Supervised learning and evaluation of KITTI's cars detector with DPM

J. Javier Yebes, Luis M. Bergasa, Roberto Arroyo and Alberto Lázaro

Abstract— **This paper carries out a discussion on the supervised learning of a car detector built as a Discriminative Part-based Model (DPM) from images in the recently published KITTI benchmark suite as part of the object detection and orientation estimation challenge. We present a wide set of experiments and many hints on the different ways to supervise and enhance the well-known DPM on a challenging and naturalistic urban dataset as KITTI. The evaluation algorithm and metrics, the selection of a clean but representative subset of training samples and the DPM tuning are key factors to learn an object detector in a supervised fashion. We provide evidence of subtle differences in performance depending on these aspects. Besides, the generalization of the trained models to an independent dataset is validated by 5-fold cross-validation.**

I. Introduction

Nowadays, vision sensors are employed in automotive industry to integrate advanced functionalities that assist humans while driving. During the last years, a big research effort has been made to design and study Advanced Driver Assistance Systems (ADAS) and autonomous vehicles that rely on cameras as sensing technology and source of data [1]. On the contrary, other sensing modalities as GPS, lidar and radar have a well-established market as on board integrated systems for navigation, active safety and primary obstacles detectors [2], [3], [4], although information fusion is an open field of research [5], [6].

The improvements in camera features, their price and size reduction, added to the progress in machine learning and computer vision approaches for intelligent vehicles, have increased the appealing of vision systems to automotive industry and researchers. Imaging devices provide a higher level of abstraction and semantic information more natural to interpret by humans compared to other sensors, e.g. light-beam [7], intelligent parking [8] and vision [9]. Furthermore, there are still many challenges on image scene understanding and object recognition to obtain more precise information for autonomous vehicles and driving assistance systems. These challenges may include and are not limited to object detection under occlusion [10], [11], estimation of objects orientation on 3D scenes [12], detection at far distances [13], determining geometric layout of the scene [14], [15], appropriate modeling and parametric learning of complex scenes [16] and large-enough and naturalistic datasets.

Indeed, a lot of research effort lies on the existence of public datasets and common evaluation metrics for advancing the performance of visual recognition systems [17]. There are many benchmarks, some of them also widening to a higher number of categories non-restricted to road environments like Caltech-101 [18], PASCAL VOC [19] and EPFL Multi-view car [20] among others.

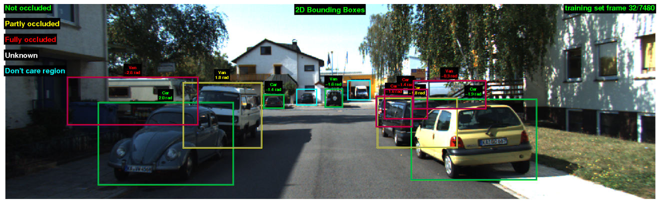


Fig. 1. Ground truth labeled samples from KITTI benchmark

In this paper, we are interested on the KITTI object evaluation challenge [21] to detect and estimate the orientation of Cars, Pedestrians and Cyclists on images from road scenes (Fig. 1). This is a topic of a great research interest [22], which implicitly requires dealing with the open tasks that have been introduced before. The KITTI Vision Benchmark Suite [23] provides a wide set of images on urban environments with ground truth labeling and multiple sensor data plus common evaluation protocols. In particular, this paper carries out a discussion on the level of supervision required to train a car detector built as a Discriminative Part-based Model (DPM) [24]. This approach has been already proposed by [23] and the contribution of our paper is not on the theoretical point of view of a new detector, but on the experimental nature of a more in-deep analysis during learning (cleanliness of the data samples and parameter tuning) and during the evaluation of predicted bounding boxes (metrics and methodology). In Section III, we provide evidence of subtle differences in performance depending on three factors: the selected evaluation method (KITTI [21] vs PASCAL [19]), the difficulty level of the training samples and the DPM internal configuration.

II. DPM and Related Works

DPM [24] classifies and locates objects at different scales based on a pyramid of appearance features. It has been successfully tested on PASCAL challenges [19] and applied to many other works and datasets. In particular, we are employing the release 4 of its open source code [25] to match the format of the pre-trained models in [21].

