Pose Estimation Errors, the Ultimate Diagnosis

Carolina Redondo-Cabrera[1], Roberto J. López-Sastre[1], Yu Xiang[2],
Tinne Tuytelaars[3] and Silvio Savarese[2]

University of Alcalá[1] Stanford University[2] KU Leuven, ESAT-PSI, iMinds[3]
carolina.redondoc@edu.uah.es, robertoj.lopez@uah.es,
yuxiang@cs.stanford.edu, tinne.tuytelaars@esat.kuleuven.be,
ssilvio@stanford.edu

Abstract. This paper proposes a thorough diagnosis for the problem of object detection and pose estimation. We provide a diagnostic tool to examine the impact in the performance of the different types of false positives, and the effects of the main object characteristics. We focus our study on the PASCAL 3D+ dataset, developing a complete diagnosis of four different state-of-the-art approaches, which span from hand-crafted models, to deep learning solutions. We show that gaining a clear understanding of typical failure cases and the effects of object characteristics on the performance of the models, is fundamental in order to facilitate further progress towards more accurate solutions for this challenging task.

Keywords: object detection, pose estimation, error diagnosis

1 Introduction

If there is one topic that has been obsessively drawing the attention of the computer vision research community, it has to be object detection. Object detectors are the heart of complex models able to interact with and understand our world. However, to enable a true interaction we need not only a precise localization but also an accurate pose estimation of the object. That is, just a bounding box does not help a robot to grasp an object: it needs to know a viewpoint estimate of the object to facilitate the inference of the visual affordance.

Since 2006, in parallel with the enormous progress in object detection, there have been appearing different approaches which go further and propose to solve the 3D generic object localization and pose estimation problem (*e.g.* [1–17]). But in this ecosystem the fauna exhibits a high level of heterogeneity. Some approaches decouple the object localization and pose estimation tasks, while some do not. There is no consensus either at considering the pose estimation as a discrete or continuous problem. Different datasets, with different experimental setups and even different evaluation metrics have been proposed along the way.

This paper wants to bring this situation under attention. We believe that to make progress, it is now time to consolidate the work, comparing different models proposed and drawing some general conclusions. Therefore, in the spirit of the work of Hoiem *et al.* [18] for the diagnosis of object detectors, we here propose a thorough diagnosis of pose estimation errors.

Our work mainly provides a publicly available diagnostic tool [1] to take full advantage of the results reported by state-of-the-art models in the PASCAL 3D+ dataset [19]. This can be considered our first contribution (Section 2). Specifically, our diagnosis first analyzes the influences of the main object characteristics (*e.g.* visibility of parts, size, aspect ratio) on the detection and pose estimation performance. We also provide a detailed study of the impact of the different types of false positive pose estimations. Our procedure considers up to five different evaluation metrics, which are carefully analyzed with the aim of identifying their pros and cons.

Our second contribution consists in offering a detailed diagnosis of four state-of-the-art models [4, 14, 19, 20] (Section 3).

We end the paper with our prescription for success. This is our last contribution and the topic of Section 4. There, we offer a comparative analysis of the different approaches, identifying the main weaknesses and suggesting directions for improvement. We even show how to use the information provided by our diagnostic tool to improve the results of two of the models, as an example.

Many studies have been proposed for the analysis of errors in object localization only [18, 21–23]. Simply in [20], some pose estimation error modes for the Viewpoints&Keypoints (V&K) model are analyzed. But this analysis is restricted to the setting where the localization and pose estimation are not analyzed simultaneously. We present here a more thorough comparison for four different approaches, and two different setting: pose estimations over the ground truth (GT) bounding boxes (BBs), and a simultaneous detection and pose estimation.

Overall, our main objective with this work is to provide a new insight into the problem of object detection and pose estimation, facilitating other researchers in the hard task of developing more precise solutions.

2 Diagnostic Tool

2.1 Dataset and object detection and pose estimation models

Different datasets for the evaluation of simultaneous object localization *and* pose estimation have been proposed, *e.g.* [2, 11–13, 24]. Although most of them have been widely used during the last decade, one can rapidly identify their most important limitation: objects do not appear *in the wild.* Other important issues are: a) background clutter is often limited and therefore methods trained on these datasets cannot generalize well to real-world scenarios; b) some of these datasets do not include occluded or truncated objects; c) finally, only a few object classes are annotated, being the number of object instances and the number of viewpoints covered with the annotation small too.

