Object Recognition with Features Inspired by Visual Cortex

Thomas Serre Lior Wolf Tomaso Poggio

Center for Biological and Computational Learning
McGovern Institute
Brain and Cognitive Sciences Department
Massachusetts Institute of Technology
Cambridge, MA 02142
{serre,liorwolf}@mit.edu, tp@ai.mit.edu

Abstract

We introduce a novel set of features for robust object recognition. Each element of this set is a complex feature obtained by combining position- and scale-tolerant edge-detectors over neighboring positions and multiple orientations. Our system's architecture is motivated by a quantitative model of visual cortex.

We show that our approach exhibits excellent recognition performance and outperforms several state-of-the-art systems on a variety of image datasets including many different object categories. We also demonstrate that our system is able to learn from very few examples. The performance of the approach constitutes a suggestive plausibility proof for a class of feedforward models of object recognition in cortex.

1 Introduction

Hierarchical approaches to generic object recognition have become increasingly popular over the years. These are in some cases inspired by the hierarchical nature of primate visual cortex [10, 25], but, most importantly, hierarchical approaches have been shown to consistently outperform flat single-template (holistic) object recognition systems on a variety of object recognition tasks [7, 10]. Recognition typically involves the computation of a set of target features (also called components [7], parts [24] or fragments [22]) at one step and their combination in the next step. Features usually fall in one of two categories: *template-based* or *histogram-based*. Several template-based methods exhibit excellent performance in the detection of a single object category, *e.g.,* faces [17, 23], cars [17] or pedestrians [14]. Constellation models based on generative methods perform well in the recognition of several object categories [24, 4], particularly when trained with very few training examples [3]. One limitation of these rigid template-based features is that they might not adequately capture variations in object appearance: they are very selective for a target shape but lack invariance with respect to object transformations. At the other extreme, histogram-based descriptors [12, 2] are very robust with respect to object transformations. The SIFT-based features [12], for instance, have been shown to excel in the re-detection of a previously seen object under new image transformations. However, as we confirm experimentally (see section 4), with such degree of invariance, it is unlikely that the SIFT-based features could perform well on a generic object recognition task.

In this paper, we introduce a new set of biologically-inspired features that exhibit a better trade-off between invariance and selectivity than template-based or histogram-based approaches. Each element of this set is a feature obtained by combining the response of local edge-detectors that are slightly position- and scale-tolerant over neighboring positions and multiple orientations (like complex cells in primary visual cortex). Our features are more flexible than template-based approaches [7, 22] because they allow for small distortions of the input; they are more selective than histogram-based descriptors as they preserve local feature geometry. Our approach is as follows: for an input image, we first compute a set of features learned from the positive training set (see section 2). We then run a standard classifier on the vector of features obtained from the input image. The resulting approach is simpler than the aforementioned hierarchical approaches: it does not involve scanning over all positions and scales, it uses discriminative methods and it does not explicitly model object geometry. Yet it is able to learn from very few examples and it performs significantly better than all the systems we have compared it with thus far.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Object Recognition with Features Inspired by Visual Cortex	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Massachusetts Institute of Technology,Center for Biological and Computational Learning,77 Massachusetts Avenue,Cambridge,MA,02139**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
The original document contains color images.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**		**7**	

Band Σ	1	2	3	4	5	6	7	8
filt. sizes s	7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37
σ	2.8 & 3.6	4.5 & 5.4	6.3 & 7.3	8.2 & 9.2	10.2 & 11.3	12.3 & 13.4	14.6 & 15.8	17.0 & 18.2
λ	3.5 & 4.6	5.6 & 6.8	7.9 & 9.1	10.3 & 11.5	12.7 & 14.1	15.4 & 16.8	18.2 & 19.7	21.2 & 22.8
grid size N^{Σ}	8	10	12	14	16	18	20	22
orient. θ	$0; \frac{\pi}{4}; \frac{\pi}{2}; \frac{3\pi}{4}$							
patch sizes n_i	$4 \times 4; 8 \times 8; 12 \times 12; 16 \times 16$ ($\times 4$ orientations)							

Table 1. Summary of parameters used in our implementation (see Fig. 1 and accompanying text).

