Model-based decoding of bottom up visual attention in the human brain

Hsin-Hung Li & Samuel Nastase
Dept. of Psychological & Brain Sciences, Hanover, NH

Machine Learning and Statistical Data Analysis
Prof. Lorenzo Torresani
Dartmouth College
2013

Introduction

The human brain effectively filters a torrent of incoming sensory data so as to select and enhance behaviorally relevant information. In the case of vision, saliency signals based on low-level properties of the visual input are rapidly and automatically computed in order to pinpoint the most important elements of complex visual scenes. This mechanism of bottom-up visual attention is tightly yoked to the motor systems governing eye movement—points in the visual field with high saliency are rapidly foveated in order to provide higher visual acuity and depth of processing. A neurally-inspired computational model of this process pioneered by Itti and Koch (2001) and further developed in Itti and Baldi (2009) has been well-supported by behavioral data; saliency maps generated by this model predict the direction of human eye gaze, across subjects, at upwards of 80% accuracy. Nonetheless, the extent to which this model captures computations carried out at the level of neural systems mediating saliency processing is an open question. The aim of the current project is to determine which neural substrates, if any, compute visual saliency in a manner analogous to that implemented by the model.

To link the saliency model to neural data, functional MRI (fMRI) was used to index neural activity while human subjects were presented with a movie, *Indiana Jones – Raiders of the Lost Ark* (Haxby et al., 2011). The saliency model was then applied to the same movie stimulus so as to produce saliency maps for each time point in the movie. The output of the model was then fit to the neural data such that for each voxel, the time course of blood oxygenation level-dependent (BOLD) responses could be predicted as a linear combination of the saliency values computed at each patch of each frame. Voxels in which the BOLD time course is well-predicted by the model can then be considered to comprise neural systems that process visual input in a way similar to the model. This method of model-based decoding has been successfully applied to low-level visual processing (Nishimoto et al., 2011) and semantic representation (Huth et al., 2012; Mitchell et al., 2008). The primary advantage of model-based decoding over more conventional neuroimaging approaches is that the model provides a principled set of a features and their relation to neural activation, thus allowing for generalization to novel stimuli.

Method and results

Saliency model

The saliency model constructed by Itti and colleagues was implemented via the *iLab Neuromorphic Vision C++ Toolkit* (*iNVT*; http://ilab.usc.edu/toolkit/). For our purposes, the model can be broken down into two functional components. The first of these consists of several banks of filters intended to simulate the receptive fields of neurons in early visual cortex (EVC). These filters mimic antagonistic center-surround receptive fields sensitive to local contrasts in several different feature spaces. The feature maps used in our implementation were color opponency, flicker, intensity, motion-energy, and orientation, all computed in a center-surround fashion at four spatial frequencies. Figure 1 presents a schematic of this model. A similar model (Serre, Wolf, Bileschi,

Riesenhuber, & Poggio, 2007) was used to accurately predict voxel time course in early visual cortex in response to movie stimuli (Nishimoto et al., 2011).

