Object Pose Dataset using Discriminatively Trained Deformable Part Models

Jinho Kim¹, Yu Xiang², and Silvio Savarese³

¹Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea
²Department of Electrical and Computer Engineering, University of Michigan Ann Arbor, MI, USA
³Computer Science Department, Stanford University Stanford, CA, USA

Abstract - Over the last couple of years computer vision has grown. While the old problem used to be object detection, we are now faced with the challenge of correctly estimating the pose of the objects. Thus in order to test algorithms for pose estimation it is important to use good datasets for training data. However the object datasets we have today are mainly for object detection. Therefore we do not have many sufficient datasets suitable for testing pose estimation algorithms. In this paper we use deformable part models and latent SVM to propose a dataset that we hope can become a good dataset for testing pose estimation algorithms.

Keywords: Deformable Part Models, latent SVM, Object Pose Estimation, ImageNet, PASCAL VOC

1 Introduction

The human ability to correctly classifying objects no matter what the pose of the object is something that cannot be replicated easily. As computer vision becomes more important, our interests are focused on how to correctly estimate the pose of objects. Thus in order to test algorithms of pose estimation it is important to use good datasets for training data. Savarese and Fei-Fei created a 3D object dataset that contained 10 categories and between 480 to 720 images per category [1]. The EPFL Car dataset has around 2300 images for 20 car instances [2]. While they continue azimuth, there is no variation to elevation or distances. Later PASCAL VOC datasets were created [3]. These datasets were intended for object detection not for pose estimation and thus had only four discrete viewpoints available, ‘front’, ‘back’, ‘left’, and ‘right’. Our goal for this project was to create a large scale dataset for pose estimation that has a) thousands of images b) contains many categories of both indoor and outdoor objects and finally c) contains a large variation of viewpoints in terms of azimuth, elevation, and distance. In order to complete this dataset, we first start by amassing a large number of images of object categories from various image databases. We combined the images from the ImageNet database and PASCAL VOC images. However these databases were not meant for object pose estimation thus we created an annotation tool based on 3D CAD models to compute the viewpoint corresponding to 2D and 3D objects. After collecting the images we made deformable part models which we used for training. In section 2.1 we will explain how we prepared our images for the dataset. In sections 2.2 and 2.3 we will briefly explain the concept of deformable part models and latent SVM respectively. Section 3 will deal with the results.

2 Methodology

2.1 Images

As stated before by just using the PASCAL VOC datasets presents us with some faults. Many of its images apart from the invariance in viewpoint present occlusion and truncation. This presents problems with trying to train the data as we want to training images to be as clean as possible. Thus when selecting images from ImageNet we attempted to select clean images with many more different viewpoints for training. We used an anchor annotator, created by Yu Xiang, to define anchor points on 3D CAD models. Annotators click on the anchor points in 2D images. After the annotator clicks on the anchor points the computer computes the viewpoint using 2D – 3D correspondence. Figure 1 shows the anchor annotator.

2.2 Deformable Part Models (DPM)

Once we have the images gathered we used models that are trained discriminatively so that they only require bounding boxes for the objects in the image. This leads to more efficient object detections. The training process returns a model that is a mixture of star models produced in the process. The testing process of DPM uses feature pyramids to detect features of the object contained within the image. From these results we test various threshold values to find which value gives the highest recall and precision. Here precision is the fraction of the reported bounding boxes that are correct detections while recall is the fraction of the objects found.
The main approach is built on pictorial structures framework [4] [5]. Pictorial structures represent objects by a collection of parts arranged in a deformable configuration. Each part captures local appearance properties of an object while the deformable configuration is characterized by spring-like connections between certain pairs of parts.

The goal for DPM is to model objects using “visual grammars”. Grammar based models [6] [7] [8] generalize deformable part models by representing objects using variable hierarchical structures. Each part in a grammar based model can be defined directly or in terms of other parts. Also grammar based models allow for, and explicitly model, structural variations. These models provide a natural framework for sharing information and computation between different object classes.

The Dalal-Triggs detector [9] used a single filter on histogram of oriented gradients (HOG) features to represent an object category. The detector determines whether or not there is an instance of the target category at the given position and scale. The first innovation DPM uses involves enriching the Dalal-Triggs model using a star-structured part-based model defined by a “root” filter plus a set of parts filters and associated deformation models [10]. Figure 2 and Figure 3 each show a star model for the bicycle and person category, respectively, from the original PASCAL dataset.

2.2 Latent SVM

To train models using partially labeled data we use a latent variable formulation of MI-SVM [11] that is called latent SVM (LSVM). In a latent SVM each example \( x \) is scored by a function of the following form,\n\[
f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z) \tag{1}\n\]

Fig. 1. The Anchor Annotator

Fig. 2. Detections obtained with a 2 component bicycle model. These examples illustrate the importance of deformations mixture models. In this model the first component captures sideways views of bicycles while the second component captures frontal and near frontal views. The sideways component can deform to match a wheel. It should be noted that this is not from our dataset but from [10].

Fig. 3. Detections obtained with a single component person model. It should be noted that this is not from our dataset but from [10].

Here \( \beta \) is a vector of model parameters, \( z \) are latent values, and the function of \((x, z)\) is a feature vector. In the case of one of our star models, \( \beta \) is the concatenation of the root filter, the part filters, and deformation cost weights, \( z \) is a specification of the object configuration, and \((x, z)\) is a concatenation of subwindows from a feature pyramid and part deformation features. It should be noted that equation (1) can handle very
general forms of latent information. To obtain high performance using discriminative training it is often important to use large training sets.

3 Results

We evaluated the performance of deformable part models on the dataset we constructed. It should be noted that this is still a trial run and we are still modifying our dataset to obtain more optimal results. Each dataset contains thousands of images of real world scenes. The datasets specify ground-truth bounding boxes for several object classes. At test time, the goal is to predict the bounding boxes of all objects of a given class in an image. Theoretically, a system will output a set of bounding boxes with corresponding scores, and we can threshold these scores at different points to obtain a precision-recall curve across all images in the test set. For a particular threshold the precision is the fraction of the reported bounding boxes that are correct detections, while recall is the fraction of the objects found.

A predicted bounding box is considered correct if it overlaps more than 50% with a ground-truth bounding box, otherwise the bounding box is considered a false positive detection. Multiple detections are penalized. If a system predicts several bounding boxes that overlap with a single ground-truth bounding box, only one prediction is considered correct, the others are considered false positives. One scores a system by the average precision of its precision-recall curve across a testset. Figure 4 shows a sample precision-recall curve for the category of cars.

In some categories our false detections are often due to confusion among classes, such as between car and bus. In other categories false detections are often due to the relatively strict bounding box criteria. The two false positives shown for the person category are due to insufficient overlap with the ground truth bounding box.

4 Conclusions

Deformable part models have been proven on various datasets. [10] Therefore we must retest it after finalizing our dataset. As of now we have determined that our dataset needs further evaluation and improvements. We realize that it is not sufficient for us to use this dataset for experimentation on object pose detection. Therefore we must add more images and discard unfit images which we are currently doing at the moment. It is also a good idea to add more categories or to divide the pre-existing categories into many separate subcategories. As of now we are exploring the idea of dividing the categories into separate subcategories. While it is important to have a big dataset for accurate testing, it is also important to select the best images. Since the deformable part model was to test our dataset we can rely on the results to show where we must improve our dataset.

References


