Scene-driven Cues for Viewpoint Classification of Elongated Object Classes

José Oramas M.
http://homes.esat.kuleuven.be/~joramas
Tinne Tuytelaars
http://homes.esat.kuleuven.be/~tuytelaa

KU Leuven, ESAT-PSI, iMinds
Leuven, Belgium

Abstract

In this paper we present a top-down, *scene-driven*, approach to classify the viewpoint of elongated objects. Viewpoint classification is achieved based on the correspondence between the image evidence acquired by a detector on one hand, and bounding boxes of object proposals generated from scene consistent cues on the other hand. We explore several methods to generate scene-driven proposals: first, by generating object proposals in the 3D scene and, using the ground plane, projecting them to the image space, and second, by sampling objects from a history set of objects previously seen with the same camera setup in the scene. Experiments on the challenging KITTI dataset show the performance of the proposed method for viewpoint classification for controlled-camera applications such as autonomous navigation.

1 Introduction

Object viewpoint classification, also referred to as object pose estimation, is a task of interest for several applications including autonomous navigation, traffic surveillance, image retrieval, etc. However, since the early days of computer vision, it has been addressed from a very "local" perspective. This perspective focuses on learning from the features on the object itself, e.g. color, texture, or gradients, to identify the different viewpoints in which an object may appear in an image [14, 16]. Lately, this trend has been extended from reasoning about local visual properties of the object in the image space to properties *in the 3D scene* [8, 10, 13, 15, 18, 19, 20].

In this work we explore the use of non-local, more scene-driven cues. In particular, we exploit a particular feature of elongated objects in that their physical extent provides a strong cue about their orientation. For example, consider the object in Fig. 1a. Even when we have no direct access to the local features of the object itself, we are able to predict, up to some level, the orientation of the underlying object (Fig. 1b). In this work we use the bounding boxes covering the objects, as a proxy to classify their viewpoint. Particularly, we are interested in how objects in specific orientations in the scene, project bounding boxes in the image space and use this as an intermediate step towards viewpoint classification. To this aim, we use the elongation orientation, direction of maximum physical length, of the object, as a cue to estimate its viewpoint. For the sake of brevity, in the rest of the paper we drop the term "orientation" and refer to elongation orientation purely as *elongation*. Moreover,

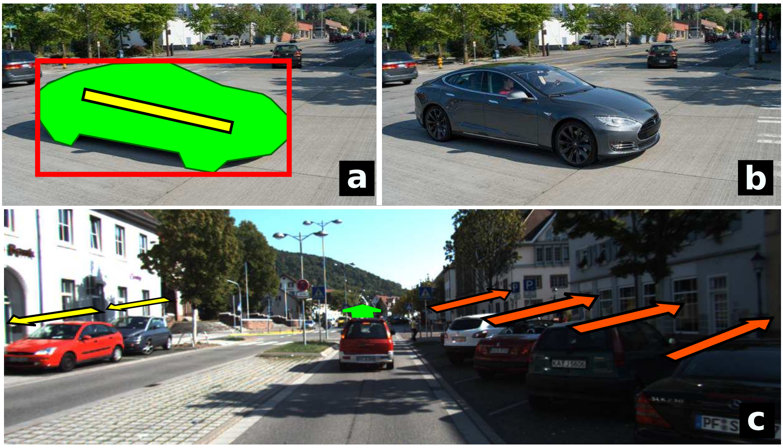

Figure 1: Note how the shape (a,b) and the location (c) of the bounding box of an object is related to its viewpoint.

in order to enforce scene-consistency in the viewpoint classification process, we define the scene not only as a space in which the objects of interest occur, but rather as a space with specific regions that are more likely to host certain objects with particular features such as class, orientation, or size. For example, note how the orientation of the objects in Fig. 1c is closely related to the regions of the scene in which they occur. Combining these two ideas, here we propose a Top-Down approach in which we first generate scene-driven object proposals in the image and select the closest ones to object hypotheses gathered with an object detector. Then, we define a correspondence descriptor between each hypothesis and its closest object proposals and perform classification to predict the elongation orientation of the object. Finally, the object viewpoint is determined by a late fusion of the elongation prediction and the initial prediction of the object detector. We explore four means to produce scene-driven object proposals, based on: a) the scene ground-plane, b) previously seen 3D annotated Objects in the scene, c) previously seen 2D annotated objects in the image, and d) previously seen 2D object hypotheses in the image, obtained by a detector.