Training. In DPM, the model of an object is a mixture of components initialized from clusters on images' aspect

*This work was supported in part by the Spanish MECD under grant AP2010-1472, by the MINECO project Smart Driving Applications (TEC2012-37104) and by the CAM project RoboCity2030 II-CM (S2009/DPI-1559). Thanks also to A. Geiger and P. Lenz in their support.

The authors are with the Department of Electronics, UAH. Alcalá de Henares, Spain. e-mail: `javier.yebes, bergasa, roberto.arroyo, alberto.lazaro@depeca.uah.es`

ratios and it is represented by a set of filter weights for the object parts and deformation weights for the spring-like star topology that connects root and part filters. These weights are learned by training a latent SVM classifier, where the latent variables are the location, scale and model component of the compositional parts. All the weights are concatenated in a high-dimensional vector β [24]. For example, considering 16 components where each of them has 1 root part (variable size depending on aspect ratio), 8 subparts of fixed size (6×6) and a normalized gradient descriptor of 32 dimensions, the total number of parameters to be learned is 170,624. A pictorial representation is on Fig. 2.

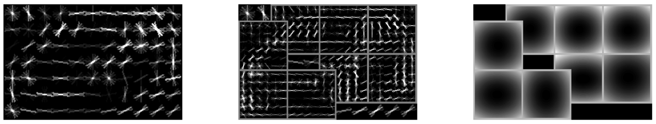


Fig. 2. Learned weights for class 'Car' in KITTI and viewpoint $5\pi/8$ rad. From left to right: root filters, part filters at twice resolution than root and 2D deformation parameters for parts placement

Detection. A feature scale pyramid is built and walked through to generate the set of hypotheses. Then, the score of one hypothesis [24] is calculated as in Eq. 1, reproduced here for clarity. Afterwards, a maximum suppression filter outputs the finally predicted bounding boxes.

$$s(z) = \sum_{i=0}^{144} F_i \cdot \phi_v(H, z) - \sum_{i=17}^{144} d_i \cdot \phi_d(dx_i, dy_i) + bias \quad (1)$$

The limits of the sums correspond to the example before. F_i represents all the learned weights of the root and part filters and d_i the learned deformation weights. H is an image scale pyramid, z are the latent variables, ϕ_v is the visual feature map containing the HOG descriptors computed from H, (dx_i, dy_i) is a relative 2D displacement of part i with respect to root filter position and ϕ_d are deformation features.

Related works. A. Geiger et al. [15], [23] made an adaptation of DPM for its testing on KITTI dataset. Basically, they discretized the number of possible object orientations, i.e. 16 bins for cars, so that, every component of the mixture model corresponded to one orientation. Besides, they enlarged small examples by factor 3 and harvested random negatives from positive images, keeping for training only those negatives with a bounding box overlapping less than 20% with a positive label. Two versions (supervised and unsupervised) were reported on [21]. We will provide further evidence of this supervised tuning in section III-C.

In [26], part-based models were evaluated for object category pose estimation where some supervised adaptations were proposed: fixing the latent component to the object pose available in the ground truth, removing bilateral symmetry and developing a modified training pipeline that regarded the coordinate descent algorithm and the selection of negatives examples from opposite views. Despite their improvement in orientation estimation tested in four different datasets, KITTI could not be compared concerning the joint challenge on detection and orientation estimation. Thus, we provide results and a discussion applying some of the suggestions from [26] to learn a car detector from KITTI.

On the other hand, a new approach (OC-DPM) for explicit occlusion reasoning [10] based on the DPM framework has recently reported increased ratios, both in object detection and orientation estimation of cars [21], but employing 12 viewpoints instead of 16. This is actually a very promising approach to overcome the missed detections and false positives of DPM over KITTI as we point out in conclusions. However, despite the benefits of occlusion modeling, it is not yet clear whether the improvements came directly from it or due to the decreased number of viewpoints.

Although the next topics are out of the scope of this paper, more complex methods have proposed a higher level of abstraction, i.e. to include a 3D cuboid model [27], in which DPM is extended in the features and filters size to learn objects 3D location and orientation from monocular images. Differently, the reduction of the high computational requirements of DPM has been studied in [28], which presented an efficient object detection with an algorithmically enhanced version of the objects image search inside DPM.