Biological visual systems as guides. Because humans and primates outperform the best machine vision systems by almost any measure, building a system that emulates object recognition in cortex has always been an attractive idea. However, for the most part, the use of visual neuroscience in computer vision has been limited to a justification of Gabor filters. No real attention has been given to biologically plausible features of higher complexity. While mainstream computer vision has always been inspired and challenged by human vision, it seems to never have advanced past the first stage of processing in the simple cells of primary visual cortex V1. Models of biological vision [5, 13, 16, 1] have not been extended to deal with real-world object recognition tasks (*e.g.,* large scale natural image databases) while computer vision systems that are closer to biology like LeNet [10] are still lacking agreement with physiology (*e.g.,* mapping from network layers to cortical visual areas). This work is an attempt to bridge the gap between computer vision and neuroscience.

Our system follows the standard model of object recognition in primate cortex [16], which summarizes in a quantitative way what most visual neuroscientists agree on: the first few hundreds milliseconds of visual processing in primate cortex follows a mostly feedforward hierarchy. At each stage, the receptive fields of neurons (*i.e.,* the part of the visual field that could potentially elicit a neuron's response) tend to get larger along with the complexity of their optimal stimuli (*i.e.,* the set of stimuli that elicit a neuron's response). In its simplest version, the standard model consists of four layers of computational units where *simple* S units, which combine their inputs with Gaussian-like tuning to increase object selectivity, alternate with *complex* C units, which pool their inputs through a maximum operation, thereby introducing gradual invariance to scale and translation. The model has been able to quantitatively duplicate the generalization properties exhibited by neurons in inferotemporal monkey cortex (the so-called view-tuned units) that remain highly selective for particular objects (a face, a hand, a toilet brush) while being invariant to ranges of scales and positions. The model originally used a very simple static dictionary of features (for the recognition of segmented objects) although it was suggested in [16] that features in intermediate layers should instead be learned from visual experience.

We extend the standard model and show how it can learn a vocabulary of visual features from natural images. We prove that the extended model can robustly handle the recognition of many object categories and compete with state-of-the-art object recognition systems. This work appeared in a very preliminary form in [18]. Our source code as well as an extended version of this paper [20] can be found at `http://cbcl.mit.edu/software-datasets`.

2 The C2 features

Our approach is summarized in Fig. 1: the first two layers correspond to primate primary visual cortex, V1, *i.e.,* the first visual cortical stage, which contains simple (S1) and complex (C1) cells [8]. The S1 responses are obtained by applying to the input image a battery of Gabor filters, which can be described by the following equation:

$$G(x, y) = \exp\left(-\frac{(X^2 + \gamma^2 Y^2)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} X\right),$$

where $X = x\cos\theta + y\sin\theta$ and $Y = -x\sin\theta + y\cos\theta$.

We adjusted the filter parameters, *i.e.,* orientation θ, effective width σ, and wavelength λ, so that the tuning profiles of S1 units match those of V1 parafoveal simple cells. This was done by first sampling the space of parameters and then generating a large number of filters. We applied those filters to stimuli commonly used to probe V1 neurons [8] (*i.e.,* gratings, bars and edges). After removing filters that were incompatible with biological cells [8], we were left with a final set of 16 filters at 4 orientations (see Table 1 and [19] for a full description of how those filters were obtained).

The next stage – C1 – corresponds to complex cells which show some tolerance to shift and size: complex cells tend to have larger receptive fields (twice as large as simple cells), respond to oriented bars or edges anywhere within their receptive field [8] (shift invariance) and are in general more broadly tuned to spatial frequency than simple cells [8] (scale invariance). Modifying the original Hubel & Wiesel proposal for building complex cells from simple cells through pooling [8], Riesenhuber & Poggio proposed a max-like pooling operation for building position- and scale-tolerant C1 units. In the meantime, experimental evidence

Given an input image I, perform the following steps:.

