# Visual Features for Invariant Coding by Face Selective Neurons

**Wilbert Zarco and Winrich Freiwald**
{wzarco, wfreiwald}@rockefeller.edu
Laboratory of Neural Systems, The Rockefeller University
1230 York Avenue New York, NY 10065 USA

## Abstract

**Complex visual objects like faces are encoded through the primate ventral visual pathway in a network of cortical patches. Neurons across the nodes display specialized tuning to faces and increasing tolerance to image transformations. However, the exact features that neurons in different nodes use to attain selectivity and tolerance remain elusive. In this paper, we first quantified the representational content of neural populations in two fMRI-identified face patches to four attributes: viewpoint, identity, expression and mirror-symmetry. We found that neural population activity is driven by compartamentalized time-varying image attributes, and that multiple variables are represented in the anterior but not the posterior face patch. We then derived maps of feature selectivity by sampling images with Gaussian apertures that linked the evoked neuronal activity and the informative image features (IFs). This allowed us to evaluate the relationship between IFs and global stimulus tuning. We report that the set of discovered IFs explain the patterns of dissimilarity for the global viewpoint tuning. The alphabet of IFs also preserves local image preferences across changes in size and position. Crucially, the derived features are interpretable, and tend to cluster on consistent image regions, providing information about the global tuning that organize the neurons into functional groups.**

**Keywords:** face processing; view-invariance; representational geometries; informative features

## Background

The quest on where, when and how neural circuits encode behavioral relevant information has proven foundational for visual neuroscience. In the ventral stream neuronal responses to oriented edges, curvature, surfaces and symmetry axes appear to form fundamental primitives on which more complex relational descriptions are built (Connor & Knierim, 2017). One ecologically relevant higher-level abstraction is the processing of faces by a cortical network of distributed modules (Freiwald & Tsao, 2010). Neurons across the network show selectivity and tolerance to stimulus transformations, nonetheless, the discriminative features learned by the neurons are still unknown. Some approaches using parameterized, artificial stimuli (Freiwald, Tsao, & Livingstone, 2009; Ohayon, Freiwald, & Tsao, 2012; Tsunoda, Yamane, Nishizaki, & Tanifuji, 2001) have revealed important tuning principles and provided cues as to what the neurons are really coding. However, there is a debate on whether the features that drive neuronal activity with natural stimuli are the same or even equivalent to the ones measured with artificial stimuli (Felsen & Dan, 2005). Here, we set out to disentangle in two cortical face selective patches, the temporal dynamics of the representational geometries about four stimulus dimensions: viewpoint, identity, expression and mirror-symmetry. Our second aim was to reveal the underlying IFs that support selectivity and tolerance across the quantified dimensions. To address this challenge, we combined fMRI-guided electrophysiology, optimal stimulus search and generation with multivariate analysis. We found a differential encoding of dimensions in the two patches and over time. The IFs are localized, tolerant and able to elicit tuned neural activity. Most importantly, as we will show, the extracted informative stimulus features are often interpretable and provide functional stimulus maps of information use that constrain theories of computation.

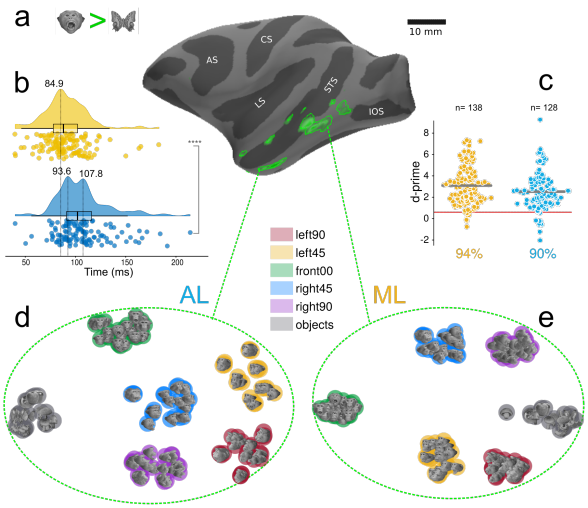


Figure 1: fMRI-guided electrophysiology and neural subpopulations characteristics. (a) Face-selective patches depicted in an inflated cortical surface for one macaque subject. The color shows the thresholded statistical value of the contrast between face images and non-face objects. Regions outside the temporal lobe have been masked out. AS, arcuate sulcus; CS, central sulcus; LS, lateral sulcus; STS, superior temporal sulcus; IOS, inferior occipital sulcus. (b) Distributions of single unit onset latencies for ML (yellow) and AL (light-blue). (c) Computed d-prime sensitivity index as a measure of face selectivity. The gray bars denote the average, and the red line a d-prime of 0.65. (d & e) High-dimensional tuning based on neural distances (1 - r) of the stimuli, visualized with t-SNE.

