On the Role of Object-specific Features for Real World Object Recognition in Biological Vision

Thomas Serre, Maximilian Riesenhuber, Jennifer Louie, and Tomaso Poggio

Center for Biological and Computational Learning, Mc Govern Institute for Brain Research, Artificial Intelligence Lab, and Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA
{serre, max, jenlouie, tp}@ai.mit.edu

Abstract. Models of object recognition in cortex have so far been mostly applied to tasks involving the recognition of isolated objects presented on blank backgrounds. However, ultimately models of the visual system have to prove themselves in real world object recognition tasks. Here we took a first step in this direction: We investigated the performance of the HMAX model of object recognition in cortex recently presented by Riesenhuber & Poggio [1, 2] on the task of face detection using natural images. We found that the standard version of HMAX performs rather poorly on this task, due to the low specificity of the hardwired feature set of C2 units in the model (corresponding to neurons in intermediate visual area V4) that do not show any particular tuning for faces *vs.* background. We show how visual features of intermediate complexity can be learned in HMAX using a simple learning rule. Using this rule, HMAX outperforms a classical machine vision face detection system presented in the literature. This suggests an important role for the set of features in intermediate visual areas in object recognition.

1 Introduction

Object recognition in the macaque has mostly been explored using idealized displays consisting of individual (or at most two) objects on blank backgrounds, and various models of object recognition in cortex have been proposed to interpret the data from these studies (for a review, see [1]). However, ultimately models of the visual system have to prove themselves in real world object recognition settings, where scenes usually contain several objects, varying in illumination, viewpoint, position and scale, on a cluttered background. It is thus highly interesting to investigate how existing models of object recognition in cortex perform on real-world object recognition tasks.

A particularly well-studied example of such a task in the machine vision literature is face detection. We tested the HMAX model of object recognition in cortex [1] on a face detection task with a subset of a standard database previously used in [3]. We found that the standard HMAX model failed to generalize to cluttered faces and faces with untrained illuminations, leading to poor detection performance. We therefore extend the original model and propose an algorithm, described in section 2, for learning object class-specific visual features of intermediate complexity. In section 3, we investigate the impact of the learned object-specific feature set on the model's performance for a face detection task. In particular, we trained and tested the same classifier (a Support

Vector Machine) on the two sets of outputs collected with the different feature sets (*i. e.,* the standard HMAX features *vs.* the new learned object class-specific features). As a benchmark, we added performances of a classical machine vision face detection system similar to [4].

2 Methods

2.1 HMAX

The model is an hierarchical extension of the classical paradigm [5] of building complex cells from simple cells. The circuitry consists of a hierarchy of layers leading to greater specificity and greater invariance by using two different types of pooling mechanisms. "S" units perform a linear template match operation to build more complex features from simple ones, while "C" units perform a nonlinear MAX pooling operation over units tuned to the same feature but at different positions and scales to increase response invariance to translation and scaling while maintaining feature specificity [2]. Interestingly, the prediction that some neurons at different levels along the ventral stream perform a MAX operation has recently been supported at the level of complex cells in cat striate cortex (Lampl, I., Riesenhuber, M., Poggio, T., and Ferster, D., *Soc. Neurosci. Abs.*, 2001) and at the level of V4 neurons in the macaque [6].

Input patterns are first filtered through a continuous layer S1 of overlapping simple cell-like receptive fields (first derivative of gaussians) of different scales and orientations. Limited position and size invariance, for each orientation, is obtained in the subsequent C1 layer through a local non-linear MAX operation over neighboring (in both space and scale) S1 cells. Response of C1 cells to typical face and background stimuli are shown in Fig. 1. Features of intermediate complexity are obtained in the next level (S2) by combining the response of 2×2 arrangements of C1 cells (for all possible combinations, giving $4^4 = 256$ different features), followed by a MAX over the whole visual field in the next layer, C2, the final pooling layer in the standard version of HMAX [1]. An arbitrary object's shape is thus encoded by an activation pattern over the 256 C2 units.

