Physiologically inspired neural model for the encoding of face spaces

Martin A. Giese[a*], David A. Leopold[b]

[a]*ARL, Dept. of Cognitive Neurology, University Clinic Tübingen, Germany*
[b]*Max Planck Institute for Biological Cybernetics, Tübingen, Germany*

Abstract

The neural principles of the encoding of face spaces in visual cortex are still unclear and multiple competing theories have been proposed. Based on new electrophysiological data from macaque area IT we test two models realizing *example-based* and *norm-referenced* encoding. Comparing the experimentally measured tuning properties with predictions from the two models we find a better agreement for the norm-referenced encoding model. This suggests that a majority of IT neurons might represent deviations from a norm-face, which is determined by an average over the distribution of typically occurring faces.

Keywords: face space, norm-referenced encoding, prototype, area IT

1 Introduction

The recognition of faces has been studied extensively in psychology and neurophysiology [8,11,12]. However, the principles underlying the neural encoding of faces remain a matter of dispute. Due to the lack of conclusive neurophysiological data a number of competing theoretical models for the encoding of face spaces have been proposed (e.g. [11,10,5]). In this paper we try to gain some new insights in the neural mechanisms of the encoding of face spaces by comparing two models that implement different encoding principles with electrophysiological data from

[1] *Corresponding author. *Email address:* martin.giese@uni-tuebingen.de

[2] We thank C. Wallraven for the generation of the face stimuli, and R. Sigala for help with the simulations. M.G. is supported by the German Volkswagen Foundation. We are grateful to H.H. Bülthoff for providing additional support.

monkey inferotemporal cortex [4]. The experimental data seems to be more compatible with the model that realizes norm-referenced encoding. This indicates that the responses of a majority of IT neurons might reflects the deviation of faces from an average stimulus, or average face.

In the following, we will first briefly review the face space concept and the two encoding principles that are implemented by the models. A detailed description of the models is given in section 4. A number of simulation results are presented in section 5, followed by some concluding remarks.

2 Face spaces

Face spaces have been a quite popular concept in psychology (e.g. [11]). It is assumed that faces are perceptually represented as points in an abstract high-dimensional metric space. Different possible neural encodings of such spaces have been discussed [8,11]. *Exemplar-based* theories assume an encoding by neural units that represent prototypical example faces, which can be parameterized by vectors $\mathbf{x}_n$ in an appropriately chosen feature space. The responses of the individual neurons are determined by the distances between the vector $\mathbf{x}_n$ of the encoded prototypical face and the vector $\mathbf{x}$ that represents the actual face stimulus. Another possible principle is a *norm-referenced* encoding, where faces are represented relatively to an average or norm face, corresponding to a vector $\mathbf{m}$ in feature space. The norm faces is typically defined by an average of a typical set of example faces. In this case the responses of the neurons depend on the difference $\mathbf{x} - \mathbf{m}$ between the actual stimulus and the average vector [3] . This implies that each neuron response depends on the overall statistics of typically occurring faces, and not only on an individual encoded example. Many psychophysical results seem compatible with both theo-

[3] This principle is mathematically equivalent to an example-based encoding of exaggerated prototypes [8]. In order to determine such exaggerations also knowledge about the average face is required.

ries. This motivates a study that links these two encoding principles quantitatively with electrophysiological data.

3 Stimulus set and physiological data

We used the same stimulus set as in the electrophysiological experiment by Leopold *et al.* [4] who recorded neurons in area IT in macaque visual cortex. The stimuli were normal and caricatured human faces that were generated using a *morphable model* [1] that allows to generate photo-realistic pictures of faces by linear combinations of 200 3D laser scans of human heads [4]. The synthesized face images can be parameterized by a weight vector $\mathbf{f}_n$, whose elements determine the weights of the scanned faces in the linear combination. Using this parametrization we computed the average of 50 female faces in the data basis [5] to define an average face vector $\bar{\mathbf{f}}$. Based on this average we generated two types of facial caricature stimuli:

Normal caricatures of a face $\mathbf{f_n}$ are defined by vectors of the form $\mathbf{f} = \bar{\mathbf{f}} + \lambda(\mathbf{f}_n - \bar{\mathbf{f}})$. The identity level λ determines the degree of caricature ($\lambda = 1$: original face, $\lambda > 1$: normal caricature with exaggerated individually-specific features, $0 < \lambda < 1$: "anti-caricature" with reduced features). Example stimuli are shown in Fig. 1.

