Physiologically inspired neural model for the encoding of face spaces

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

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

Abstract

Due to the lack of electrophysiological data the neural basis of the encoding of faces is largely unclear. We present a model for the neural encoding of face spaces that is based on new electrophysiological results. It reproduces important properties of the physiological data and shows that faces might be encoded exploiting a *norm-based* rather than an *example-based* neural representation. This implies that the encoding exploits an internal representation of an average (norm) face that might be derived by averaging over the previous stimulus history.

1 Introduction

The recognition of faces has been an important topic in visual psychophysics. A variety of experiments have been collected, and a number of theoretical models for the encoding of faces have been proposed [9,8,5]. However, due to the lack of conclusive neurophysiological data, so far no conclusive answer about the nature of the underlying neural representation has been found.

We present a model for the neural encoding of face spaces that is based on electrophysiological results on the representation of caricatures of human faces in monkey *inferotemporal cortex*. The model reproduces several critical properties of the phys-

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

[2] We thank 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.

Preprint submitted to Elsevier Science *23 January 2004*

iological data and provides quantitative framework for exploring possible computational implementations. This analysis leads to the conclusion that the experimental data seems to support a *norm-based* encoding of face stimuli, rather than an encoding in terms of prototypes that are corresponding to fixed points in an appropriately chosen feature space. Norm-based encoding exploits an internal representation of an average, or norm face that represents the expectation (average) of typically occurring face stimuli.

The following two sections describe briefly the idea and the key results of the electrophysiological experiment which forms the basis of our model. The model is described in section 4, and a number of simulation results are discussed in section 5. Implications of the proposed model for further experiments are discussed in the concluding section.

2 Face spaces

In the psychology of face perception the idea of *metric face spaces* has been very influential (e.g. [9]). Among others, two theories for the encoding of faces in such high-dimensional spaces have been discussed frequently. *Exemplar-based* encoding theories assume that the face space is parameterized in terms of absolute coordinates, which are valid independently from the encoded set of faces. Each face is parameterized by a vector $\mathbf{x}_n$ in this absolute coordinate system. Contrasting with this idea are *norm-based theories* that assume that faces are encoded relatively to a common reference or *normative* face, which is typically assumed to represents the average or expectation of the encoded set of faces. This average is given by a vector $\mathbf{m}$ in an appropriately chosen multidimensional space. Critical for the encoding is thus the difference vector $\mathbf{x}_n - \mathbf{m}$ rather than the absolute position $\mathbf{x}_n$ in face space. It has turned out to be difficult to dissociate between the two theories based on psychophysical data, since with appropriate extensions, both theories are consistent with many existing experimental observations. Electrophysiological data might be helpful for narrowing down the set of possible underlying computational principles.

3 Physiological data

Our model is based on an electrophysiological experiment by Leopold *et al.* [4] who tested neurons in area TE of macaque monkeys with normal and caricatured human faces. The stimuli were generated using a *morphable model* that was based on 200 3D laser scans of human heads [1]. The morphable model allows to synthesize photo-realistic new face images by linear combination [3] of the stored example faces. Each generated face is parameterized by a weight vector $\mathbf{f}_n$, whose elements determine the weights of the example faces to the linear combination.

To define caricatures we first computed the average of all 200 faces in the data basis [4] in this parameterization resulting in the average face vector $\bar{\mathbf{f}}$. Based on this average two types of facial caricature stimuli were generated:

Normal caricatures of a face $\mathbf{f_n}$ were determined by vectors of the form $\mathbf{f} = \bar{\mathbf{f}} + \lambda(\mathbf{f}_n - \bar{\mathbf{f}})$. The parameter λ determines the caricature level. The value $\lambda = 1$ corresponds to the original face. Values $\lambda > 1$ define caricatures, and values $0 < \lambda < 1$ result in anti-caricatures. Example stimuli are shown in Fig. 1.

Lateral caricatures of a face $\mathbf{f_n}$ were determined by interpolation on curved paths that connect the example face $\mathbf{f_n}$ with other faces $\mathbf{f_m}$. For this purpose the lengths of the vectors $\mathbf{f_n}$ and $\mathbf{f_m}$ and their directions (given by the angle in the plane defined

[3] The color and 3D shape information parameterized as high-dimensional polygon model are linearly combined after the heads have been brought in three-dimensional correspondence.