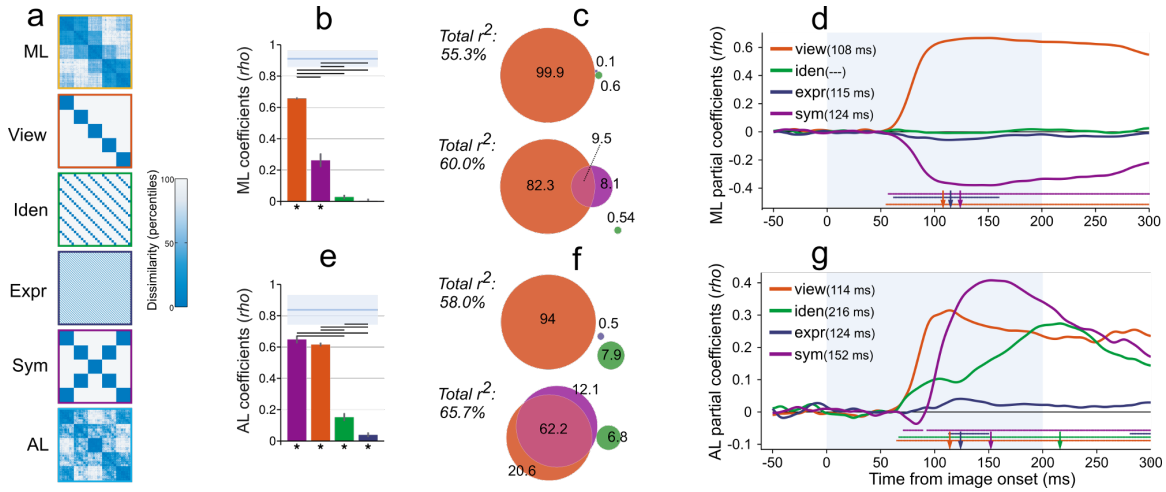


Figure 2: Neural information encoding estimated with representational similarity analysis (RSA). (a) Neural (ML & AL) and categorical predictors modeling four stimuli characteristics. (b & e) Correlations between the neural dissimilarity matrices (RDM) and the predictors RDMs. Asterisks (*) indicate $p<0.01$ for model-specific stimuli-label randomization tests, while top horizontal bars indicate $p<0.01$ for pair-wise stimulus conditions bootstrap tests between models; p-values are FDR-corrected across both types of comparisons. The light-blue top bar reflects the upper and lower bounds of the noise ceiling, indicating the expected explainable similarity given the noise in the data. Error bars show the SEM based on bootstrap resampling of the stimulus set. (c & f) Euler diagrams showing the proportions of explained unique and shared variance for two arrangements of three models predictors. Unique (non-overlapping sets) and shared (overlapping sets) variances are expressed as percentages of the total variance explained by all models combined. (d & g) Models performance evaluated by partial correlations across time. Color coded lines below the traces indicate time points with effects significantly exceeding baseline (nonparametric cluster-correction; cluster inclusion and significance level $p<0.01$). Arrowheads denote the first time point of peak correlation for each corresponding trace.