To overcome these limitations, the PASCAL 3D+ dataset [19] has been proposed. It is a challenging dataset for 3D object detection, which augments 11 rigid categories of PASCAL VOC 2012 [25] with 3D annotations. Furthermore, more images are added for each category from ImageNet [26], attaining on average

[1] https://github.com/gramuah/pose-errors

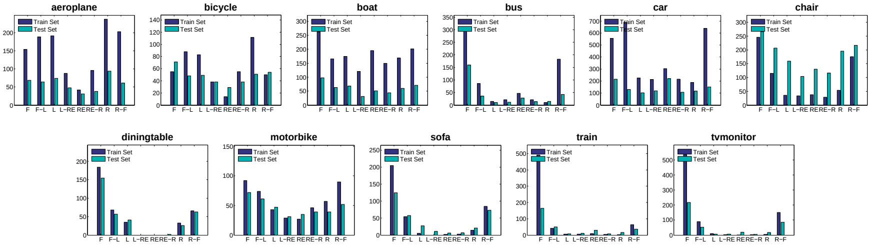


Fig. 1: Viewpoint distribution (in terms of azimuth). F: frontal. F-L: frontal-left. L: Left. L-RE: left-rear. RE: rear. RE-R: rear-right. R: right. R-F: right-frontal.

more than 3,000 object instances per class. Analyzing the viewpoint annotation distribution for the *training* and *test* sets, shown in Figure 1, it can be observed that the dataset covers all the viewpoints, although the annotation seems to be biased towards frontal poses. Since its release in 2014, the PASCAL 3D+ has experienced a great acceptance by the research community (*e.g.* [1, 5, 14, 16, 20]). We can affirm that it is rapidly becoming the *de facto* benchmark for the experimental validation of object detection and pose estimation methods.

We apply our diagnostic tool to four different approaches [4, 14, 19, 20]. All these models or provide the code or have officially submitted the results to PASCAL 3D+. In any case, these solutions have been selected not only because they define the state-of-the-art on PASCAL 3D+, like V&K [20], but also because they are representative for different approaches towards the pose estimation problem, allowing a variety of different interesting analyses: hand-crafted features based [4, 14, 19] vs. deep learning models [20]; Hough Forest (HF) voting models [14] against deformable part models (DPM) [4, 19] and template models [20].

We have two DPM based approaches. VDPM [19] simply modifies DPM such that each mixture component represents a different viewpoint. For DPM-VOC+VP [4] a structured labeling problem for the learning is proposed, where a viewpoint variable for each mixture component of the DPM is used. We also include in the study the Boosted Hough Forest (BHF) model [14]. It is a Hough voting approach able to perform a simultaneous object detection and pose estimation. As it is usual [27, 28], we incorporate a verification step using the faster R-CNN model [29] trained on the PASCAL VOC 2007, in order to re-score the detections of BHF and augment its recall. Finally, we diagnose V&K [20], a CNN based architecture for the prediction of the viewpoint. For the object localization, V&K relies on the R-CNN [30] detector. This is the only method, that does not perform a simultaneous object detection and pose estimation.

2.2 Diagnosis details and evaluation metrics

We offer a complete diagnosis which is split into **two analyses**. The first one focuses only on the viewpoint estimation performance, assuming the detections are given by the GT bounding boxes. In the second one, the performance for the simultaneous object detection and viewpoint estimation task is evaluated.

Our diagnostic tool analyzes the frequency and impact of different types of false positives, and the influence on the performance of the main object characteristics. Analyzing the different types of false pose estimations of the methods, we can gather very interesting information to improve them. Since it is difficult to characterize the error modes for generic rotations, we restrict our analysis to only the predicted azimuth. We discretize the azimuth angle into K bins, such that the bin centers have an equidistant spacing of $\frac{2\pi}{K}$. Thus, we define the following types of error modes. *Opposite viewpoint error*, which measures the effect of flipped estimates (*e.g.* confusion between frontal and rear views of a car). *Nearby viewpoint errors.* Nearby pose bins are confused due to they are very correlated in terms of appearance. Finally, the *Other* rotation errors, which include the rest of false positives.

