OBJECT RECOGNITION BY LEARNING INFORMATIVE, BIOLOGICALLY INSPIRED VISUAL FEATURES

Yang Wu, Nanning Zheng, Qubo You, Shaoyi Du

Institute of Artificial Intelligence and Robotics
Xi'an Jiaotong University, Xi'an, 710049, P. R. China
ywu@aiar.xjtu.edu.cn, nnzheng@mail.xjtu.edu.cn, qbyou@aiar.xjtu.edu.cn, sydu@aiar.xjtu.edu.cn

ABSTRACT

This paper presents a novel, effective way to improve the object recognition performance of a biologically-motivated model by learning informative visual features. The original model has an obvious bottleneck when learning features. Therefore, we propose a circumspect algorithm to solve this problem. First, a novel information factor was designed to find the most informative feature for each image, and then complementary features were selected based on additional information. Finally, an intra-class clustering strategy was used to select the most typical features for each category. By integrating two other improvements, our algorithm performs better than any other system so far based on the same model.

Index Terms— object recognition, feature learning, visual cortex, biologically-inspired model, Caltech-101 database

1. INTRODUCTION

Since humans and other primates have great object recognition power that well outperforms any machine vision system, building a system that emulates object recognition in visual cortex has always been an attractive idea [1]. However, it's not easy to design a robust recognition system based on both state-of-the-art neurophysiologic findings and powerful machine learning technologies.

Serre et al. [1] proposed a biologically-motivated framework for robust object recognition, which used a hierarchical image representation expanded from the standard model of object recognition in primate cortex [2]. This framework alternately performs template matching (tuning) and max pooling operations to achieve a good trade-off between selectivity and invariance. Its built-in gradual shift- and scale-tolerance allowed it to outperform most contemporaneous complex computer vision systems.

Serre et al.'s system was a successful attempt to bridge the gap between computer vision and neuroscience, but it's still very simple. Many aspects of this framework could be modified to improve it. Wolf et al. [3] focused on discussing different perception strategies in hierarchical vision systems instead of Serre et al.'s feed-forward framework. Mutch and Lowe [4] proposed many useful modifications that further improved recognition performance on multi-class experiments. These efforts resulted in much higher classification rates, but they were still not as good as some newly proposed, well-connected computer vision systems [5]. Much more work should be done to find more powerful improvements.

In fact, Serre et al. [1] paid too little attention to a very important bottleneck: the learning of visual features (prototypes). They only sampled these features randomly from training images or random natural images. Though Mutch and Lowe [4] introduced an SVM-based supervised feature selecting technique to find more discriminative features from the randomly sampled prototype patches, it still depended on the casual features which were very weak and desultory. Furthermore, it was a totally supervised learning approach, which disobeyed Serre et al.'s unsupervised learning assumption. Instead, in this paper we propose an effective way to learn many more informative visual features in an unsupervised way based on biological evidences. Moreover, an intra-class clustering strategy is introduced to reduce the redundancy of candidate features. Meanwhile, some reasonable improvements such as lateral inhibition and limitation of feature invariance presented in [4] are also used here to improve recognition performance. In the multi-class recognition experiments, we present better results than any of the systems so far based on the same biologically motivated model.

2. BIOLOGICALLY MOTIVATED FRAMEWORK

Our framework is mostly based on that of Serre et al. [1], with some small changes according to the base model described in [4]. The model generates final features by alternating "S" (simple cell) and "C" (complex cell) layers [6], which account for the tuning and invariance properties respectively. It can be summarized as follows:

Given an input image, we first convert it to grayscale and scale the shorter edge to 140 pixels while maintaining the as-

This work was supported by the National Basic Research Program of China under Grant No. 2006CB708303, and the National High-Tech Research and Development Plan of China under Grant No. 20060101Z1059.

