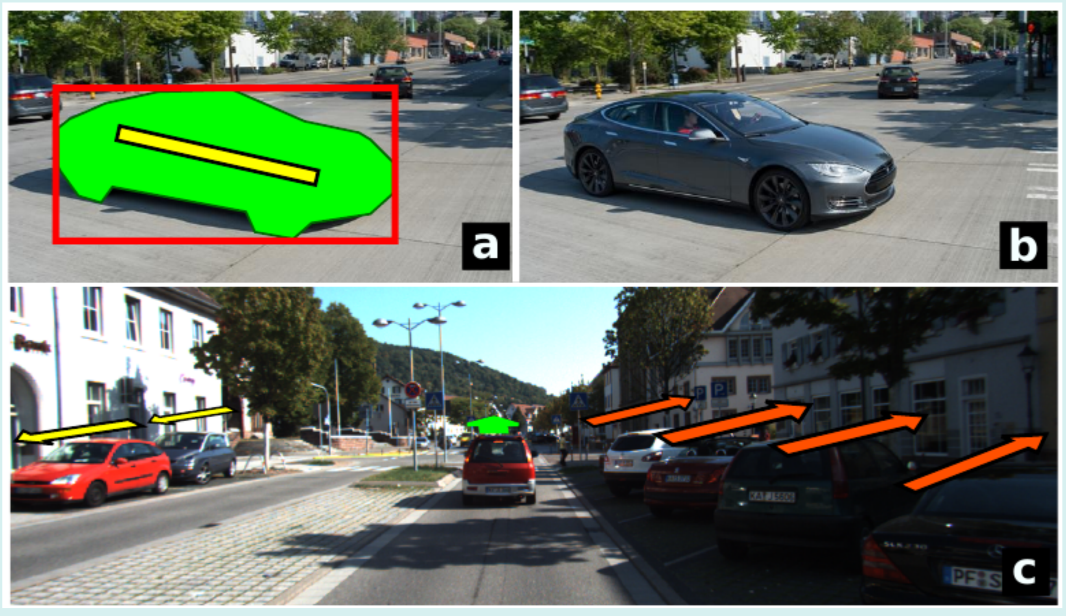
Scene-driven Cues for Object Viewpoint Estimation

José Oramas M. , Tinne Tuytelaars

Abstract

- Generate scene consistent object proposals to validate the appearance-based hypotheses predicted by an object detector.
- Use object elongation orientation classification as a intermediate step prior to object viewpoint estimation.



Motivation

- Object viewpoint/pose estimation has been traditionally addressed from a very local perspective based on object-driven features.
- There are certain regions in the scene that are more likely to host object with specific features such as class, size, viewpoint, etc.
- The orientation of the elongation of the object gives a strong cue about its viewpoint.

Algorithm

- **a)** Object detection.
- **b)** Scene-driven object proposal generation.
- **c)** Object hypotheses - proposals matching.
- **d)** Object elongation orientation classification.
- **e)** Object viewpoint classification.

a) Object Detection

Detection:
Using the detectors from [1,2,3], we collect a set of object hypotheses:

$$o = \{o_1, o_2, \ldots, o_n\}$$

where

$$o_i = (s_i, b_i, \alpha_i)$$
$$b_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i})$$

s_i : detection score.
α_i : object viewpoint.
b_i : object bounding box.

b) Scene-driven object proposal generation

Generate a set of object proposals:

$$o' = \{o'_1, o'_2, \ldots, o'_n\} \quad \text{where} \quad o' = \Omega(scene)$$

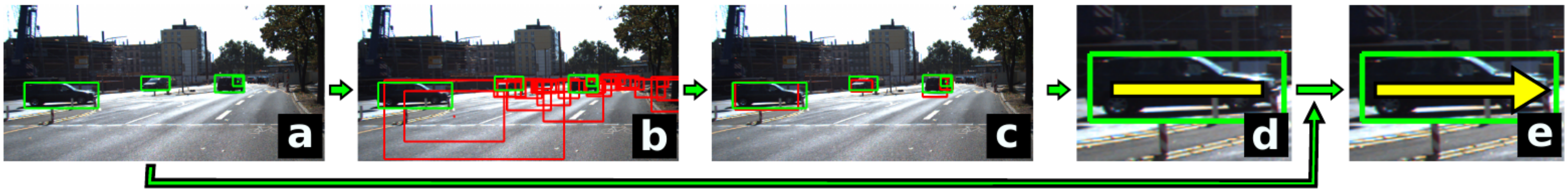
Scene-consistency is enforced in 4 ways:

1) Ground-plane
- There is a ground-plane that supports the objects of interest. (Inspired by [4]).
- Generate 3D proposals that can be physically in the scene.