S1: Apply a battery of Gabor filters to the input image. The filters come in 4 orientations θ and 16 scales s (see Table 1). Obtain $16 \times 4 = 64$ maps $(S1)_\theta^s$ that are arranged in 8 *bands* (*e.g.,* band 1 contains filter outputs of size 7 and 9, in all four orientations, band 2 contains filter outputs of size 11 and 13, *etc*).

C1: For each *band*, take the max over scales and positions: each band member is sub-sampled by taking the max over a grid with cells of size N^Σ first and the max between the two scale members second, *e.g.,* for band 1, a spatial max is taken over an 8×8 grid first and then across the two scales (size 7 and 9). Note that we do not take a max over different orientations, hence, each band $(C1)^\Sigma$ contains 4 maps.

During training only: Extract K patches $P_{i=1,\ldots K}$ of various sizes $n_i \times n_i$ and all four orientations (thus containing $n_i \times n_i \times 4$ elements) at random from the $(C1)^\Sigma$ maps from all training images.

S2: For each C1 image $(C1)^\Sigma$, compute: $Y = \exp(-\gamma||X - P_i||^2)$ for all image patches X (at all positions) and each patch P learned during training for each band independently. Obtain S2 maps $(S2)_i^\Sigma$.

C2: Compute the max over all positions and scales for each S2 map type $(S2)_i$ (*i.e.,* corresponding to a particular patch P_i) and obtain shift- and scale-invariant C2 features $(C2)_i$, for $i = 1 \ldots K$.

Figure 1. Computation of C2 features.

in favor of the max operation has appeared [6, 9]. Again pooling parameters were set so that C1 units match the tuning properties of complex cells as measured experimentally (see Table 1 and [19] for a full description of how those filters were obtained).

Fig. 2 illustrates how pooling from S1 to C1 is done. S1 units come in 16 scales s arranged in 8 bands Σ. For instance, consider the first band $\Sigma = 1$. For each orientation, it contains two S1 maps: one obtained using a filter of size 7, and one obtained using a filter of size 9. Note that both of these S1 maps have the same dimensions. In order to obtain the C1 responses, these maps are sub-sampled using a grid cell of size $N^\Sigma \times N^\Sigma = 8 \times 8$. From each grid cell we obtain one measurement by taking the maximum of all 64 elements. As a last stage we take a max over the two scales, by considering for each cell the maximum value from the two maps. This process is repeated independently for each of the four orientations and each scale band.

In our new version of the standard model the subsequent S2 stage is where learning occurs. A large pool of K

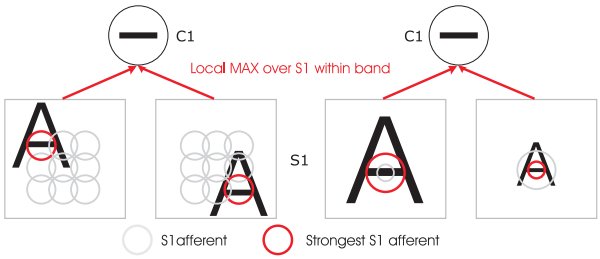


Figure 2. Scale- and position-tolerance at the complex cells (C1) level: Each C1 unit receives inputs from S1 units at the same preferred orientation arranged in bands Σ, i.e., S1 units in two different sizes and neighboring positions (grid cell of size $N^\Sigma \times N^\Sigma$). From each grid cell (left) we obtain one measurement by taking the max over all positions allowing the C1 unit to respond to an horizontal edge anywhere within the grid (tolerance to shift). Similarly, by taking a max over the two sizes (right) the C1 unit becomes tolerant to slight changes in scale.