The contributions of this work are: a) the introduction of an intermediate step, towards viewpoint classification, object elongation orientation classification, and b) a top-down approach that produces scene consistent results for viewpoint classification outperforming results that are obtained in a purely local fashion. This paper is organized as follows: Section 2 positions our work with respect to similar work. In Section 3 we present the details of our method. Section 4 introduces the evaluation protocol used, followed by experimental results and discussion (Section 5). In Section 6 we conclude this paper.

2 Related Work

There exists a significant amount of work addressing the problem of object viewpoint classification. For the sake of brevity, we position our work based on two, closely related, groups of work:

Local viewpoint classification : The problem of object viewpoint classification has been traditionally addressed under the assumption that the visual features, e.g. color, texture or gradients, projected by an object on an image differ between viewpoints. The problem is then to define proper descriptors to represent such visual features and to find a method to distinguish between the descriptors from each viewpoint. Following this trend methods have evolved, from reasoning about features of the object in the image space [14, 16] to reasoning about object parts in the 3D space [10, 13, 18]. Similar to the recent group of work, we explore how the scene and the features of the 3D objects in the scene shape the 2D object evidence we perceive in an image. We focus on the size feature of the objects of interest. Specifically, we use features derived from the bounding boxes that circumscribe the objects in the image. This removes the requirement of, more detailed, CAD models at the cost of producing a coarse object representation.

Scene-driven viewpoint classification: This group covers work that exploits the full scene to estimate the viewpoint of the objects of interest. Methods from this group enforce geometric consistency of the objects with the scene and/or consistency between objects in the scene. Aiming to enforce object-scene consistency, [5] proposed a deformable 3D cuboid model composed of faces and parts that can deform with respect to their anchors in the 3D bounding box. Then, after learning a viewpoint-invariant appearance of each face, a sliding 3D bounding box approach is used for localization. A complementary approach was introduced by [15] which exploits pairwise relations between 3D objects in the scene from an object-centered perspective. They show how reasoning about relations between objects in the scene could serve as a cue for the classification of the object pose. Parallel to this, [19] presented a spatial layout model that enforced scene consistency based on the 3D aspectlets of individual objects with object-object consistency in the form of occlusion reasoning. This combination not only improved 3D object detection but also produced accurate oriented object hypotheses. Very recently, and in parallel to our work, [20] uses a fine detail shape representation based on CAD models. This representation improved the reasoning about object support on the ground-plane and mutual occlusion. Similar to this group of works, we derive our object representation from features from the 3D object in the scene. However, we employ much simpler features derived from the bounding boxes of the objects and not from, more complex, CAD models as in [20]. We enforce scene consistency by either assuming that the objects of interest are located on the scene-ground plane as proposed in [9] in the context of object detection; or by assuming that the object evidence extracted by the detector should align with objects previously seen with the same camera setup. Finally, different from [15, 19, 20] we do not reason about object relations. Furthermore, our method operates in still images and does not require image sequences as in some SfM-based approaches [1, 2].

3 Proposed Method

In a nutshell our method can be summarized in five steps (see Fig. 2): First, we run a viewpoint-aware object detector in order to collect a set of hypotheses $o = \{o_1, o_2, ..., o_n\}$. Then, based on a proposal generation function Ω we generate a set of scene-driven object proposals $o' = \{o'_1, o'_2, ..., o'_n\}$. In the next step we estimate a correspondence descriptor d_i between each object hypothesis o_i and its closest scene-driven object proposal o'_i. Then, we estimate the elongation of the initial object hypothesis o_i via multiclass classification of the descriptor d_i. Finally, the viewpoint of the objects is estimated by the fusion of the responses of the viewpoint-aware local object detector and the scene-driven elongation classifier. Now we take a deeper look in the different stages of the proposed method.