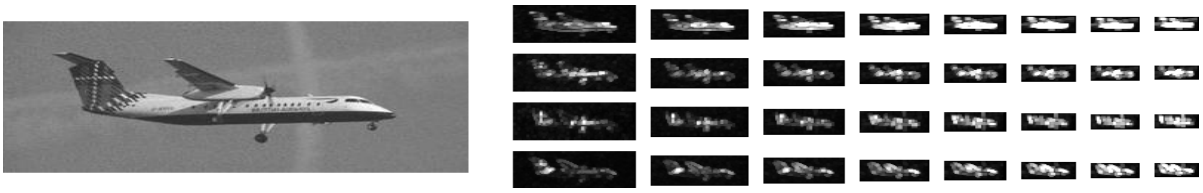


Fig. 1. A sample gray image and its C1 band maps (based on the true sizes). Rows indicate 4 orientations and columns show the 8 bands of different scales.

pect ratio, as in [4]. Then, four steps are performed according to the four hierarchical layers on the preprocessed image:

S1: A battery of Gabor filters is applied to the input image, which can be described by:

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

where $X = x\cos\theta + y\sin\theta$ and $Y = -x\sin\theta + y\cos\theta$. There are, in total, 64 different well-chosen filters with 4 orientations θ and 16 scales s that were listed clearly in [1]. Every 2 scales form a band. In total, eight band maps with four orientations are generated as S1 responses.

C1: This complex cell layer pools from the S1 layer by performing an experimentally effective max operation over a group of S1 cells within the neighborhood of each band. It creates slight position and scale invariance. Fig. 1 shows an example of a gray image and its corresponding C1 band maps.

Feature Learning Stage (during training only): Extract K patches $P_{i=1,\cdots,K}$ of various sizes $n_i \times n_i$ as a vocabulary of prototype features. Each patch still has 4 orientations for keeping the directional information.

S2: Apply each patch (prototype) as a filter to the C1 band maps to generate responses according to the Gaussian radial basis function:

$$R(X,P) = \exp\left(-\frac{\|X-P\|^2}{2\sigma^2\alpha}\right).$$

This function was proposed in [4] as an improvement over the original model, where σ was set to 1 and α was a normalizing factor for different patch sizes. It measures the similarity between the S2 cells and a trained patch.

C2: For each patch, C2 performs the max operation over all the eight bands (across positions and scales) to get a final maximum response. By doing this, it generates a K-dimensional vector as the shift- and scale-invariant C2 feature representation of the input image.

After normalization to each dimension of the C2 features, the C2 feature vector is ready for classification. Usually, an SVM classifier is used to do the final analysis.

3. LEARNING INFORMATIVE FEATURES

The above framework compromises selectivity and invariance by alternating tuning and max pooling operations, but obviously the important learning process that generates proper vocabulary of prototypes for the two upper layers has been treated cursorily. To improve it, we designed a circumspect algorithm to learn informative features (prototypes) from the training C1 band maps. It includes the following three steps.

3.1. Finding the most informative features

We consider all the possible predefined patch sizes as long as they do not exceed 50% of the shorter edge of the C1 band. For each patch size, we examine patches in positions placed on a regular grid with a step of 1/3 patch size. For every patch, an information factor is computed to evaluate its importance:

$$IF(P) = \frac{S\left(P^{\max}\right)}{\beta n^2} + \frac{S\left(P^{\max} - P^{\min}\right)}{\beta n^2}$$

where P is the candidate patch with size $n \times n$, β is a balancing factor for patches with different sizes, and

$$S(I) = \sum_i \sum_j I(i,j)^2$$
$$P^{\max} = \arg \max_{P^k \in \Phi(P)} S\left(P^k\right)$$
$$P^{\min} = \arg \min_{P^k \in \Phi(P)} S\left(P^k\right)$$

with $\Phi(P)$ denotes the set of different orientation parts of patch P. The first part of $IF(P)$ indicates the average energy or response strength of the most notable orientation, while the latter describes the distinctness of the direction information. It is biologically reasonable that the bigger the information factor $IF(P)$ is, the more informative the patch is. Thus, for each training image, the patch with the highest information factor score is selected as the maximal informative feature (prototype). Note that this definition of information factor is trying to characterize the informativeness of local contrast and regular directional information of the patches. It may not be the optimal measure for selecting the most discriminative features towards classification.

3.2. Selecting complementary features

If some features have been extracted, the next feature to be selected should deliver the maximal amount of additional information with respect to previously selected features. Therefore, an update stage was designed to refresh the band where

the last selected feature stayed by setting all the pixels covered by that feature to zero. Then, the same extracting rule presented in Section 3.1 is applied to select optimal complementary features one by one until the number of selected features has reached a predefined limit or the additional information factor score has fallen beneath a certain percent of the summed information factor score of the former features.