With respect to the impact of the main object characteristic, we use the definitions provided in [18]. In particular, the following characteristic are considered in our study: occlusion/truncation, which indicates whether the object is occluded/truncated or not; object size and aspect ratio, which organizes the objects in different sets, depending on their size or aspect ratio; visible sides, which indicates if the object is in frontal, rear or side view position; and part visibility, which marks whether a 3D part is visible or not. For the object size, we measure the pixel area of the bounding box. We assign each object to a size category, depending on the object's percentile size within its object category: extra-small (XS: bottom 10%); small (S: next 20%); large (L: next 80%); extra-large (XL: next 100%). Likewise, for the aspect ratio, objects are categorized into extra-tall (XT), tall (T), wide (W), and extra-wide (XW), using the same percentiles.

Finally, we consider essential to incorporate into the diagnostic tool an adequate evaluation metric for the problem of simultaneous object localization and pose estimation. Traditionally, these two tasks have been evaluated separately, an aspect which complicates the development of a fair and meaningful comparison among the competing methods. For instance, a method with a very low average precision (AP) in detection, can offer an excellent mean angle error in the task of viewpoint estimation. How can we then compare these models?

In order to overcome this problem, our diagnostic tool considers three metrics, all evaluating simultaneously the pose estimation and object detection performance. They all have an associated precision/recall (prec/rec) curve. First, for the problem of detection and *discrete* viewpoint estimation, we use *Pose Estimation Average Precision* (PEAP) [3]. PEAP is obtained as the area under the corresponding prec/rec curve by numerical integration. In contrast to AP, for PEAP, a candidate detection can only qualify as a true positive if it satisfies the PASCAL VOC [25] intersection-over-union criterion for the detection *and* provides the correct viewpoint class estimate. We also use *Average Viewpoint Precision* (AVP) [19], which is similar to PEAP, with the exception that for the recall of its associated prec/rec curve, it uses the true positives according to the *detection* criterion only. The third metric is the *Average Orientation Similarity*

Table 1: Pose estimation with GT. Viewpoint threshold is $\frac{\pi}{12}$ for AVP and PEAP.

Method	aero	bicycle	boat	bus	car	chair	table	mbike	sofa	train	tv	AVG
AOS/AVP/PEAP												
RAND	51.3/11.2/1	54.8/12.1/1.1	55.4/12.8/1.3	48/9.5/0.8	50.8/9.4/0.7	50.2/8.4/0.5	54.4/11.4/1.3	55.1/10.5/0.8	51.6/10.3/0.9	51.4/10.7/1	50.6/12/1.2	52.2/10.8/1
BHF [14]	67.7/23.3/4.1	66.2/25.9/5.1	65/18.8/2.6	83.5/45.1/16.8	59.7/24.1/4.5	64.1/17.2/2.5	83.5/31.3/8.6	64.4/17.7/2.6	84.4/35.2/9.7	81/42.9/15.6	92.8/45.5/17.2	73.8/29.7/8.1
VDPM [19]	81.3/36.1/10.1	87.6/44.5/15.5	55.9/10.5/1	77/71.6/37	70.8/40.7/12.3	70.5/25.9/5.3	74.2/31.6/8.3	85.8/43.1/15.4	74.2/40.8/12.9	81.5/68.1/35	90.9/50.4/20	77.2/42.1/15.7
V&K [20]	94.9/63.1/30.7	92/61.7/29.6	80.6/44.9/15.7	97/81.6/55	93.8/72.5/41	92.7/60.5/28.4	85.1/45.3/18	93.4/61.8/31	93.4/54.6/23.9	88.4/72.2/41.2	96.1/54.6/23.9	**91.6/61.2/30.8**
MAE/MedError												
RAND	90.8/94.3	90.6/85.7	85.5/84	97.5/104	89.4/88.6	91.1/91.7	90.9/89	88.8/84.4	90.2/92.4	92.3/91.5	89.2/87.6	90.6/90.3
BHF [14]	76.2/65.6	78.9/77.1	79.1/73.2	47.9/29.4	85.9/83.3	75.6/70.3	39.6/35	73.6/68.5	45.1/33.3	48.2/26.2	35.4/22	62.3/53.1
VDPM [19]	60.6/43.7	54.8/27.8	83.4/80.8	73.1/38.6	75.2/55	70.2/63.5	58.9/60	55.5/25	64.3/55.5	63.8/20.2	44.7/22.5	64/44.8
V&K [20]	31/16.7	40.1/17.5	59.4/27.8	26.6/10.4	36.1/13.6	36.2/17.1	36.9/17.1	35.7/16.5	31.8/17.1	41/13.4	29.8/17	**36.8/16.7**

