Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2004-074
AIM-2004-026
CBCL-243

November 14, 2004

A new biologically motivated framework for robust object recognition

Thomas Serre, Lior Wolf, and Tomaso Poggio

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Abstract

In this paper, we introduce a novel set of features for robust object recognition, which exhibits outstanding performances on a variety of object categories while being capable of learning from only a few training examples. 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 – motivated by a quantitative model of visual cortex – outperforms state-of-the-art systems on a variety of object image datasets from different groups. We also show that our system is able to learn from very few examples with no prior category knowledge. The success of the approach is also a suggestive plausibility proof for a class of feed-forward models of object recognition in cortex. Finally, we conjecture the existence of a *universal* overcomplete dictionary of features that could handle the recognition of all object categories.

Copyright ©Massachusetts Institute of Technology, 2004

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.

1 Introduction

Most state-of-the-art object recognition systems appear to be hand-crafted and heuristically optimized for one object category, *i.e.,* faces [1–3], cars [2] or pedestrians [4]. Typically these systems require a large number of segmented training examples. This is in sharp contrast with primates' ability to learn to categorize objects from only very few unsegmented examples.

Recently, systems have been presented that can learn to recognize many objects (one at a time) using an unsegmented training set [5, 6]. These methods recognize highly informative object components and their spatial relations called *constellations*. Not only did those methods achieve good performance but were also shown to work with a very small training set containing few positive examples [6]. Yet another striking difference between these recent systems and the state-of-the-art single-object-recognition systems (*e.g.,* MERL's AdaBoost-based face and pedestrian detection systems, or MobileEye's SVM-based car detection system) is that they use generative (Bayesian) algorithms.

In this work we present a system that is simpler than constellation models [5, 6]: it uses discriminative methods and does not make use of any local object geometry. Yet it is able to learn from very few examples and to perform significantly better than all other systems we have tested. Our system first computes a set of biologically-inspired C2 features learned from the positive training set. We then run a standard classifier on the vector of features obtained from the input image. We report results using both linear SVM and gentleBoost. Since the source codes are readily available, our results should be easy to reproduce and extend.

Other existing features. Hierarchical approaches to generic object recognition have become increasingly popular over the years. They have been shown to outperform non-hierarchical single template (holistic) object recognition approaches on a variety of object recognition tasks (*e.g.,* face-detection [7]). Recognition is usually done in two steps: target features (also called components [4, 7], parts [8] or fragments [9]) are first computed and then passed to a combination classifier for final analysis. Constellation models using generative methods have been proposed in [5, 6, 8]. A robust face-detection system was built with a two-layer SVM system in [7] and variants of boosting algorithms were presented for fast face-detection [3] and multi-class [10] object recognition approaches.

One limitation of those template-matching-based features is that they do not capture adequately variations in the 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 [11, 12] have been shown to be very robust with respect to object transformations. The SIFT features [11], for instance, have been shown to excel in the re-detection of a previously seen object under new image transformations.

However, as we confirmed experimentally (see section 4), with such degree of invariance, it is very unlikely that those features could perform well on a generic object recognition task. We here propose a new set of features that exhibit just the right trade-off between invariance and selectivity. They are much more flexible than components [4] or fragments [9] and more selective than local descriptors. Though they are not strictly invariant to rotation, invariance to rotation can be introduced via the training set (*e.g.,* by introducing rotated versions of the original input).

Biological visual systems as guides. Because humans and primates outperform in almost any measure the best machine vision systems, 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 very first stage of processing in the simple cells of $V1$. Models of biological vision [13–16] have not been extended to deal with real-world object recognition tasks and tested on them.

The standard model of visual cortex. Our system follows *the standard model* of object recognition in primate cortex [17]. The model itself attempts to summarize in a quantitative way what most visual neuroscientists generally agree on: the first few hundred milliseconds of visual processing in primate cortex follows a mostly feed-forward hierarchy. At each stage, the receptive field of the neuron (*i.e.,* the part of the visual field that could potentially elicit a neuron's response) tends to get larger along with the complexity of its preferred stimuli (*i.e.,* the set of stimuli that are susceptible to elicit a neuron's response).

In its simplest form, the standard model consists of four layers of computational units where *simple* S units alternate with *complex* C units. The S units combine their inputs with Gaussian-like tuning to increase object selectivity. The C units pool their inputs through a maximum operation, thereby introducing invariance to scale and translation. The standard model has been able to duplicate quantitively 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 range of scales and positions.

The standard model in its simplest version [15] used a very simple *static* dictionary of features. It was suggested that features from the third and higher layer in the model should instead be learned from visual experience. We have extended the standard model by showing how to learn a vocabulary of visual features from images and applying it to the recognition of real-world object categories.