patches of various sizes at random positions are extracted from a target set of images at the C1 level for all orientations, *i.e.,* a patch P_i of size $n_i \times n_i$ contains $n_i \times n_i \times 4$ elements, where the 4 factor corresponds to the four possible S1 and C1 orientations. In our simulations we used patches of size $n_i = 4, 8, 12$ and 16 but in practice any size can be considered. The training process ends by setting each of those patches as *prototypes* or *centers* of the S2 units which behave as radial basis function (RBF) units during recognition, *i.e.,* each S2 unit response depends in a Gaussian-like way on the Euclidean distance between a new input patch (at a particular location and scale) and the stored prototype. This is consistent with well-known neuron response properties in primate inferotemporal cortex and seems to be the key property for learning to generalize in the visual and motor systems [15]. When a new input is presented, each stored S2 unit is convolved with the new $(C1)^\Sigma$ input image at all scales (this leads to $K \times 8$ $(S2)_i^\Sigma$ images, where the K factor corresponds to the K patches extracted during learning and the 8 factor, to the 8 scale bands). After taking a final max for each $(S2)_i$ map across all scales and positions, we get the final set of K shift- and scale-invariant C2 units. The size of our final C2 feature vector thus depends only on the number of patches extracted during learning and not on the input image size. This C2 feature vector is passed to a classifier for final analysis.[1]

An important question for both neuroscience and computer vision regards the choice of the unlabeled target set from which to learn – in an unsupervised way – this vocabulary of visual features. In this paper, features are learned from the positive training set for each object category (but see [20] for a discussion on how features could be learned from random natural images).

[1] It is likely that our (non-biological) final classifier could correspond to the task-specific circuits found in prefrontal cortex (PFC) and C2 units with neurons in inferotemporal (IT) cortex [16]. The S2 units could be located in V4 and/or in posterior inferotemporal (PIT) cortex.

Figure 3. Examples from the MIT face and car datasets.

Datasets		Bench.	C2 features	
			boost	SVM
Leaves (Calt.)	[24]	84.0	**97.0**	95.9
Cars (Calt.)	[4]	84.8	99.7	**99.8**
Faces (Calt.)	[4]	96.4	**98.2**	98.1
Airplanes (Calt.)	[4]	94.0	**96.7**	94.9
Moto. (Calt.)	[4]	95.0	**98.0**	97.4
Faces (MIT)	[7]	90.4	**95.9**	95.3
Cars (MIT)	[11]	75.4	**95.1**	93.3

Table 2. C2 features vs. other recognition systems (Bench.).

3. Experimental Setup

We tested our system on various object categorization tasks for comparison with benchmark computer vision systems. All datasets we used are made up of images that either contain or do not contain a single instance of the target object; The system has to decide whether the target object is present or absent.

MIT-CBCL datasets: These include a near-frontal ($\pm 30^\circ$) face dataset for comparison with the component-based system of Heisele *et al.* [7] and a multi-view car dataset for comparison with [11]. These two datasets are very challenging (see typical examples in Fig. 3). The face patterns used for testing constitute a subset of the CMU PIE database which contains a large variety of faces under extreme illumination conditions (see [7]). The test non-face patterns were selected by a low-resolution LDA classifier as the most similar to faces (the LDA classifier was trained on an independent 19×19 low-resolution training set). The full set used in [7] contains 6,900 positive and 13,700 negative 70×70 images for training and 427 positive and 5,000 negative images for testing. The car database on the other hand was created by taking street scene pictures in the Boston city area. Numerous vehicles (including SUVs, trucks, buses, *etc*) photographed from different view-points were manually labeled from those images to form a positive set. Random image patterns at various scales that were not labeled as vehicles were extracted and used as the negative set. The car dataset used in [11] contains 4,000 positive and 1,600 negative 120×120 training examples and 3,400 test examples (half positive, half negative). While we tested our system on the full test sets, we considered a random subset of the positive and negative training sets containing only 500 images each for both the face and the car database.

The Caltech datasets: The Caltech datasets contain 101 objects plus a *background* category (used as the negative set) and are available at `http://www.vision.caltech.edu`. For each object category, the system was trained with $n = 1, 3, 6, 15, 30$ or 40 positive examples from the target object class (as in [3]) and 50 negative examples from the background class. From the remaining images, we extracted 50 images from the positive and 50 images from the negative set to test the system's performance. As in [3], the system's performance was averaged over 10 random splits for each object category. All images were normalized to 140 pixels in height (width was rescaled accordingly so that the image aspect ratio was preserved) and converted to gray values before processing. These datasets contain the target object embedded in a large amount of clutter and the challenge is to learn from unsegmented images and discover the target object class automatically. For a close comparison with the system by Fergus *et al.* we also tested our approach on a subset of the 101-object dataset using the exact same split as in [4] (the results are reported in Table 2) and an additional leaf database as in [24] for a total of five datasets that we refer to as the *Caltech* datasets in the following.