Lateral caricatures of a face $\mathbf{f_n}$ were determined by interpolation on curved paths that connect it with other example faces $\mathbf{f_m}$. To generate these paths we interpolated the lengths of the vectors $\mathbf{f_n}$ and $\mathbf{f_m}$ and their directions in the high dimensional space separately. The lateral caricature level μ parameterizes these paths ($\mu = 0$ corresponding to the point $\mathbf{f_n}$, and $\mu = 1$ to the point $\mathbf{f_m}$, see Fig. 1). We selected four female faces from the data basis for the generation of the lateral caricatures that were perceived as maximally dissimilar by human subjects in a previous study.

[4] The 3D shapes of different heads, parameterized as high-dimensional polygon models that have been brought in correspondence, are linearly combined. Our stimuli had an average texture that was computed from all 200 heads in the data basis.

[5] Dimensionality was reduced using a PCA retaining 100 principal components.

These stimuli were presented to a monkeys during a fixation task. 157 neurons were recorded in the anterior part of area TE in inferotemporal cortex. For comparison with the model we used the average spike rates in an interval $200 - 300$ ms after stimulus onset. Monkeys had prior experience with human faces, but not with this specific stimulus set.

A detailed description of the experimental data is given in [4]. In this paper we only briefly review two important key results:

(1) Many neurons shows gradual and monotonic tuning with the identity level λ. The majority of tuning curves has positive slopes and no extrema for intermediate levels of λ (Fig. 2a).

(2) Tuning with the lateral caricature parameter μ is typically smooth and a significant fraction of neurons shows maximum responses for intermediate levels, i.e. for $0 < \mu < 1$ (Fig. 2b).

4 Neural model

Both implemented models consists of a hierarchy of neural feature detectors (Fig. 3). The first hierarchy levels extract local orientation and more complex form features. This part is identical for both models and consistent with many other physiologically inspired object recognition models (e.g. [7,9]). The highest hierarchy level contains the face-selective neurons whose activities are compared to the electrophysiological data from area IT.

4.1 Preprocessing

The first level of the processing hierarchy consists of Gabor filters with physiologically realistic parameters [3] that model simple cells in primary visual cortex. The gain of these filters increases with their preferred spatial frequency in order to compensate for the $1/f$-dependence of the frequency spectrum of natural images.

We used 8 preferred orientations and three different spatial scales (0.125, 0.25 and 0.5 deg per cycle). The receptive fields for each scale were strongly overlapping.

The next level of the hierarchy consists of neural detectors with larger receptive fields (bigger by factor 5) that pool the responses of the local orientation detectors on the previous level separately for each orientation and spatial frequency. Pooling is accomplished by MAXIMUM computation. The pooled responses show partial position invariance and higher robustness [9]. Maximum computation has been observed electrophysiologically for orientation-selective neurons in area V4 [2]. The receptive fields of the orientation detectors on the second hierarchy level were strongly overlapping. Only 10 % of the outputs of this hierarchy level were transmitted to the next higher level. This subset was determined by the requirement that the variance of the output signal over a representative set of faces had to exceed a certain threshold value, implementing a simple form of feature selection.

The third hierarchy level is a linear neural network that extracts significant complex form features from the outputs from the previous level. The weights of this network were determined using a principle component analysis. Principle components can be learned with multiple physiologically plausible learning rules. For the simulations we retained 20 principle components that explain about 84 % of the variability. The qualitative trends of our results were highly robust against strong variations of the parameters of all preprocessing levels (e.g. number and size of spatial scales, number of principle components, variance threshold).

4.2 Face-selective neurons

The highest hierarchy levels of our models consist of neurons that are selective for faces. The level was modelled in two different ways in order to realize the two different encoding principles. For modelling *example-based coding* we use radial basis functions. Consistent with other learning-based recognition models (e.g. [9]),

these units were trained with 49 random faces from the data basis, which were chosen to be disjoint from set that was used to generate the caricatures. This assumption reflects that the monkey had substantial opportunity to memorize other human faces before the start of the experiment [6]. The tuning function of the face-selective neurons in this implementation is given by a gaussian function:

$$v_k = C_1 \exp\left(-|\mathbf{u} - \mathbf{u}_k|^2/(2\eta^2)\right) \tag{1}$$

The 20-dimensional vectors $\mathbf{u}$ and $\mathbf{u}_k$ signify the responses of the previous layer for the actual face and the training face of model neuron k. The parameter η determines the selectivity of the neurons. It was adjusted in order to match the fraction of activated neurons in the physiological data.