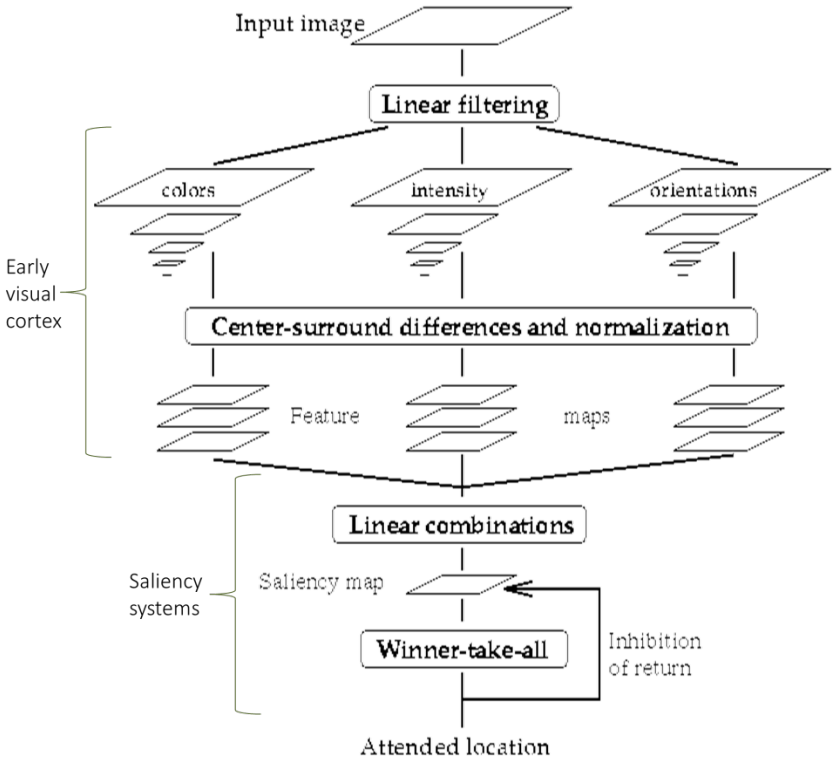

Figure 1. Schematic of the saliency model. Feature maps are computed via antagonistic center-surround filters. A saliency metric is then computed across all feature maps and the maximally salient point in visual space is selected. The model can be divided into two functional modules—the early visual component and the saliency component—as illustrated.

The second functional component of the model computes, over each of these features maps, a saliency metric based on Bayesian surprise (roughly the Kullback-Liebler divergence) from frame to frame and selects the maximum over all feature maps. This portion of the model is intended to reflect the selection process of bottom-up visual attention. In addition to the fitting the output of the final saliency computation to the neural data, we also separately analyzed the output of the EVC model component as a control. Due to the low temporal resolution of fMRI, the saliency maps were computed for each frame then collapsed into 2.5 s chunks. The model was applied to the entire movie, frame by frame, resulting in a time series of saliency values at 33 × 60 patches. For the purpose of illustration, the output of the early visual component of the model for several frames of the movie stimulus is presented in Figure 2.

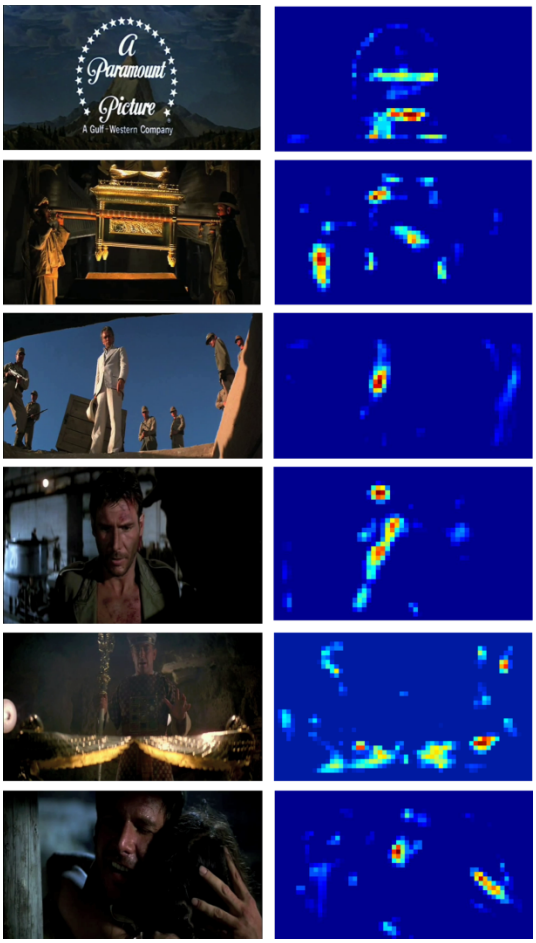

Figure 2. Output of the early visual component of the model for selected movie frames.

