Face Prediction from fMRI Data During Movie Stimulus: Strategies for Feature Selection

†Jukka-Pekka Kauppi, †Heikki Huttunen, †Heikki Korkala, ‡Iiro P. Jääskeläinen, ‡Mikko Sams, and †Jussi Tohka

†Tampere University of Technology, Dept. of Signal Processing, Tampere, Finland
‡Aalto University School of Science, Dept. of Biomedical Engineering and Computational Science, Espoo, Finland

Abstract. We investigate the suitability of the multi-voxel pattern analysis approach to analyze diverse movie stimulus functional magnetic resonance imaging (fMRI) data. We focus on predicting the presence of faces in the drama movie based on the fMRI measurements of 12 subjects watching the movie. We pose the prediction as a regression problem where regression coefficients estimated from the training data are used to estimate the presence of faces in the stimulus for the test data. Because the number of features (voxels) exceeds the number of training samples, an emphasis is placed on the feature selection. We compare four automatic feature selection approaches. The best results were achieved by sparse regression models. The correlations between the face presence time-course predicted from fMRI data and manual face annotations were in the range from 0.43 to 0.62 depending on the subject and pre-processing options, i.e., the prediction was successful. This suggests that proposed methods are useful in testing novel research hypotheses with natural stimulus fMRI data.

Keywords: Natural stimulation, brain imaging, regression

1 Introduction

Functional magnetic resonance imaging (fMRI) studies based on continuous, complex stimuli such as movies allow investigation of the brain functions in more natural contexts compared with the traditional, highly controlled experimental designs [13]. Because the resulting functional brain data is difficult to interpret due to complex neural spatio-temporal interactions in the brain during the stimulus, powerful and flexible analysis approaches are necessary to better understand this type of data. Multi-voxel pattern analysis (MVPA), aiming at modeling the multivariate relationship between fMRI measurements and the stimulus using machine learning algorithms, is a popular approach in predicting and decoding brain states [4, 10]. Typically, MVPA is applied to fMRI data sets acquired during strictly controlled experiments.

In this paper, we investigate the suitability of the MVPA approach to analyze diverse movie stimulus fMRI data. We focus on predicting the presence of faces

in the drama movie "Crash" (Paul Haggis, Lions Gate Films, 2005) based on the fMRI measurements of the subjects watching the movie. Compared with the data from controlled experiments, movie data is considerably more complicated to analyze because it embeds a subject to an extremely diverse environment with rapidly changing scenes and emotional content, activating many brain regions simultaneously.

We concentrate on predicting the presence of faces in the movie because this question is already well-established in previous studies using MVPA in controlled experiments (see, e.g., [2, 4]) and it is important to investigate if the findings in these experiments can be generalized to more natural conditions [13]. Novel research questions may be investigated using similar tools if the face prediction can be performed reliably. This is not the first time when MVPA is used for natural stimulus fMRI data. In the 2006 and 2007 Pittsburgh Brain Activity Interpretation Competitions (PBAIC) (`http://www.lrdc.pitt.edu/ebc/2006/competition.html`, `http://www.lrdc.pitt.edu/ebc/2007/competition.html`), the task was to predict the mental states of subjects based on the fMRI data collected during watching movie clips (2006 PBAIC) or a virtual reality task (2007 PBAIC). The prediction of the presence of faces was a sub-task in both competitions. However, our "Crash" data is more naturalistic than the 2006 PBAIC data, in which each session consisted of consecutive clips of short movie segments. Moreover, unlike in the PBAIC 2006, where subjects themselves rated the presence of faces, we use the same manually collected face annotation for each subject. Also, the virtual reality task in the 2007 PBAIC is different from movie viewing.

We predict the prevalence of the face in the movie based on the fMRI signal activity from several voxels at a single time point. Because the number of available features P (the number of voxels) exceeds the number of samples N (the number of time points), the most critical part of the prediction algorithm is dimension reduction as in many other fMRI-applications of MVPAs (e.g. [9]). To this aim, we compare four feature selection methods: *stepwise regression* (SWR) [3], *simulated annealing* (SA) (see e.g.,[8]), *Least Absolute Shrinkage and Selection Operator* ($LASSO$) [3, 14] and *Least Angle Regression* ($LARS$) [1, 3].