III. Experiments

Generalizing trained models to an independent dataset requires a cross-validation that assesses on the best performing algorithm or configuration. Indeed, four of the current entries in [21] published results based on DPM [15], [24], but they lack of a deeper analysis on the experiments carried out. In our work, the comparative results are based upon 5-fold cross-validation. Firstly, we review the evaluation criteria, then, we give an insight on clean training data samples and we conclude reporting results after tuning DPM.

A. Discussion on the evaluation criteria

Evaluation metrics. Geiger et al. [23] employed the Average Precision (AP) and proposed Average Orientation Similarity (AOS) as common evaluation metrics based upon [29]. The predicted bounding boxes are sorted in decreasing order of confidence (s) and precision (p) and recall (r) are computed from the cumulative distribution of True Positives (TP), False Positives (FP) and False Negatives (FN). Then, AP and AOS are obtained as the Area under the Curve (AuC).

Evaluation algorithm. Despite the common metrics above, counting TP, FP and FN differs from PASCAL [29] to KITTI [23]. In fact, given a set of different experiments and the corresponding sets of predicted bounding boxes, the gradients in AP between the experiments yielded opposite signs and the AP values differed up to 20 points in KITTI vs PASCAL evaluations. Therefore, there is a high risk of extracting misleading conclusions from the experiments depending on the evaluation protocol. We bring here a detailed analysis of KITTI vs PASCAL evaluation approaches because there is no reference in the literature concerning this issue. Next, the common aspects are presented:

- Intersection over Union (IoU) [29] measures the overlap between predicted and ground truth bounding boxes.
- Every TP is the highest scoring detection with the highest overlap. The remaining overlapped detections are FP.
- AP is obtained as the AuC from the "p-r curve".

Additionally, KITTI follows these premises[1]:

- 'DontCare' regions (usually far away and some occluded objects) do not count as TP or FP when detected or as FN when missed. Besides, their overlap is treated differently, dividing by the area of the predicted bounding box instead of the union. This favors partial overlapped predictions around these ground truth regions of a relatively small size (95% of them are below 50 pixels of height, which is a 13.3% of the image height representing the scene).
- Neighbouring classes (e.g. 'Van' for class 'Car', 'Cyclist' for class 'Pedestrian') do not count as TP, FP or FN.
- Three difficulty levels are evaluated ('easy', 'moderate' and 'hard') [21]. The detections overlapping ground truth objects of a difficulty higher than the one under evaluation do not count as TP or FP. Similarly, they do not count as FN when missed.
- The detections lower than 30 pixels in height are not evaluated because at this scale they are more prone to error, being a source of FP.
- To compute the final "p-r curve", the recall points are approximated to a linear function, being built from a subsampled version of the sorted scores from TP list. By default, KITTI computes 41 points and we observed small variations in AP for higher number of points.

Attending to the first three premises, a detector is not rewarded for detecting those labeled objects, but also not penalized. Simply discarding the indicated ground truth regions, does not count them as TP or FN. Indeed, these training samples are marked as *ignored* such that predicted bounding boxes fulfilling the minimum overlap constraint do not count as FP either. This is the main source of variation between the AP estimated by PASCAL vs KITTI. In general, the KITTI evaluation [23] will lead to higher precision estimates because of the FP subtraction. This filtering of ground truth and detected samples during evaluation is also supported by a recent pedestrian detection survey [30].

Minimum overlap requirement. Typically, most of the works and datasets on object recognition [29], [30] impose a minimum overlap requirement of 50% between ground truth and predicted bounding boxes. In particular, KITTI [21] imposes 70% for cars. Table I compares AP and AOS for the same experiment evaluated with two distinct overlaps on the 5th fold of a randomly balanced split of training cars.

The results are divided in the three evaluation categories proposed in [21]. One of the experiments employs the pre-trained LSVM-MDPM-sv for cars [21] and the other has been trained using the remaining 4 folds of the cross-validation on a selection of easy samples. As can be seen, all cases yielded a boost in precision when reducing the minimum required overlap, which comes from a reduction of FN to a couple of miss-labeled ground truth 'easy' samples ($AP \simeq 98\%$) and also due to a notable decrease in FP for 'moderate' and 'hard' categories (FN is still significant in these categories due to smaller and/or occluded samples). Thus, supervising the evaluation protocols and establishing commonalities greatly influences the possible bias in the conclusions obtained from the results.