Norm-referenced encoding was implemented using neurons whose responses depend explicitly on a vector $\mathbf{m}$ that represents the average or expectation of the input vectors from the previous layer over a representative set of randomly chosen faces. Interestingly, this vector is very similar to the input vector $\mathbf{u}$ that arises during presentation of the average face. The expectation of the input vector can be estimated using simple neural mechanisms, e.g. by a slow leaky integrator that averages over many stimulus presentations. The tuning function of the norm-referenced encoding units are given by a product of two terms [7]:

$$v_k = C_2|\mathbf{u} - \mathbf{m}| \left(\frac{(\mathbf{u} - \mathbf{m})^T \mathbf{n}_k}{2|\mathbf{u} - \mathbf{m}|} + \frac{1}{2} \right)^{\nu} \tag{2}$$

The first term depends only on the distance between the actual input vector $\mathbf{u}$ and the average vector $\mathbf{m}$. The second term is responsible for the tuning with respect to facial identity. It depends on the cosine of the angle between the difference vector

[6] Electrophysiological experiments show that neurons in area IT can learn to respond selectively to novel stimuli, showing tuning properties that are consistent with radial basis functions [6].

[7] For $\nu = 1$ an approximation of this function can be implemented with a simple two layer linear threshold network. This implementation was used for the simulations.

$\mathbf{u} - \mathbf{m}$ and a unit vector $\mathbf{n}_k$. This unit vector is given by a fixed set of weights that characterizes the neuron, and which is learned from training faces. The cosine of this angle is proportional to the scalar product $(\mathbf{u} - \mathbf{m})^T \mathbf{n}_k$ and realizes a direction tuning in the high dimensional input space. The positive parameter ν determines the width of this direction tuning. Reasonable fits of the data were obtained for values $\nu = 1...1.5$.

5 Results

Our models were trained with 49 randomly chosen male faces from the Max Planck data basis [1] and tested with 199 facial caricatures. During training no caricatured stimuli were presented. The stimulus set was identical with the one used in the electrophysiological experiment.

Fig. 4 show population responses from the two models implementing norm-referenced encoding (panels a and c), and example-based encoding (panels b and d). The averages were computed by ordering the responses for each neuron according to the face that elicits maximum, second-maximum, third-maximum, etc. response (taking the maximum over all identity levels of the same face). The "rank-ordered" responses were then averaged over the whole population of neurons. Panels a and b show the means and the standard errors of these average responses as function of the identity level λ. The responses of the norm-referenced encoding model (panel a) vary monotonically with the identity level and show a strong dominance of positive slopes, consistent with the physiological data (Fig. 2a). The responses of the example-based encoding model (panel b) show a dominance of negative slopes.

These observations are confirmed by a more detailed statistical analysis where the slopes were fitted using linear regression. 76% of the tuning curves of the norm-referenced model show positive slopes, comparable to the experimental data (78%), whereas only 29% of the tuning curves of the example-based model have posi-

tive slopes. We computed also the percentage of the variance of the tuning curves that can be predicted with the two models. The norm-referenced encoding model predicts 63% of the variability of the experimentally measured tuning curves. The example-based model produces huge prediction errors resulting in an error variance that exceeds the variability of the tuning curves by 270%. This model this does not predict the experimental data very well. By fitting quadratic functions we also quantified the percentage of tuning functions that have extrema for intermediate levels of λ. This percentage is closer to the experimental data for the norm-referenced model (26%) than for the example-based model (37%, experimental data 19%).

A similar analysis can be applied for the tuning functions with respect to the lateral caricature parameter μ. F1 ... F4 indicate for each neuron the face that elicits minimum, second-minimum, third-minimum, etc. responses for normal caricatures. The Fig. 4c and d show the averages and standard errors of the responses for different lateral caricature levels between F1 and all other faces, and F4 and all other faces. In this case the averaged tuning curves of the two models (panels c and d) are rather similar. A regression analysis shows that both models show significant, but higher fractions of tuning curves with intermediate extrema (36 respectively 38%) than observed for the experimental data (18%). Both models explain a significant amount of the variance (64% for the norm-referenced and 62% for the example-based model) of the average tuning curves (Fig. 2b).