4 Results

Table 2 contains a summary of the performnace of the C2 features when used as input to a linear SVM and to gentle Ada Boost (denoted *boost*) on various datasets. For both our system and the benchmarks, we report the error rate at the equilibrium point, *i.e.,* the error rate at which the false positive rate equals the miss rate. Results obtained with the C2 features are consistently higher than those previously reported on the Caltech datasets. Our system seems to outperform the component-based system presented in [7] (also using SVM) on the MIT-CBCL face database as well as a fragment-based system implemented by [11] that uses template-based features with gentle Ada Boost (similar to [21]).

Fig. 4 summarizes the system performance on the 101-object database. On the left we show the results obtained using our system with gentle Ada Boost (we found qualitatively similar results with a linear SVM) over all 101 categories for 1, 3, 6, 15, 30 and 40 positive training examples (each result is an average of 10 different random splits). Each plot is a single histogram of all 101 scores, obtained using a fixed number of training examples (*e.g.,* with 40 examples the system gets 95% correct for 42% of the object categories). On the right we focus on some of the same object categories as the ones used by Fei-Fei *et al.* for illustration in [3]: the C2 features achieve error rates very

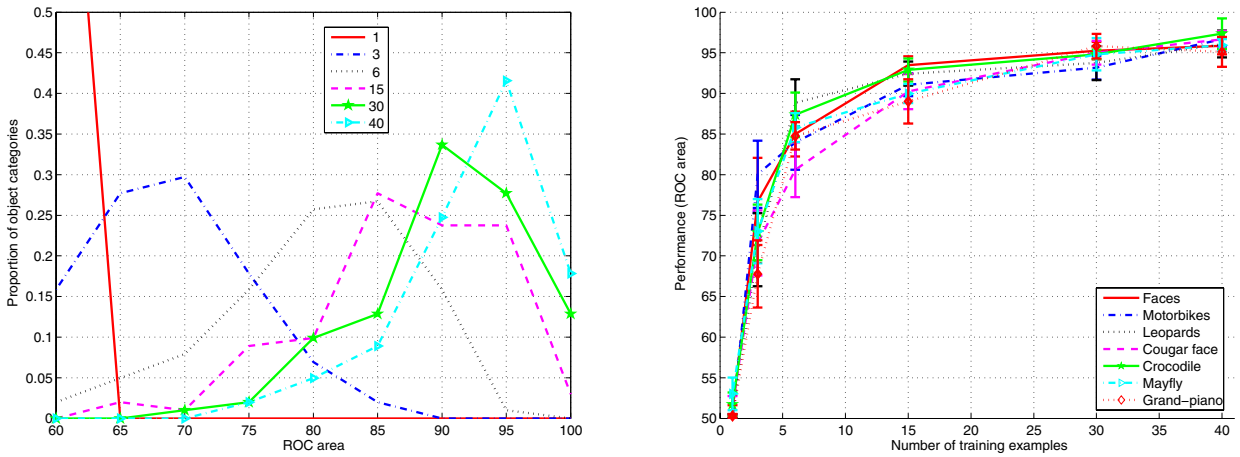


Figure 4. C2 features performance on the 101-object database for different numbers of positive training examples: (left) histogram across the 101 categories and (right) performance on sample categories, see accompanying text.