2 Methods and Materials

2.1 Regression Model and Feature Selection

The face prediction problem could be posed as a classification problem, where the classes are "face" and "non-face". However, we predict a continuous-valued face annotation using linear regression because this approach is well-suited to movie stimulus fMRI data due to rapid and unexpected changes of the movie scenes. For example, continuous-valued annotation allows interpreting the extent of the face prevalence during each TR (repetition time) despite of the relatively low temporal resolution of the fMRI signal. In addition, although the face is present in the movie frame, it may be partially occluded, shadowed or far away, thus

decreasing the "face response" measured by the fMRI. There are also numerous other stimuli in the movie which compete from the attention of the subject.

The training consists of estimating the unknown parameters $\boldsymbol{\beta}$ of the regression model, which is then used to predict the face-prevalence on independent test data. The simplest approach for estimation is the ordinary least squares (OLS) regression. Denote the annotation time-course (convolved with expected hemodynamic response) by $\mathbf{y} = (y_1, y_2, \ldots, y_N)^T$ and the fMRI time series from the k'th voxel ($k = 1, 2, \ldots, P$) by $\mathbf{x}_k = (x_{1k}, x_{2k}, \ldots, x_{Nk})^T$. The linear model for the annotations explained by the fMRI measurements is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with the $N \times P$ matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_P)$, the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_P)^T$ and the residual term $\boldsymbol{\varepsilon} \in \mathbb{R}^N$. The OLS estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ is found by minimizing the training error $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$. However, the solution assumes that $N \geq P$, which does not hold in our case. Therefore, the OLS has to be accompanied by a feature selection algorithm that chooses at most P meaningful features and discards the remaining columns from the matrix $\mathbf{X}$. We will next review the four feature selection methods assessed in this work.

Sequential forward selection (SFS) is an intuitive search heuristic for feature selection [3]. SFS starts with no features in the model and adds features to it until a stopping criterion is satisfied or the whole feature set is included in the model. In each iteration, the best ranked feature is added to the model according to a predefined model selection criterion. In addition to forward steps, the method also performs backward elimination in case some of the features become unnecessary after more features are included in the model. When applied specifically for regression, the SFS is called *Stepwise Regression* (SWR) [3]. SWR begins with an empty model and sequentially tries to add and remove those features that increase the explanatory power the most. In both steps, statistical hypothesis testing regarding the variable coefficients is performed using the F-statistic. The drawback of SWR is that the method typically converges to a local optimum due to iterative selection procedure. Also the statistical inference procedure faces the multiple comparisons problem because several dependent F-tests are applied to the same data. This complicates interpretation of the p-values.

Simulated Annealing (SA) has been used for feature selection [8] to avoid local minima. SA is a randomized search heuristic with roots in condensed matter physics, where slowed-down cooling is used to reduce the defects of the material by allowing the molecule configuration to reach its global minimum state. The method has been successfully used in various optimization problems with multiple local extrema. SA feature selection is a *wrapper method*, which iteratively evaluate the performance of candidate feature subsets by training a regression model. SA starts with the empty feature subset, and at each iteration attempts to add or remove a random feature from the set. The change in cross-validated prediction error is then used to determine whether the new subset is accepted. All improved results are accepted, while worse solutions are accepted at random with probability $\exp(-\Delta_{\text{error}}/T)$, where $\Delta_{\text{error}} = \varepsilon_{\text{new}} - \varepsilon_{\text{old}}$ is the change in error and T is the simulated temperature. The temperature is initialized to a

high value where almost all configurations are accepted, and it is decreased at each iteration according to the rule $T \leftarrow \alpha T$ with $\alpha < 1$. We added an extra penalty term for the number of features to favor simple solutions. The size of the penalty and the cooling parameter α were selected by cross-validation.

An alternative to the wrapper approach is to embed the feature selection into the performance criterion, resulting in *embedded feature selection methods*, which introduce a penalty for the number of nonzero coefficients in the regression model. LASSO *(Least Absolute Shrinkage and Selection Operator)* regression method enforces sparsity via l_1 -penalty by optimizing the constrained LS criterion [14]:

$$\min_{\beta} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{P} |\beta_j| \leq t, \tag{1}$$

where $t \geq 0$ is a tuning parameter. When t is large, this is identical to the OLS solution. With small values of t, the solution becomes shrunken and sparse version of the OLS solution where only a few of the coefficients β_j are non-zero.