6 Conclusion

We have quantitatively compared two neural models that implement norm-referenced and example-based encoding of face spaces with electrophysiological data from area IT in monkey cortex. In particular for normal caricatures the models make different predictions, which were robust against parameter changes on the different levels of the processing hierarchy. The monotonicity of the tuning curves with

respect to the identity level and the dominance of positive slopes could be better reproduced with the norm-referenced encoding model. The behavior of the two models for lateral caricatures is very similar and consistent with the data. This makes it potentially difficult to distinguish norm-referenced example-based encoding based on lateral caricatures.

Obviously, more fine tuning between the proposed models and physiological data has to be accomplished, and other alternative models have to be tested. However, the proposed framework offers a quantitative link between real images and electrophysiological data that might be quite useful to unravel underlying encoding principles, and to explore the coding efficiency of different possible physiological implementations of face spaces.

Figure legends

(1) Face space and facial caricatures generated by the morphable model with caricature levels λ and μ for normal and lateral caricatures.

(2) Schematic illustration of the most important electrophysiological results: (a) Monotonic tuning with respect to the identity level λ. (b) Gradual tuning with lateral caricature parameter μ with extrema for intermediate levels.

(3) Overview of the hierarchical model (including numbers of neurons and potentially corresponding cortical areas).

(4) Simulation results: Mean and standard errors over the whole population after rank-ordering responses by faces that induce maximum responses (see text) as functions of the caricature levels λ and μ. Panels a and c show the responses for the norm-referenced encoding model, and panels b and d responses of the example-based model. Significance of the response variation with the caricature parameters was assessed with ANOVAs ($** \; p < 0.01$, $* \; p < 0.05$).

References

[1] V. Blanz, T. Vetter, A morphable model for the synthesis of 3D faces, *SIGGRAPH'99 Conference Proceedings* (1999) 187-194.

[2] T.J. Gawne, J.M. and Martin, Responses of primate visual cortical V4 neurons to simultaneously presented stimuli, *J. Neurophysiol.* 88 (2002) 1128-1135.

[3] T.S. Lee, Image representation using 2D Gabor wavelets. *IEEE Transactions Pattern. Anal. Mach. Intell.* 18 (1996) 959-971.

[4] D.A. Leopold, I.V. Bondar, M.A. Giese, N.K. Logothetis, Prototype-referenced encoding of faces in monkey inferotemporal cortex. (submitted).

[5] M.B. Lewis, R.A. Johnston, A unified account for the effect of caricaturing faces, *Visual Cognition* 6 (1996) 1-41.

[6] N.K. Logothetis, J. Pauls, T. Poggio, Shape representation in the inferior temporal cortex of monkeys. *Curr Biology* 5 (1995) 552-63.

[7] D.I. Perrett, M.W. Oram, Neurophysiology of shape processing, *Imag. Vis. Comput.* 11 (1993) 317-333.

[8] G. Rhodes, S. Brennan, S. Carrey, Identification and ratings of caricatures: implications for mental representations of faces. *Cognititive Psychology* 19 (1987) 473-497.

[9] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. *Nature Neuroscience* 2 (1999) 1019-1025.

[10] J.W. Tanaka, V.B. Simon, Caricature recognition in a neural network, *Visual Cognition* 3 (1996) 305-324.

[11] T. Valentine, A unified account of the effects of ditinctiveness, inversion and race in face recognition, *Quart. J. Exp. Psych. A* 43 (1991) 519-554.

[12] M.P. Young, S.Yamane, Sparse population coding of faces in es in the inferotemporal cortex. *Science* 256 (1992) 1327-1331.