2 The C2 features

It is important to stress that biology imposes strong constraints on our system architecture: consistent with the standard view in neuroscience, our architecture is feed-forward and does not involve image scanning over all positions and sizes, the standard approach in computer vision. It also limits the basic operations that can be performed by individual units.

Our system 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 [18]. The S1 responses are obtained by applying to the input image a battery of Gabor filters, which can be described by the following equation:

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

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

We adjusted the four filters parameters, *i.e.*, orientation θ, aspect ratio γ, effective width σ, and wavelength λ, so that S1 units tuning profiles match those of V1 parafoveal simple cells. This was done by first sampling the space of the parameters and then generating a large number of filters. We applied those filters to stimuli commonly used to assess V1 neurons' tuning properties [18] (*i.e.*, gratings, bars and edges). After removing filters that were incompatible with biological cells [18], we were left with a final set of 16 filters at 4 orientations (see table 1).

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 [18] (shift invariance) and tend to be more broadly tuned than simple cells [18] (scale invariance). Modifying the original Hubel & Wiesel proposal for building complex cells from simple cells through pooling, Riesenhuber & Poggio proposed a max-like pooling operation for building position and scale tolerant C1 units. In the meantime, experimental evidence in favor of the max operation has appeared [19, 20]. Again, parameters governing this pooling operation were set so that C1 units match complex cells' tuning properties as measured experimentally (see table 1).

Given an input image, 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 $(S1)^s_\theta$ that are are arranged in 8 *bands* (*e.g.*, band 1 contains filters outputs of size 7 and 9, in all four orientations).

C1: For each *band*, we 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 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: We do not take a max over different orientations, hence, each band $(C1)^\Sigma$ contains 4 orientation 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) from the $(C1)^\Sigma$ maps from all training images.

S2: For image patches X at all positions from C1 image $(C1)^\Sigma$, compute: $Y = \exp(-\gamma||X - P_i||^2)$ for each band and each P_i independently.
Obtain the S2 maps $(S2)^\Sigma_i$.

C2: Compute the max over all positions and scales for each patch P_i and obtain shift and scale invariant C2 features $(C2)_i$, for $i = 1 \ldots K$.

Figure 1: Computing the C2 features.

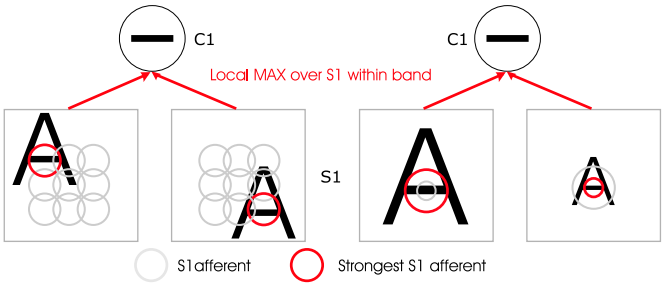

Figure 2: How scale and position tolerance is gained at the C1 level: Each C1 unit receives inputs from S1 units at the same orientation (*e.g.*, 0^o) arranged in *bands*. For each orientation, a band Σ contains S1 units in two different sizes and various positions (grid cell of size $N^\Sigma \times N^\Sigma$). From each grid cell (see left side) we obtain one measurement by taking the maximum over all positions: this allow the C1 unit to respond to an horizontal bar anywhere within the grid, thus providing a translation-tolerant representation. Similarly, taking a max over the two sizes (see right side), enables the C1 unit to be more tolerant to changes in scale.

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

Band Σ	1	2	3	4	5	6	7	8
filters sizes s	7 & 9	11 & 13	15 & 17	19 & 21	23 & 25	27 & 29	31 & 33	35 & 37
effective width 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
wavelength 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
orientation 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)							

Fig. 2 illustrates how pooling from S1 to C1 is done. For instance, consider the first band: $\Sigma = 1$. For each orientation, it contains two S1 maps: the one obtained using a filter of size 7, and the 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 done for each of the four orientations and each scale band independently.

In our new version of the standard model the subsequent S2 stage is where learning occurs. A large pool of *patches* of various sizes and at random positions are extracted from a target set of images at the level of the C1 layer for all orientations, *i.e.*, a patch P of size $n \times n$ contains $n \times n \times 4$ elements. The training process ends by setting each of those patches as *prototypes* or *centers* of the S2 units (at each position and scale) which behave as radial basis function (RBF) units during recognition. This is consistent with well-known neurons' response properties in primate inferotemporal cortex [21] and seems to be the key property for learning to generalize in the visual and motor systems **??**. Each S2 unit response depends in a Gaussian-like way on the Euclidean distance between a new input and the stored prototype.