The *Least Angle Regression* algorithm (LARS) is another feature selection method applicable for sparse regression problems with very similar regularization path to that of LASSO [1]. LARS iteratively adds the features that are most correlated with the residual of the current model. It increases the coefficient of the selected features towards the OLS solution until the feature no longer has the highest correlation with the residual. Then, LARS chooses a new direction equiangular between the first and the most highly correlated feature and continues moving along this direction until another feature with higher correlation is found. The procedure is continued until all predictors are included in the model. Finally, the resulting $\boldsymbol{\beta}$ is selected along the regularization path by cross-validation. Also the LASSO solution can be found by modifying the basic LARS algorithm. We used LARS and LASSO implementations from [11].

2.2 Functional MRI Experiment

Functional MRI was measured from 12 subjects while they watched the last 36 minutes of the drama movie "Crash" (Lions Gate Films, 2005; for details see [5]). Preprocessing of the data included motion correction, detrending, spatial smoothing, and registration of the functional images to a stereotactic template (for details, see [7]). The data for each subject was acquired in two sessions, lasting 14 ($N = 244$ with TR = 3.4 s) and 22 minutes ($N = 382$), respectively. For each subject, we used the first session for training and the second session for evaluating the prediction performance. In the feature selection phase, 10-fold cross-validation was used to select the parameter t for LASSO and LARS, p-values for SWR, and cooling and regularization parameters for SA. The test session was not used in any way during the training or feature selection.

The face annotation was collected by manually detecting one or more faces during the periods of 1 s. Because the sampling rate of the face annotation was higher than that of the fMRI measurements, we integrated the face annotations

over a single fMRI measurement. We convolved the resulting annotation with the double gamma model of the hemodynamic response for several hemodynamic lags, and evaluated prediction performances separately for each lag.

Before automatic feature selection, we reduced the number of features (voxels) in the original large feature set including over 200000 voxels across the whole brain. Two different initial feature subsets were formed. The first subset contained the measurements across fusiform cortex, which is expected to be important for the face detection [6]. The region of interest (ROI) was localized using Harvard-Oxford probabilistic cortical atlas from the FSL software package [12] with 50 % threshold. It contained $P = 601$ features. As face detection in the brain may be distributed [2], we chose a second feature subset including voxels across cortex. We further reduced the number of features by preserving only measurements showing the pairwise averaged inter-subject correlations (ISCs) larger than 0.30 across the 12 subjects (see [7]). This reduced the number of features down to 1480.

3 Results

Figures 1 and 2 present the prediction performances of the four methods. One box plot contains the results of 12 subjects, and several box plots are shown for different hemodynamic lags. Correlation coefficient between the face annotation and the model estimate was used as the performance measure in accordance with PBAIC. The best test predictions were obtained using the LARS (Fig. 1) having the median correlation coefficients over 0.50 for both fusiform and cortex feature sets (see e.g. results with the hemodynamic lags of 7-9 s in Figs. 1C-D). The correlations were slightly higher for the fusiform than for the cortex features, even though correlations for the cortical features were higher for the training session (Figs. 1A-B). The highest correlations for the training session were obtained using the hemodynamic lag of 6 s whereas the lag of 8 s provided the best results for the test session.

Figure 2 shows the test prediction performance using LASSO, SWR, and SA with the fusiform feature set. The results of LASSO were almost as good as the results of LARS. Feature analysis between LASSO and LARS indicated that both methods found some common features for all the subjects but there were also differences between the solutions, i.e., the solution of the LASSO contained features not present in the solution of LARS and vice versa. The performances of SWR and SA were slightly lower than LASSO and LARS, but they resulted in simpler models with fewer features. Table 1 summarizes the prediction results for the test session using the lag of 8 s, which provided the best prediction performance.