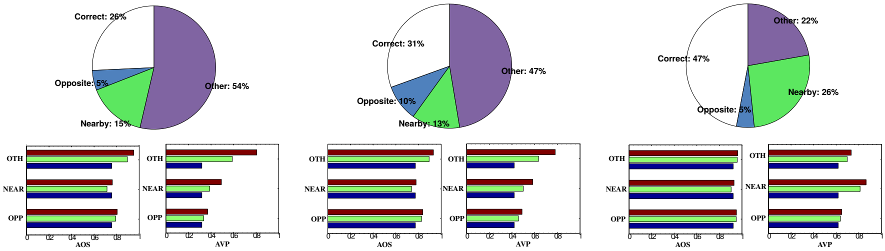
(AOS) [24], which corrects the precision with a cosine similarity term, using the difference in angle between the GT and the estimation. Finally, we report the *mean angle error* (MAE) and *median angle error* (MedError) [2, 20], which do not consider the object localization performance.

3 Diagnostic

3.1 Pose estimation performance over the GT

We start the diagnosis of the different models by analyzing their performance estimating the viewpoint of an object when the GT BBs are given. We run the models over the cropped images, using the detector scores to build the detection raking. The main results are shown in Table 1, which clearly demonstrates that the deep learning method [20] performs significantly better than all hand-crafted features based models [14, 19]. VDPM exhibits a performance, in terms of AOS and MedError, slightly better than BHF. However, BHF achieves better MAE than VDPM. If we now compare these models using AVP or PEAP, VDPM is clearly superior. This fact reveals that VDPM is able to report a higher number of accurate pose estimations. These results also reveal that AVP and PEAP are more *severe* than AOS, penalizing harder the errors on the pose estimation. This aspect makes the other metrics more appropriate than AOS to establish meaningful comparisons between different approaches. Interestingly, this conclusion is reinforced when a random pose assignment is used and evaluated in terms of AOS, see Table 1 RAND model. This approach reports a very high AOS of 52.2, compared to the 10.8 or 1.0 for AVP and PEAP, respectively.

Types of false pose estimations Figure 2 shows the frequency and impact on the performance of each type of false positive. For this figure a pose estimation is considered as: a) *correct* if its error is $< 15°$; b) *opposite* if its pose error is $> 165°$; c) *nearby* if its pose error is $\in [15°, 30°]$; d) *other* for the rest of situations. The message of this analysis is clear: errors with opposite viewpoints are not the main problem for any of the three models, being the highest confusions with *others*. However, here we show that the DPM-based methods are more likely to show opposite errors, as it has been shown in [3]. Overall, the large visual similarity between opposite views for some classes, and the unbalancedness of the training set (see Fig. 1) have a negative impact on DPM-based models.



(a) BHF [14] (b) VDPM [19] (c) V&K [20]

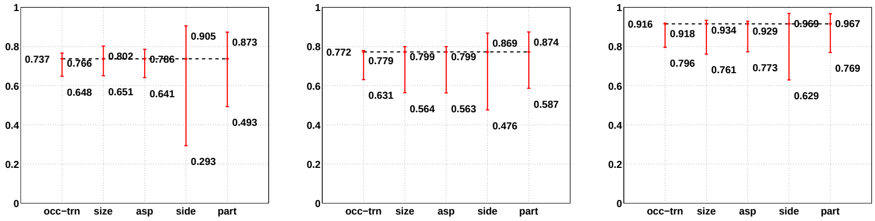
Fig. 2: Pie Chart: percentage of errors that are due to confusions with Opposite, Nearby or Other viewpoints, and Correct estimations. Bar Graphs: pose performance in terms of AOS (left) and AVP (right). Blue Bar displays the overall AOS or AVP. Green Bar displays AOS or AVP improvement by removing all confusions of one type: OTH (other errors); NEAR (nearby viewpoints); OPP (opposite viewpoints). Brown Bar displays AOS or AVP improvement by correcting all estimations of one type: OTH, NEAR or OPP.

The deep learning model, V&K [20], seems to exhibit the hights confusion between *nearby viewpoints*. This error type is above 25% for V&K, while for the other methods [14, 19] it does not exceed the 15%. This fact, a priori, may seem to be a good property of V&K, since its error distribution is concentrated on small values. However, these nearby errors are treated as false positives for the AVP and PEAP metrics, hence reducing the performance.