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 the remainder of this paper, features are learned from the positive training set for each object, but the reader can refer to section 6 for a discussion on how features can be learned from natural images.

Our final set of shift and scale invariant C2 responses is computed by taking a global max over all scales and positions for each S2 type at each position on the S2 lattice. This results in as many C2 features as patches we extracted during the learning stage. Finally, in the computer system described here, the C2 responses to a new input image are passed to a classifier for final analysis*.

3 Experimental Setup

To demonstrate the quality of the C2 features, we compared their performances – when used as inputs to a classifier – with other systems on a variety of databases. Datasets that were available online at `www.vision.caltech.edu` include five (Caltech) databases (*i.e.*, frontal-face, motorcycle, rear-car and airplane datasets from [5] and a leaf dataset from [8]) and a 101-object database from [6]. We also considered two more challenging datasets: a near-frontal ($\pm 30°$) face dataset from [7] provided by Heisele *et al.* and a new multi-view car dataset that we collected. Fig. 4 shows some sample image patterns taken from the car and the face dataset.

For the Caltech datasets, positive training and test sets were generated using the splits provided by Fergus *et al.* . The negative training and test sets were randomly generated from the same background images as in [5]. All results we report for the 101-object category were generated with 10 random splits each using 50 training and 50 test negative examples from the same background image category as in [6]. For testing, we also used 50 positive test examples and experimented with different training set sizes (1, 3, 15, 30). All splits for the near-frontal face database were identical to the ones used in [7].

The face dataset contains about 6,900 positive and 13,700 negative images for training and 427 positive and 5,000 negative images for testing. The car dataset contains 4,000 positive and 1,600 negative training examples and 1,700 test examples (both positive and negative). Although *benchmark algorithms* were trained on the full sets and the results reported accordingly, our system only used a subset of the training sets (500 examples of each class only).

*While it would be straightforward to match our final classifier with prefrontal (PFC) cortex and C2 units with anterior inferotemporal (AIT) cortex [15, 22], it is more difficult to commit to a brain area for S2 units. Considering their size and complexity, they could be located in V4 and/or posterior inferotemporal (PIT) cortex. This reflects the lack of a precise characterization for neurons in intermediate brain areas, *i.e.*, between primary visual cortex (S1 and C1 layers) and AIT (C2 layer).

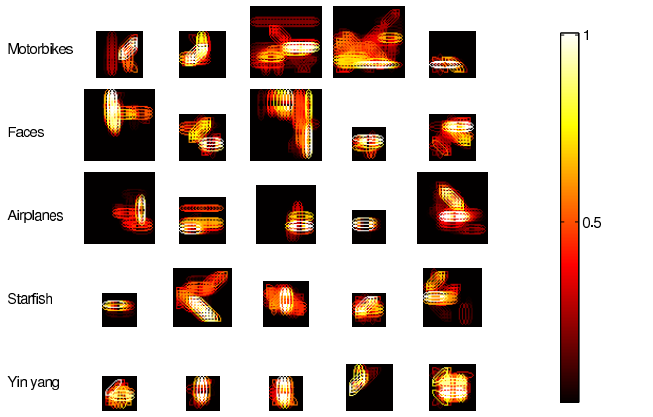

Figure 3: Examples of learned features (first 5 features returned by gentleBoost for each category). Each ellipse characterizes a C1 afferent at matching orientation, while color encodes for response strength.

It is important to point out that those two datasets are challenging. The face patterns used for testing are 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 car database was created by taking street scene pictures in the Boston area. Numerous vehicles (including SUVs, trucks, buses, *etc*) were manually labeled from those images to form a positive test set. Random image patterns at various scales that were not labeled as vehicles were extracted and used as a negative test set. For benchmarking this dataset, we implemented a fragment-based gentleBoost algorithm as in [10], as well as a gray-value single-template linear SVM.

As a preprocessing step to our system, we normalized images in size: all images from the Caltech web site were rescaled to be 140 pixels in height (width was rescaled accordingly so that the image aspect ratio was preserved) and converted to gray values. Images from the face database [7] were all 70×70 pixels and images from the car database were scaled down to 120×120 pixels.

In the remainder of this paper, to make past and future comparisons with other systems easier, we report two accuracy measures for our system: the Receiver Operator Characteristic area (*Area* in Fig. 5) that corresponds to the area under the curve and the error rate at equilibrium point (*Eq pt* in Fig. 5), *i.e.,* when the false positive rate equals the miss rate, since both measures are reported in the literature equally frequently.

