Analysis by Synthesis: 3D Object Recognition by Object Reconstruction

Mohsen Hejrati
University of California, Irvine
shejrati@ics.uci.edu

Deva Ramanan
University of California, Irvine
dramanan@ics.uci.edu

Abstract

We introduce a new approach for recognizing and reconstructing 3D objects in images. Our approach is based on an analysis by synthesis strategy. A forward synthesis model constructs possible geometric interpretations of the world, and then selects the interpretation that best agrees with the measured visual evidence. The forward model synthesizes visual templates defined on invariant (HOG) features. These visual templates are discriminatively trained to be accurate for inverse estimation. We introduce an efficient “brute-force” approach to inference that searches through a large number of candidate reconstructions, returning the optimal one. One benefit of such an approach is that recognition is inherently (re)constructive. We show state of the art performance for detection and reconstruction on two challenging 3D object recognition datasets of cars and cuboids.

1. Problem and Proposed Method

We focus on the task of recognizing and reconstructing objects in images. Specifically, we describe a single model that simultaneously detects instances of general object categories, and reports a detailed 3D reconstruction of each instance. Our approach is based on an analysis by synthesis strategy. A forward synthesis model constructs possible geometric interpretations of the world, and then selects the interpretation that best agrees with the measured visual evidence. One benefit of such an approach is that recognition is inherently (re)constructive.

Challenges: Though attractive, an “inverse rendering” approach to computer vision is wildly challenging for two primary reasons. (1) It is difficult to build accurate generative models that capture the full complexity of the visual world. (2) Even given such a model, inverting it is difficult because the problem is fundamentally ill-posed (different reconstructions may generate similar images) and full of local minima (implying local search will fail).

Our approach: Our approach addresses both difficulties. (1) Instead of generating pixel values, we use forward models to synthesize visual templates defined on invariant (HOG) features. These visual templates are discriminatively trained to be accurate for inverse estimation. (2) We describe a “brute-force” approach to inference that efficiently searches through a large number of candidate reconstructions, returning the optimal one (or multiple likely candidates, if desired).

2. Results


Synthesis strategies: We explored numerous strategies for constructing a set of 3D shapes. First, our Exemplar model uses the shapes encountered in the training set of annotated images, augmented with synthetic camera transla-
Our shape model should capture nonrigid shape variation within a shape parameter space. We make use of morphable models to model any 3D shape instance as a linear combination of 3D basis shapes. To do so, we make use of morphable shape model that can be learned from a large set of candidate 3D reconstructions. Our 3D shape model approaches are similar to work that relies on a non-parametric method, such as Hough transforms. However, our approach differs from past work in that we use a geometric model to synthesize 3D shapes from a particular object category (such as cars). We will then sample from this parametric family to define a large set of candidate 3D reconstructions. Our 3D shape model uses efficient correspondence search approaches (since many views may be needed). Exemplar and parametric synthesis have been used for many years. Recent work has brought a larger set of keypoint annotations.

Figure 1. We describe a method for synthesizing a large set of discriminative templates, each associated with a candidate 3D reconstruction of an object (in this case, cars). Our model makes use of a generative 3D shape model to synthesize a large collection of 2D landmarks, which in turn specify rules for composing 2D templates out of a common pool of parts.

Figure 2. We describe a method for synthesizing a large set of discriminative templates, each associated with a candidate 3D reconstruction of an object (in this case, cars). Our model makes use of a generative 3D shape model to synthesize a large collection of 2D landmarks, which in turn specify rules for composing 2D templates out of a common pool of parts.

Figure 3. Detection (left) and reconstruction accuracy (right) versus running time of our method and other baselines, including DPMs [1], supervised-tree models [2], and multi-view star models [2]. Points correspond to different (constant-time) baselines, while curves correspond to our models. Because our models can process a variable number of synthesized templates, we sweep over $K \in \{20, 50, 100, 500, 1000, 4000\}$ templates to generate the curves. Our box detection and reconstruction results (43% and 48%) nearly double the best previously-reported performance from Xiao et al [4] (24% and 38%), while being 10X faster.

**Exemplar Synthesis:**

Exemplar Synthesis augments this set with additional exemplar shapes. We implement this strategy by learning a model with a subset of training images, but using the larger (full) set of keypoint annotations. This mimics scenarios where we have access to a limited amount of image data, but a larger set of keypoint annotations. Parametric Synthesis constructs a shape set by discretely enumerating $\theta_0$ over bounded parameter ranges. Finally, Oracle Synthesis uses shapes extracted from annotated test-data. We use this upper bound on performance (given the “perfect” synthesis strategy) for additional analysis.

**Benchmark results:** Fig. 3 plots performance for box detection and localization. Exemplars almost double the best previously-reported numbers in [4], in terms of detection (43% vs 24%) and landmark reconstruction (48% vs 38%). Interestingly, the tree model of [2] outperforms [4], perhaps due to its modeling of local part mixtures. Our models even surpass [2], while directly reporting 3D reconstructions and while being 10X faster. Exemplar and Parametric Synthesis perform similarly for low numbers of templates, but Exemplars do better with more templates, particularly with respect to reconstruction accuracy. These results suggest that our parametric model is not capturing true shape statistics. For example, people may take pictures of certain objects from iconic viewpoints. Such dependencies are not modeled by Parametric Synthesis, but are captured by Exemplars and Exemplar Synthesis. We find similar results for car detection and reconstruction, but refer the reader to our full paper [3] for additional discussion.

**Anytime recognition/reconstruction:** Our models have a free parameter $K$, the number of enumerated shapes. Both performance and run-time computation increase with $K$. When comparing to baselines with fixed run-time costs, we plot performance as a function of run-time, measured in terms of seconds per image. All methods are run on the same physical system (a 12-core Intel 3.5 Ghz processor). Our plots reveal that a simple re-ordering of shapes in a coarse-to-fine fashion (with hierarchical clustering) can be used for any-time analysis. For example, after enumerating the first $K = 50$ coarse shapes, one can still obtain 37% box landmark reconstruction accuracy (which in turns improves as more shapes are enumerated).

**References**