TABLE I
EVALUATING MINIMUM OVERLAP REQUIREMENT

		70%		50%	
		AP %	AOS %	AP %	AOS %
LSVM-MDPM-sv [21]	easy	72.02	64.95	98.07	88.45
	mod.	55.95	51.01	78.87	70.70
	hard	40.89	37.47	63.54	56.77
Easy train (ours)	easy	83.56	81.88	98.16	96.06
	mod.	47.79	45.52	66.08	63.80
	hard	35.91	34.89	51.91	49.95

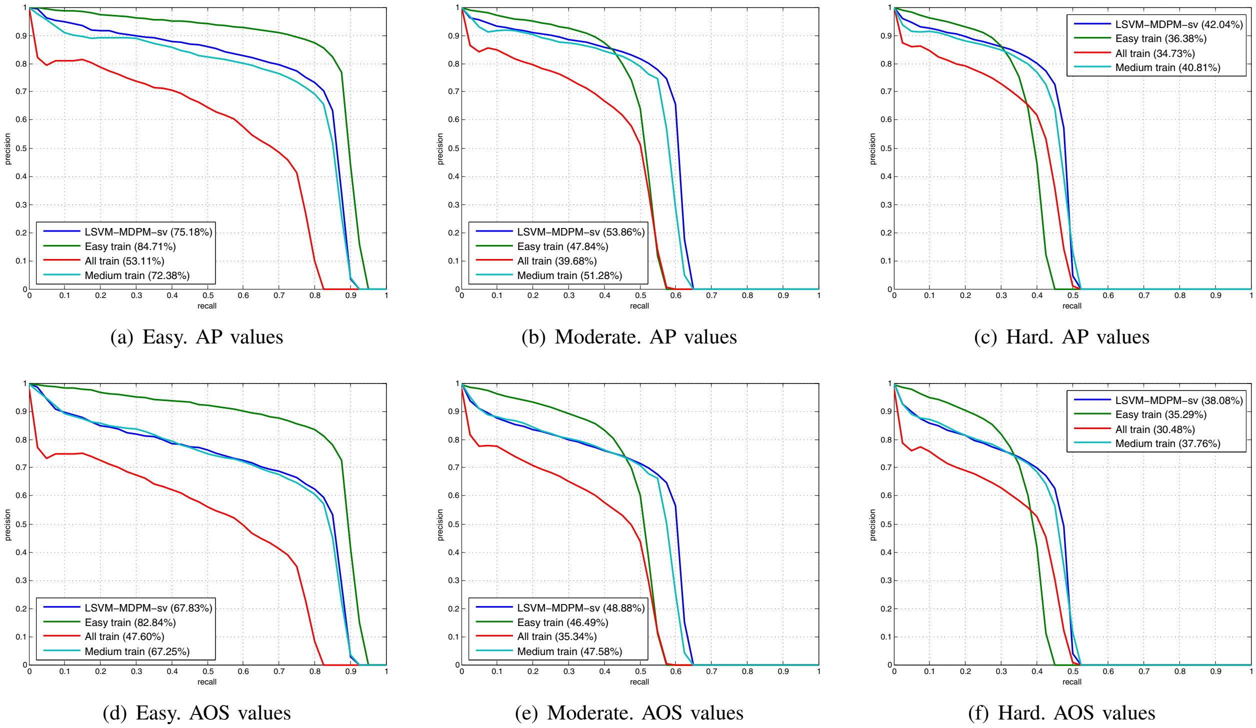
B. Data cleanliness

Supervised training regards the selection of the training samples such that, the cleaner data the better model learning. However, it also depends on the complexity grade that the model is designed to represent [16]. DPM is able to model an object category at multiple scales, under small partial occlusions, illumination changes and it is relatively flexible to intraclass variability. Hence, to account for the performance variability, we have carried out a set of experiments increasing the complexity of the training samples.

Four training modalities: *'LSVM-MDPM-sv'* [21], *'Easy'* (cars labeled with height > 40 pixels, fully visible and truncation < 15%), *'All'* (all labeled cars) and *'Medium'* (same as 'easy' plus 25 < height < 40 pixels and partly occluded samples). Besides, we discretize in 16 viewpoints initializing the model components in a supervised fashion. The results are averaged from 5-fold cross-validation experiments and evaluated into three categories [21] (columns in Fig. 3), whilst the rows refer to AP and AOS calculation respectively.