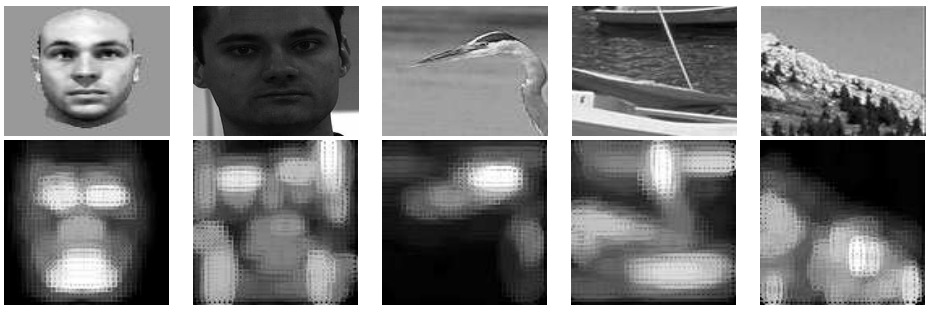

Fig. 1. Typical stimuli and associated responses of the C1 complex cells (4 orientations). The orientation of the ellipses matches the orientation of the cells and intensities encode response strength. For simplicity, only the response at one scale (std 2.75–3.75 pixels, 6×6 pooling range) is displayed. Note that an individual C1 cell is not particularly selective either to face or to non-face stimuli.

2.2 Classification Stage

We wish to compare the impact of two different representations on HMAX's performance on a benchmark face detection task: (i) the representation given by hardwired features from the standard HMAX, and (ii) the representation given by the new learned object class-specific features. A standard technique in machine vision to compare feature spaces is to train and test a given classifier on the data sets produced by projecting the data into the different representation spaces. It is still unclear how categorization tasks are learned in cortex [7] (but see the accompanying BMCV paper by Knoblich *et al.*). We here use a Support Vector Machine [8] (SVM) classifier, a learning technique that has been used successfully in recent machine vision systems [4, 3]. It is important to note that this classifier was not chosen for its biological plausibility, but rather as an established classification back-end that allows us to compare the quality of the different feature sets for the detection task independent of the classification technique.

2.3 Face Detection Task

Each system (*i. e.,* standard HMAX, HMAX with feature learning, and the "AI" system (see below)) was trained on a reduced data set similar to [3] consisting of 200 synthetic frontal face images generated from 3D head models [9] and 1,000 non-face image patterns randomly extracted from larger background images. After training, we tested each system on a test set of 1,300 face images (denoted "all faces" in the following) containing: (i) 900 "cluttered faces" and (ii) 400 "difficult faces". The "cluttered faces" were generated from 3D head models [9] that were different from training but were synthesized under similar illumination conditions. The "difficult faces" were real frontal faces presenting untrained extreme illumination conditions. The negative test set consisted of 1,845 difficult background images[1]. Examples for each set are given in Fig. 2.

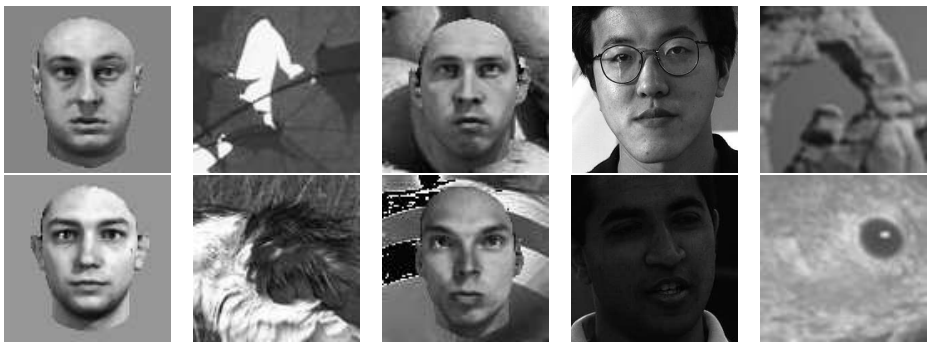

Fig. 2. Typical stimuli used in our experiments. From left to right: Training faces and non-faces, "cluttered (test) faces", "difficult (test) faces" and test non-faces.

