3D Object Detection with a Deformable 3D Cuboid Model

Sanja Fidler
TTI Chicago
fidler@ttic.edu

Sven Dickinson
University of Toronto
sven@cs.toronto.edu

Raquel Urtasun
TTI Chicago
rurtasun@ttic.edu

Abstract

This paper addresses the problem of category-level 3D object detection. Given a monocular image, our aim is to localize the objects in 3D by enclosing them with tight oriented 3D bounding boxes. We propose a novel approach that extends the deformable part-based model [1] to reason in 3D. Our model represents an object class as a deformable 3D cuboid composed of faces and parts, which are both allowed to deform with respect to their anchors on the 3D box. We model the appearance of each face in fronto-parallel coordinates, thus effectively factoring out the appearance variation induced by viewpoint. We train the cuboid model jointly and discriminatively. In inference we slide and rotate the box in 3D to score the object hypotheses. We evaluate our approach in indoor and outdoor scenarios, and show that our approach outperforms the state-of-the-art in both 2D [1] and 3D object detection [4].

1. Introduction

Estimating semantic 3D information from monocular images is an important task in applications such as autonomous driving and personal robotics [7, 6]. Let's consider for example, the case of an autonomous agent driving around a city. In order to properly react to dynamic situations, such an agent needs to reason about which objects are present in the scene, as well as their 3D location, orientation and 3D extent. Likewise, a home robot requires accurate 3D information in order to navigate in cluttered environments as well as grasp and manipulate objects.

In this paper we extend DPM to reason in 3D. Our model represents an object class with a deformable 3D cuboid composed of faces and parts, which are both allowed to deform with respect to their anchors on the 3D box (Fig 1). We introduce a *stitching point*, which enables the deformation between the faces and the cuboid to be encoded efficiently. We model the appearance of each face in fronto-parallel coordinates, thus effectively factoring out the appearance variation due to viewpoint. We train the cuboid model jointly and discriminatively. In inference, our model outputs 2D along with oriented 3D bounding boxes around the objects. This enables the estimation of object's viewpoint which is a continuous variable in our representation. We evaluate our approach in indoor [4] and outdoor scenarios [3], and show that our approach significantly outperforms the state-of-the-art in both 2D [1] and 3D object detection [4]. The details of our method are in [2].

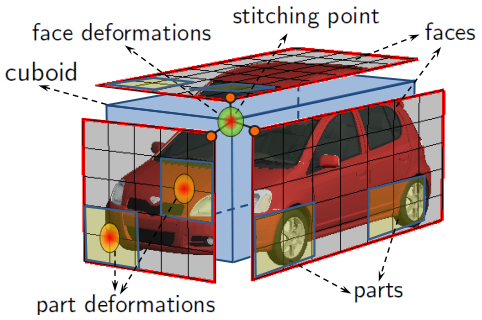

Figure 1. Our deformable 3D cuboid model.

2. A Deformable 3D Cuboid Model

Given a single image, we aim to estimate the 3D location and orientation of the objects present in the scene. We represent an object class as a deformable 3D cuboid, which is composed of 6 deformable faces, i.e., their locations and scales can deviate from their anchors on the cuboid. The model for each cuboid's face is a 2D template that represents the appearance of the object in view-rectified coordinates, i.e., where the face is frontal. Additionally, we augment each face with parts, and employ a deformation model between the locations of the parts and the anchor points on the face they belong to. We assume that any viewpoint of an object in the image domain can be modeled by rotating our cuboid in 3D, followed by perspective projection onto the image plane. Thus inference involves sliding and rotating the deformable cuboid in 3D and scoring the hypotheses.

For any viewpoint θ, at most 3 faces are visible in an image. Topologically different visibility patterns define different *aspects* [5]. Our model reasons about the occurring aspects of the object class of interest, which we estimate from training data. Fig. 2 shows estimated aspects for beds.

In order to make the cuboid deformable, we introduce a *stitching point*, which is a point on the box that is common to all visible faces for a particular aspect. We incorporate a quadratic deformation cost between the locations of the faces and the stitching point to encourage the cuboid to be as rigid as possible. We impose an additional deforma-

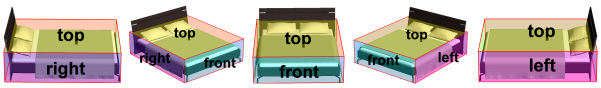

Figure 2. Aspects (computed from train. data) for beds used in our model.

tion cost between the visible faces, ensuring that their sizes match when we stitch them into a cuboid hypothesis. The appearance templates and the deformation parameters in the model are defined for each face in a canonical view where that face is frontal. We thus score a face hypothesis in the fronto-parallel coordinates.

