Visual cortex allows prediction of perceptual states during ambiguous structure-from-motion

Supplementary Methods and Figures

23-10-2006

G.J. Brouwer, R. van Ee
Helmholtz Institute, University of Utrecht

Corresponding Author
G.J. Brouwer
Helmholtz Institute, University of Utrecht
Princetonplein 5
3584 CC Utrecht
The Netherlands
G.J.Brouwer@phys.uu.nl

Number of Supplementary Pages: 10
Number of Supplementary Figures: 5
Supplementary Methods

In addition to the Support Vector Machine algorithm using a linear kernel, we also used both quadratic and cubic kernels (see below) and two other implementations of multivariate classifiers to establish that the results we obtained are not specific to our particular choice of classifier and/or kernel method.

Support Vector Machines

Support Vector Machines or SVMs (Vapnik, 1988; Burges, 1998) are a class of optimization algorithms generally used for classification purposes. The algorithm finds the margins which contains a hyperplane that best separates the data. Assuming binary classification, let $x_i \in \mathbb{R}^n, i = 1, ..., N$ denote $N$ number of examples each having $n$ features, and an associated class label $y_i \in \{-1, 1\}$. Assuming that the data is linearly separable (Fig. 1a), let $w$ be the weight vector $\perp$ to the separating hyperplane and $b$ be the shortest distance of the hyperplane from the origin. The equation to the hyperplane can thus be written as:

$$w \cdot x + b = 0 \quad (1)$$

The points lying below and above the margins of the hyperplane form constraints that can be formulated in the following way:

$$w \cdot x_i + b \geq +1 \quad (2)$$
$$w \cdot x_i + b \leq -1$$

Combining these constraints (equalities and inequalities) into one equation, we get:

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad (3)$$

The margins are now defined as the minimum distance to the points that are nearest to the hyperplane. For all positive examples, normalizing the equations with the weight vector i.e. $||w||$, the distance of the margin from the hyperplane is $|1 - b| / ||w||$. Similarly, the distance of the margin for negative examples is $|1 + b| / ||w||$. Therefore, the total width of the margin is given by $2/||w||$. The SVM attempts to maximize the width of the margin in order to separate a maximum number of training data points. The points that define such a margin are referred to as the support vectors. Optimization occurs by minimizing $||w||^2/2$, subject to certain constraints. However, generally data is not linearly separable because of outliers. Taking into account the effect of outliers, cost variables can be introduced and equation (3) becomes:
Optimization now involves both minimizing the weight vector as well as the total penalty induced by the outliers. Mathematically, the minimization problem takes the form

\[ \| w \|^2 / 2 + C \sum_{i=1}^{N} \xi_i = \] (5)

The SVM algorithm thus tries to find a tradeoff by maximizing the margin to separate a maximum number of points while at the same time bounding the training error induced by the outliers. When data is non-linearly separable, instead of linearly separable, SVM uses the kernel trick method. This method converts a linear classifier algorithm into a non-linear one, by mapping the original observations into a higher-dimensional non-linear space so that linear classification in the new space is equivalent to non-linear classification in the original space. This is done using Mercer’s condition, which states that any positive semi-definite kernel \( K(x, y) \) can be expressed as a dot product in a high-dimensional space (Aizerman et al., 1964). Nonlinear kernels are typically polynomials, although most nonlinear functions can be used.