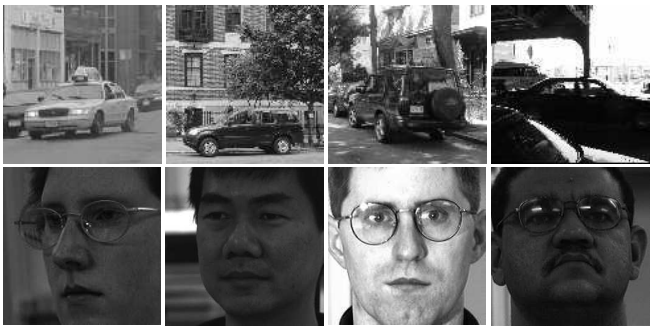

Figure 4: Examples taken from our difficult multi-view car dataset and the difficult face datasets used in [7].

4 Results

Figure 5 contains a summary of the performance exhibited by the C2 features used in conjunction with linear SVM (*C2 + (linear) SVM*) and gentleBoost [3] (*C2 + gentleBoost*) for various datasets, along with published results from other systems (*Benchmark algorithms*). Results obtained with the C2 features are consistently higher than those previously reported on all the datasets we tested: the leaf database [8], rear-car, frontal-face, motorcycle and airplane datasets [5], as well as two more challenging datasets, that is, the near-frontal ($\pm 30°$ rotation) face dataset [7] and a newly introduced multi-view car database.

As Fig. 6 (left) shows, after a critical number of C2 features (about 100), performances do not depend strongly on the number of features. For this experiment we first created a set of $10,000$ features total and randomly selected subsets of various sizes. The results shown are the average of 10 independent runs. As evident, performance could still be improved when allowing more features (*e.g.,* the whole set of 10,000), but reasonable performance can be obtained even with 50 features.

Our system seems to outperform the component-based system presented in [7] using a hierarchy of SVMs on the difficult face database. It also seems to outperform a system similar to [10] based on Ullman's features [9] and gentleBoost on the difficult car database (though it was trained using a much smaller training set). For illustration, we show on Fig. 6 (right), the ROC curves for both systems (C2-based and fragment-based with gentleBoost) and a single-template linear SVM.

Fig. 7 and 8 summarize the results we obtained on the 101-object database. For each object category, we generated positive training sets of sizes 1, 3, 6, 15 and 30 as in [6] (10 random splits for each). The negative training sets, and the test sets (both positive and negative) all contained 50 examples randomly selected. Fig. 7 (left) shows the C2 features-based system's performances (with gentleBoost) on the same datasets as the ones used by Li *et al.* for illustration.

Datasets	Benchmark algorithms			C2 + gentleBoost		C2 + (linear) SVM	
	Ref.	Area	Eq pt	Area	Eq pt	Area	Eq pt
Leaves (Caltech)	[8]	NA	84.0	99.4	97.0	99.5	95.9
Cars (Caltech)	[5]	NA	84.8	100.0	99.7	100.0	99.8
Faces (Caltech)	[5]	NA	96.8	99.8	98.2	99.8	98.1
Airplanes (Caltech)	[5]	NA	94.0	99.6	96.7	98.8	94.9
Motorcycles (Caltech)	[5]	NA	95.0	99.8	98.0	99.7	97.4
Faces	[7]	96.0	90.4	99.3	95.9	99.2	95.3
Cars	(*)	83.3	75.4	98.8	95.1	97.7	93.3

Figure 5: Sample results obtained by classifying the C2 features (1,000) with both a linear SVM (*C2 + (linear) SVM*) and gentleBoost (*C2 + gentleBoost*) and comparison with existing systems (*Benchmark algorithms*). We report both the ROC area (*Area*) and the error rate at equilibrium point (*Eq pt*). (*) corresponds to a system we implemented that uses Ullman's features [9] and gentleBoost as in [10].