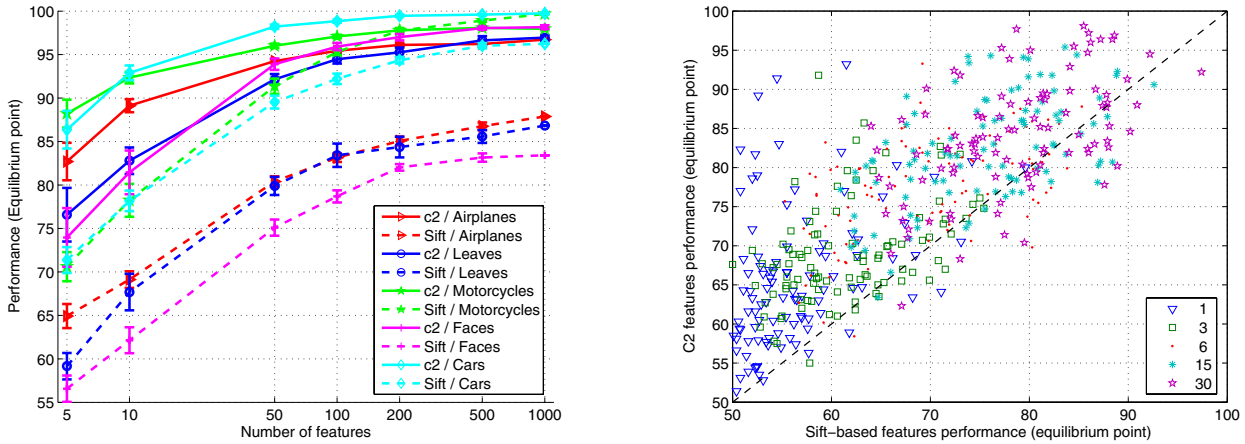
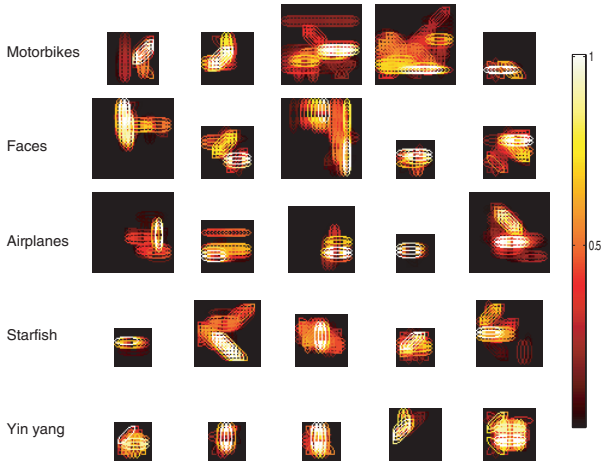


Figure 5. Superiority of the C2 vs. SIFT-based features on the Caltech datasets for different number of features (left) and on the 101-object database for different number of training examples(right).

similar to the ones reported in [3] with very few training examples.

We also compared our C2 features to SIFT-based features [12]. We selected 1000 random reference key-points from the training set. Given a new image, we measured the minimum distance between all its key-points and the 1000 reference key-points, thus obtaining a feature vector of size 1000 (for this comparison we did not use the position information recovered by the algorithm). While Lowe recommends using the ratio of the distances between the nearest and the second closest key-point as a similarity measure, we found that the minimum distance leads to better performance than the ratio on these datasets. A comparison between the C2 features and the SIFT-based features (both passed to a Gentle Ada boost classifier) is shown in Fig. 5 (left) for the Caltech datasets. The gain in performance obtained by using the C2 features relative to the SIFT-based features is obvious. This is true with gentle Ada Boost – used for classification on Fig. 5 (left) – but we also found very similar results with SVM. Also, as one can see in Fig. 5 (right), the performance of the C2 features (error at equilibrium point) for each category from the 101-object database is well above that of the SIFT-based features for any number of training examples.

Finally, we conducted initial experiments on the multiple classes case. For this task we used the 101-object dataset. We split each category into a training set of size 15 or 30 and a test set containing the rest of the images. We used a simple multiple-class linear SVM as classifier. The SVM applied the all-pairs method for multiple label classification, and was trained on 102 labels (101 categories plus the background category, *i.e.,* 102 AFC). The number of $C2$ features used in these experiments was 4075. We obtained above 35% correct classification rate when using 15 training examples per class averaged over 10 repetitions, and 42% correct classification rate when using 30 training examples (chance below 1%).