## Methods

### fMRI-guided electrophysiology

We used a block design face localizer with eight image categories to map the face patches in three awake macaques, the full methods can be found elsewhere (Schwiedrzik & Freiwald, 2017). ML and AL were then targeted for electrophysiological sampling aided by a neuro-navigation software to plan the penetration trajectories (Ohayon & Tsao, 2012). We compiled a stimulus set (VIE) of macaque faces with 6 identities, 5 viewpoints (L90, L45, F00, R45 and R90), and 4 expressions (neutral, feargrin, lipsmack and threat), plus 24 non-face objects for a total of 144 images (see Figure 1 d & e). Subjects performed a rapid serial visual presentation (RSVP) task while eye fixating within a $2^{\circ}$x $2^{\circ}$ window on the screen. On each session, the electrode (1-3 M$\Omega$) was advanced to reach the face patch, from there on every neuron spaced 200 $\mu$m along the trajectory was measured. Three main experiments were run. In Experiment 1, receptive field mapping and neuron image preference was determined. Experiment 2, aimed to derive IF maps for the cell-specific image tuning and transformations. Finally, Experiment 3 evaluated the efficacy of recovered IFs to drive neural responses.

### Experiment 1. Measuring neuronal selectivity for stimulus dimensions

For each neuron, the center of the receptive field was manually identified using a reduced version of the VIE image set. To assess image preference the full VIE stimuli was displayed at 5 Hz no gap, for at least 10 repetitions per image (this timing prescription was followed in all the experiments), and the mean responses were rank ordered. For subsequent experiments, receptive field mapping was estimated by reverse-correlating the sampled positions of the preferred image on the screen. Feature matrices were constructed z-scoring the average responses per neuron across stimuli, and used to compute pairwise dissimilarities (1-r) to reveal the representational geometry in each neural population (Kriegeskorte & Kreiman, 2012). Different categorical predictor matrices were modeled based on the stimuli characteristics (Figure 2a). To disentangle the independent contribution of each model to the elicited neural representation, we applied commonality analysis to the data (Groen et al., 2018). We conducted a separate time-resolved RSA, in which each model similarity to the neural RDMs was evaluated partialling out all but one model at the time (Figure 2 d & g).

### Experiment 2. Spike-triggered informative features (STIF)

Neuronal driving features were estimated using an unbiased random sampling of the stimulus with a single 2D Gaussian aperture ($\sigma$=16 pixels) weighted by the integrated evoked response and averaged, after latency correction (Schyns, Jentzsch, Johnson, Schweinberger, & Gosselin, 2003). Apertures jumped every 200 ms for $\geq$900 iterations, and full im-

ages were scaled (range: 3°- 8°) to fit inside the receptive field. Inferential analysis to extract the IFs were done with a cluster test specifically designed for smooth Gaussian fields (Chauvin, Worsley, Schyns, Arguin, & Gosselin, 2005). STIF was run for the preferred image and it's corresponding 5 head orientations in separate blocks. The preferred image was also tested with STIF at two positions (˜25% of image size displacement) and two sizes (66% and 150% of the original size) inside of the receptive field.

### Experiment 3. Probing the efficacy of the recovered informative feature

From the informative features recovered in previous experiments, we generated a new set of images (IFE) including the 5 head orientations for the preferred full image, 5 corresponding outlines filled with spectral noise, the 5 IFs in isolation, the 5 IFs embedded in the spectral noise outlines. The idea of filling the outlines with similar face amplitude spectra is to equate for stimulus energy that may account for some variance in the response.

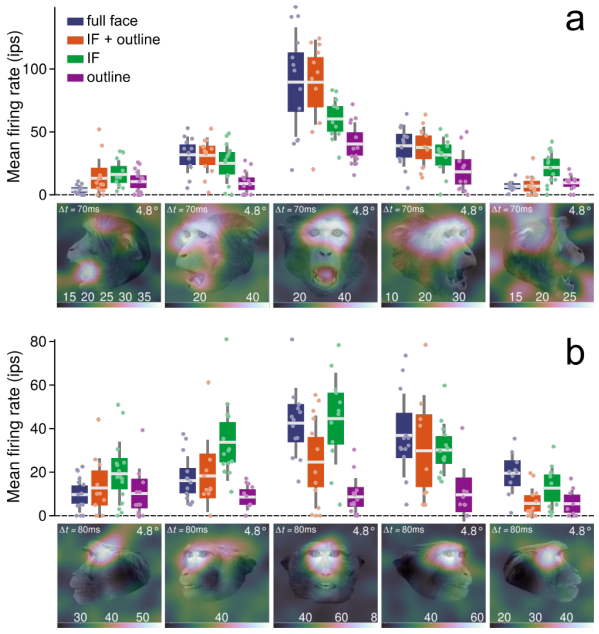


Figure 3: Examples of IFs that drive selective and tuned neural responses. (a) Top: ML neuron tuned to full front view faces (central blue boxes). Bottom: informative feature image maps recovered with STIF, where the color code indicates firing rate. (b) AL neuron also tuned to full front faces and the corresponding informative feature image maps. ips, impulses per second; white central line is the average; rectangles are SEM, and whiskers the standard deviation.