3.3. Clustering intra-class features

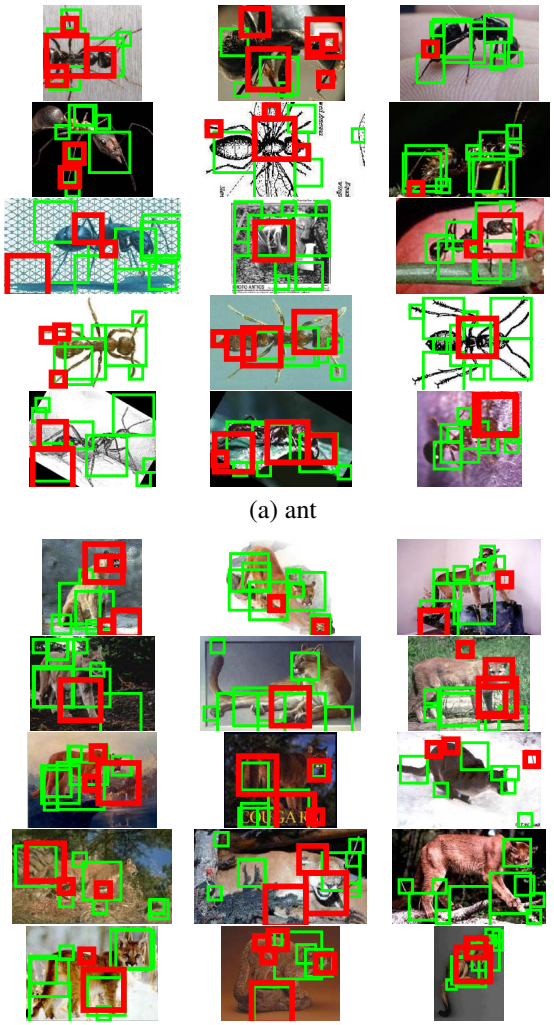
To control the dimensionality of C2 features, the number of prototype features needs to be fixed in a certain range. However, the number of training images might be so large that the average number of prototypes assigned to each image is quite small. For avoiding losing useful information of each training image, we choose to extract adequate features from each image using the above approach and then perform a k-means clustering on each category to get typical intra-class features. To respect the features extracted from real training images, a nearest neighbor scheme is used to find the features nearest to the cluster centers. Different patch sizes are treated separately. This strategy balances information coming from different images within a category while at the same time reducing the redundancy, and it can perform better if there are more training images. However, the intra-class clustering is supervised. We think it's reasonable that some limited supervised learning may exist in primary visual cortex.

4. OTHER IMPROVEMENTS

Two of the more useful improvements proposed in [4] are used here to get better performance. We inhibit C1 outputs by suppressing relatively smaller responses to make the directional information more obvious. At the same time, limiting position and scale invariance of S2 features can make use of the statistical information when the training and testing set have their contained objects arranged at similar positions and scales. If the system doesn't have this property, the second improvement should be removed or changed.

5. EXPERIMENTS

For a close and equitable comparison with other systems based on the same biologically motivated model, we chose to do our experiments on the Caltech-101 database [7], which has become a de facto standard for evaluating algorithms for multi-class category-level recognition, though it lacks in several important sources of intra-class variability [8]. We ran our model on the entire set containing 102 categories. 10 independent repetitions were performed on both 15 and 30 training examples per class respectively to get average recognition rates. As in [4], up to 100 of the rest of the images in each class were randomly sampled to work as the test set.



(a) ant

(b) cougar_body

Fig. 2. Extracted candidate patches (green and red) and finally selected patches after intra-class clustering (red) of two hard categories when 15 images are trained, best viewed in color. (a) ant, (b) cougar_body.

To get a similar feature dimension to [1], 40 patches (features) per category (except the background) were learned to form a total prototype vocabulary of 4040. Instead of using only band 2 for patch extracting, three bands (bands 2, 4, and 6) with proportional spacing were chosen to make the learned patches scale-invariant. Patch sizes were also fixed to 4, 8, 12, and 16, but the numbers of patches of each size were made adaptive to the patches' performance. The balancing factor β is set to $\frac{1}{n}$, avoiding bias to the smallest size. We experientially extracted 10 candidate patches per image before clustering, regardless of how many images per category were used for training. The inhibition level was set to 0.5 and the invariance of the generated C2 features was fixed to $\pm 5\%$ of the possible position and ± 1 of the scale as were in [4].