fMRI preprocessing

The BOLD data used in the study were from a single participant who viewed the entire movie over the course of eight functional runs. The data were preprocessed in AFNI (http://afni.nimh.nih.gov/; Cox, 1996) according to the standard pipeline. Volumes over all runes were spatially registered to a single reference volume. Slice-timing alignment corrected for temporal disparities in interleaved slice acquisition within a single volume. BOLD data were then despiked and bandpassed in order to minimize the effects of statistical outliers and low frequency drift in the MR signal. Head motion parameters returned by the initial spatial registration were regressed out of the BOLD time series to minimize effects of head movement. A 4 mm spatial smoothing kernel was applied to the data in order to increase signal-to-noise ratio. Fully preprocessed images were then multiplied by a binary mask tightly conforming to the edges of the brain in order to negate signal in voxels corresponding to the skull and adjacent non-cortical tissue. Finally, due to computational constraints and the high dimensionality of fMRI data (2,718 time points × 71,773 voxels for the current participant), signal values in the volume domain were projected to the surface and spatially down-sampled using FreeSurfer (Fischl, Sereno, & Dale, 1999) and AFNI's SUMA utility. Effectively, this ignores all voxels corresponding to subcortical tissue and white matter and projects cortical voxels onto a cortical sheet in the surface domain. This reduced our data to signal values at 2,562 nodes (the surface equivalent of voxels) in a single (left) hemisphere. The 2,718 signal values for each 2.5 s time point at each of 2,562 cortical nodes were the time series to which the output of the saliency model were fitted. After complete preprocessing, these data were imported into MATLAB where voxel-wise model fitting was performed.

Voxel-wise model fitting

To characterize the saliency information carried by the brain, we tested for each node whether the response time course could be estimated by the saliency values corresponding to each time point in the movie (see Figure 3). The spatial resolution of the saliency maps was 33 × 60 patches per frame, and the length of the movie was 2,718 time points. The saliency values at these 33 × 60 patches, or pixels, for each time point comprise the features used in optimization. To account for the lag of the hemodynamic response (essentially the delay between a neural event and the corresponding blood oxygenation), we introduced temporal smoothing and a 5 s delay to the time series of saliency maps without changing the length of the movie. This resulted in a 2,718 (time point) × 1,980 (features/pixels) matrix X for the regression problem $Xw^i = y^i$ in which y^i is the time series of activity of one surface node i.

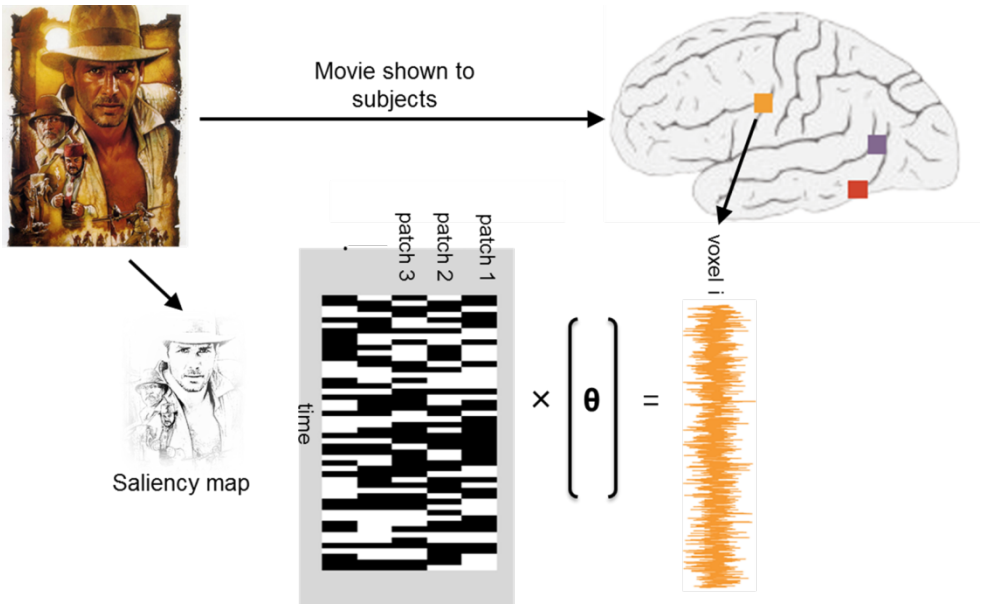

Figure 3. Schematic of the voxel-wise model fitting procedure. First, the movie is presented to human subjects while BOLD responses are recorded. The saliency model is then applied to the movie stimulus to produce a time series of saliency maps. Model parameters of the saliency maps are optimized so as to best predict single-voxel BOLD time course.