If we focus now the attention on the evaluation metrics, *i.e.* the bar graphs in Fig. 2, we observe that the nearby errors have a negative impact on the AOS metric. If we proceed to remove all the estimations of this type (see green bar), the performance decreases. Furthermore, if we correct these errors (see brown bar), AOS does not significantly improve for any method. In contrast, the AVP metric always improves when any error type is removed or corrected.

Impact of object characteristics Figure 3 provides a summary of the sensitivity to each characteristic and the potential impact on improving pose estimation robustness. The worst-performing and best-performing combinations for each object characteristic are averaged over the 11 categories. The difference between the best and the worst performance indicates sensitivity; the difference between the best and the overall indicates the potential impact. Figure 3 shows that the three methods are very sensitive to occlusion and truncation properties of the objects, but the impact is very small. The reduction of performances for VDPM and V&K are higher than for BHF, indicating that they perform worse for occluded objects than BHF. Remember that BHF is a part-based approach, an aspect that increases the robustness to occlusions and truncations.

All models show sensitivity to the object size. BHF is trained cropping and rescaling all training objects to the same size. Therefore, this model is not very robust to changes in the size of the test objects, but it works well with (extra)



(a) BHF [14]

(b) VDPM [19]

(c) V&K [20]

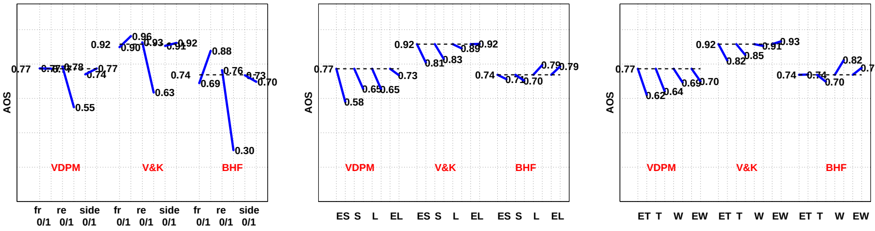
Fig. 3: Summary of Sensitivity and Impact of Object Characteristics. We show AOS of the highest performing and lowest performing subsets within each characteristic (occ-trn: occlusion/truncation, size: object size, asp: aspect ratio, side: visible sides and part: part visibility). Dashed line is overall AOS.

large objects (see Fig. 4(b)). As it is described in [20], the effect of small and extra small objects on V&K is very significant. The worst performance is exhibited by the VPDM model, which has difficulties to work with both (extra) small and (extra) large objects.

All models are sensitive to the aspect ratio of the objects. Since the mixture component concept of VDPM is closely related to the aspect ratio, this characteristic has more negative effect on this approach. V&K and VDPM present difficulties to work with tall and extra tall objects, while BHF does not (see Figure 4(c)). VDPM works poorly for wide and extra wide categories, while BHF is the only one that improves its performance working with these aspect ratios. Note that these aspect ratios are the most common on the training set (82% of the training objects).

Part visibility exhibits a very high impact (roughly 0.102 for VDPM, 0.136 for BHF and 0.051 for V&K). Due to the learning process based on local object parts, BHF is the most sensitive to this object property. In general (see Fig. 5), we have observed that the parts that are most likely to be visible, have a positive impact over the pose performance, and they present a negative effect when they are barely visible. But a high level of visibility does not imply that these parts are going to be the most discriminative. For instance, in the sofa class, the seat bottom parts (*p2* and *p3*) are the most visible, but the models are more sensitive to the back parts (*p5* and *p6*). For car, the wheels (the first 4 parts) and the frontal lights (*p9* and *p10*) are the parts least visible, but while the wheels seem not to affect the performance, the frontal lights do. There are some exceptions between the behavior of the different models. For aeroplanes, the wings (*p2* and *p5*) are not important for VDPM and V&K, while they are for BHF. BHF and V&K vary in a similar way with the parts of the diningtable.

One interpretation of our results is that the analyzed estimators do well on the most common modes of appearance (*e.g.* side or frontal views), but fail when a characteristic such as viewpoint is unusual. All models show a strong preference



(a) (b) (c)

Fig. 4: Effect of Object Characteristics. Dashed lines are overall AOS. (a) Effect of visible sides. fr: Frontal, re: Rear and side: Side. ‘1’ visible; ‘0’ no visible. (b) Effect of object sizes. (c) Effect of aspect ratios.