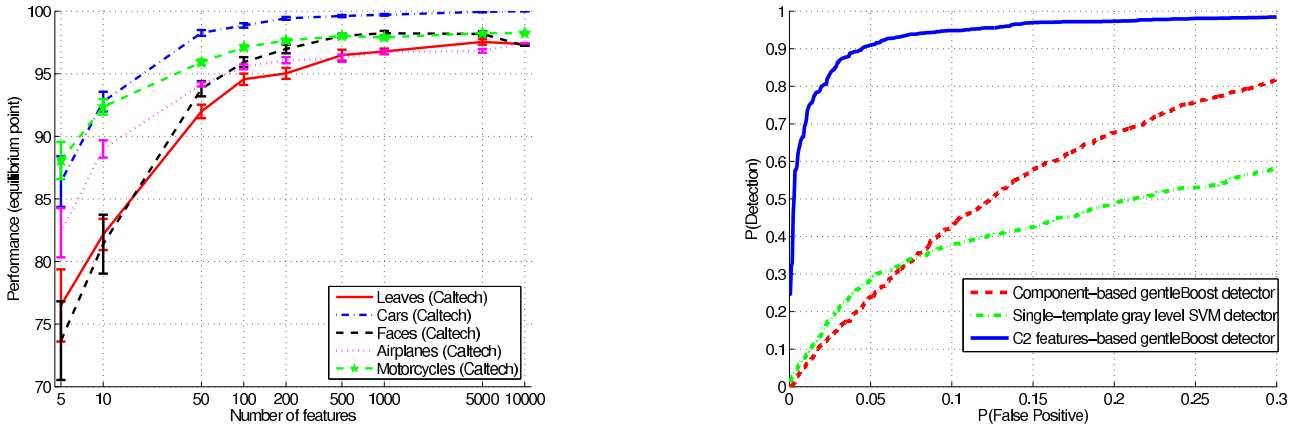

Figure 6: (left) C2 features performances (with gentleBoost) on various Caltech datasets [5] for different numbers of features and (right) ROC curves obtained with the C2 features on the difficult car dataset for comparison with a component-based (gentleBoost) system similar to [10] and a single-template SVM.

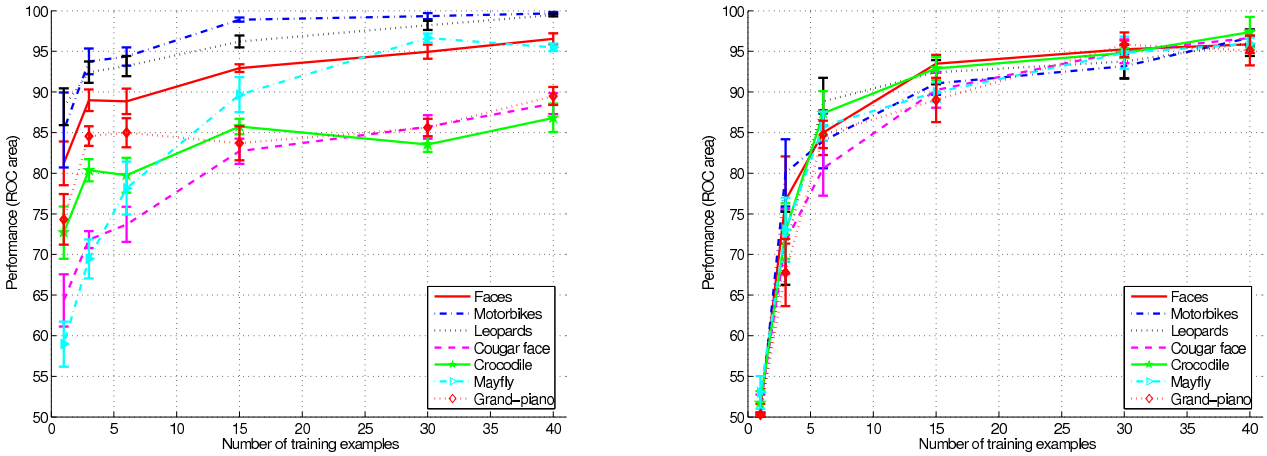

Figure 7: C2 features performances with linear SVM (left) and gentleBoost (right) on sample categories [6] from the 101-object database for different numbers of positive training examples.