## Results and discussion

For the Experiment 1, we recorded from 138 ML single units which included 85 and 53 units in the left and right hemisphere of monkey Y and monkey O. 128 AL single units, 65 and 63

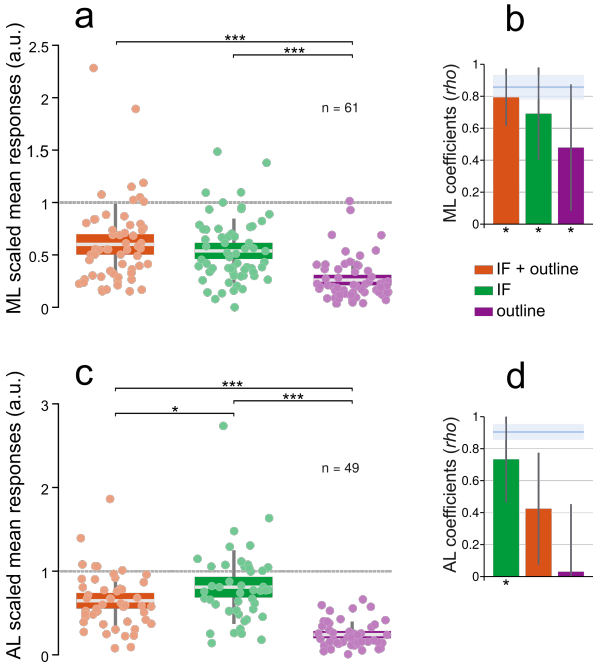


Figure 4: Population efficacy analysis of elicited neural responses by the IFE image set. (a & c) Comparative gain for IFs neural responses relative to the full preferred image (horizontal dotted unity line). (b & d) RSA evaluating the similarity patterns of three conditions for viewpoint tuning to the reference full images RDM. Statistics as in Figure 2.

units in the left and right hemisphere of monkey Y and monkey M, respectively. The number of neurons in the other experiments depended on the full completion of intended conditions. Face selectivity was estimated with d-prime, yielding 94% in ML and 90% in AL above 0.65, equivalent to a two fold response from non-face objects (Aparicio, Issa, & DiCarlo, 2016). Median onset latencies in ML and AL were 88 and 102.5 ms (Figure 1), and the distributions in the two groups differed significantly ($U$-test, $p < 0.0001$). Classical receptive field sizes were not statistically different with medians of 7.8°in ML and 8.1°in AL. Using RSA to compare the models' RDMs from the VIE dimensions (Experiment 1), to ML or AL RDMs, we found evidence that information about viewpoint is strongly encoded in ML, with a collinearity for mirror-symmetry (Figure 2b). Notably, in AL the four models were significant with precedence for mirror-symmetry (Figure 2e). To better estimate the unique and shared variance accounted for by each predictor, we performed commonality analysis on the RDMs, depicted as Euler diagrams in Figure 2c & f. The computation was split in two combinations: one excluding mirror-symmetry and the other expression. This unveiled a dissociation of collinearities among predictors, showing the dominance of viewpoint and the negligible contribution of identity and expression in ML. In contrast, we observed a representation of multiple dimensions in AL, with the highest total

variance explained ($r^2$= 65.7%). Next, we investigated how each of these representational geometries unfolds over time by computing partial Spearman correlations between models and face areas at each time point. ML showed a sustained viewpoint representation with a "knee point" at 108 ms, and a deflection anticorrelated in time for mirror-symmetry (Figure 2d). For AL we found an early peak for viewpoint (114 ms), followed by a small but significant elevation for expression (124 ms). Mirror-symmetry paralleled viewpoint dynamics (20 ms delay) with an initial negative dip, reaching a global maximum at 152 ms. This temporal signature suggest a critical role of local recurrent processing for the emergence of mirror-symmetry. Whereas identity information peaked at 216 ms, likely indicating feedback from higher visual areas.