We used smooth support vector regression (SSVR) to find the optimal model weights w^i for each node. In the method proposed by Lee, Hsieh, and Huang (2005; http://dmlab8.csie.ntust.edu.tw/ssvmtoolbox.html), the ε-insensitive SSVR minimizes the following unconstrained problem by the Newton-Armijo Algorithm

$$\min_{(w,b)} \frac{1}{2}(w^T w + b^2) + \frac{C}{2}\sum[Xw^i + b - y^i]^2.$$

Parameter C is the tradeoff between the fitting error and the flatness of the weight vector. Estimated activity can be computed as $\hat{y} = Xw^i + b$. High correlation between this predicted time series of activity and actual neural activity indicates that the cortical node carries saliency information; that is, the cortical node responds to the movie stimulus in a way that is accurately captured by the saliency model. We repeated this same procedure using only the output of the early visual component of the model, prior to saliency computation, to provide grounds for comparison. Results from both the early visual cortex model component and the subsequent saliency component

are presented. These correlation values for each voxel were exported from MATLAB and projected back onto the cortical surface for the purpose of visualization using SUMA. These topographies are depicted in Figure 4.

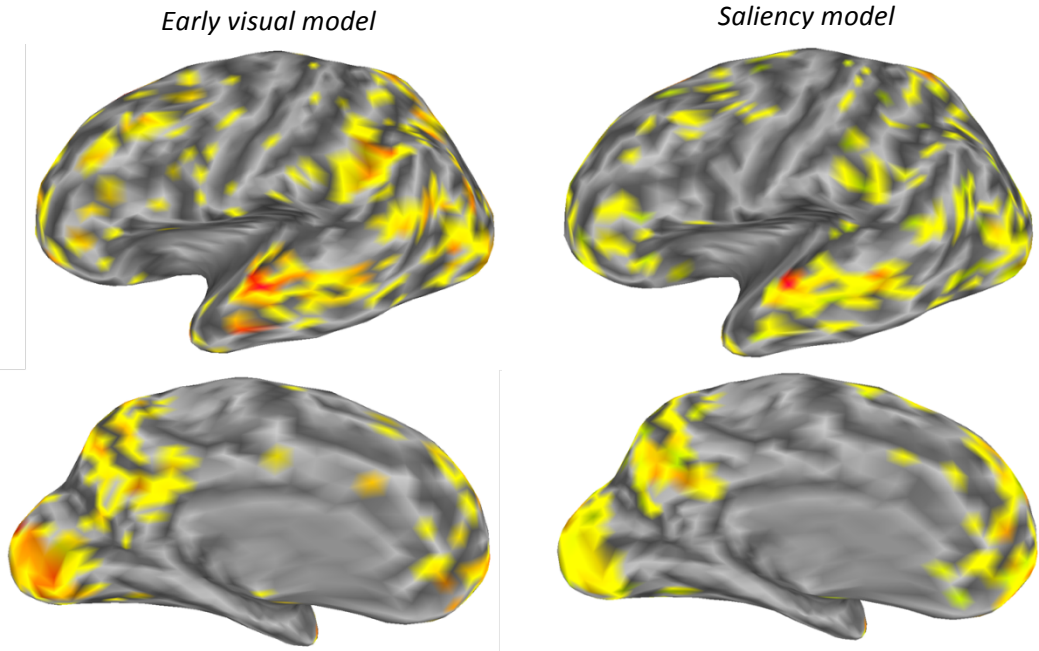

Figure 4. Cortical nodes with highest correlation between predicted and actual voxel time course for both the early visual component of the model and the output of saliency computation. Correlation strength depicted as increasing from yellow to red.