[1] Both 400 difficult frontal faces and background images were extracted from the larger test set used in [3]. Background patterns were previously selected by a low-resolution classifier as most similar to faces.

2.4 Feature Learning

The goal of the feature learning algorithm was to obtain a set of object class-specific features. Fig. 3 shows how new S2 features are created from C1 inputs in the feature learning version of HMAX: Given a certain *patch size* p, a feature corresponds to a $p \times p \times 4$ pattern of C1 activation $\mathbf{w}$, where the last 4 comes from the four different preferred orientations of C1 units used in our simulations. The precise learned features or prototypes $\mathbf{u}$ (the number of which was another parameter, n) were obtained by performing vector quantization (VQ, using the k-means algorithm) over randomly chosen patches of size $p \times p \times 4$ of C1 activation obtained from extraction at random position over 200 face images (also used in training the classifier). Choosing m patches per face image therefore leaded to $M = 200 \times m$ total patches for training. In all simulations, p varied between 2 and 20, n varied between 4 and 3,000, m varied between 1 and 750 and M varied between 200 and 150,000. S2 units behave like gaussian RBF-units and compute a function of the squared distance between an input pattern and the stored prototype: $f(x) = \alpha \exp -\frac{||x-\mathbf{u}||^2}{2\sigma^2}$, with α chosen to normalize the value of all features over the training set between 0 and 1.

2.5 The "AI" (Machine Vision) System

As a benchmark we added performances of a classical machine vision face detection system similar to [4]. Detection of a face was performed by scanning input images at different scales by use of a search window. At each scale and for each position of the window, gray values were extracted and pre-processed as in [4] to feed a second-degree polynomial SVM.[2] All systems (*i. e.,* standard HMAX, HMAX with feature learning, and the "AI" system) were trained and tested on the same data sets (see section 2.3).

3 Results

3.1 Performance of Standard HMAX

As evident from Fig. 4, performance of the standard HMAX system on the face detection task is pretty much at chance: The system didn't generalize well to faces with similar illumination conditions but set into background ("cluttered faces") or to faces with less clutter (indoor scenes) and untrained illumination conditions ("difficult faces"). This indicates that the object class-unspecific dictionary of features in standard HMAX is insufficient to perform robust face detection. This is easily understood, as the 256 features cannot be expected to show any specificity for faces *vs.* background patterns. In particular, for a specific image containing a face on a background pattern, the activity of C2 model units (which pool over S2 units tuned to the same feature but having different receptive field locations) will for some C2 units be due to image patches belonging to the face. But, for other S2/C2 features, a part of the background might cause a stronger activation than any part of the face, thus interfering with the response that would have been caused by the face alone. This interference leads to poor generalization performances, as borne out in Fig. 4.

[2] Using a linear SVM yielded comparable detection performance.