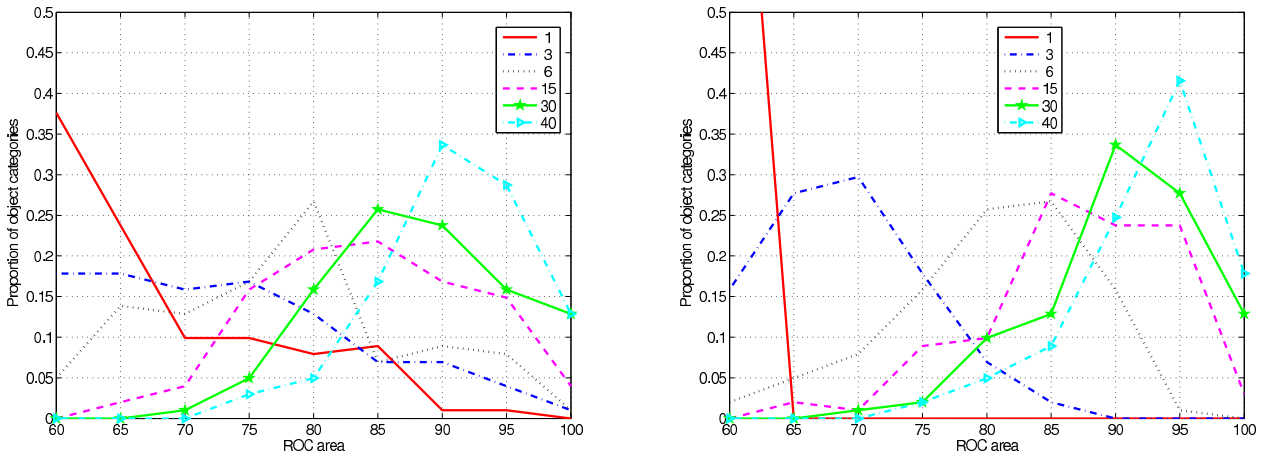

Figure 8: Overall peformances (histograms) accross the 101-object categories for the C2 features with (left) linear SVM and (right) gentleBoost and different numbers of positive training examples.

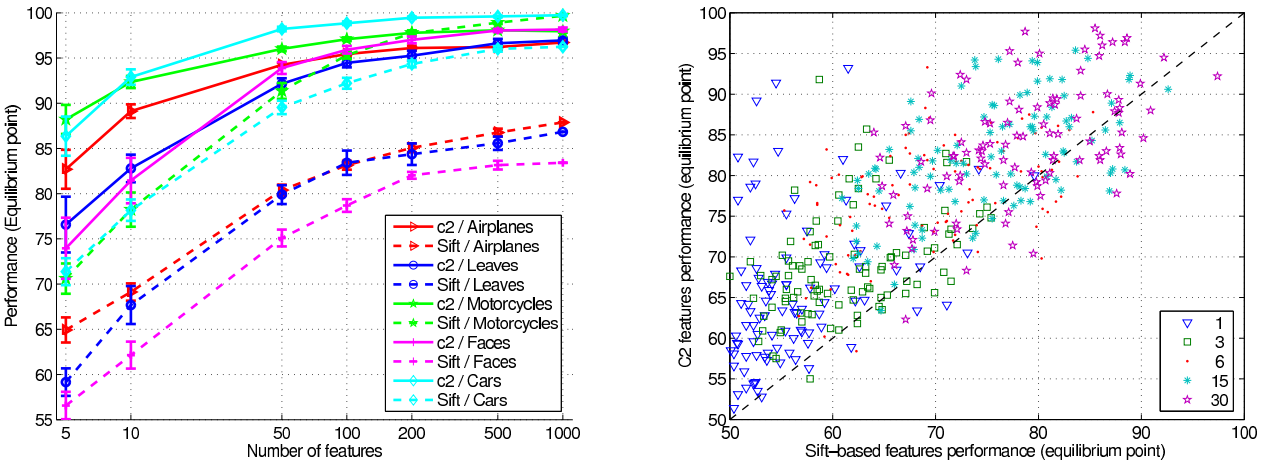

Figure 9: Superiority of the C2 *vs.* SIFT-based features on (left) various (Caltech) datasets for different numbers of features and (right) on the 101-object database for different numbers of positive training examples as in [6].

Our gentleBoost system achieves error rates similar to the ones reported in [6], with very few training examples (from 3 to 6) and tends to do better with more examples. It seems that SVM avoids overfitting even for one example – see Fig. 7 (left) – and outperforms gentleBoost for *one-shot learning* (learning from one example). However, since SVM does not seem to be able to *select* the relevant features, its performance is lower than gentleBoost as the number of training examples increases (see 15 and 30 examples). Fig. 8 shows the performances of the gentleBoost and SVM classifiers used with the C2 features over all categories and for various training set sizes (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 gentleBoost-based system gets 95% correct for 42% of the object categories.

We also compared our C2 features to a system based on Lowe's SIFT features [11]. For this comparison, we neglected all position information recovered by Lowe's algorithm. 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. Note that Lowe recommends using the ratio of the distances between the nearest and the second closest key-point as a similarity measure. We found instead that the minimum distance leads to better performances than the ratio.

On Fig. 9 (left) we compared the SIFT-based features and the C2 features on various Caltech datasets (leaf, motorcycle, airplane, car and face). The gain in performance obtained by using the C2 features relative to the SIFT-based features is obvious. This is true with gentleBoost – used for classification on Fig. 9 (left) – but we also found very similar results with a linear SVM. Also, as one can see in Fig. 9 (right), the C2 features performances (error at equilibrium point) for each category from the 101-object database is well above those of the SIFT-based features for any number of training exam-