Principal component analysis

After fitting the regression model on all the voxels, each voxel can be describe by its estimated model weights. We define a model weight matrix

$$M = \begin{matrix} w_1^T \\ w_2^T \\ \vdots \end{matrix}.$$

We applied principal component analysis (PCA) to M. The purpose of this analysis is to reduce the dimensionality of the model weights returned by the regression. Specifically, we are interested in the topography of the values for the first several principal components (PCs) when projected back onto the cortex. The values of i^{th} component is computed as $p^i = M * e^j$, in which e^j is the eigenvector of $M^T M$. The values of the first several PCs are projected onto the cortical surface in Figure 5. We also performed the same analysis for the output of the early visual model.

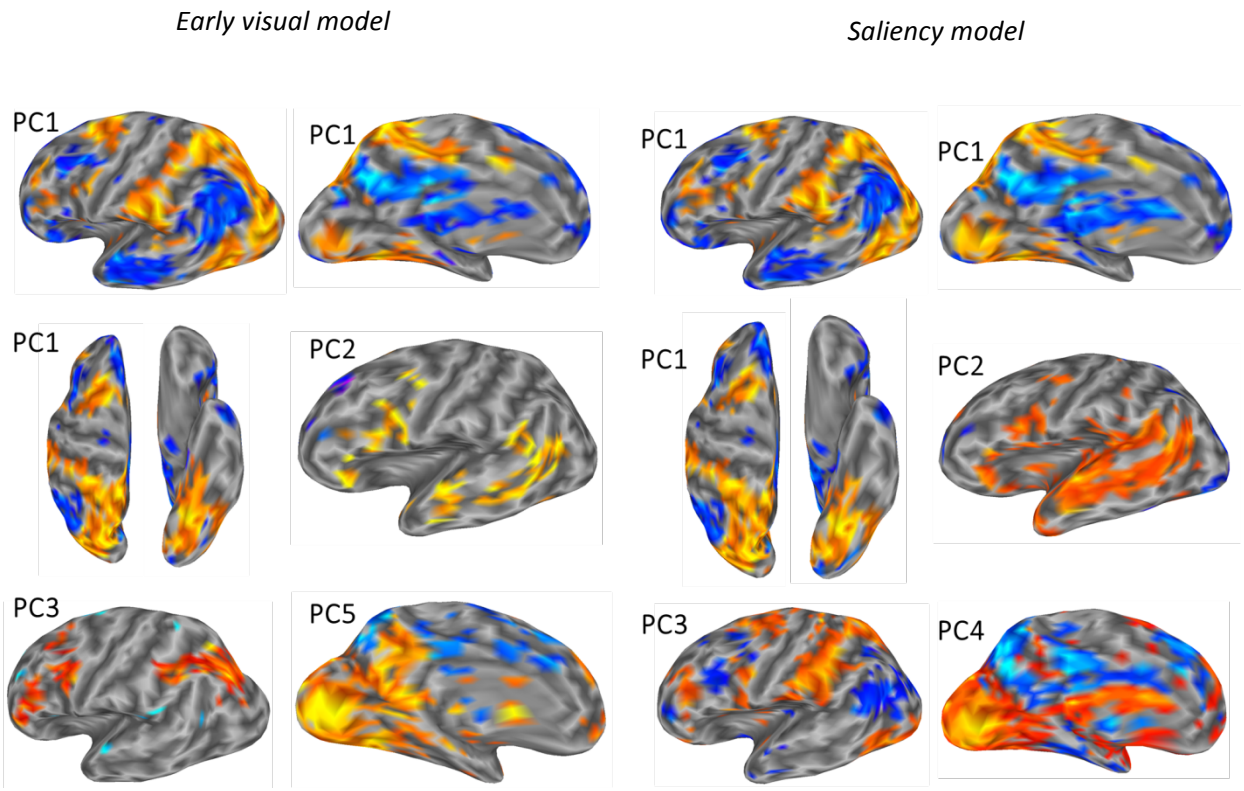

Figure 5. Principle components of model weight matrix for both early visual and saliency model components projected onto the cortical surface. Positive and negative PC values are distinguished by hot and cold colors.