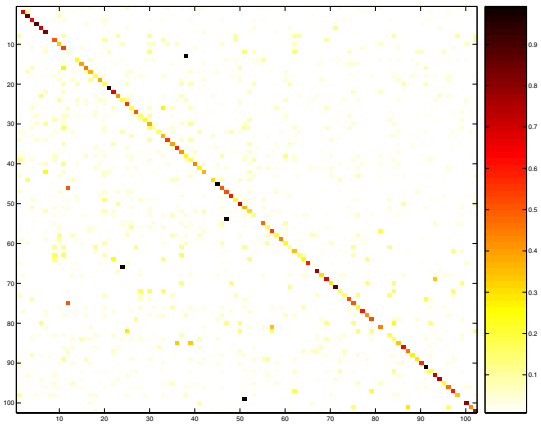


Figure 6. (left) Sample features learned from different object categories (i.e., first 5 features returned by gentle Ada Boost for each category). Shown are S2 features (centers of RBF units): each oriented ellipse characterizes a C1 (afferent) subunit at matching orientation, while color encodes for response strength. (right) Multiclass classification on 101 object database with a linear SVM.

5 Discussion

This paper describes a new biologically-motivated framework for robust object recognition: Our system first computes a set of scale- and translation-invariant C2 features from a training set of images and then runs a standard discriminative classifier on the vector of features obtained from the input image. Our approach exhibits excellent performance on a variety of image datasets and compete with some of the best existing systems.

This system belongs to a family of feedforward models of object recognition in cortex that have been shown to be able to duplicate the tuning properties of neurons in several visual cortical areas. In particular, Riesenhuber & Poggio showed that such a class of models accounts quantitatively for the tuning properties of view-tuned units in inferotemporal cortex (tested with idealized object stimuli on uniform backgrounds), which respond to images of the learned object more strongly than to distractor objects, despite significant changes in position and size [16]. The performance of this architecture on a variety of real-world object recognition tasks (presence of clutter and changes in appearance, illumination, *etc*) provides another compelling plausibility proof for this class of models.

While a long-time goal for computer vision has been to build a system that achieves human-level recognition performance, state-of-the-art algorithms have been diverging from biology: for instance, some of the best existing systems use geometrical information about the constitutive parts of objects (constellation approaches rely on both appearance-based and shape-based models and component-based system use the relative position of the detected components along with their associated detection values). Biology is however unlikely to be able to use geometrical information – at least in the cortical stream dedicated to shape processing and object recognition. The system described in this paper is respects the properties of cortical processing (including the absence of geometrical information) while showing performance at least comparable to the best computer vision systems.

The fact that this biologically-motivated model outperforms more complex computer vision systems might at first appear puzzling. The architecture performs only two major kinds of computations (template matching and max pooling) while some of the other systems we have discussed involve complex computations like the estimation of probability distributions [24, 4, 3] or the selection of facial-components for use by an SVM [7]. Perhaps part of the model's strength comes from its built-in gradual shift- and scale-tolerance that closely mimics visual cortical processing, which has been finely tuned by evolution over thousands of years. It is also very likely that such hierarchical architectures ease the recognition problem by decomposing the task into several simpler ones at each layer. Finally it is worth pointing out that the set of C2 features that is passed to the final classifier is very redundant, probably more redundant than for other approaches. While we showed that a relatively small number of features (about 50) is sufficient to achieve good error rates, performance can be increased significantly by adding many more features. Interestingly, the number of features needed to reach the ceiling (about 5,000 features) is much larger than the number used by current systems (on the order of 10-100 for [22, 7, 21] and 4-8 for constellation approaches [24, 4, 3]).

Acknowledgments

We would like to thank the anonymous reviewers as well as Antonio Torralba and Yuri Ivanov for useful comments on this manuscript.

This report describes research done at the Center for Biological & Computational Learning, which is in the McGovern Institute for Brain Research at MIT, as well as in the Dept. of Brain & Cognitive Sciences, and which is affiliated with the Computer Sciences & Artificial Intelligence Laboratory (CSAIL).