[4] Dimensionality was reduced using a principal components analysis retaining 100 principle components.

by $\mathbf{f_n}$, $\mathbf{f_m}$ and $\bar{\mathbf{f}}$) were separately interpolated. The caricature level μ parameterizes the curve, with $\mu = 0$ corresponding to the point $\mathbf{f_n}$, and $\mu = 1$ to the point $\mathbf{f_m}$ (cf. Fig. 1). For the generation of the lateral caricatures we selected six female faces from the data basis that were perceived as maximally dissimilar in a previous study by human subjects.

The stimuli were presented to the monkeys during a fixation task. About 85 neurons were recorded in the anterior part of area TE in inferotemporal cortex. Our model was compared to the average responses $200 - 300$ ms after stimulus onset. The monkeys had prior experience with human faces. The key results from the electrophysiological study were:

(1) Most neurons that are tuned to faces show a very gradual tuning with respect to the caricature levels λ and μ.

(2) A majority of the neurons shows a monotonic variation of their responses with the identity parameter λ. Only few neurons have higher responses for intermediate caricature levels or the average face. In particular, most neurons show higher responses for caricatures of faces than for the example faces. The activity of most neurons increases with λ.

(3) The responses of some neurons peak for intermediate levels of the parameter μ (i.e. for $0 < \mu < 1$).

4 Neural model

Like many other physiologically inspired models for object recognition (e.g. [6,7]) our model consists of a hierarchy of neural feature detectors (Fig. 2). The first levels of the model extract orientation and form features from images. These preprocessing steps are consistent with many other models. The highest level of the hierarchy contains the face-secetive neurons that are the focus of this study.

4.1 Preprocessing

The first level of the processing hierarchy consists of 84.888 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 frequeny spectrum of natural images. 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 responsea are more robust and show partial position invariance [7]. Maximum pooling has been observed electrophysiologically in orientation-selective neurons in area V4 [2]. The receptive fields of the 8368 orientation detectors on the second hierarchy level were strongly overlapping. From the responses of the detectors on this level only a subset was transferred to the next higher hierarchy level. This subset was determined by the requirement that the variance of the detector over the training set of faces had to exceeded a minimum value. The responses of about 835 (10%) of the neural detectors on this level were transmitted to the next higher hierarchy level.

The next hierarchy level computes a reduced set of maximally independent features by a simple linear neural network. The weights of this network were determined using a principle component analysis. These principle components, in principle, can be learned by a number of physiologically plausible learning rules. For the simulations in this paper we retained 20 principle components of the responses, explaining about 84 % of the variability. Results were robust against the choice of this this parameter.

4.2 Face-selective neurons

The innovative part of our model is the highest hierarchy level that consists of neurons that are selective for faces, and which receive a 20-dimensional input vector $\mathbf{u}$ from the previous level. Property (2) from the electrophysiological results suggests that the tuning of many face-selective neurons fulfills a specific symmetry with respect to the average face. To model this symmetry property we assume that the tuning of the face units in our model depends explicitly on a vector $\mathbf{m}$ that represents the average of all input vectors $\mathbf{u}$ over the training set. This vector $\mathbf{m}$ corresponds to the normative face in norm-based models.

The activity of the face-selective neurons in our model is given by a product of two terms. The first term depends only on the radial distance between the input vector $\mathbf{u}$ and the average vector $\mathbf{m}$. The second term is responsible for the selectivity of the tuning for individual faces. It depends on the cosine of the angle between the difference vector $\mathbf{u} - \mathbf{m}$ in feature space and a preferred direction vector $\mathbf{n}_k$ of the neuron. This unit vector is given by a fixed set of weights that characterizes the neuron. The cosine of this angle is proportional to the scalar product $(\mathbf{u} - \mathbf{m})^T \mathbf{n}_k$. The output activity of neuron k is given by

$$v_k = C_1 |\mathbf{u} - \mathbf{m}| \left(\frac{(\mathbf{u} - \mathbf{m})^T \mathbf{n}_k}{2|\mathbf{u} - \mathbf{m}|} + \frac{1}{2} \right)^{\nu} \tag{1}$$

C_1 is a positive constant and ν is a positive parameter that determines the selectivity of the neuron. (Reasonable fits of the data were obtained for values $\nu = 1...1.5$). In our model we introduced one neural unit for each training face. The unit vectors were set to the values $\mathbf{n}_k = (\mathbf{u}_k - \mathbf{m})/|\mathbf{u}_k - \mathbf{m}|$, where $\mathbf{u}_k$ is the input vector that corresponding to training face k.