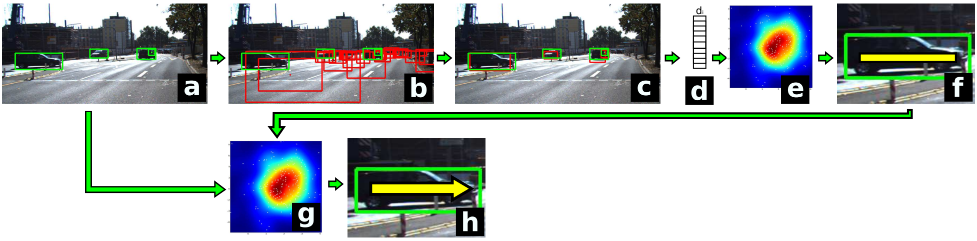

Figure 2: Algorithm Pipeline: a) Object Detection, b) Scene-driven Object Proposal Generation c) Object-hypotheses - Object-Proposal Matching, d) Correspondence descriptor extraction, e) Elongation Classification, f) Elongation Estimate, g) Viewpoint Classification, and h) Viewpoint Estimate.

3.1 Scene Representation and Object Detection

This work is inspired by the work of Hoiem et al. [9] in the sense that we consider the idea of a scene concept behind the image. Whereas [9] defined the scene as a ground-plane and focused on the task of object detection, we explore different scene representations and exploit these ideas for the problem of object viewpoint estimation. Here we define a scene as the area in which the objects of interest occur. This scene can be defined either in 3D or in 2D. Furthermore, we explore an extension of the original idea of [9] where the scene can serve as a prior for the location of the objects of interest, to consider the scene as a space with specific regions that are more likely to hold certain objects with particular features like class, orientation and/or size.

To guide the process, we locate regions of the image that appear to host the objects of interest based on appearance. To this end, we run a standard object detector which produces a set of 2D object hypotheses $o = \{o_1, o_2, ..., o_n\}$, where each object hypothesis $o_i = (s_i, b_i, \alpha_i)$ is defined by its confidence score s_i, its bounding box coordinates $b_i = (x_{1i}, y_{1i}, x_{2i}, y_{2i})$, and its viewpoint α_i, for the case of viewpoint-aware detectors such as the ones presented in [6] and [14].

3.2 Scene-driven Object Proposal Generation

Once we have spotted a set of regions in the image, i.e. the object hypotheses o_i, which are likely to host the objects of interest the next step is to recover scene-driven object proposals o'_i (i.e. Fig. 2.b), that will serve to validate the evidence o_i collected by the detector. We generate a set of scene-driven object proposals as $o' = \Omega(scene)$, where Ω is an object proposal generation function defined over the scene. We define Ω using one of the following methods:

a) Ground-Plane: This approach is heavily inspired from the work of Hoiem et al. [9]. We model the geometry of the scene by assuming the existence of a ground plane that supports the objects of interest. Given the ground plane, we densely generate a set of 3D object proposals $O = \{O_1, O_2, ..., O_m\}$ resting on the ground plane for each of the discrete orientations $\theta_k = \{\theta_1, \theta_2, ..., \theta_K\}$. Each 3D object proposal, $O = (X, Y, Z, L, W, H, \theta)$, is defined by its 3D location (X, Y, Z), its physical length, width and height (L, W, H) and its orientation θ in the scene. In our work we define the length, width and height (L, W, H) of the proposed 3D object proposals O based on statistics from real world objects. We drop the 3D location coordinate Y since all the 3D object proposals are assumed to be supported by the ground plane, hence $Y = 0$ for all the proposals. Then, once we have generated all the 3D objects that can physically be in the scene, using the camera parameters we project each of the 3D object proposals O to the image space, assuming a perspective camera model, producing a set of 2D object proposals o'. Specifically, each 2D proposal o' is obtained by projecting each of the corners of the 3D proposal O, using a perspective camera model and selecting the 2D points that enclose the rest. The viewpoint α of each 2D object proposal o' is estimated as a function of the 3D orientation θ and location (X,Y,Z) of the 3D object proposal O that generated it.

b) History of 3D objects: This approach is an extension of the previous scenario, in which we assume that there are some regions of the 3D space that are more likely to support objects with particular features. For this purpose, we start from a set of previously seen ground truth 3D Objects and sample from the distribution defined by $p(X,Z,\theta)$ a subset of $O = \{O_1, O_2, ..., O_m\}$. Following the same procedure as before, assuming a perspective camera model, we project the 3D objects O to the image space producing a set of object proposals o'. Note that for history-based methods to be informative, a similar camera setup is required at test time.