Results analysis. As can be seen in the first column of Fig. 3, training on *'Easy'* yields outstanding improvements. Nevertheless, in the subsequent graphs (b, c, e, f), its performance clearly degrades for more complex samples showing higher precision at low recalls but plummeting precision at medium recall. This is caused by a higher number of both FN and FP, the latter one accentuated when increasing recall. On the other hand, training on *'All'* obtains the poorest curves, although showing less FN for heights within $25 - 40$ pixels and/or partially occluded cars. This low performance is due to the lack of cleanliness in training data: too small cars, severe occlusions and truncations, which are an important handicap for parameter learning. Hence, increasing the amount of data does not always produce better results, unless the object model and training methodology could learn complex part-based topologies and adapt to high intraclass variability.

[1] Some premises of the KITTI evaluation protocol are in a text file inside its development kit [21], but others are directly in the source code.



(a) Easy. AP values

(b) Moderate. AP values

(c) Hard. AP values

(d) Easy. AOS values

(e) Moderate. AOS values

(f) Hard. AOS values

Fig. 3. Precision-recall curves, AP and AOS values for cars detection and orientation estimation after 5-fold cross-validation. Every column corresponds to one evaluation category [21]. Four different training modalities are compared on each plot: *'LSVM-MDPM-sv'* [21], *'Easy'*, *'All'* and *'Medium'*. These graphs show the importance of selecting a clean dataset, but general enough to represent naturalistic urban scenes. *'All'* yields the worst results (red line), while *'Easy'* (green line) outperforms only on the easy samples and downgrades for the remaining difficulty levels.

Attending to the distribution of difficult samples in terms of height, occlusion and truncation, the majority of FN have a truncation lower than 10%. On the contrary, small cars (<40 pixels in height and sometimes under hard illumination conditions) and small to medium occlusions are the source of many missed cars. In addition, FP image patches (Fig. 4) typically include cars viewed from the back, multiple cars parked on the street, cars occluded by other cars in parkings or traffic jams, parts of cars, loose fitting around the car and a few samples of scene background.

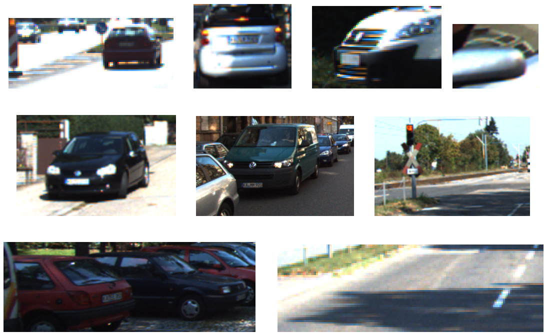


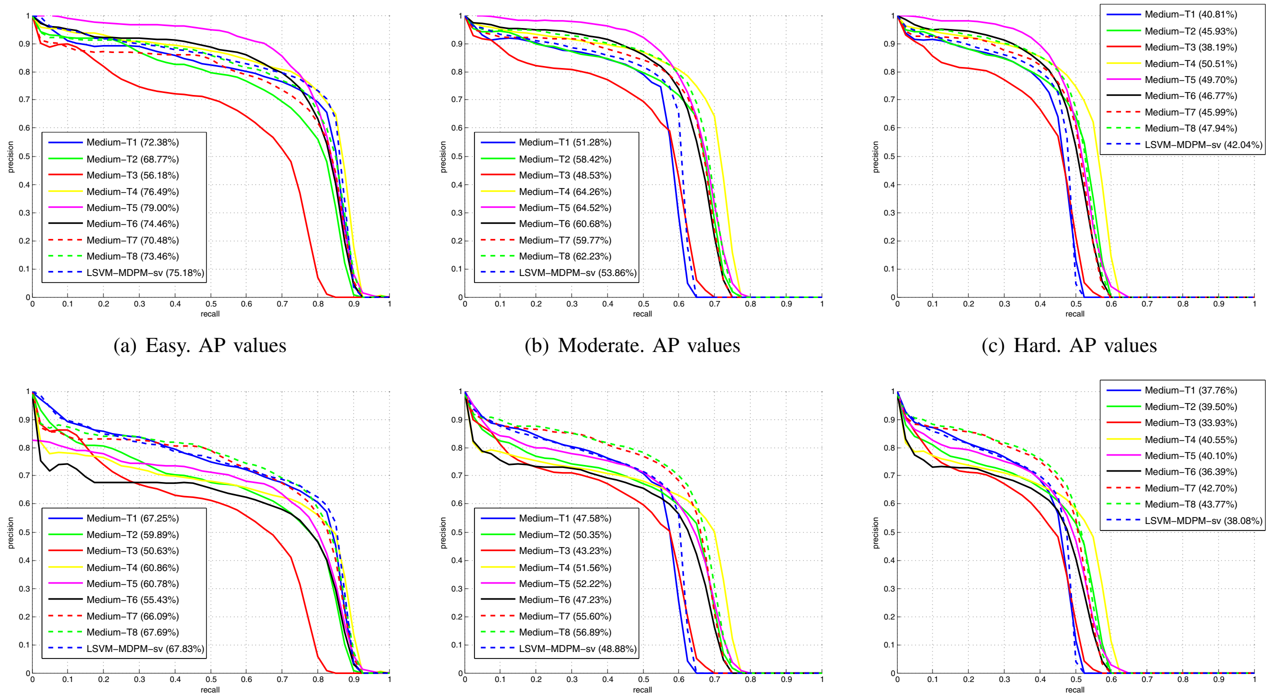
Fig. 4. Examples of false positives for class car.