The vector $\mathbf{m}$ can be estimated by averaging over the previous stimulus history. A simple neural circuit that yiels such an estimate by computing a *population vector* from the neural responses is given by the following simple neural dynamics:

$$\tau \dot{\mathbf{m}}(t) = \sum_k v_k \mathbf{n}_k \tag{2}$$

It can be shown that, if the time constant τ is sufficently large, the variable $\mathbf{m}(t)$ converges [5] against the expectation value of the input vectors $\mathbf{u}_k$.

For comparison we implemented a second type of face-sensitive neurons that realizes a prototype or example-based encoding of faces. In this case the activity of the neuron was simply given by a Gaussian radial basis function: $v_k = C_1 \exp(\frac{|\mathbf{u}-\mathbf{u}_k|^2}{2\eta^2})$. η determines the selectivity of the neurons, and was adjusted in order to match the fraction of activated neurons.

5 Results

The model was trained with 49 randomly chosen male faces from the Max Planck data basis [1]. During training no caricatured stimuli were presented. The model was then tested with 199 facial caricatures that were generated in the form discussed in section 3. The stimuli were identical to the ones used in the electrophysiological experiment.

The upper panel Fig. 3 shows the activity of a typical model neurons as a function of the caricature parameter λ. The different curves indicate the six different test faces. Most neurons in our model show a smooth monotonic increase of the activity with the identity parameter, consistent with properties (1) and (2) of the physiological data. This is also evident from the average plot in the lower panel. The averages were computed after the responses of each neuron had been rank ordered according to the face that elicited maximum, second maximum, etc. response. Beyond the

[5] For sufficintly slow variation of $\mathbf{m}(t)$ the equilibrium state of the dynamics is given by the equation $\mathbf{m}^* = E\{\mathbf{u}_k\} + \mathbf{R}^{-1}\mathbf{s}E\{|\mathbf{u}_k - \mathbf{m}^*|\}$ with $\mathbf{R} = \sum_k \mathbf{n}_k \mathbf{n}_k^T$ and $\mathbf{s} = \sum_k \mathbf{n}_k$. If $\mathbf{s}$ is sufficintly close to zero the last term can be neglected.

strongly monotonic trend, the plot shows also that neurons that are strongly responsive to a face, on average aso respond stronger to caricatures of the face than to the original face. This is consistent with property (2).

For comparison we tested the RBF model for the protype-based encoding with the same stimuli. Responses vary smoothly and monotonically with the caricature level, but they tend to show a decay with λ. The prototype-based encoding model seems thus to be less consistent with the electrophysiological data than the norm-based model.

We also tested the response of the model for lateral caricatures. The results for the norm-based and the radial basis function model are shown in Fig. 4. The discs in the figure indicate the six tested example faces and the brightness of the gray connection linear provides a color code for the activity of the neurons. The small black dots indicate the tested values of the parameter μ (with $0 \leq \mu \leq 1$). For both models the activity varies very smoothly with this parameter. Also there are some model neurons that are more activated for intermediate values of μ (arrow in Fig. 4), matching property (3).

6 Conclusion

We have presented a first skeleton of a neural theory for the encoding of face spaces that is motivated by physiological data. We have demonstrated that the proposed model reproduces several key properties of the electrophysiological data, and seems to suggest that faces are cortically encoded exploiting a norm-based rather than in a prototype-based encoding strategy.

Undoubtly, much more fine tuning between model and the physiological data remains to be done, and its computational limits need to be explored. Also precise neural circuits for the implementation of the individual model componenents have to be developed. However, the model provides a quantitative framework that links the behavior of individual neurons to real visual stimuli. This distinguishes our model from many previous models for the neural encoding of faces, and makes it a suitable framework for testing more specific computational hypotheses on the basis of physiological data, and the real statistical properties of face images.