c) History of 2D objects: This is the 2D counterpart of the previous approach where we assume that the scene is defined over the image space, hence, we start from a set of ground truth 2D objects from a training set. Here we obtain the set of object proposals o' by sampling the distribution defined by $p(x,y,w,h,\alpha)$, where x and y are the 2D location coordinates of the object in the image, w and h define the bounding box size and α its viewpoint. Note that since this approach operates in the image space, its annotation cost is relatively lower than its 3D counterpart.

d) History of 2D object hypotheses: This approach is very similar to the previous in the sense that the scene is defined over the image space. It differs in that it starts from the output of an object detector. The set of object proposals o' is obtained by sampling the distribution defined by a set of previously detected object hypotheses.

In order to model the history-based distributions we gather a specific set of features from the object instances present in the training set. Depending on the scene representation, we collect either (X,Z,θ) for 3D objects or (x,y,w,h,α) for the 2D objects. Then, based on oKDE [12] we construct each distribution as a Gaussian mixture model. We uniformly sample this distribution producing a total of 1000 object proposals o'.

As final step of this stage (see Fig 2.c), for each of the hypotheses o_i spotted by the object detector, we select its closest scene-driven object proposal o'_i using the PASCAL VOC criterion [4]. Note that due to the box representation, objects with opposite orientations (orientation difference=π) will project the same 2D bounding boxes. For this reason we will focus on a smaller set of $K/2$ discrete viewpoints. This subset of viewpoints measures the orientation of the maximum extent of the objects (i.e. Fig. 2.f), which we refer in this paper as the *elongation* orientation ε of the object. Given an object o_i its elongation ε_i is defined as $\varepsilon_i = (\alpha_i \bmod \pi)$.

3.3 Elongation Classification

So far, for each of the object hypotheses o_i reported by the detector, we have spotted its closest 2D object proposal o'_i. For each hypothesis-proposal pair (o_i, o'_i) we compute the correspondence descriptor $d_i = (rw, rh, rx, ry, \alpha')$ where $rw = |\frac{w'}{w}|$, $rh = |\frac{h'}{h}|$, $rx = |\frac{x'-x}{w}|$, $ry = |\frac{y'-y}{h}|$ and the viewpoint α' of the closest object proposal.

In this work, we take advantage of the physical extent of elongated objects and how the elongation of an object is related to its viewpoint perceived in the image space. To this end, we first classify the elongation $\hat{\varepsilon}$ of an object, and then use this prediction to improve the viewpoint prediction $\hat{\alpha}$ given by the detector based on appearance features.

Training: Given a set of training images containing objects with annotated bounding boxes b_i and elongation ε_i, for each annotated object o_i we obtain its matching object proposal o_i' and compute its correspondence descriptor d_i following the procedure described above. Then, using the pair (d_i, ε_i) we model each discrete elongation ε by a probability density function (pdf). We use the odKDE method from [11] to model these pdfs. This method has the advantage of providing compression mechanisms in which each pdf is approximated by a Gaussian Mixture Model (GMM), reducing the number of components required to model each distribution.

Inference: Given a new image, we compute the set of correspondence descriptors d between the object hypotheses o and the object proposals o' generated from $\Omega(scene)$. Then, the elongation $\hat{\varepsilon}_i$ of each object o_i is the MAP estimate by applying the Bayes rule:

$$\hat{\varepsilon}_i = \arg_{\varepsilon_k} \max p(\varepsilon_k|d_i) = \arg_{\varepsilon_k} \max p(d_i|\varepsilon_k)p(\varepsilon_k), \quad (1)$$

where the class likelihoods $p(d_i|\varepsilon_k)$ are computed using Kernel Density Estimation (odKDE) and the priors $p(\varepsilon_k)$ are obtained from the occurrence of the elongation ε_k on the training data.