It is worth noting that the values p derived in the PCA above, are equivalent to the eigenvectors of $M^T M$. The values projected to the cortical surface in Figure 5 depict the nodes that were highly correlated (or anti-correlated) when each node was represented in terms of its model weights. Thus, the results displayed in Figure 5 can be viewed as the most important networks involved in processing the saliency of the movie stimulus. A critical observation is that the cortical topography of PC values in Figure 5 is distributed and smooth—one network (i.e., topography of PC values) is distributed across different brain areas, and neighboring nodes tend to have similar values and constitute local clusters. Future work is needed to quantify the smoothness and clustering of the PC values.

As a control, we also ran an identical PCA on the BOLD data, without reference to the models or weights. The purpose of this analysis was to demonstrate that the cortical topographies captured by model weight PCs in fact differ from the networks inherent to the brain. Results of this analysis are presented in Figure 6.

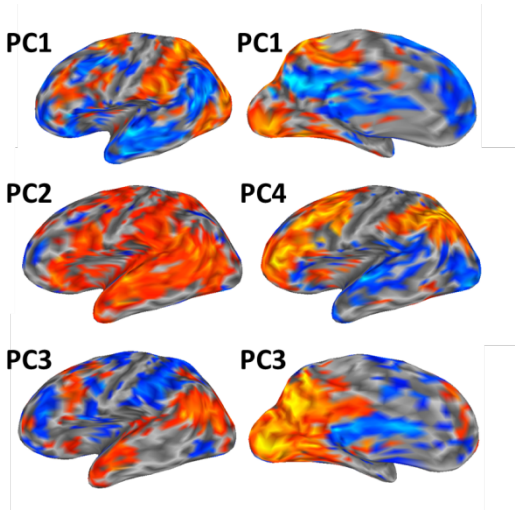

Figure 6. Principle components of BOLD data projected onto cortical surface.

We tested whether the PCs derived from model weight matrix *M* were similar to the PCs of the time series of saliency maps *X.* If this were the case, the patterns of activity would be a simple replica of the time series of saliency values. We applied PCA to the saliency time series—the *X* matrix in SVR—and used the weights to explain the variance in model weight matrix *M*. The results are plotted in Figure 7.

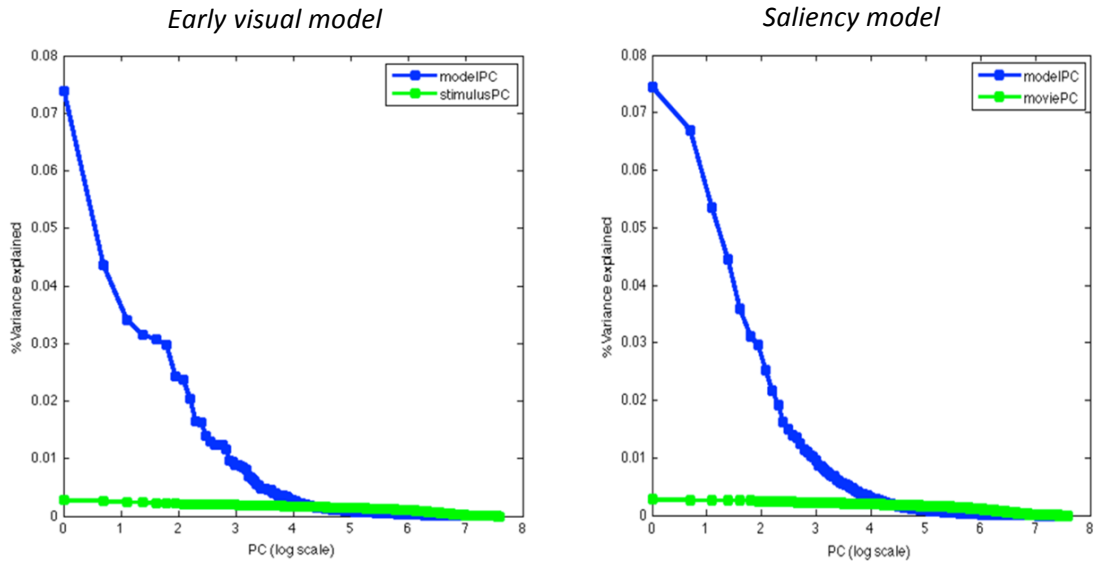

Figure 7. Variance of model weight matrix explained by model weight PCs as compared to unfitted saliency values for both early visual and saliency model components.