Figure 1

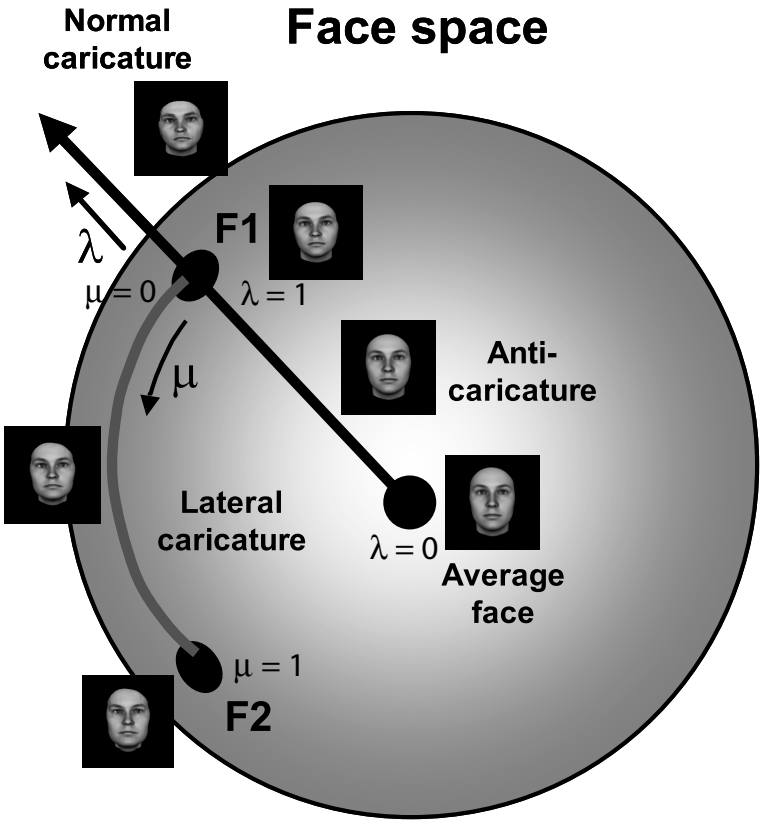

Face space
Normal caricature
λ
μ = 0
F1
λ = 1
μ
Anti-caricature
Lateral caricature
λ = 0
Average face
μ = 1
F2