3.4 Viewpoint Classification

We estimate the viewpoint of the object o_i by late fusion of the response of the viewpoint-aware object detector and the response of our object elongation classifier (Fig. 2.g). Given the two responses, we define the coupled response $r_i = [\alpha_i, s_i, \varepsilon_i, \lambda_i]$ where (α_i, s_i) are the responses of viewpoint and score from the object detector, and $(\varepsilon_i, \lambda_i)$ are the responses of elongation and score of our elongation classifier. In this case, the score λ_i is the posterior $p(\varepsilon_i|d_i)$ estimated in Eq. 1. Finally, to classify the viewpoint $\hat{\alpha}_i$ of an object o_i we perform MAP inference for the coupled response r_i over the discrete viewpoint classes in a similar fashion as Eq. 1:

$$\hat{\alpha}_i = \arg_{\alpha_k} \max p(r_i|\alpha_k)p(\alpha_k). \quad (2)$$

Here, the class likelihoods $p(r_i|\alpha_k)$ are computed performing Kernel Density Estimation (odKDE) considering the response of the object detector and the elongation classifier on a validation set. The priors $p(\alpha_k)$ are obtained from the occurrence of the viewpoint α_k on the validation data.

4 Evaluation

4.1 Experimental Settings

We perform experiments on the KITTI benchmark [7], specifically, on the object detection dataset. Since the test set is not available, to provide quantitative results, we split the training set into three separated sets using the sequences and time labels attached to each image. We report results in two subsets of ground truth objects from the testing set. The first set, coined *fullSet*, contains all the objects from the testing set while the second set, coined *easySet*, contains all the objects whose bounding box height is larger than 50 pixels. We run experiments starting from one of three off-the-shelf standard detectors used to collect the initial object hypotheses. We employ the deformable parts model (DPM) detector (release 5) [17] and two extensions of DPM, [6] and [14], modified to predict 8 viewpoints. These detectors were trained

	Geiger et al. [6]		Lopez et al. [14]		Felzenswalb et al. [17]	
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

Table 1: Object Elongation Classification Performance. Mean Precision on Pose Estimation (MPPE).

on the PASCAL VOC 2007 [3], the Karlsruhe urban [6], and the EPFL cars [16] datasets, respectively. We present results for five methods: 1) *local detector*, the isolated output of one of the object detectors ([6, 14, 17]), i.e. not using any scene-level information; 2) *GroundPlane*, when assuming the scene defined by its ground-plane (Sec. 3.2(a)); 3) *Hist3DObjects*, when considering the 3D regions of the scene that are more likely to host objects with particular features (Sec. 3.2(b)), 4) *Hist2DObjects*, the 2D counterpart of *Hist3DObjects*, where we start from a history of 2D ground-truth objects (Sec. 3.2(c)), and 5) *Hist2DHypotheses*, where we start from a history of 2D hypotheses collected with an object detector (Sec. 3.2(d)).

4.2 Experiment: Object Evidence Extraction

This experiment aims to show the performance of the object detectors for recovering the object bounding boxes. We evaluate detection of 2D objects following the PASCAL VOC criterion [4] (50% intersection-over-union) and report mean Average Precision (mAP) as performance metric. The local detectors [6, 14, 17] produced a total of 3986, 9713 and 10525 hypotheses, respectively, during the detection stage. On the *easySet*, the selected detectors achieved a performance of 52%, 35% and, 44% mAP, respectively. This performance dropped to 34%, 20% and 32% mAP, respectively, when the *fullSet* was considered. The superior performance of the detector from [6] can be attributed to the fact that it was trained on the Karlsruhe urban dataset [6], which more closely resembles the settings from the KITTI dataset used for evaluation. We provide these performance measurements to indicate the volume of data processed in the following experiments.

4.3 Experiment: Elongation Classification

The elongation of an object is a feature closely related to its viewpoint. For this reason, we use a performance metric that has been traditionally used in previous work for measuring the performance of pose/viewpoint estimation. In particular, we adopt the Mean Precision in Pose Estimation (MPPE) as performance metric. MPPE is computed as the average of the class-normalized confusion matrix of the pose/viewpoint classifier. It is computed from hypotheses that are assumed correct based on the PASCAL VOC criterion [4], as in prior work [8, 14, 18]. For the evaluation of this experiment, we derived elongation annotations, from the original viewpoint annotations of the dataset, producing four possible discrete elongation values in the range $[0,\pi)$. Furthermore, we run experiments using the object detector from [17] to collect object hypotheses, and using our method we extend its output to provide elongation estimates. Note that [17] is purely an object detector and does not provide viewpoint estimates.