**Perceptrons**

Implementation of the perceptron model (Rosenblatt, 1961) was done using the Neural Networks toolbox of Matlab (Mathworks). In its simplest form the perceptron model is a single-layer network, whose weights and biases can be trained to produce a correct target vector when presented with the corresponding input vector, see Supp. Fig. 1b. Each external input element \( p_i \) is weighted with an appropriate weight \( w_i \), and the sum of the weighted inputs is thresholded by a hard-limit transfer function \( \theta \), yielding an output of either 0 or 1:

\[ \theta \left( \sum_{i=1}^{N} p_i w_i \right) \] (6)

We associated an input vector (normalized voxel intensities for a particular volume) with a binary output vector \( (\theta = CCW, 1 = CW) \) and trained the algorithm using the perceptron learning rule:

\[ \Delta w = \alpha(t - a)p \] (7)

where \( a \) is the response of the perceptron unit, \( t \) the desired response, \( w \) the weight vector associated with input vector \( p \). This rule can be put in words as follows: determine both the
actual and desired output of the unit and subtract these. This value (which can be 0, 1, or -1) is then multiplied with the input vector $p$ and a learning parameter $a$, and added to the weight vector.

**Differential Mean**

For the differential mean approach, we determined the mean intensity of voxels for all volumes associated with either perceptual state, yielding two weight vectors (one for each perceptual state). Each test-pattern (voxel intensities per volume) was multiplied with both weight vectors in turn and summed (dot product). The two resulting values were subtracted (CW-CCW) and thresholded. Positive values indicated a CW prediction for a particular volume, negative values a CCW prediction for a particular volume.
Supplementary Results

Convolution of perceptual timecourses with the haemodynamic response function
An important step in associating each fMRI measurement (voxel intensities per volume) with the dynamics of perception is to account for the haemodynamic delay between changes in the Blood-Oxygen Level Dependent (BOLD) signals and reported alternations in perceived states. To achieve optimal correspondence, we convolved the perceptual states with a canonical haemodynamic response function used by the SPM2 (http://www.fil.ion.ucl.ac.uk/spm), identical to the method used by (Haynes and Rees, 2005b). The main effect of convolution is to shift the transitions between perceptual states to a later point in time. Justification for the usage of the particular HRF used is provided by cross-correlating the graded output of the classifiers with the convolved perceptual states (Supp. Fig. 2a, right-top panel). The highest correlation is found around a shift of zero volumes, indicating that convolving perceptual states with this particular HRF is optimal for associating voxel intensities with perceptual states. In addition, we changed the shape of the HRF (varying the parameters determining delay and dispersion of the positive response) to determine whether the particular shape of the canonical HRF was indeed optimal. Changing the shape of the HRF away from its canonical form reduces accuracy, indicating that the canonical HRF is indeed optimal (Supp. Fig. 2b).

Relationship between the number of voxels and prediction accuracy
If prediction accuracy does indeed depend on the combined signals of multiple voxels, accuracy should be a function of the number of voxels that are used to classify in that prediction accuracy will increase gradually when more and more voxels are added. That this is indeed that case can be observed in Fig. 3a: the accuracy of prediction increases gradually and asymptotes at around 50 percent.

A comparison of classification functions
One of the benefits of the support vector machine algorithm is that it can incorporate different kernels for separating classes (see supplementary methods). For our analysis we used a linear kernel. Here we compare the accuracy provided by this linear kernel with a quadratic and cubic kernel. In addition, we also compare the SVM algorithm with other classification
functions, the perceptron model and a differential mean approach (Supp. Fig. 3b). The linear and cubic kernels are comparable in performance; both attain a similar level of acceptable accuracy. The quadratic kernel, however, is unable to reach similar levels, although it does perform slightly above chance level. Supp. Fig. 3b appears to provide evidence that the classifiers do not differ significantly in the accuracy they attain. However, direct comparison of individual prediction accuracies reveals that the SVM with a linear kernel systemically outperforms the other classifiers \((w\text{-parameter})\). In addition, the accuracies produced by both the perceptron and differential mean classifiers correlate strongly with the accuracies produced by the SVM classifier (perceptron model: 0.72, \(p < 0.01\), differential mean 0.81, \(p < 0.01\)). This suggests that all three functions essentially use the same structure within the data, although the SVM is most accurate in extracting it.