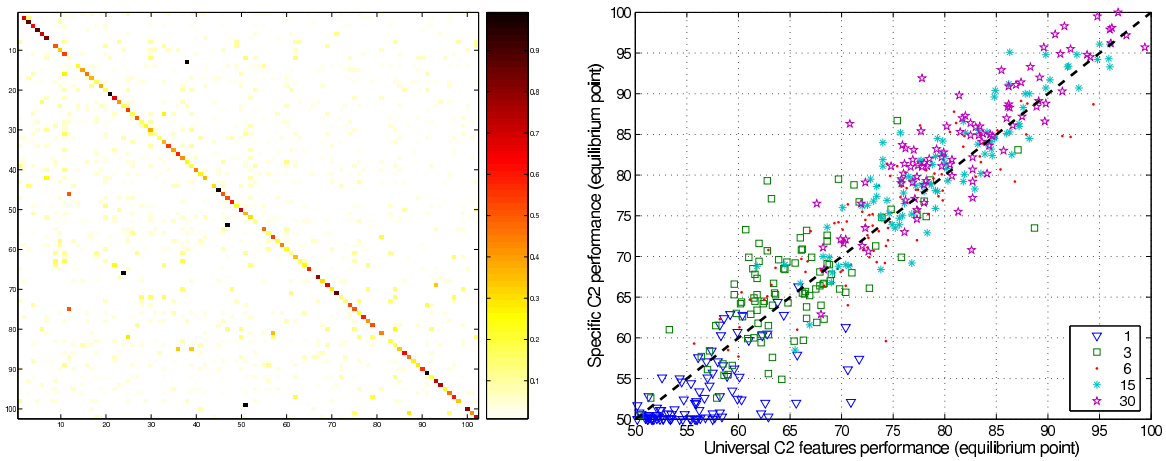

Figure 10: (Left) Multiclass classification on 101-object database with linear SVM and (right) Object-specific *vs.* universal features on the 101-object database.

ples. This difference was significant (paired t-test over all training sizes, $p = 10^{-78}$).

Fig. 3 shows examples of features we obtained after training for motorcycle, face, airplane, starfish and yin-yang images. This figure shows the five first features selected by the gentleBoost algorithm. Recall that each feature contains $n \times n \times 4$ elements, where n is the number of S1 afferents or *patch size* and 4 corresponds to the 4 orientations. For visualization, we collapsed all orientations onto a single map, *i.e.*, each ellipse characterizes a S1 afferent at matching orientation, while color encodes for its response strength. One should keep in mind that this simplified representation is inaccurate : C1 units are translation and scale tolerant *i.e.*, their *preferred stimulus is not unique*. For simplicity, we represent an ellipse in the center of each unit but in practice its exact location may vary. As one can see from Fig. 3, features that were chosen by the boosting algorithm also vary widely from one category to another (both in size and shape).

Finally, we conducted initial experiments on the multiple class case. For this task we used the 101-object dataset. We split each category into a training set of size 15 or 30 and and a test set containing the rest of the images. We used a multiple-class linear SVM to train a classifier. The SVM applied the all-pairs method for multiple labels classification, and was trained on 102 labels (101 categories plus the background category). The number of $C2$ features used in these experiment was 4075. The results we obtained, averaged over 10 repetitions of the experiments, were 35% correct classification rate when using 15 training examples per class, and 42% correct classification rate when using 30 training examples. Fig. 10 shows the confusion matrix for the 101-object categories.

5 Implications for Object Recognition in Cortex

Our approach is biologically motivated and our system belongs to a family of feed-forward models of object recognition in cortex that have been shown to be able to duplicate neurons' tuning properties 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 (IT) which respond to images of the learned object more strongly than to distractor objects, despite significant changes in position and size [22]. Riesenhuber & Poggio reported performance of the model only on idealized stimuli such as paperclips on a uniform background [23] (no real-world image degradation such as change in illumination, clutter, *etc*). The success of our extension of their original model on a variety of real-world object databases provides a compelling plausibility proof for this class of feed-forward models.

A long-time goal for computer vision has always been to build a system that achieves human-level recognition performance. Until now, biology had not suggested a good solution. In fact, the superiority of human performances over the best artificial recognition systems has continuously lacked a satisfactory explanation. The computer vision approaches had also diverged from biology: for instance, some of the best existing computer vision systems use geometrical information about objects constitutive parts whereas biology is unlikely to be able to use it - at least in the cortical stream dedicated to shape processing and object recognition. The system described in this paper may be the first counter-example to this situation: it is based on a model of object recognition in cortex [15], it respects

10 best		10 worst	
metronome	100.0	chair	73.0
inline skate	99.5	barrel	72.1
scissors	98.3	ibis	72.1
pagoda	98.1	octopus	71.6
trilobite	97.9	cup	71.3
faces	97.3	cannon	71.1
accordion	97.2	wheelchair	70.8
minaret	96.2	lamp	70.6
faces easy	95.7	flamingo	68.4
car side	95.7	ewer	62.9

Table 2: 10 best and 10 worst categories (*Eq pt*) from the 101-object database.

the properties of cortical processing (including the absence of geometrical information) while showing performance at least comparable to the best computer vision systems.

