Object Class Recognition and Localization Using Sparse Features with Limited Receptive Fields

Jim Mutch · David G. Lowe

Abstract We investigate the role of sparsity and localized features in a biologically-inspired model of visual object classification. As in the model of Serre, Wolf, and Poggio, we first apply Gabor filters at all positions and scales; feature complexity and position/scale invariance are then built up by alternating template matching and max pooling operations. We refine the approach in several biologically plausible ways. Sparsity is increased by constraining the number of feature inputs, lateral inhibition, and feature selection. We also demonstrate the value of retaining some position and scale information above the intermediate feature level. Our final model is competitive with current computer vision algorithms on several standard datasets, including the Caltech 101 object categories and the UIUC car localization task. The results further the case for biologically-motivated approaches to object classification.

Keywords Object class recognition · Ventral visual pathway · Sparsity · Localized features

1 Introduction

The problem of recognizing multiple object classes in natural images has proven to be a difficult challenge for computer vision. Given the vastly superior performance of human vision on this task, it is reasonable to look to biology for inspiration. Recent work by Serre et al. (2005) used a computational model based on our knowledge of visual cortex to obtain promising results on some of the standard classification datasets. Our paper builds on their approach by incorporating some additional biologically-motivated properties, specifically, sparsity and localized intermediate-level features. We show that these modifications further improve classification performance, strengthening our understanding of the computational constraints facing both biological and computer vision systems.

Within machine learning, it has been found that increasing the sparsity of basis functions (Figueiredo 2003; Krishnapuram et al. 2005) (equivalent to reducing the capacity of the classifier) plays an important role in improving generalization performance. Similarly, within computational neuroscience, it has been found that adding a sparsity constraint is critical for learning biologically plausible models from the statistics of natural images (Olshausen and Field 1996). In our object classification model, one way we have found to increase sparsity is to use a lateral inhibition step that eliminates weaker responses that disagree with the locally dominant ones. We further enhance this approach by matching only the dominant orientation at each position within a feature rather than comparing all orientation responses. We also increase sparsity during final classification by discarding features with low weights and using only those that are found most effective. We show that each of these changes provides a significant boost in generalization performance.
While some current successful methods for object classification learn and apply quite precise geometric constraints on feature locations (Fergus et al. 2003; Bouchard and Triggs 2005), others ignore geometry and use a “bag of features” approach that ignores the locations of individual features (Csurka et al. 2005; Opelt et al. 2006). Intermediate approaches retain some coarsely-coded location information (Agarwal et al. 2004) or record the locations of features relative to the object center (Leibe et al. 2004; Berg et al. 2005). According to models of object recognition in cortex (Riesenhuber and Poggio 1999), the brain uses a hierarchical approach, in which simple, low-level features having high position and scale specificity are pooled and combined into more complex, higher-level features having greater location invariance. At higher levels, spatial structure becomes implicitly encoded into the features themselves, which may overlap, while explicit spatial information is coded more coarsely. The question becomes one of identifying the level at which features have become complex enough that explicit spatial information can be discarded. We investigate retaining some degree of position and scale sensitivity up to the level of object detection, and show that this provides a significant improvement in final classification performance.

We test these improvements on the large Caltech dataset of images from 101 object categories (Fei-Fei et al. 2004). Our results show that there are significant improvements to classification performance from each of the changes. Further tests on the UIUC car database (Agarwal et al. 2004) and the Graz-02 datasets (Opelt et al. 2006) demonstrate that the resulting system can also perform well on object localization. Our results further strengthen the case for incorporating concepts from biological vision into the design of computer vision systems.

2 Models

The model presented in this paper is a partial implementation of the “standard model” of object recognition in cortex (as summarized by Riesenhuber and Poggio 1999), which focuses on the object recognition capabilities of the ventral visual pathway in an “immediate recognition” mode, independent of attention or other top-down effects. The rapid performance of the human visual system in this mode (Potter 1975; Thorpe et al. 1996) implies mainly feedforward processing. While full human-level classification performance is almost certain to require feedback, the feedforward case is the easiest to model and thus represents an appropriate starting point. Within this immediate recognition framework, recognition of object classes from different 3D viewpoints is thought to be based on the learning of multiple 2D representations, rather than a single 3D representation (Poggio and Edelman 1990).

2.1 Previous Models

Our model builds on that of Serre et al. (2005), which in turn extends the “HMAX” model of Riesenhuber and Poggio (1999). These are the latest of a group of models which can be said to implement parts of the standard model, including neocognitrons (Fukushima 1980) and convolutional networks (LeCun et al. 1998). All start with an image layer of grayscale pixels and successively compute higher layers, alternating “S” and “C” layers (named by analogy with the V1 simple and complex cells discovered by Hubel and Wiesel 1959).

- Simple (“S”) layers apply local filters that compute higher-order features by combining different types of units in the previous layer.
- Complex (“C”) layers increase invariance by pooling units of the same type in the previous layer over limited ranges. At the same time, the number of units is reduced by subsampling.

Recent models have moved towards greater quantitative fidelity to the ventral stream. HMAX was designed to account for the tuning and invariance properties (Logothetis et al. 1995) of neurons in IT cortex. Rather than attempting to learn its bottom-level (“S1”) features, HMAX uses hard-wired filters designed to emulate V1 simple cells. Subsequent “C” layers are computed using a hard max, in which a C unit’s output is the maximum value of its afferent S units. This increases feature invariance while maintaining specificity. HMAX is also explicitly multiscale: its bottom-level filters are computed at all scales, and subsequent C units pool over both position and scale.

Serre et al. (2005) introduced learning of intermediate-level shared features, made additional quantitative adjustments, and added a final SVM classifier to make the model useful for classification.

2.2 Our Base Model

We start with a “base” model which is similar to (Serre et al. 2005) and performs about as well. Nevertheless, it is an independent implementation, and we give its complete description here. Its differences from (Serre et al. 2005) will be listed briefly at the end of this section. Larger changes, representing the main contribution of this paper, are described in Sect. 2.3.

The overall form of the model (shown in Fig. 1) is very simple. Images are reduced to feature vectors, which are

---

1 Source code and related documentation for our model may be downloaded at http://www.mit.edu/~jmutch/fhlib.html.