Finally, *'LSVM-MDPM-sv'* and *'Medium'* showed the best stability at all evaluation categories. Our *'Medium'* training curves are very close to the baseline *'LSVM-MDPM-sv'* because we employed a very similar training subset, but without the additional modifications proposed in LSVM-MDPM-sv [21]. Therefore, these modifications, which do not increase model complexity, do not seem to provide a boost in performance. However, a correct level of supervision can provide subtle differences while training DPM, as it will be shown in following experiments.

It must be noted that observing Fig. 3, an increasing gap between AP and AOS appears when increasing the complexity of the training subset. This gap is around 1.5% for *'Easy'* (green lines) and 4-7% for the remaining plots depending also on the evaluation category. This loss of precision in orientation estimation can be motivated by the less informative features extracted from distant (small samples) and partially occluded cars. These errors use to belong to miss-classifications in neighboring viewpoints, which could be mitigated by reducing the number of orientation bins, although this also influences AOS values by definition [23].

C. DPM tuning

In this section, we report results (Fig. 5) on incremental DPM modifications with the aim of tuning the parameter learning to increase AP and AOS figures and get a better knowledge of DPM strengths and weaknesses. Overfitting is prevented by employing 5-fold cross-validation and *'Medium'* cars are employed as positive labeled samples, based upon the results from previous section. It must be noted that each experiment (5 trainings) can take 100-170 hours on an i7 CPU machine, depending on the experiment configuration described next.



(a) Easy. AP values

(b) Moderate. AP values

(c) Hard. AP values

(d) Easy. AOS values

(e) Moderate. AOS values

(f) Hard. AOS values

Fig. 5. Precision-recall curves, AP and AOS values for cars detection and orientation estimation after DPM tuning and 5-fold cross-validation on *'Medium'* samples. Every column corresponds to one evaluation category [21]. Moreover, 8 different experiments are carried out, as described in the text.