2) History of 3D Objects
- Start from a set of ground-truth 3D objects.
- Uniformely sample the distribution $p(X, Z, \theta)$ defined by their location and pose.
- Project 3D proposals to the image space.

3) History of 2D Objects
- Start from a set of ground-truth 2D objects.
- Uniformely sample the distribution $p(x, y, w, h, \alpha)$ defined by its 2D location, width, height and viewpoint.

4) History of 2D Hypotheses
- Similar to previous case.
- Start from a set of hypotheses collected with an object detector.



c) Object hypotheses - proposals matching

For each hypothesis o_i we find its closest proposal o'_i using the intersection over union criterion from PASCAL VOC .

d) Object elongation orientation classification

For each pair (o_i, o'_i) , compute the descriptor:

$$d_i = (rw, rh, rx, ry, \alpha') \quad \text{where} \quad rw = |\tfrac{w'}{w}| \quad rh = |\tfrac{h'}{h}| \quad rx = |\tfrac{x'-x}{w}| \quad ry = |\tfrac{y'-y}{h}|$$

We focus on a smaller set of K/2 viewpoints
-> Elongation orientation ε_i of an object

$$\hat{\varepsilon}_i = \arg_{\varepsilon_k} \max p(\varepsilon_k | d_i) = \arg_{\varepsilon_k} \max p(d_i | \varepsilon_k) p(\varepsilon_k)$$

e) Object viewpoint classification

The viewpoint of an object is estimated as the late fusion of:

Object detector response: (α_i, s_i)
Elongation classifier response: $(\varepsilon_i, \lambda_i)$

coupled response

$$r_i = [\alpha_i, s_i, \varepsilon_i, \lambda_i]$$

the viewpoint of the object is classified as:

$$\hat{\alpha}_i = \arg_{\alpha_k} \max p(r_i | \alpha_k) p(\alpha_k)$$

Experiments

Settings

Dataset :
- KITTI object detection benchmark [5].

Local Classifiers
- Viewpoint-aware DPM detector from [1,2] (8 viewpoints) .
- DPM detector from [3] (No viewpoint information) .

Elongation Orientation Classification

	Geiger et al. [1]		Lopez et al. [2]		Felzenswalb et al. [3]	
Method	**Easy Set**	**Full Set**	**Easy Set**	**Full Set**	**Easy Set**	**Full Set**
Local detector	**0.69**	**0.68**	0.42	0.46	—	—
GroundPlane	0.50	0.52	**0.69**	**0.57**	**0.72**	**0.64**
Hist3DObjects	0.39	0.41	0.57	0.53	0.54	0.53
Hist2DObjects	0.40	0.39	0.46	0.43	0.53	0.51
Hist2DHypotheses	0.41	0.42	0.61	0.49	0.34	0.34

Viewpoint Classification

	Geiger et al. [1]		Lopez et al. [2]	
Method	**Easy Set**	**Full Set**	**Easy Set**	**Full Set**
Local detector	0.38	0.43	0.38	0.36
GroundPlane	0.44	**0.45**	0.50	0.42
Hist3DObjects	**0.46**	0.43	**0.55**	**0.48**
Hist2DObjects	0.42	0.39	0.43	0.38
Hist2DHypotheses	0.45	0.42	0.47	0.39

Conclusions

- Scene-driven object elongation orientations can assist purely appearance-based viewpoint classifiers.
- There are relatively simple cues in the scene that can bring improvement for object viewpoint estimation.
- Coarse 3D scene-level reasoning, apart from context, is beneficial.

References

[1] A. Geiger et al., NIPS 2011.
[2] R. Lopez et al., WS@ICCV 2011.
[3] P. Felzenszwalb et al., TPAMI 2010.
[4] D. Hoiem et al., CVPR 2006.
[5] A Geiger et al., CVPR 2012.