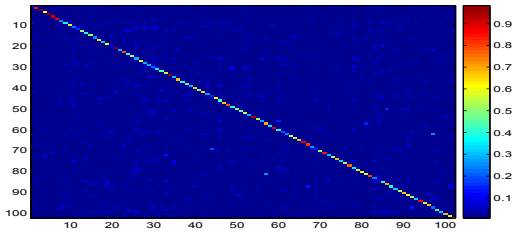


Fig. 3. Confusion matrix of the Caltech-101 database with 15 training images per class.

Fig. 2 shows the candidate patches and finally selected patches of two hard categories: "ant" and "cougar_body," when using 15 training images. These two categories are hard to handle for most systems because they vary greatly in pose, view, background, and quality. However, our learning algorithm can focus on the most informative areas, catching the patches with strong directional information and high contrast. The clustering strategy then selects the most typical ones that are most common to appear. From the "ant" training samples we can see that the legs and typically shaped body parts are more easily selected than the corners and the backgrounds. Of course, some similarly shaped backgrounds, or those with obviously strong edges, may also be falsely chosen.

Model	**15 training images/cat.**	**30 training images/cat.**
Mutch & Lowe (base) [4]	33	41
Serre et al. [1]	35	42
Mutch & Lowe (final) [4]	51	56
Wolf et al. (best result) [3]	51.18($\pm$1.2)	
Our system	**52.16**($\pm$1.0)	**60.23**($\pm$0.8)

Table 1. Correctness rates of systems based on the same biologically motivated model for the Caltech-101 database (in percentage, and std dev. where available).

The final tested results are summarized in Table 1, comparing our system to other systems. Our system performs best by using relatively simpler improvements. The confusion matrix for 15 training images is presented in Fig. 3. If the parameters are optimized by tuning carefully as was done in [4], our system might perform even better.

6. CONCLUSION AND FUTURE WORK

This paper presents an effective solution for the important learning bottleneck of a biologically-motivated model. Instead of using randomly sampled features or an SVM-based supervised selecting strategy, we designed our learning algorithm mainly according to the unsupervised learning assumption in the primary visual cortex. A novel information factor was proposed to extract the most informative features and the optimal complementary features based on their additional information. An intra-class clustering strategy was also used to select the most typical ones from the extracted features, while at the same time reducing the information redundancy of all the training images. Together with two other improvements, our system performs better than any other system based on the same biologically-inspired model on the Caltech-101 database, though it hasn't exceeded the best published results reached by other computer vision or pattern recognition based systems on this database recently.

Future work could be done on two issues: one is finding proper schemes for dealing with heavy clusters; the other is introducing rotation-invariant descriptors or effective algorithms to adapt to object rotations, view changes, and even deformations to some extent.

7. REFERENCES

[1] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Proc. CVPR*. IEEE, 2005.

[2] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nat. Neurosci.*, vol. 2, no. 11, pp. 1019–1025, 1999.

[3] L. Wolf, S. Bileschi, and E. Meyers, "Perception strategies in hierarchical vision systems," in *Proc. CVPR*. IEEE, 2006.

[4] J. Mutch and D. G. Lowe, "Multiclass object recognition with sparse, localized features," in *Proc. CVPR*. IEEE, 2006.

[5] A. Frome, Y. Singer, and J. Malik, "Image retrieval and classification using local distance functions," in *Proc. NIPS*, 2006.

[6] T. Serre, M. Kouh, C. Cadieu, U. Knoblich, G. Kreiman, and T. Poggio, "A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex," TR 082, MIT-CSAIL, 2005.

[7] 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*. IEEE, 2004.

[8] Jean Ponce, T. L. Berg, M. Everingham, D. Forsyth, M. Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, C. Russell, A. Torralba, C. Williams, Jianguo Zhang, and Andrew Zisserman, "Dataset issues in object recognition," in *Towards Category-Level Object Recognition*. Springer, 2006, to appear.