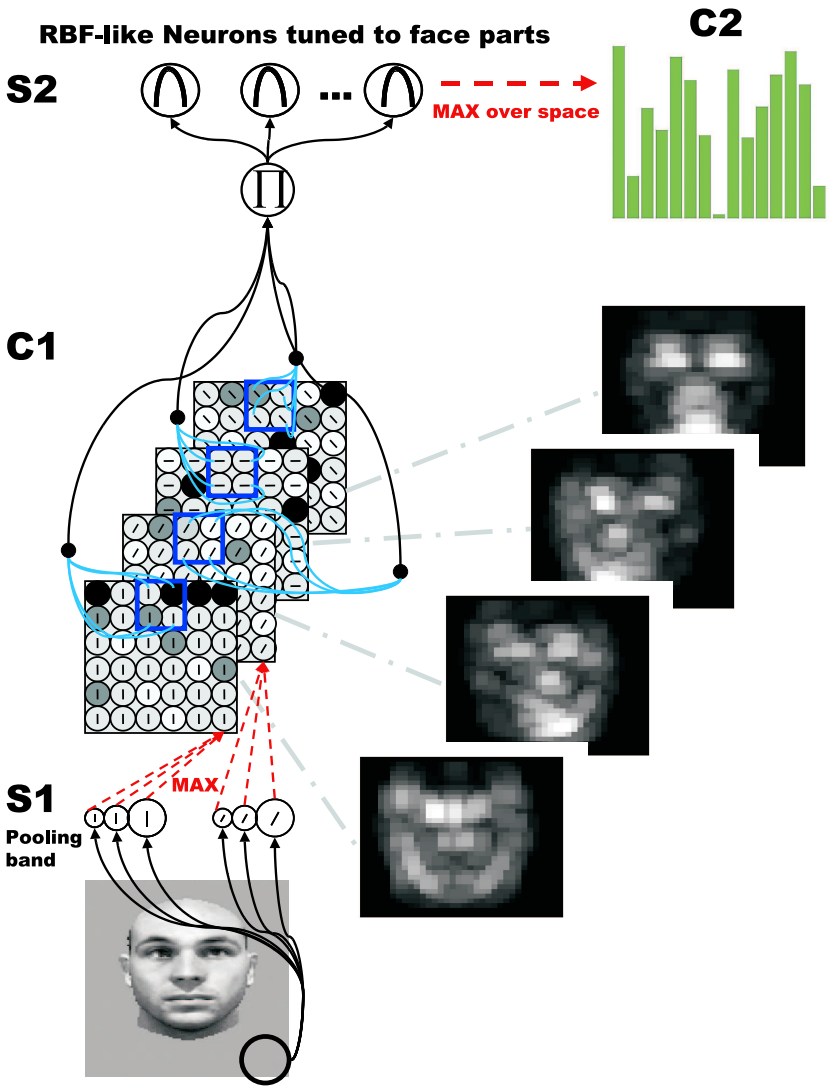

Fig. 3. Sketch of the HMAX model with feature learning: Patterns on the model "retina" are first filtered through a continuous layer S1 (simplified on the sketch) of overlapping simple cell-like receptive fields (first derivative of gaussians) at different scales and orientations. Neighboring S1 cells in turn are pooled by C1 cells through a MAX operation. The next S2 layer contains the RBF-like units that are tuned to object-parts and compute a function of the distance between the input units and the stored prototypes ($p = 4$ in the example). On top of the system, C2 cells perform a MAX operation over the whole visual field and provide the final encoding of the stimulus, constituting the input to the SVM classifier. The difference to standard HMAX lies in the connectivity from C1→S2 layer: While in standard HMAX, these connections are hardwired to produce 256 2×2 combinations of C1 inputs, they are now learned from the data.