Figure 2

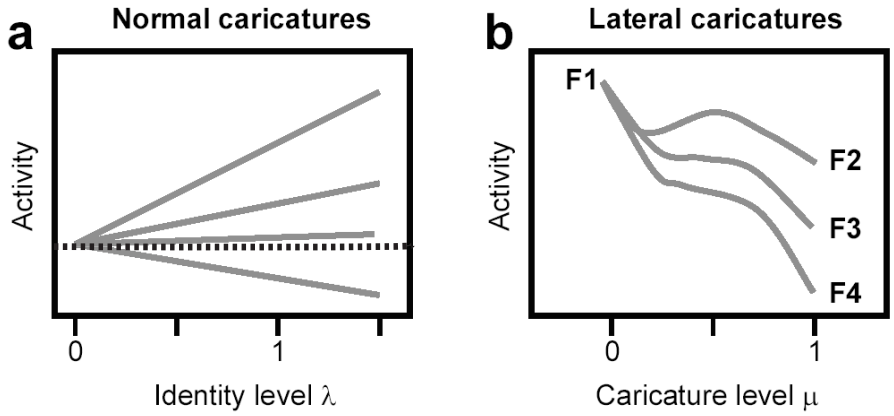


Figure 3

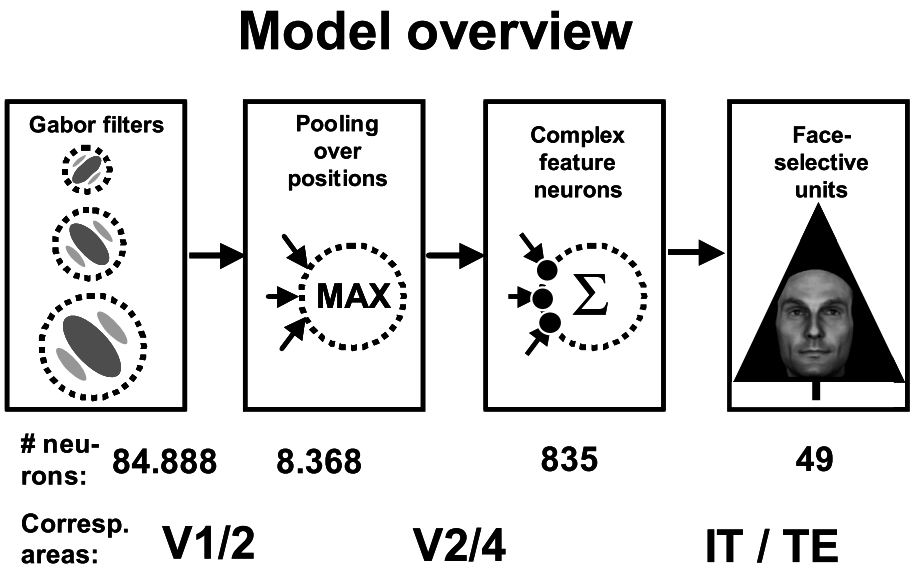
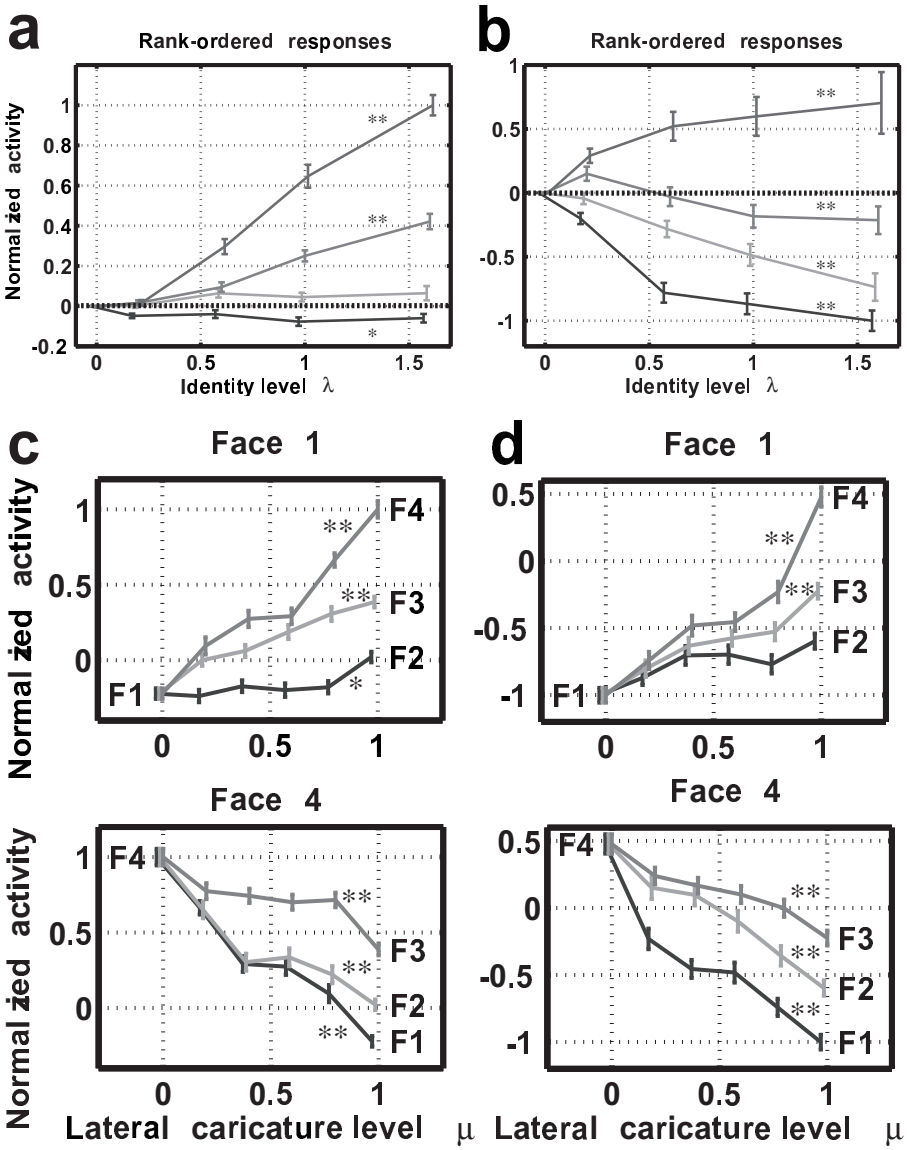


Figure 4


a
Rank-ordered responses
Normalized activity
Identity level λ
b
Rank-ordered responses
Identity level λ
c
Face 1
F4
F3
F2
F1
Face 4
Normalized activity
Lateral caricature level μ
d
Face 1
Face 4
Lateral caricature level μ