Discussion

In the current project, we examined the brain regions involved in processing two different types of visual information. The first corresponds to low-level visual information processed by early visual cortex, while the second includes a downstream saliency computation. The same analysis pipeline was applied to both representations. We went on to compare the neural systems for which the early visual model and the saliency model best predict voxel time series. We then applied PCA to the

fitted model weight matrices for both model components and projected these PCs onto the cortical surface to examine their topography.

In comparing the correlation values between predicted and actual patterns of activation for both the EVC and saliency models (see Figure 4), we see that the cortical topographies are qualitatively very similar. This suggests that saliency computation performed by the model does not correspond to a computation carried out in radically different neural systems than the low-level visual computations. A possible explanation for this is that the saliency computation truly does not map onto to separate systems, but is embedded in the early visual systems at the level of neuronal circuitry. Processes carried out at this fine-grained anatomical scale are largely inaccessible to fMRI because its spatial resolution averages over thousands of neurons within a single voxel.

Nonetheless, some qualitative differences can be seen between the correlation maps for the two model components. Notably, the early visual model seems to better capture activity in the cuneus and lingual gyrus, two of the earliest structures in the visual pathway. This suggests that the EVC model better captures very rudimentary visual processing and retinotopy better than the saliency model. Furthermore, activity in inferomedial frontal and medial parietal cortex (i.e., precuneus) appears to be better predicted by the saliency model. This is a promising result because these areas are implicated in higher-level task-related processing, and particularly the basal forebrain is responsible for supplying the acetylcholine driving attentional enhancement effects in sensory cortices (Sarter, Bruno, & Turchi, 1999). Finally, both models tend to accurately predict activity in superior temporal sulcus and nearby lateral temporal areas. These areas are responsible for biological motion processing and likely reflect the brain's strong sensitivity to the motion energy of intentional agents prevalent in the film (Thompson, Clarke, Stewart, & Puce, 2005).

The first PC returned by the PCA corresponds to very similar cortical topographies when compared across the EVC and saliency model outputs. This suggests that the PC accounting for the most variance in the model weight space is largely identical between the models. This is unsurprising, given that the cortical topography of this PC reflects almost the entire secondary visual system, capturing both dorsal and ventral visual pathways (Mishkin, Ungerleider, & Macko, 1983). As can be seen in the ventral view of the first PC topography, the fusiform face area (FFA) is very well-predicted by both models. This likely reflects the fact that much of the movie consists of faces presented in the center of the display. The second PC maps onto lateral temporal cortex for both models, and as above likely reflects the processing of biological motion.

Interestingly, after the first two PCs, the model weight spaces appear to diverge. While PC3 of the EVC model is not easily interpretable, PC3 of the saliency model cleanly captures both the postcentral gyrus and the entire superior parietal lobule (SPL), and the frontal eye fields (FEF). This particularly notable because these regions comprise the putative fronto-parietal attention network (Corbetta & Shulman, 2002). FEF controls eye movements and is likely better approximated by the model's winner-take-all saliency component; that is, the saliency model predicts eye gaze, which is largely driven by FEF. Furthermore, the identified parietal areas are thought to represent coordinate maps of the visual field anchored to different origins (e.g., the head, an effector), which may be better captured by the saliency model's more localized output (Sereno, Pitzalis, & Martinez, 2001).

Finally, Figure 4 demonstrates that PCs for both models capture early visual activity, but that this occurs for PC5 for the EVC model and PC4 for the saliency model. The fact that PCs of roughly similar cortical topography are returned in different orders demonstrates that the weight spaces for each fitted model differ significantly. Importantly, the control PCA applied to the BOLD data

revealed that, although the first two PCs are highly consistent and likely general to inherent brain function, the other PCs are qualitatively divergent. This suggests that the weight space for both models does not simply replicate inherent brain networks, but that in fact both models capture different neural systems.