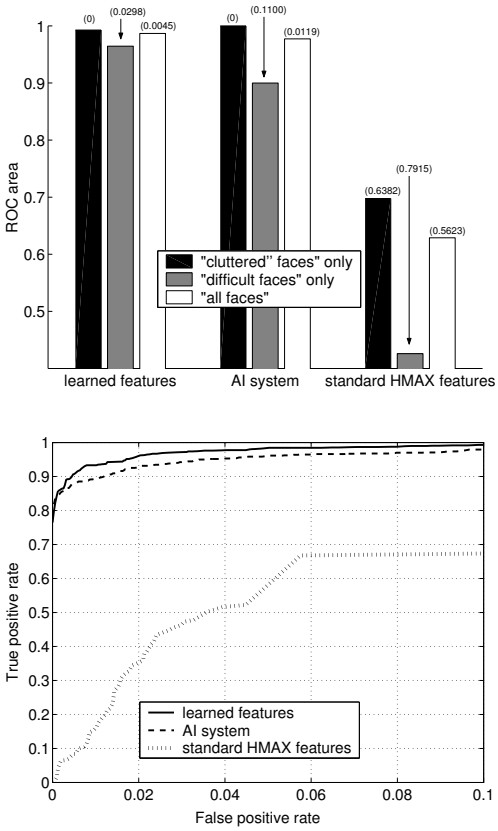

Fig. 4. Comparison between the new extended model using object-specific learned features ($p = 5$, $n = 480$, $m = 120$, corresponding to our best set of features), the "AI" face detection system and the standard HMAX. Top: Detailed performances (ROC area) on (i) "all faces", (ii) "cluttered faces" only and (iii) "difficult faces" only (background images remain unchanged in the ROC calculation). For information, the false positive rate at 90% true positive is given in parenthesis. The new model generalizes well on all sets and overall outperforms the "AI" system (especially on the "cluttered" set) as well as standard HMAX. Bottom: ROC curves for each system on the test set including "all faces".

3.2 Feature Learning

As Fig. 5 makes clear, the challenge is to learn a set of features in the S2 layer that reliably permits the system to detect image patches belonging to a face and not be confused by non-face patterns, even though objects from the two classes can cause very similar activations on the C1 level (Fig. 1). In general, the learned features already show a high degree of specificity for faces and are not confused by simultaneously appearing backgrounds. They thus appear to offer a much more robust representation than the features in standard HMAX.

Using the learned face-specific features leads to a tremendously improved performance (Fig. 4), even outperforming the "AI" system. This demonstrates that the new features reliably respond to face components with high accuracy without being confused by non-face backgrounds.

3.3 Parameter Dependence

The results in Fig. 4 were obtained with a dictionary of $n = 480$, $m = 120$ and $p = 5$ features. This choice of parameters provided the best results. Fig. 6 (bot-

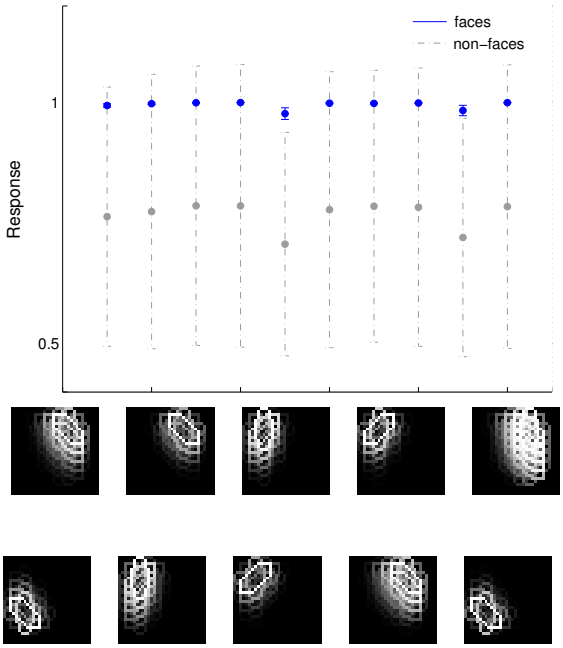

Fig. 5. Mean response over the training face and non-face stimuli, resp., for the top 10 learned features. The top ten features were obtained by ranking the learned features obtained with $p = 5$, $n = 480$, $m = 120$ according to their individual ROC value and by selecting the ten best. Individual features are already good linear separators between faces and non-faces. Bottom: Corresponding features. The orientation of the ellipses matches the orientation of the cells and intensities encode response strength.

tom) shows the dependence of the model's performance on patch size p and the percentage of face area covered by the features (the area taken up by one feature p^2 times the number of patches extracted per faces m divided by the area covered by one face). As the percentage of the face area covered by the features increases, the overlap between features should in principle increases. Features of intermediate sizes work best[3]: First, compared with large features, they probably have more flexibility in matching a greater number of faces. Second, compared to smaller features they are probably more selective to faces. Those results are in good agreement with [10] where gray-value features of intermediate sizes where shown to have higher mutual information. Similarly, performance as a function of the number of features n show first a rise with increasing numbers of features due to the increased discriminatory power of the feature dictionary. However, with large features, overfitting may occur. Fig. 6 (top) shows performances for $p = 2, 5, 7, 10, 15, 20$ and $n = 100$.

