Enriching Object Detection with 2D-3D Registration and Continuous Viewpoint Estimation

Christopher Bongsoo Choy[†], Michael Stark[†‡], Sam Corbett-Davies[†], Silvio Savarese[†]
[†]Stanford University, [‡]Max Planck Institute for Informatics

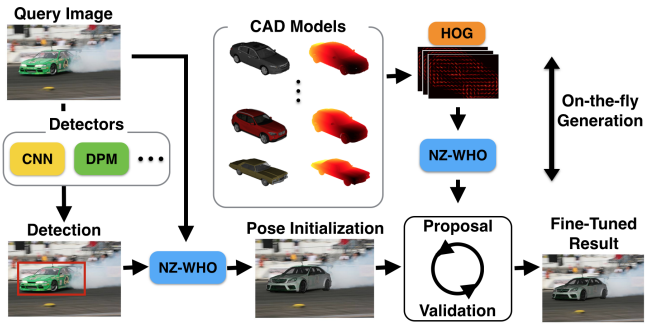

Figure 1: Using a database of 3D CAD models, we generate NZ-WHO templates which can be used to either detect objects directly or enrich the output of an existing detector with high-quality, continuous pose and 3D CAD model exemplar.

Figure 2: Example enriched bounding boxes. Given R-CNN detection bounding boxes, our method predicts the best 2D-3D registration given the bounding box input. Blue boxes are R-CNN output and purple boxes are the tightest bounding box enclosing predicted CAD model.

A large body of recent work on object detection has focused on providing more information than the bounding box of an object. One set of methods use 3D CAD model databases to provide viewpoint estimation. These approaches work by aligning exact 3D models to images using templates generated from renderings of the 3D models at a set of discrete viewpoints [1].

However, training templates for all viewpoints and CAD models is not feasible since it requires a huge amount of training data to be acquired. Instead, it has been realized that template-based exemplar detectors based on HOG features can be trained analytically, by replacing the standard SVM with an LDA classifier w_{x_t} for a template x_t [2].

The result is a whitened feature representation, termed WHO (Whitened Histogram of Orientations). This development makes it feasible to train a large number of mid-level patch detectors for recognition.

Though generating and calibrating LDA templates (via matrix decomposition with Gaussian Elimination) achieved remarkable improvements, they are computationally expensive and use a prohibitive amount of memory and storage. In addition, viewpoint discretization hampers pose estimation performance.

Overview. In this paper, we propose a novel method for 2D-3D alignment of exemplar CAD models to real-world images that circumvents the need for calibration and greatly enhances the scalability of WHO. As a result, we can render novel views and train corresponding exemplar models *on-the-fly*, without the need for offline processing. We call these Non-Zero Whitened Histograms of Orientations (NZ-WHO) templates. To our knowledge, our method constitutes the first attempt to simultaneously render and train exemplar detectors on-the-fly.

First, we adapt the whitening to the specific case of rendered images. We show how to speed up the whitening by two orders of magnitude for high-resolution templates using Conjugate Gradient method. Also, we improve the evaluation of our 3D exemplar template detectors at test time by performing convolutions in the frequency domain.

These components make template generation and evaluation computationally inexpensive. Thus, we are able to efficiently explore the continuous parameter space to find the best object pose, scale, 3D CAD model type and camera focal length. An overview of our pipeline can be seen in Fig. 1.

Finally, our experimental study demonstrates the effectiveness of the approach on several standard benchmarks for object detection and viewpoint estimation. We also demonstrate that our method can enrich the output of an existing object class detector, such as DPM or R-CNN, with additional 3D information.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

Non-Zero Whitened Histograms of Orientations. Many works that use rendering do not take the background into account when synthesizing and whitening to generate an LDA template [1]. Our NZ-WHO removes the background region while whitening thus preventing irrelevant regions from corrupting the foreground.

In addition, our NZ-WHO method uses Conjugate Gradient to greatly speeds up the LDA generation. We show that our uncalibrated NZ-WHO template performs on par with a calibrated WHO template while being 100x faster to generate.

Fine-Tuning using MCMC. Due to the speed of our template synthesis procedure, we are able to perform joint optimization of scale, translation, continuous rotation, and focal length using the Metropolis-Hastings algorithm.

We model the probability of an object with the parameter θ in the test image $\mathcal{I}$ as $P(\theta|\mathcal{I}) \sim e^{\max_s w(\theta) * \mathcal{T}_s(\mathcal{I})}$. We approximate the MAP solution for θ by drawing samples from the distribution $P(\theta|\mathcal{I})$, using the Metropolis-Hastings algorithm.

Experiments First, we verify that our NZ-WHO method delivers performance that is at least on par with the original WHO formulation [2] in terms of accuracy, while at the same time resulting in large computational savings.

Second, we demonstrate that our method can be used for multi-view object class detection in isolation. It can be applied in a sliding window fashion to deliver 2D bounding boxes and viewpoint information. Our method is competitive with the state-of-the-art in this case.

Finally, we show that our method can be used to complement the detections provided by an existing object class detector, such as DPM or R-CNN. In this case, we show a considerable performance improvement compared to previous work in the task of joint object class detection and viewpoint estimation. Example outputs of our model are in Fig. 2.

Acknowledgement. We acknowledge the support of NSF CAREER grant (N1054127), Ford-Stanford Innovation Alliance Award, DARPA, Korea Foundation for Advanced Studies, Fulbright New Zealand and the Max Planck Center for Visual Computing & Communication.

[1] Mathieu Aubry, Daniel Maturana, Alexei Efros, Bryan Russell, and Josef Sivic. Seeing 3d chairs: exemplar part-based 2d-3d alignment using a large dataset of cad models. In *CVPR*, 2014.

[2] Bharath Hariharan, Jitendra Malik, and Deva Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012.