Armed with this knowledge, we asked: to what extent are these neural representations explained by a reduced encoding of discriminative image features? Our STIF experiments revealed that both ML and AL neurons fire maximally to localized image regions, such regions are highly consistent across cells with similar viewpoint tuning (Figure 3). For example, close to 70% of the frontal tuned cells preferred the eye region, consistent with what Issa and DiCarlo (2012) reported for the posterior lateral face patch. Virtually all the IFs in both areas were tolerant to position and size manipulations[1]. The recovered features tracked the facial landmarks dependent of the tuning across head orientations, indicating that the IFs may arise from spatio-temporal associative learning biased by early retinotopy. Our data from the IFE RSVP demonstrate that most of the features in the context of the outline, or as in AL the feature alone are good enough to elicit a discriminative response (Figure 4a & c). RSA on the IFE conditions further evidence that for ML the outline still provides additional information as reported by Freiwald et al. (2009); in contrast the IF alone is the only significant comparison in AL, suggesting a more compact representation in this node (Figure 4b & d). In conclusion, these results provide evidence that a reduced set of local informative features can account for the tuning and tolerance previously observed for full faces along the ventral stream. This finding provide an alternative interpretation to the notion that neurons are selective to full faces, but are instead coding for optimized face features.

## Acknowledgments

This research was supported by the Center for Brains, Minds and Machines, the NSF STC award CCF-1231216, the NEI-NIH grant R01 EY021594, the New York Stem Cell Foundation (Robertson Investigator) to W.F. The Pew Charitable Trusts Fellows program and the NIH (International Neuroscience Fellowship F05MH094113) to W.Z.

## References

Aparicio, P. L., Issa, E. B., & DiCarlo, J. J. (2016, December). Neurophysiological Organization of the Middle Face Patch in Macaque Inferior Temporal Cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(50), 12729–12745.

Chauvin, A., Worsley, K. J., Schyns, P. G., Arguin, M., & Gosselin, F. (2005, October). Accurate statistical tests for smooth classification images. *Journal of Vision*, *5*(9), 659–667.

Connor, C. E., & Knierim, J. J. (2017, November). Integration of objects and space in perception and memory. *Nature Neuroscience*, *20*(11), 1493–1503.

Felsen, G., & Dan, Y. (2005, December). A natural approach to studying vision. *Nature Neuroscience*, *8*(12), 1643.

Freiwald, W. A., & Tsao, D. Y. (2010, November). Functional Compartmentalization and Viewpoint Generalization Within the Macaque Face-Processing System. *Science*, *330*(6005), 845–851.

Freiwald, W. A., Tsao, D. Y., & Livingstone, M. S. (2009, September). A face feature space in the macaque temporal lobe. *Nature Neuroscience*, *12*(9), 1187–1196.

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018, March). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, *7*, e32962.

Issa, E. B., & DiCarlo, J. J. (2012, November). Precedence of the eye region in neural processing of faces. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(47), 16666–16682.

Kriegeskorte, N., & Kreiman, G. (2012). *Visual Population Codes: Toward a Common Multivariate Framework for Cell Recording and Functional Imaging*. MIT Press. (Google-Books-ID: GoI5hxBEjooC)

Ohayon, S., Freiwald, W. A., & Tsao, D. Y. (2012, May). What makes a cell face selective? The importance of contrast. *Neuron*, *74*(3), 567–581.

Ohayon, S., & Tsao, D. Y. (2012, March). MR-guided stereotactic navigation. *Journal of Neuroscience Methods*, *204*(2), 389–397.

Schwiedrzik, C. M., & Freiwald, W. A. (2017, September). High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. *Neuron*, *96*(1), 89–97.e4.

Schyns, P. G., Jentzsch, I., Johnson, M., Schweinberger, S. R., & Gosselin, F. (2003, September). A principled method for determining the functionality of brain responses. *Neuroreport*, *14*(13), 1665–1669.

Tsunoda, K., Yamane, Y., Nishizaki, M., & Tanifuji, M. (2001, August). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, *4*(8), 832–838.

[1] measured as the normalized cross-correlation of the STIF maps.