Figure legends

(1) Overview of the hierarchical model.

(2) Face space and facial caricatures generated by the morphable model with caricature levels λ and μ.

(3) Results from examples of model face neurons (*top panels*), and the population response to face that elicits strongest, second strongest etc. response (*bottom panels*). Errorbar indicate SD over all neurons.

(4) Activities of two typical neurons for lateral caricatures. Arrows indicate morphs for which a lateral caricature induces a higher response than the example faces. Brightness of the connection lines indicate the neural acitvity level (white: high, and black: low activity).

References

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

[2] T.J. Gawne, J.M. and Martin (2002) Responses of Primate Visual Cortical V4 Neurons to Simultaneously Presented Stimuli. *J. Neurophysiol.* 88, 1128-1135.

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

[4] D.A. Leopold, I.V. Bondar, M.A. Giese, N.K. Logothetis (2003) Prototype-reference encoding of faces in monkey inferotemporal cortex. *Soc. of Neurosciences Abstracts*, 590.7.

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

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

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

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

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

Figure 1

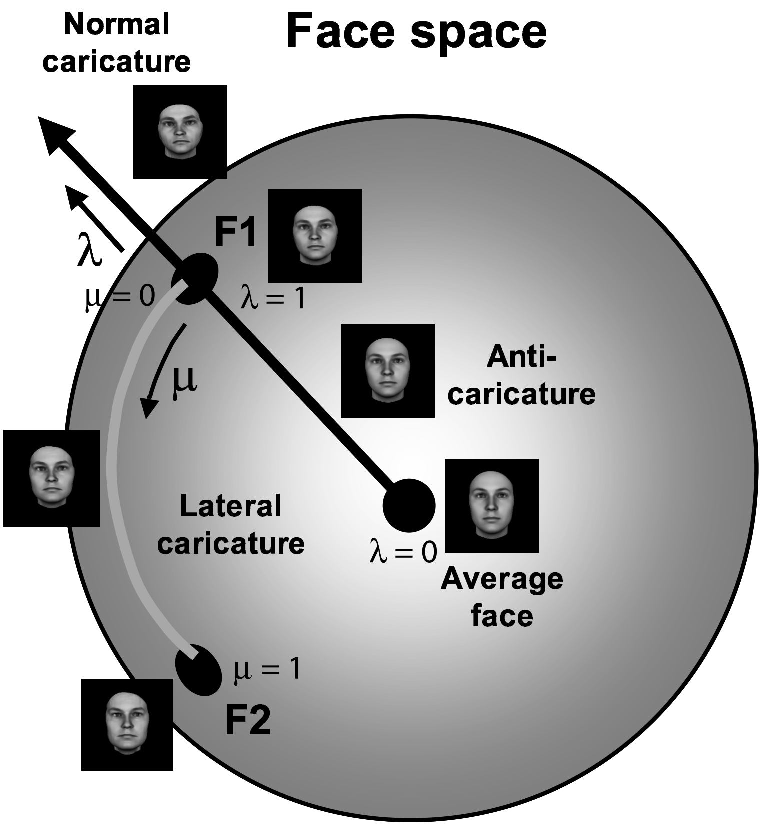


Figure 2

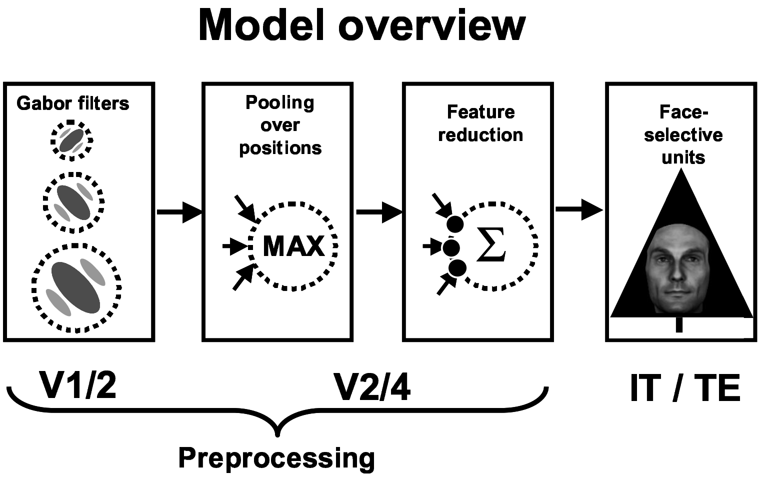


Figure 3

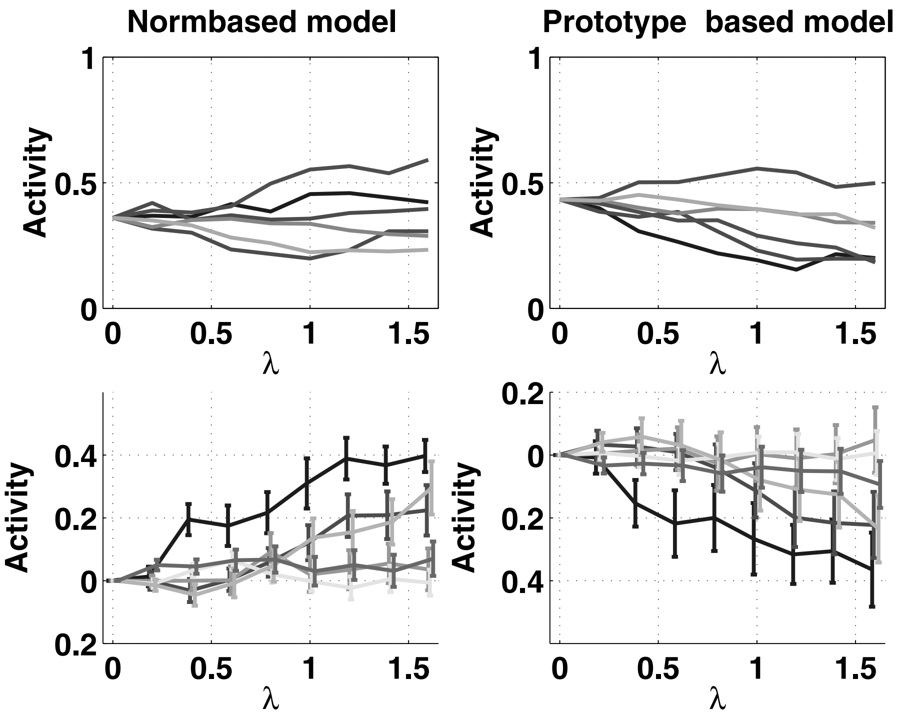
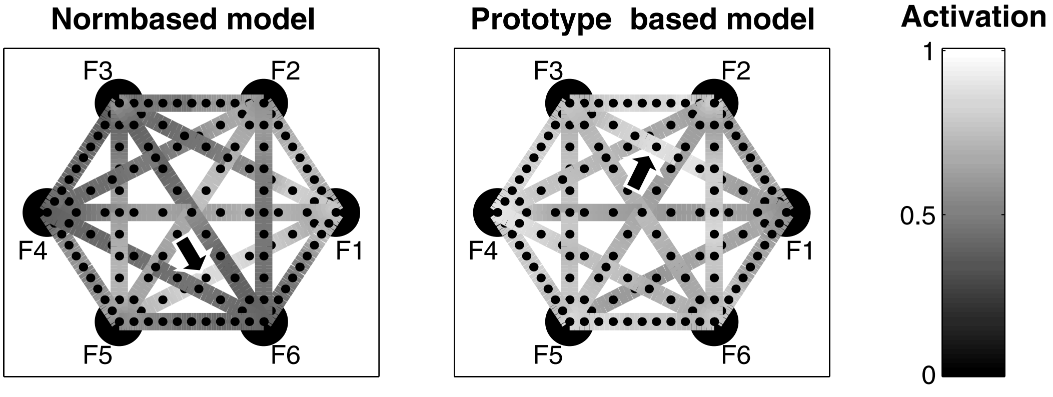


Figure 4


Normbased model
Prototype based model
Activation
F1
F2
F3
F4
F5
F6
1
0.5
0