4 Discussion

In this paper, we have applied a model of object recognition in cortex to a real-world object recognition task, the detection of faces in natural images. While HMAX has been shown to capture the shape tuning and invariance properties from physiological experiments, we found that it performed very poorly on the face

[3] 5×5 and 7×7 features for which performances are best correspond to cells' receptive field of about a third of a face.

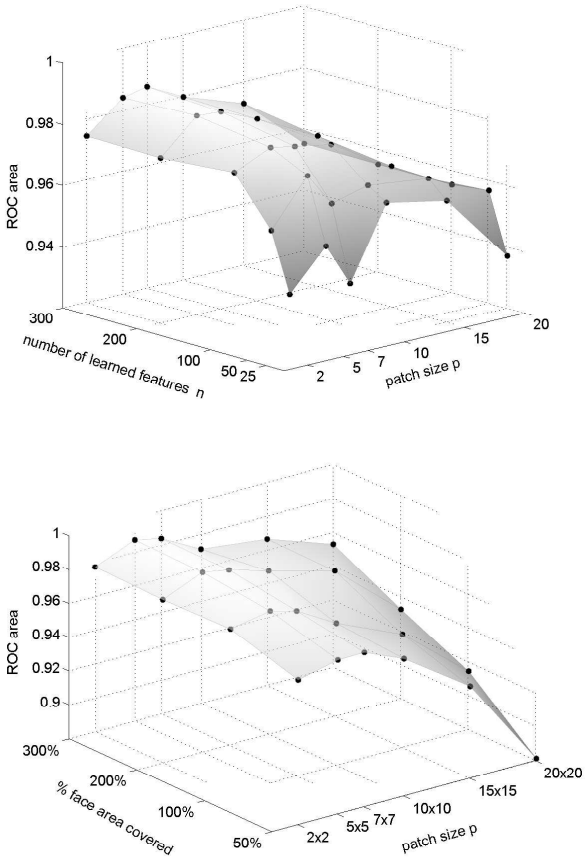

Fig. 6. Investigating prototypes tuning properties. Top: Performances (ROC area) with respect to the number of learned features n (fixed $p = 5$ and $m = 100$). Performances increase with the number of learned features to a certain level and larger patches start overfit. Bottom: features overlap (or equivalently % of the face area covered): overlapping intermediate features perform best. Best performances were obtained with $p = 5$, $n = 480$, $m = 120$.

detection task. Because visual features of intermediate complexity in HMAX were initially hardwired, they failed to show any specificity for faces *vs.* background patterns. In particular, for an image containing a face on a background pattern, the activity of some (C2) top units could be due to image parts belonging to the face. For others, a part of the background could elicit a stronger activation than any part of the face thus interfering with the response that would have been caused by the face alone. This led to poor generalization performance.

Extending the original model, we proposed a biologically plausible feature learning algorithm and we showed that the new model was able to outperform standard HMAX as well as a benchmark classical face detection system similar to [4]. Learned features therefore appear to offer a much more robust representation than the non-specific features in standard HMAX and could thus play a crucial role in the representation of objects in cortex.

Interestingly, we showed that features that were chosen independently of any task (*i. e.,* independently of their ability to discriminate between face and non-face stimuli or between-class discrimination) produced a powerful encoding scheme. This is compatible with our recent theory of object recognition in cortex [2] in which a general, *i.e.,* task-independent object representation provides input