[Supplementary Figure 3 about Here]

Decoding
In the light of accurate prediction, it is crucial to determine the underlying pattern responsible for the generation of such accuracy. First, we determined the correlation between the timecourses of each voxel and perceptual states (after convolution with the HRF). It is this correlation that is used in conventional univariate (GLM) fMRI statistics: the higher the correlation, the higher the associated t-statistic will be for a given voxel. We examined the relationship between this correlation and the \(w\text{-parameter}\) associated by the SVM with each voxel. The \(w\text{-parameter}\) of a voxel signifies the direction orthogonal to the support vector of that voxel. If the \(w\text{-parameter}\) is 0, no support vector was assigned to that voxel/dimension. The more positive or negative the \(w\text{-parameter}\) becomes, the more separable the data is for that voxel/dimension and the more accuracy the SVM can derive from the voxel. Complete orthogonality is expressed by either -1 or 1, with the sign of the \(w\text{-parameter}\) determining on which side of the hyperplane the data belonging to each class (perceptual state) is positioned. More specifically, a positive \(w\text{-parameter}\) for a particular voxel informs us that CW volumes are associated with higher BOLD signal levels than CCW volumes and vice versa. Note that this simplification only holds when one considers SVMs with linear kernels. Supp Fig. 4a plot the correlation for all voxels of each subjects' visual area V7. As can be seen, the more a voxel is correlated in time with the expected BOLD signal changes as a result of perceptual states, the higher its assigned \(w\text{-parameter}\). This correlation of 0.48 was significant \((p < 0.01)\).
Furthermore, an equal amount of voxels seems to code for CW compared to the CCW perceptual state, as can be seen by the symmetric scattering of data points on the abscissa.

**Univariate vs. Multivariate Statistics**

In a final analysis, we compared the multivariate prediction accuracy with a conventional univariate GLM-approach. Analyzing the same voxels that were used to train our classifiers, we determined the t-statistic of the GLM contrast between phases associated with CW perceived rotation and those with CCW perceived rotation in all subjects for areas V1, V3A and V7. Although small significant activation was found in this manner for two subjects, comparing these t-statistics with the accuracy of prediction revealed that the univariate contrast and the multivariate approach were not correlated: a high t-statistic (even when significant) does not guarantee high predictive accuracy. **Supp. Fig. 4b** even shows the opposite to be true: the higher accuracies are associated the lowest t-statistics. Although both methods were sensitive to similar signals (i.e. those that are correlated with perception, see above), the use of univariate contrasts (e.g. \( CW > CCW \)) will typically not reveal any consistent activation. As we have pointed out, the region-of-interest with highest mean accuracy (i.e. V7) contains an equal amount of voxels that show a positive correlation with CW perceptual states compared to voxels showing positive correlation with the CCW perceptual state. This explains why univariate methods fail to reveal significant activation, even if both approaches rely on the same structure in the data. While this distribution of positive and negative correlations can be use by multivariate classifiers, it is not consistent with either of the univariate contrasts (CW > CCW and vice versa). When using such a specific contrast (e.g. \( CW \) perceptual states > \( CCW \) perceptual states), we find that half of the voxels will support one contrast (voxels positively correlated to the CW perceptual states), while the other half (voxels positively correlated to the CCW perceptual state) do not. Hence, the area as a whole will show no or only weakly significant activation for our particular contrast (as well as the reverse contrast).

[Supplementary Figure 4 about Here]
References

- Burges, CJC (1998). A Tutorial on Support Vector Machine for Pattern Recognition, Data Mining and Knowledge Discovery, Vol. 2:121-167.

- Haynes, JD, and Rees, G (2005b). Predicting the stream of consciousness from activity in human visual cortex. Curr. Biol. 15:1301-1307.

- Rosenblatt, F, Principles of Neurodynamics. (Spartan Press, Washington DC).

- Vapnik, VN (1998). Statistical Learning Theory. (Wiley, New York).