- **Medium-T1.** Initialization to 16 components corresponding to discretized car orientations. Bilateral symmetry assumption disabled because most of the car views are asymmetric. Besides, L-SVM regularization constant C=0.001. Default root filter area limited to $3,000-5,000$ pixels.
- **Medium-T2.** Analyzing previous results, several small cars are missed, then, we propose to allow smaller root filters (area > 1,000 pixels). This impacts during latent search on the image scale pyramid and we observe a detection improvement for the difficult samples (Fig. 5.b and 5.c). However, this presents the shortcoming of some smaller model components with lower level of detail, thus AP and AOS decrease for easy samples (Fig. 5.a). Nevertheless, there is a better orientation estimation at higher recalls for the difficult samples (Fig. 5.e and 5.f).
- **Medium-T3.** Considering the comments above, we propose to also enlarge the upper limit to 6,000 pixels to favor detection of 'easy' samples. Besides, we impose a loose fit for latent parts training in order to give more flexibility to the model, moving their overlap requirement from 70% to 60%. As a result of this move, the learned parameters are not representative enough causing a loss of precision for all cases (continuous red plots in Fig. 5).
- **Medium-T4.** Consequently, we opt to fix a tighter constraint, i.e. 80% overlap during latent parts search. This yields a medium gain for easy samples, but an important boost for the difficult ones. However, the orientation estimation shows a slight gain in precision (actually below previous curves at low and medium recalls) and AOS falls a 7% for easy samples (yellow plots in Fig. 5).
- **Medium-T5.** Assuming the naturalistic features of the KITTI urban dataset, most of the images have at least one labeled car. Hence, the DPM internal restriction to only negatives images for data mining is a handicap for the learning process. We include further tuning to DPM cropping the hard negatives during data mining, from strictly positive images. These negatives must not overlap more than 20% with a ground truth sample (like in LSVM-MDPM-sv [21]). The first bootstrapping step of DPM remains harvesting random negatives from strictly negative images. In spite of the increase in training time, we achieve an enhanced precision for all evaluation levels (magenta lines in Fig. 5) thanks to the increased number of background samples found during hard negatives mining. However, AP values are similar to previous experiment due to an earlier drop of precision at upper recalls. Although AOS replicates previous observations, the gap between AP and AOS is still too wide. Viewpoint discrimination is benefited for difficult samples but not for easy ones given the modifications carried out so far.
- **Medium-T6.** In order to generate more samples for all viewpoints, which could favor model learning, we duplicate the dataset by mirroring every sample and clustering them on the corresponding mirrored viewpoint with respect to $\pi/2$ and $-\pi/2$. Besides, we fix latent car viewpoints during the merge of the model components relying on the ground truth labels. Additionally, we mark 'DontCare' labeled regions as potential positives during hard negatives mining. However, we observe a lower performance for all cases (black plots in Fig. 5).

- **Medium-T7** Interestingly, using *Medium-T6* and reducing back the overlap for latent parts to the original 70%, a superior precision in orientation estimation is achieved (red dashed plots in Fig. 5), but lower AP values.
- **Medium-T8** Furthermore, tightening the same constraint to 75%, we obtain a moderate AP and AOS increase at all levels, which is better than the baseline LSVM-MDPM-sv [21] for evaluation levels 'moderate' and 'hard'.

IV. Conclusions and Future Works

This paper has presented a wide set of 5-fold cross-validation experiments to train DPM [24] in a supervised fashion for KITTI challenge [21]. In particular, this work has shown the convenience of choosing a well-defined evaluation protocol to correctly measure and analyze detection results. We have compared PASCAL [29] vs KITTI [23] evaluation methods that rely on the same metrics but different underlying algorithms. Besides, 3 training modalities, regarding the cleanliness of training data, have been compared with the baseline LSVM-MDPM-sv [21], in which *'Medium'* yielded similar "p-r curves". In addition, we have proposed several modifications during the learning of DPM weights, with the aim of achieving higher AP and AOS while gaining a better knowledge of DPM behavior. Three features proved to be the most relevant: the overlap requirement during latent search (75% as the best tradeoff); the harvesting of hard negative samples from strictly positive images and fixing the latent viewpoint during model components merging. After this tuning, we observed a precision boost in both detection and orientation estimation for the evaluated categories 'moderate' and 'hard', i.e. up to 10% for AP and 5% for AOS.

As future guidelines, we support the recent approaches on DPM extensions to 3D [11], [12], [27], in order to increment the level of supervision but also the complexity of the models that could reduce false positives and could also provide more accurate and non-discretized estimates of the objects orientation. Similarly, difficult samples, i.e. small and/or occluded ones, will require more cues to be detected, in the form of better input features or more flexible models.

References

[1] S. Sivaraman and M. M. Trivedi, "A review of recent developments in vision-based vehicle detection," in *IEEE Intelligent Vehicles Symposium (IV)*, 2013, pp. 310–315.

[2] E. Lissel, H. Rohling, and W. Plagge, "Radar sensor for car applications," in *IEEE 44th Vehicular Technology Conference*, 1994, pp. 438–442 vol.1.

[3] Autoliv, "Active Safety," www.autoliv.com/ProductsAndInnovations/ActiveSafetySystems/Pages/RadarSystems.aspx, Last viewed: January 2014.

[4] M. Montemerlo, S. Thrun, et. al., "Junior: The Stanford Entry in the Urban Challenge," *Journal of Field Robotics*, vol. 25, no. 9, pp. 569–597, 2008.