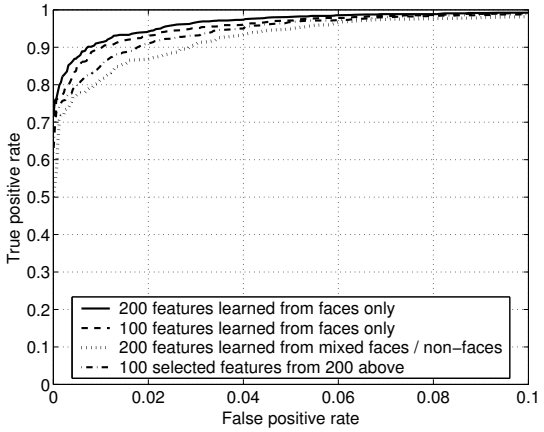

Fig. 7. Preliminary results on unsupervised feature learning: Comparison (ROC area) between features ($p = 5$) learned from face parts only ($n = 200$ and $n = 100$ features) and features learned from both face and non-face parts ($n = 200$, from an equal number of positive and negative examples). Performances on the "unsupervised" representation can be improved by selecting face-specific features for the detection task.

to task-specific circuits. We would expect the same features to be useful for recognition tasks at different levels (*i.e.,* identification), possibly with different weights, and we intend to explore these questions further.

Our proposed mechanisms for learning object-specific features is partially supervised since features are only extracted from the target object class. However, preliminary results using unsupervised learning ($n = 200$ features, $p = 5$, learned from 10,000 face parts and 10,000 non-face parts) have produced encouraging results. As Fig. 7 makes clear, a system using the features learned with k-means over face and non-face stimuli performs slightly worse than the systems using features extracted from face parts only. However, weeding out non-selective features by keeping only the 100 most discriminant features (as given by their ROC value) is enough to bring the system at a higher level. It is worth emphasizing that selecting features based on their mutual information produced similar results. We are currently exploring how this feature selection can be done in a biologically plausible way.

Modelling the biological mechanisms by which neurons acquire tuning properties in cortical areas was not the scope of the present paper. Rather, we focused on the type of computation performed by cortical neurons. We proposed a 2-step learning stage where an object representation is first learned and then a strategy is selected. For simplicity, we chose the (non-biological) k-means algorithm to learn features that provide a suitable representation independently of any task. While it is unlikely that the cortex performs k-means clustering, there are more plausible models of cortical self-organization that perform very similar operations in a biologically more plausible architecture. It should be easy to replace with a more biologically plausible linear classifier the SVM classifier we have used here, while accounting well for the sharp class boundary exhibited by some category-specific neurons in prefrontal cortex [7].

Acknowledgments

We thank T. Vetter for providing us with the 3D head models used to generate the synthetic faces. This report describes research done at the Center for Biological & Computational Learning, which is affiliated with the Mc Govern Institute of Brain Research and with the Artificial Intelligence Laboratory, and which is in the Department of Brain & Cognitive Sciences at MIT.

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. N00014-00-1-0907, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR) Contract No. IIS-0112991, National Science Foundation (KDI) Contract No. DMS-9872936, and National Science Foundation Contract No. IIS-9800032.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Daimler Chrysler AG, Compaq/Digital Equipment Corporation, Eastman Kodak Company, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd., and The Whitaker Foundation.

References

1. M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–25, 1999.
2. M. Riesenhuber and T. Poggio. Models of object recognition. *Nature Neuroscience*, 3 supp.:1199–1204, 2000.
3. B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 657–62, Hawaii, 2001.
4. K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition*. PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.
5. D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phys.*, 160:106–54, 1962.
6. T.J. Gawne and J.M. Martin. Response of primate visual cortical V4 neurons to simultaneously presented stimuli. *To appear in J. Neurophysiol.*, 2002.
7. D.J. Freedman, M. Riesenhuber, T. Poggio, and E.K. Miller. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, 291:312–16, 2001.
8. V. Vapnik. *The nature of statistical learning*. Springer Verlag, 1995.
9. T. Vetter. Synthesis of novel views from a single face. *International Journal of Computer Vision*, 28(2):103–116, 1998.
10. S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermediate complexity and their use in classification. *Nat. Neurosci.*, 5(7):682–87, 2002.