This research was sponsored by grants from: Office of Naval Research (DARPA) Contract No. MDA972-04-1-0037, Office of Naval Research (DARPA) Contract No. N00014-02-1-0915, National Science Foundation (ITR/IM) Contract No. IIS-0085836, National Science Foundation (ITR/SYS) Contract No. IIS-0112991, National Science Foundation (ITR) Contract No. IIS-0209289, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218693, National Science Foundation-NIH (CRCNS) Contract No. EIA-0218506, and National Institutes of Health (Conte) Contract No. 1 P20 MH66239-01A1. Additional support was provided by: 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., Eugene McDermott Foundation, Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sony MOU, Sumitomo Metal Industries, Toyota Motor Corporation, and WatchVision Co., Ltd.

References

[1] Y. Amit and M. Mascaro. An integrated network for invariant visual detection and recognition. *Vision Research*, 43(19):2073–2088, 2003.

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 2002.

[3] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR, Workshop on Generative-Model Based Vision*, 2004.

[4] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, volume 2, pages 264–271, 2003.

[5] K. Fukushima. Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.*, 36:193–201, 1980.

[6] T.J. Gawne and J.M. Martin. Response of primate visual cortical V4 neurons to simultaneously presented stimuli. *J. Neurophysiol.*, 88:1128–1135, 2002.

[7] B. Heisele, T. Serre, M. Pontil, T. Vetter, and T. Poggio. Categorization by learning and combining object parts. In *NIPS*, Vancouver, 2001.

[8] D. Hubel and T. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.*, 28:229–89, 1965.

[9] I. Lampl, D. Ferster, T. Poggio, and M. Riesenhuber. Intracellular measurements of spatial integration and the max operation in complex cells of the cat primary visual cortex. *J. Neurophysiol.*, 92:2704–2713, 2004.

[10] Yann LeCun, Fu-Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of CVPR'04*. IEEE Press, 2004.

[11] B. Leung. Component-based car detection in street scene images. Master's thesis, EECS, MIT, 2004.

[12] D.G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, pages 1150–1157, 1999.

[13] B.W. Mel. SEEMORE: Combining color, shape and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9(4):777–804, 1997.

[14] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in images by components. In *PAMI*, volume 23, pages 349–361, 2001.

[15] T. Poggio and E. Bizzi. Generalization in vision and motor control. *Nature*, 431:768–774, 2004.

[16] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neurosci.*, 2(11):1019–25, 1999.

[17] H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *CVPR*, pages 746–751, 2000.

[18] T. Serre, J. Louie, M. Riesenhuber, and T. Poggio. On the role of object-specific features for real world recognition in biological vision. In *Biologically Motivated Computer Vision, Second International Workshop (BMCV 2002)*, pages 387–97, Tuebingen, Germany., 2002.

[19] T. Serre and M. Riesenhuber. Realistic modeling of simple and complex cell tuning in the hmax model, and implications for invariant object recognition in cortex. Technical Report CBCL Paper 239 / AI Memo 2004-017, Massachusetts Institute of Technology, Cambridge, MA, July 2004.

[20] T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. Technical Report CBCL Paper 243 / AI Memo 2004-026, Massachusetts Institute of Technology, Cambridge, MA, November 2004.

[21] A. Torralba, K.P. Murphy, and W. T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. In *CVPR*, 2004.

[22] S. Ullman, M. Vidal-Naquet, and E. Sali. Visual features of intermdediate complexity and their use in classification. *Nature Neuroscience*, 5(7):682–687, 2002.

[23] P. Viola and M. Jones. Robust real-time face detection. In *ICCV*, volume 20(11), pages 1254–1259, 2001.

[24] M. Weber, M. Welling, and P. Perona. Unsupervised learning of models for recognition. In *ECCV*, Dublin, Ireland, 2000.

[25] H. Wersing and E. Korner. Learning optimized features for hierarchical models of invariant recognition. *Neural Computation*, 15(7), 2003.