Figure 3 displays the prediction curve of the LARS for the best predicting subject across a part of the test session. The prediction followed the fluctuations of the faces in the movie accurately, indicating that the found feature set was responsible for the face detection during the movie. The anatomical locations of some of the automatically selected features are also shown in Fig. 3. For this

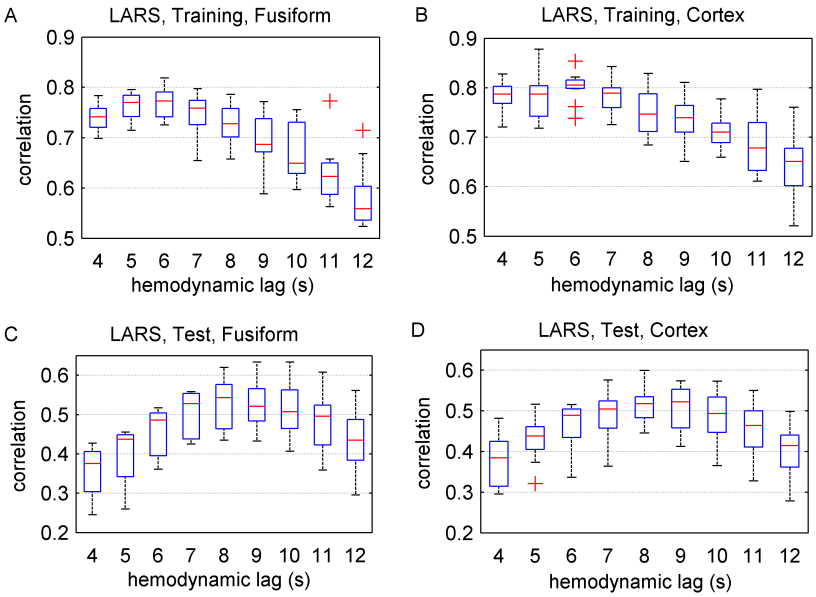

Fig. 1. Correlation coefficients between the face annotation and the model estimate using LARS for 12 subjects: (A) training session results for the fusiform features, (B) training session results for the cortical features, (C) test session results for the fusiform features, and (D) test session results for the cortical features.

Table 1. Test prediction performance and numbers of selected features in the final models. The hemodynamic lag was 8 s. LASSO and LARS provided the best predictions with nearly equal results, but SA and SWR resulted in simpler models with fewer features.

		Prediction Performance				*Number of Features*			
		SWR	SA	LASSO	LARS	SWR	SA	LASSO	LARS
Fusiform	mean	0.49	0.48	0.53	0.53	12.1	10.6	38.3	31.0
	max	0.59	0.60	0.62	0.62	19	20	60	49
	min	0.33	0.42	0.45	0.43	7	5	19	9
	std	0.08	0.06	0.06	0.06	3.2	4.7	14.7	12.9
Cortex	mean	0.44	0.43	0.52	0.51	18.8	12.3	47.6	42.9
	max	0.54	0.53	0.59	0.60	31	25	67	83
	min	0.35	0.30	0.45	0.45	11	1	21	16
	std	0.07	0.07	0.05	0.04	5.9	6.9	15.2	18.0

subject, we found 25 features located in five brain regions: temporal occipital fusiform cortex (TOFC), occipital pole (OP), inferior lateral occipital cortex (LOC), occipital fusiform gyrus (OFG), and lingual gyrus (LiG).

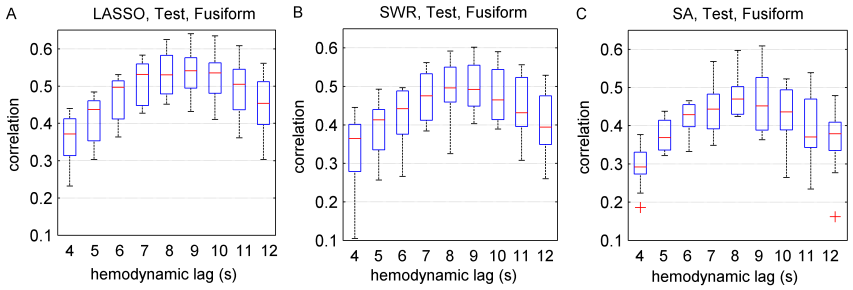

Fig. 2. Test prediction performance for the 12 subjects using the fusiform features.