Discussion: We can see in Table 1 that all the proposed methods for elongation classification have a performance clearly above chance levels (25% for the case of 4 elongation

	Geiger et al. [6]		Lopez et al. [14]	
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

Table 2: Object Viewpoint Classification Performance. Mean Precision on Pose Estimation (MPPE).

orientations). This shows that indeed our methods are encoding some useful cue for elongation classification. It is also remarkable that this can be achieved without having direct access to the local data of the objects, the pixels inside the bounding boxes. Furthermore, the difference in performance between detectors is more evident. For the case of [14], there is more room for improvement and our elongation classifiers achieve a mean improvement of 16.3 and 4.5 percentage points (pp) for the *easySet* and the *fullSet*, respectively. In addition, starting from this particular detector, the method defined by *GroundPlane* leads with improvements of 23 and 11 pp in the respective image sets. On the opposite, for the case of [6], none of the proposed methods improves over the local detector for the task of elongation classification. For all cases, it is notable how the methods based on a 2D scene representation (*Hist2DObjects*, *Hist2DHypotheses*), which require significant lower annotation effort, have a comparable performance to some of the 3D-based methods. Also, we can see that for the pure detector [17], we are able to predict its elongation to a significant level, clearly above chance levels.

4.4 Experiment: Viewpoint Classification

We now measure the performance of the proposed method for the task of viewpoint classification, that is, the classification of 8 discrete viewpoints. For this evaluation we will again use MPPE. Similarly, we report performance results on the same five methods: *local detector*, *GroundPlane*, *Hist3DObjects*, *Hist2DObjects*, *Hist2DHypotheses*. However, notice that since the viewpoint of an object is defined by the combination of the responses of the detector and the elongation classifier (Sec. 3.4), we can only report viewpoint classification results on the detectors from [6] and [14], which provide viewpoint-related information in their response. Please see Fig. 3 for some qualitative results.

Discussion: For the task of viewpoint classification we notice a drop in most of the performance values (Table 2). This is to be expected as it involves more classes than the one of elongation. However, all the proposed methods are again well above chance levels (12.5% for 8 viewpoint classification). For the case of [14] our proposed methods achieve mean improvements of 10.75 and 5.75 pp, for the *easySet* and *fullSet*, respectively. Experiments using this detector are lead by the *Hist3DObjects* method, with the respective improvement of 17 and 12 pp. Different to the elongation classification task, for viewpoint classification our methods do bring an improvement also over [6] namely, 6.25 pp on the *easySet*. This may hint that some of the failure cases for the detector from [6] are caused by opposite viewpoints. In addition, for this detector, best results are obtained by the methods based on a 3D scene representation (*GroundPlane*, *Hist3DObjects*), producing mean improvements of 6 and 1 pp on the corresponding image sets. Similarly to the elongation experiments (Sec. 4.3), the methods defined on a 2D scene representation have a comparable performance to the methods starting from a 3D scene representation. In addition, for the case of viewpoint classification, at least for the relatively larger objects of the *easySet*, the cues from the 2D scene always bring improvements over the purely local methods ([6, 14]).

5 Discussion

In this paper we have explored several ways to extract cues from the scene with the objective of estimating the viewpoint of the objects of interest while enforcing scene consistency. We have seen that by taking into account scene-driven cues, viewpoint classification results can be improved relative to those obtained when using only local information. Recently, other methods that perform scene-driven viewpoint estimation have been proposed, [19] presented a few months ago and [20] developed parallel to this work. However, both these methods focus purely on the ground plane assumption as means to enforce consistency with the scene. Furthermore, they require a calibrated camera and CAD models, plus the traditional object annotations on the image set. Both depend on fine-part detection during the object detection stage, which makes them inappropriate for low resolution images. Finally, they have a strong link between the methods to enforce scene consistency and to perform object detection. This complicates the integration of future, possibly improved, object detectors in their methods for enforcing scene consistency. To their advantage, by learning from CAD models, they are able to predict object polygons or wireframe models that are more pleasing to the eye and closer to the original object shape than our bounding box predictions. Furthermore, they are able to predict continuous object orientation values.