We finally show results suggesting that it is possible to perform robust object recognition from C2 features learned from natural images. In Fig. 10, we compare the performances of two sets of features on the 101-object database: (1) a set of *object-specific* features that were learned from the training set of the target object category (20 features per training image); and (2) a *universal* set of 10,000 features that were learned from a general set of natural images (downloaded from the web). While the *object-specific* set performs significantly better with enough training examples ($p = 3.7 \cdot 10^{-7}$, paired t-test for 30 training examples), the situation is reversed for small training sets ($p = 7.5 \cdot 10^{-12}$, paired t-test for 1 training example).

This apparent superiority of the *universal* set over the object-specific one for small training sets is somewhat counter-intuitive and very interesting. First, the *universal* feature set is less prone to overfitting with few training examples (recall that both the features learning and classifier training is performed on the same set in the *object-specific* case). Second, the size of the *universal set* is constant regardless of the number of training examples (10,000), while the size of the *object-specific* set is much smaller (20 times the number of training images).

We believe that this represents a relevant and intriguing result on its own - towards the holy grail of finding the elusive *universal dictionary of visual shapes*. Our results also suggest that it should be possible for biological organisms to acquire a basic vocabulary of features early in development while refining it with more specific features during adulthood. The latter point is consistent with reports of plasticity in inferotemporal cortex from adult monkey [22, 24] (our C2 features complexity and sizes are consistent with neurons receptive field in posterior IT [24]).

6 Discussion

In the present paper we described a new framework for robust object recognition: our system first computes a set of biologically-inspired scale- and translation-invariant C2 features from a training set of images. We then run a standard classifier on the vector of features obtained from the input image. We showed that our approach exhibits outstanding performances on a variety of image datasets.

A biologically-inspired state-of-the-art approach. While significantly simpler than other state-of-the-art systems, our approach nonetheless exhibits consistently better results than all systems we have compared it to. For instance, the systems described in [5, 6, 8] involve the estimation of probability distributions; [7] uses a hierarchy of SVM classifiers and requires correspondences between positive training images (3D head models).

We also showed that a relatively small number of features (about 50) is sufficient to achieve reliable object recognition. However, performances can be increased significantly by adding more features. Interestingly, the number of features needed to reach the plateau (about 5,000 features) is much larger than the number used by current systems (on the order of 10-100 for [7, 9, 10] and 4-8 for constellation approaches [5, 6, 8]).

On the role of relative geometry for generic object recognition. It is also important to point out that, contrary to recent trends — but consistent with neurophysiology constraints – we do not model local object geometry. The constellation approaches [5, 6, 8] rely on a probabilistic shape model; in [7] the position of the facial components is passed to a combination classifier (along with their associated detection values); in [10] object parts are searched only in their approximated expected position. We should emphasize that the absence of shape information in our approach follows directly the standard model; it was therefore guided by what we know about properties of visual processing within the sequence of visual areas comprising the ventral stream, which is responsible for object recognition.

Use of prior *vs.* use of negative examples. In recent years, generative models have gained popularity in object recognition applications. In particular, it was recently shown that generative models combined with the use of prior category information could produce systems able to learn from few examples [6]. Our system does not exploit any prior, but instead uses a training set which contains negative examples. Negative examples provide extremely useful information to our classifier with little cost (negative examples are easy to obtain). Note that in the tests reported here, we did not tune any parameter to obtain optimal performance. Instead, we tuned the parameters to match what is known about the primate visual system.

The quest for universal features. We finally showed preliminary results suggesting that it is possible to perform robust object recognition with a *universal* set of C2 features learned from natural images (see section 5). We plan on making this universal feature set available to the community on our web site soon. As those features were learned from randomly selected images, they might not all be useful for classification; we are now studying which features, out of this large set, are indeed informative.

References

[1] 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.

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

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

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

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

[6] 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.

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

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

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

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

[17] M. Riesenhuber and T. Poggio. How visual cortex recognizes objects: The tale of the standard model. *The Visual Neurosciences*, 2:1640–1653, 2003.

[18] 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.

[19] 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.

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

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

[22] N. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.*, 5:552–63, 1995.

[23] M. Riesenhuber and T. Poggio. Neural mechanisms of object recognition. *Curr. Neurosci.*, 12:162–68, 2002.

[24] K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Shape representation in the inferior temporal cortex of monkeys. *J. Neurophysiol.*, 66:170–89, 1991.