[5] S. Matzka, A. M. Wallace, and Y. R. Petillot, "Efficient resource allocation for attentive automotive vision systems," *IEEE Trans. on Intelligent Transportation Systems*, vol. 13, no. 2, pp. 859–872, 2012.

[6] F. Erbs, B. Schwarz, and U. Franke, "From stixels to objects - A conditional random field based approach," in *IEEE Intelligent Vehicles Symposium (IV)*, 2013, pp. 586–591.

[7] P. Alcantarilla, L. Bergasa, P. Jiménez, I. Parra, D. F. Llorca, M. A. Sotelo, and S. S. Mayoral, "Automatic LightBeam Controller for driver assistance," *Machine Vision and Applications*, vol. 22, no. 5, pp. 819–835, 2011.

[8] Toyota, "Intelligent Parking System," www.toyota-global.com/innovation/safety_technology/safety_technology/ parking/, Last viewed: January 2014.

[9] BMW, "Intelligent vision," www.bmw.com/com/en/insights/technology/connecteddrive/2013/driver_assistance/intelligent_vision.html, Last viewed: January 2014.

[10] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Occlusion Patterns for Object Class Detection," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3286–3293.

[11] M. Hejrati and D. Ramanan, "Analyzing 3D Objects in Cluttered Images," in *Advances in Neural Information Processing Systems*, 2012, pp. 602–610.

[12] B. Pepik, P. Gehler, M. Stark, and B. Schiele, "3D2PM - 3D Deformable Part Models," in *Eur. Conf. on Computer Vision (ECCV)*, 2012, pp. 356–370.

[13] D. Park, D. Ramanan, and C. Fowlkes, "Multiresolution Models for Object Detection," in *Eur. Conf. on Computer Vision (ECCV)*, 2010, pp. 241–254.

[14] C. Wojek, S. Walk, S. Roth, K. Schindler, and B. Schiele, "Monocular Visual Scene Understanding: Understanding Multi-Object Traffic Scenes," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 35, no. 4, pp. 882–897, 2013.

[15] A. Geiger, C. Wojek, and R. Urtasun, "Joint 3D Estimation of Objects and Scene Layout," in *Advances in Neural Information Processing Systems*, vol. 24, 2011, pp. 1467–1475.

[16] X. Zhu, C. Vondrick, D. Ramanan, and C. C. Fowlkes, "Do we need more training data or better models for object detection?" in *British Machine Vision Conf. (BMVC)*, 2012, pp. 1–11.

[17] A. Mogelmose, M. M. Trivedi, and T. B. Moeslund, "Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey," *IEEE Trans. on Intelligent Transportation Systems*, vol. 13, no. 4, pp. 1484–1497.

[18] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). Workshop on Generative Model Based Vision*, vol. 12, 2004, pp. 1–9.

[19] PASCAL VOC, "The Pattern Analysis, Statistical modeling and Computational Learning Visual Object Classes," http://pascallin.ecs.soton.ac.uk/challenges/VOC/, 2012.

[20] M. Ozuysal, V. Lepetit, and P.Fua, "Pose Estimation for Category Specific Multiview Object Localization," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[21] The KITTI Vision Benchmark Suite, "Object Detection and Orientation Estimation Benchmark," www.cvlibs.net/datasets/kitti/eval_object.php, 2012.

[22] ICCV Workshop, "Reconstruction Meets Recognition Challenge," http://ttic.uchicago.edu/˜rurtasun/rmrc/index.php, December 2013.

[23] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[24] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[25] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Discriminatively Trained Deformable Part Models, Release 4," http://people.cs.uchicago.edu/ pff/latent-release4/, 2010.

[26] R. J. López-Sastre, T. Tuytelaars, and S. Savarese, "Deformable part models revisited: A performance evaluation for object category pose estimation," in *ICCV Workshops*, 2011, pp. 1052–1059.

[27] S. Fidler, S. Dickinson, and R. Urtasun, "3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 620–628.

[28] I. Kokkinos, "Rapid Deformable Object Detection using Bounding-based Techniques," INRIA, Tech. Rep. RR-7940, 2012. [Online]. Available: http://hal.inria.fr/hal-00696120

[29] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *Intl. J. of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[30] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 34, no. 4, pp. 743–761, 2012.