- Friston, KJ, Holmes, AP, Worsley, KJ, Poline, JP, Frith, CD, and Frackowiak, RSJ (1995). Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. 2:189–210.

- Aizerman, M, Braverman, E, and Rozonoer, L (1964). Theoretical foundations of the potential function method in pattern recognition learning. Automation and Remote Control, 25:821-837.
Figure Legends

Supplementary Figure 1: Multivariate Classifiers. (A) A schematic representation of the support vector machine algorithm for a binary classification in 2-dimensional space using a linear kernel. The SVM finds the hyperplane that maximizes separation between classes in all dimensions while at the same time it minimizes the error induced by outliers. (B) Perceptron model. In its simplest form, the perceptron multiplies an input vector \( p \) with weights vector \( w \) and a bias \( b \). The resulting vector is summated and thresholded (dot-product), yielding an output of 0 or 1.

Supplementary Figure 2: Convolution of perceptual timecourses with the haemodynamic response function. (A) In order to accurately associate measurements (voxel intensities per volume) with perceptual states, perceptual timecourses were convolved with the canonical haemodynamic response function (HRF) of the SPM2 package. This convolution accounts for the haemodynamic delay between actual signal changes accompanying perceptual transitions. Justification for the usage of the canonical HRF comes from considering the cross-correlation between graded SFM outputs and convolved perceptual timecourses. The highest correlation is found at a time lag of zeros volumes, indicating the convolution provided an optimal shift of the perceptual transitions to account for the haemodynamic delay. (B)Variations in the shape of the canonical HRF (changes in the delay and dispersion of the positive response of the HRF) decreased the accuracy of the classification. This indicates that, in addition to the shift produced by the HRF, the shape of the canonical HRF was also optimal in accounting for the specific haemodynamic effects of perceptual transitions on BOLD signal changes. Shown are the accuracies of prediction using the data taken from area V7 of two subjects. The center of the graph represents the canonical HRF, as used by SPM and is associated with the highest accuracy.

Supplementary Figure 3: Comparisons of classifications methods. (A). Prediction accuracy (subject AK, area V7) as a function of the percentage of voxels used. Prediction accuracy (gray lines: individual runs, black line: average over all 10 runs) increases gradually and asymptotes at around 50 percent of voxels used. Error bars represent SEM. (B) Mean accuracy of the different classification algorithms used. SVMs with linear kernels outperform SVMs with quadratic and cubic kernel, as well as the perceptron model and the differential mean. (C) Correlation, per prediction, produced by linear SVM classifier and those produced by the other classifiers (quadratic, cubic SVM, perceptron and differential mean). Covariation between methods is observed, as well as the fact that the linear SVM consistently outperforms all other classifiers. This suggests that all three classification methods use the same underlying pattern in classification and prediction, but that the linear SVM is most successful in extracting this information.
Supplementary Figure 4: Decoding of the SVMs. (A) We calculated the fit between voxel and convolved timecourses and correlated these fits to the $w$-parameter assigned to that voxel by the SVM algorithm. The $w$-parameter in a particular voxel signifies the separability of the data (i.e. perceptual states) for that voxel. The observed correlation indicates that SVM accuracy is determined by voxels whose timecourses correlate with perceived rotation, meaning both univariate and multivariate methods are sensitive to the same signals: those correlating with perception. Curves show the symmetric distribution of voxels in terms of correlations and $w$-parameters: half of the voxels show increased BOLD signals correlating with perceiving CW rotation, half of the voxels with perceiving CCW. (B) We used a conventional univariate GLM approach to determine the significance of differential activation between CW and CCW states in areas V1, V3A and V7. This produced weakly significant data in two subjects. More importantly, there was no correlation between the significance of the univariate contrast and the accuracy of the multivariate approach. This indicates that an univariate analysis of V3A and V7 would not have revealed any significant activation, while the multivariate methods can predict perceptual accuracy with quite some accuracy.