Let p_i be a random variable encoding the location and scale of a box's face in a rectified HOG pyramid, and $\{p_{i,j}\}_{j=1,\cdots,n}$ be a set of its parts. We define the compatibility score between the parts and the face as in a DPM:

$$\text{score}_{\text{parts}}(\mathbf{p}_i, \theta) = \sum_{j=1}^{n} \left(w_{ij}^T \cdot \phi(p_{i,j}) + w_{ij,def}^T \cdot \phi_d(p_i, p_{i,j}) \right)$$

We define the score of a cuboid hypothesis to be the sum of scores of each face and its parts, and the deformation of each face with respect to the stitching point and the deformation of the faces with respect to each other as follows

$$\text{score}(x, \theta, \mathbf{s}, \mathbf{p}) =$$

$$\sum_{i=1\ldots6} V(i,a) \left(w_i^T \cdot \phi(p_i, \theta) + w_{a,i}^{stitch} \cdot \phi_d^{stich}(p_i, \mathbf{s}, \theta) \right) +$$

$$\sum_{i>ref} V(i,a) \cdot w_{i,ref}^{face} \cdot \phi_d^{face}(p_i, p_{ref}, \theta) +$$

$$\sum_{i=1\ldots6} V(i,a) \cdot \text{score}_{\text{parts}}(\mathbf{p}_i, \theta) \; + b_a$$

where $V(i,a)$ is a binary variable encoding whether face i is visible under aspect a. We use ref to index the first visible face in the aspect model, and $\phi_d(p_i, p_{i,j}, \theta)$ are the quadratic part deformation features, computed in the rectified image of face i implied by the 3D angle θ. Here, $\phi_d^{stich}(p_i, \mathbf{s}, \theta)$ are the quadratic deformation features between the face p_i and the stitching point $\mathbf{s}$. The deformation cost $\phi_d^{face}(p_i, p_k, \theta)$ between the faces is a function of their relative dimensions, enforcing the common edge between the two faces to be of similar length.

Inference: We compute $\max_{\theta,\mathbf{s},\mathbf{p}} \mathbf{w}_a \cdot \Phi_a(x, a, \mathbf{p})$, which can be solved exactly via dynamic programming. We first compute the score for each face in its rectified view as in DPM [1]. We then use distance transforms to compute the deformation scores for each face and the stitching point, and the deformation scores between the faces and the reference. Finally, we reproject the scores to the image coordinate system and sum them to get the final score.

Learning: We assume that we have 3D box annotations available in training. To train our model (weights w) we use a latent SVM formulation [1].

3. Experiments

We evaluate our approach on the bed dataset of [4]. We first evaluate our model in 2D detection. We compute the

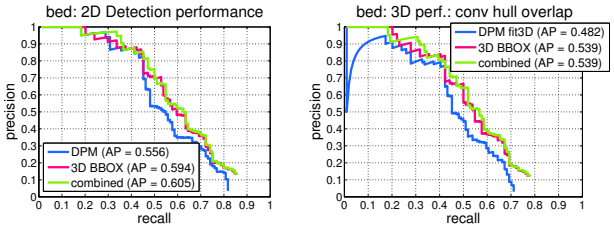

Figure 3. Precision-recall curves for (**left**) 2D overlap (**right**) convex hull.


Figure 4. Detection examples of beds obtained with our model.


Figure 5. KITTI: examples of car detections. (**top**) GT, (**bottom**) Our 3D detections augmented with best fitting CAD models depicting inferred 3D orientations.

2D bounding boxes by fitting a 2D box around the convex hull of the projection of the predicted 3D box. We report average precision (AP) using the 50% IOU criteria. The (rigid) cuboid model of Hedau et al. [4] achieves 51.3%, the DPM [1] gets 55.6%, while our deformable cuboid achieves 59.4%. This is notable, as to the best of our knowledge, this is the first time that a 3D approach outperforms the DPM. Examples of detections are shown in Fig. 4. To evaluate 3D performance, we use the convex hull overlap measure as in [4]. The precision-recall curves are shown in Fig. 3.

We also conducted preliminary tests on KITTI [3]. Examples of detections are shown in Fig.5. To show predicted viewpoint we insert a CAD model inside each inferred 3D box. One can see that our 3D detector is able to predict the viewpoints of the objects well, as well as the type of car.

References

[1] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 1, 2

[2] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, 2012. 1

[3] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? In *CVPR*, 2012. 1, 2

[4] V. Hedau, D. Hoiem, and D. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, volume 6, pages 224–237, 2010. 1, 2

[5] J. Koenderink and A. van Doorn. The singularities of the visual mappings. *Bio. Cyber.*, 24(1):51–59, 1976. 1

[6] J. Liebelt, C. Schmid, and K. Schertler. Viewpoint-independent object class detection using 3d feature maps. In *CVPR*, 2008. 1

[7] M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3d object classes. In *CVPR*, 2009. 1