As a sanity check, we also examined the amount of variance of the weight space for both EVC and saliency models accounted for by the model weight PCs as compared to the unfitted maps output by both model components. The fact that the fitted model weight PCs capture radically more variance in the weight space simply proves that the fitted model weights are not simply a replication of the actual unfitted maps.

In future work, we hope to introduce statistical tests to compare the sets of parameters estimated for the two model components. Such a difference analysis will point out more specifically the brain regions that diverge according to the two model components. Statistical testing of the correlation between predicted and actual BOLD time series can be accomplished by bootstrapping for each voxel. By permuting the actual time series of activation for y^i, we can simulate the distribution of the correlation between y^i and $\hat{y}$. The simulated distribution would enable us determine at which voxels the two model outputs perform differently in predicting actual BOLD time series. We implemented this analysis in MATLAB, but were unsuccessful in mapping it onto the cortical surface in an interpretable way.

Within a single subject, it is difficult to use bootstrapping to test the parameters estimated by PCA because the decomposition of $M^T M$ is time consuming. However, it is possible to validate the principle component in a held-out validation dataset for each subject. In the future, by including more subjects in the analysis and functionally aligning their brains into a common space (Haxby et al., 2011), we should be able perform these analyses across subjects and determine in what ways subjects differ in terms of the correlation of between predicted and actual neural data and the cortical topographies of the model weight PCs.

External codes and material:

1. We used SSVR script developed by Yuh-Jye Lee, Wen-Feng Hsieh and Chien-Ming Huang (2005) (modified by S.Y. Huang based on original authors' SSVR_M code)
http://dmlab1.csie.ntust.edu.tw
2. We used the iLab Neuromorphic Vision C++ Toolkit to process the movie, and generate the outputs of early visual model and saliency model.
http://ilab.usc.edu/toolkit/downloads-virtualbox.shtml

fMRI data was pre-processed by Python codes developed in Jim Haxby's lab, and two set of neuroimaging toolbox:
AFNI: http://afni.nimh.nih.gov/afni
SUMA: http://afni.nimh.nih.gov/afni/suma
These fMRI processing are run under Python

*important command for iLab Neuromorphic Vision C++ Toolkit, AFNI and SUMA are listed in pre_processing.m
*see final_code.zip file and the README file for more information

References

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 215-229.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162-173.

Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis. *Neuroimage, 9*(2), 195-207.

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., . . . Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron, 72*(2), 404-416.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A Continuous Semantic Space Describes the Representation of Thousands of Object and Action Categories across the Human Brain. *Neuron, 76*(6), 1210-1224.

Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research, 49*(10), 1295-1306.

Itti, L., & Koch, C. (2001). Computational modeling of visual attention. *Nature Reviews Neuroscience, 2*(3), 194-203.

Lee, Y.-J., Hsieh, W.-F., & Huang, C.-M. (2005). epsilon-SSVR: A Smooth Support Vector Machine for epsilon-Insensitive Regression. *IEEE Transactions on Knowledge and Data Engineering*, 678-685.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences, 6*, 414-417.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K. M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191-1195.

Nishimoto, S., Vu, A. T., Naselaris, T., Benjamini, Y., Yu, B., & Gallant, J. L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology, 21*(19), 1641-1646.

Sarter, M., Bruno, J. P., & Turchi, J. (1999). Basal forebrain afferent projections modulating cortical acetylcholine, attention, and implications for neuropsychiatric disorders. *Annals of the New York Academy of Sciences, 877*(1), 368-382.

Sereno, M., Pitzalis, S., & Martinez, A. (2001). Mapping of contralateral space in retinotopic coordinates by a parietal cortical area in humans. *Science, 294*(5545), 1350-1354.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 411-426.

Thompson, J. C., Clarke, M., Stewart, T., & Puce, A. (2005). Configural processing of biological motion in human superior temporal sulcus. *The Journal of Neuroscience, 25*(39), 9059-9066.