Generic Object Recognition with Probabilistic Models and Weak Supervision

Yatharth Saraf
University of California, San Diego

ysaraf@cs.ucsd.edu

Abstract

Real world images of objects belonging to a particular class typically show large variability in shape, appearance, scale, degree of occlusion, etc. Thus, a major challenge for generic object recognition is to develop object models that are flexible enough to accomodate these large intra-class variabilities. Such powerful models, in turn, require large amounts of training data to be effective and it becomes imperative to reduce the degree of human supervision required to a minimum. In this project, parameterized probabilistic models will be used to explicitly model different object attributes and these parameters will be estimated by maximizing the likelihood of training data. The training data needs to be labeled but does not require segmentation or any other kind of manual processing. This project is inspired by the work of Fergus et al. (2003).

1. Introduction

This project will deal with recognizing classes of objects based on weakly supervised learning. Training data will be used to build representative models for the classes and a Bayesian approach will be employed to categorize a new, previously unseen image as belonging to one of the learned classes. The project will explore the approach proposed in [3]. The model proposed seems well suited to capture the large intra-class variabilities that arise during the task of generic object recognition. The model used is a "constellation of parts" model that relies upon an entropy-based feature detector to detect regions of interest within an image. These regions of interest are used to represent the images both during learning and recognition. The results presented in the paper suggest that the model used was quite successful for object categorization.

2. Questions

The method is heavily reliant on the performance of the feature detector. The authors used the detector of Kadir and Brady [5] because it was found to be stable across different scales and the number of features detected was easily controllable. However, there have been new feature detectors developed since the time at which this work was done. Therefore, one question we may ask is:

- Can we improve the performance of the algorithm by using a different feature detector? One possibility could be using a detector based on Maximally Stable Extremal Regions (MSER).

One serious drawback of the method seems to be that it entails fairly high computational complexity and does not scale very well with the number of parts detected. The problem is that the set of hypotheses (for the valid assignment of detected parts to the object model) is very large and this entire hypothesis space must be explored in order to compute the likelihood function during both learning and recognition. The problem of slow recognition is more serious than that of learning (since learning typically happens off-line). If we can reduce the size of the hypothesis set, we can speed up the algorithm significantly. To address this, we can examine the following question:

- How much will some degree of manual supervision (during recognition) help in reducing the size of the hypothesis set that needs to be explored, e.g. if the user indicates which regions belong to the object or the test images are presented with a bounding box around the object of interest?

In addition, we can examine the assumptions of the model. For example, it is assumed that the appearances of the various parts are independent of each other and that the appearance of each part is independent of the shape distribution. This would suggest, for example, that for objects of fixed shapes that have a narrow shape distribution (e.g. faces, bicycles) any set of random features arranged in the shape of that object might yield a false positive. Therefore, we can explore the following question:

- Can we come up with pathological test images that would expose some of the assumptions of the models and result in classification errors? This might help in identifying the most common points of failure and which assumptions are causing errors.

Finally, once the implemented algorithm is working satisfactorily and giving good results, we can look at extending it by trying to cluster data in a fully unsupervised environment. The feasibility of this depends on how much time the code takes to run as well.

These are a tentative set of questions that might be useful to explore during the project.

3. Milestones

A tentative approach for this project can be expressed in terms of the following milestones:

- Complete surveying related papers. These include [3], [4], [2] and [1]. (by Friday, Jan. 12)

- Decide upon and obtain relevant datasets. (by Friday, Jan. 12)

- Develop or request for Matlab code from the author (*http://cs.nyu.edu/∼fergus/research/cm/constellation_model.html*). as appropriate. Requested code would be used as a reference to help in demystifying specific implementation details. (by Friday, Feb. 9)

- Implement any additional code or pre-processing needed for examining the questions posed earlier. (by Friday, Feb. 23)

- Execute and evaluate the results on the chosen datasets. Analyze the results in the light of the questions posed. (by Friday, Feb. 30)

- Prepare report draft (by Mar. 3) based on the results and final version of the report (by Mar. 17).

4. Datasets

Since the learning proceeds with minimal supervision, we just need large numbers of images of any category. The images are used after grayscaling so colour information is discarded in this approach. The datasets used in the original paper [3] such as the Caltech Cars (Rear), Caltech Motorbikes, etc. will be used for validating and comparing results. These can be obtained from *http://www.robots.ox.ac.uk/∼vgg/data*. In addition, new datasets may be used such as the Cal-IT2 underwater video data *http://web.mac.com/atulnayak/iWeb/VisionProject/Photos.html* to measure how the method works in a totally different setting such as in water.

The training and test images could be obtained by splitting the datasets in some proportion depending on the number of images available.

5. Experience

Relevant courses that I have taken at UCSD include CSE 252A (Computer Vision I), CSE 252B (Computer Vision II) and MATH 102 (Applied Linear Algebra). For the current quarter, I am enrolled in ECE 271A (Statistical Learning I), CSE 250A (Artificial Intelligence I) and CSE 252C (Object Recognition). Apart from class projects for my graduate courses, I worked on a project during my undergraduate days that dealt with extraction of skeletal structure (or medial axis) from 2D object shapes by defining a potential field within the object and then following ridge lines of the resulting vector field.

References

[1] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Cognitive Vision Systems, Editor: Nagel, H.H., Kluwer Academic Publishers, in press.* 2

[2] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision (in print).* 2

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Proc. of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR)*, 2003. 1, 2

[4] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. *Proc. of the IEEE Conf on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[5] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision. 45 (2):83-105.* 1