors thank A. Harner and S. Murata for technical assistance, and T. Beck and M. Takemiya for manuscript editing. This research was supported in part by SRPBS, MEXT, the NICT-KARC, the Nissan Science Foundation, and the SCOPE, SOUMU.

References

3 586-98

[22] Poggio T and Bizzi E 2004 Nature 431 768-74