On the opposite, we have presented several ways to enforce scene consistency with different levels of annotation cost. Additionally, as demonstrated in our experiments, we are able to integrate any object detector in our method. This last feature allows our method to improve the box representation of its predictions by integrating more advanced detectors, e.g. the ones used in [19] and [20], as long as they produce viewpoint information in their responses. In its current state, our method does not have the requirement of high resolution images for proper performance. This is again handled by the flexibility of the method for the integration of any object detector. This flexibility also makes our method useful to extend pure detectors, such as [17], to produce elongation estimates. Note that for some applications, such as obstacle detection on the roads, object elongation prediction might be enough. Finally, the similar camera setup requirement of our history-based methods may be seen as a strong constraint. However, there are many scenarios that resemble this setting, e.g. dashcams and backup cameras attached on cars; inspection in manufacturing, and fixed security cameras found on streets, airports, shopping centers and several areas of interest where human activity takes place.

6 Conclusions

In this paper we have introduced scene-driven object elongation orientation classification as an intermediate step prior to viewpoint classification. Our experiments show how considering object elongation estimates brings improvements over purely appearance-based viewpoint-aware object detectors. In addition, we have presented several approaches to perform scene-driven object viewpoint classification at different levels of annotation cost. The proposed method is flexible enough to allow the integration of future, more advanced, viewpoint-aware detectors. To conclude, this work complements very recent work, by sending the message

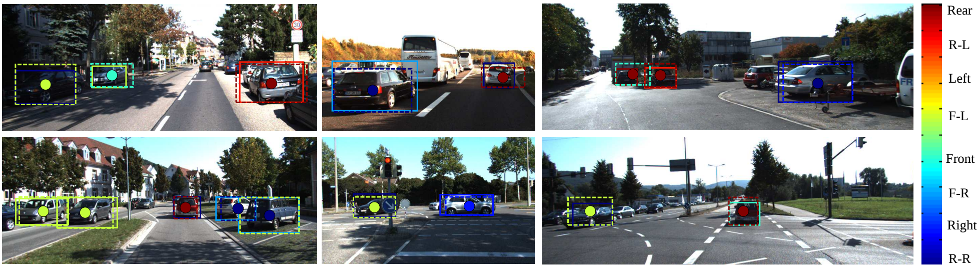

Figure 3: Viewpoint classification results encoded in jet scale. Continuous line, local detector prediction; Dashed line, scene-driven object proposals. Circle, ground-truth viewpoint. For better visualization please refer to the supplementary material (Best viewed in color).

that there are relatively simple cues in the scene that can bring improvements for the task of object viewpoint classification.

Acknowledgments: This work is supported by the FP7 ERC grant 240530 COGNIMUND, the KU Leuven OT Project VASI and the DBOF Research Fund KUL 3E100864.

References

[1] S. Y. Bao and S. Savarese. Semantic structure from motion. In *CVPR*, 2011.

[2] S. Y. Bao, M. Bagra, Y. Chao, and S. Savarese. Semantic structure from motion with points, regions, and objects. In *CVPR*, 2012.

[3] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, .

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, .

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

[6] A. Geiger, C. Wojek, and R. Urtasun. Joint 3d estimation of objects and scene layout. In *NIPS*, 2011.

[7] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[8] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In *CVPR*, 2011.

[9] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, 2006.

[10] D. Hoiem, C. Rother, and J. M. Winn. 3d layoutcrf for multi-view object class recognition and segmentation. In *CVPR*, 2007.

[11] M. Kristan and A. Leonardis. Online discriminative kernel density estimator with gaussian kernels. *Cybernetics, IEEE Transactions on*, 2014.

[12] M. Kristan, A. Leonardis, and D. Skočaj. Multivariate online kernel density estimation with gaussian kernels. *Pattern Recognition*, 2011.

[13] J. Liebelt and C. Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, 2010.

[14] R. J. Lopez-Sastre, T. Tuytelaars, and S. Savarese. Deformable part models revisited: A performance evaluation for object category pose estimation. In *ICCV WS*, 2011.

[15] J. Oramas M., L. De Raedt, and T. Tuytelaars. Allocentric pose estimation. In *ICCV*, 2013.

[16] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[17] D. McAllester P. Felzenszwalb, R. Girshick. Cascade object detection with deformable part models. In *CVPR*, 2010.

[18] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3d geometry to deformable part models. In *CVPR*, 2012.

[19] Y. Xiang and S. Savarese. Object detection by 3d aspectlets and occlusion reasoning. In *3ddr@ICCV*, 2013.

[20] M. Z. Zia, M. Stark, and K. Schindler. Are cars just 3d boxes? - jointly estimating the 3d shape of multiple objects. In *CVPR*, 2014.