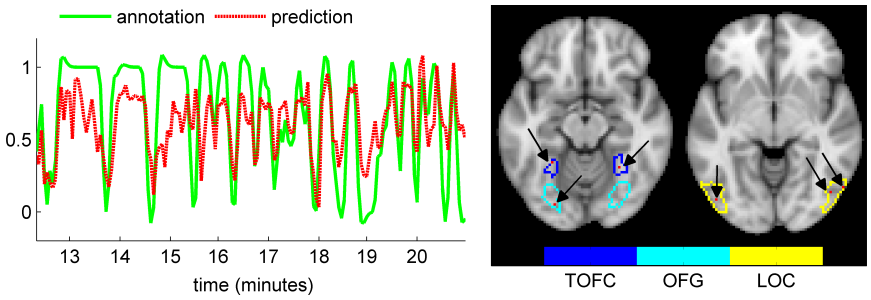

Fig. 3. The best prediction curve shown for the last part of the test session using LARS with the anatomical locations of some of the found cortical features (hemodynamic lag 8 s, correlation coefficient 0.60). Two axial slices show the location of 6 features in three brain regions.

We analyzed whether exactly the same cortical features were selected across the subjects. Using the criterion that the same feature needs to be found at least across 3 out of 12 subjects, we found features in the following brain regions (the number of features in the given brain region is shown in parentheses for LASSO/LARS/SA/SWR): superior LOC (2/3/0/0), inferior LOC (7/8/0/2), LiG (1/2/0/0), TOFC (5/10/0/0), OFG (3/1/0/0), planum temporale (3/1/0/0), and OP (3/7/0/0). Hence, LASSO and LARS solutions were more consistent across subjects than SA and SWR solutions.

4 Conclusion

In this work, we successfully predicted the presence of faces in the eventful drama movie based on the fMRI data using multi-voxel linear regression. Two different preprocessing methods (ROI-based, ISC-based) were used to select initial voxel sets for prediction, and the final subsets were optimized using four automatic feature selection methods (SWR, SA, LARS, and LASSO). The best predictions were obtained with sparse regression models LARS and LASSO, but also SWR

and SA provided good results. Importantly, the prediction was successful using the voxel set that initially contained voxels across several cortical regions, i.e., specific anatomical prior knowledge was not used in the initial selection. This suggests that proposed methods can be useful when testing novel research hypotheses with natural stimulus fMRI data. From neuroscientific point of view, our results support the view that face detection is distributed across the visual cortex, albeit the fusiform cortex has a strong influence on the face detection.

Acknowledgments. Supported by the Academy of Finland, (grants 129657, Finnish Programme for Centres of Excellence in Research 2006-2011, and 130275) and the aivoAALTO project funding of the Aalto University.

References

1. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Annals of Statistics 32(2), 407–499 (2004)
2. Hanson, S.J.J., Schmidt, A.: High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. NeuroImage 54(2), 1715–1734 (2011)
3. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: Data mining, inference, and prediction. Springer Series in Statistics, Springer (2009)
4. Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P.: Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293(5539), 2425–2430 (2001)
5. Jääskeläinen, I.P., Koskentalo, K., Balk, M.H., Autti, T., Kauramäki, J., Pomren, C., Sams, M.: Inter-subject synchronization of prefrontal cortex hemodynamic activity during natural viewing. The open neuroimaging journal 2, 14–19 (2008)
6. Kanwisher, N., McDermott, J., Chun, M.M.: The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. neurosci. 17(11), 4302–4311 (1997)
7. Kauppi, J.P., Jääskeläinen, I.P., Sams, M., Tohka, J.: Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. Frontiers in neuroinformatics 4:5 (2010)
8. Lin, S.W., Lee, Z.J., Chen, S.C., Tseng, T.Y.: Parameter determination of support vector machine and feature selection using simulated annealing approach. Appl. Soft Comput. 8, 1505–1512 (2008)
9. Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S.: Learning to Decode Cognitive States from Brain Images. Machine Learning V57(1), 145–175 (2004)
10. Norman, K., Polyn, S., Detre, G., Haxby, J.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10(9), 424–430 (2006)
11. Sjöstrand, K.: Matlab implementation of LASSO, LARS, the elastic net and SPCA (2005), `http://www2.imm.dtu.dk/pubdb/p.php?3897`
12. Smith, S.M., et al: Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 Suppl 1, S208–S219 (2004)
13. Spiers, H., Maguire, E.: Decoding human brain activity during real-world experiences. Trends in Cognitive Sciences 11(8), 356–365 (2007)
14. Tibshirani, R.: Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B 58, 267–288 (1994)