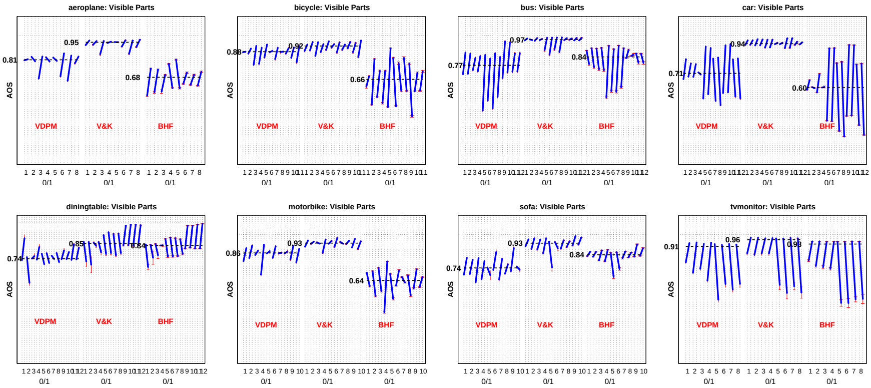


Fig. 5: Effect of Visible Parts. Visible Parts: ‘1’ = visible; ‘0’ = no visible.

for the frontal views. The main problem for all the approaches seems to be how to achieve a precise pose estimation when the rear view is visible. Overall, VDPM and V&K are more robust than BHF to the bias towards frontal viewpoints.

3.2 Simultaneous object detection and pose estimation

It is time now for our second diagnosis: joint object localization and pose estimation. Table 2 shows a detailed comparison of all the methods. Note we report now the AP for the detection, and then the AOS, AVP and PEAP metrics.

V&K [20] again reports the best average performance. Interestingly, AVP and PEAP reveal that all methods exhibit lower loss of pose estimation performance than working with GT BBs (see Tables 2 and 1). This indicates that all the models are able to report more accurate pose estimations when the detections are given by a detector approach, instead of with the GT annotations. In other words: the good pose estimations seem to be associated to *clear or easy* detections. Take

Table 2: Simultaneous object detection and pose estimation comparison.

Method	Metrics	aero	bicycle	boat	bus	car	chair	table	mbike	sofa	train	tv	AVG
BHF [14]	AP	29	28.9	3.9	35.6	15	4.1	9.5	15.6	4.2	19	8.8	15.8
	AOS	23.3	22	2.9	31.1	10.9	2.5	6.9	11.4	3.8	14.7	8.5	12.5
	AVP ($\frac{\pi}{12}$)	10.2	11.2	1.3	16.4	6.1	1.1	1.8	3.7	2	5.8	4.1	5.8
	PEAP ($\frac{\pi}{12}$)	3	3.4	0.4	6.3	2	0.2	0.3	0.9	1	2	1.8	1.9
DPM-VOC+VP [4]	AP	36	45.9	5.3	54	42.3	8.1	5.4	34.8	11	28.2	27.3	27.1
	AOS	33.7	44.3	4.1	52.3	37.4	7.6	3.9	33.5	10.7	25.2	26.7	25.4
	AVP ($\frac{\pi}{12}$)	18	26.3	2.6	51.2	32.7	5.7	2.7	20.5	7.3	22.7	19.2	19
	PEAP ($\frac{\pi}{12}$)	8.4	13.8	1.2	43.5	22	3.6	1.3	11.6	4.6	16.9	12.3	12.7
VDPM [19]	AP	42.2	44.4	6	53.7	36.5	12.7	11.2	35.5	17.1	32.7	33.6	29.6
	AOS	39.4	42.8	3	50.8	28.8	10.3	8.4	33.9	16.3	28.8	32.7	26.8
	AVP ($\frac{\pi}{12}$)	17.7	24.6	0.4	49.6	21.1	5.9	4.6	16.5	13.6	26.3	18.5	18.1
	PEAP ($\frac{\pi}{12}$)	6.7	11.5	0.04	40	10.3	2.4	1.8	7.3	9.6	19.6	10.1	10.8
R-CNN [30]	AP	72.5	68.7	34	73	62.7	33.3	36.7	70.8	50	70.1	57.2	57.2
	AOS	69.9	64.8	27.6	71.1	59.5	30.6	30.7	67.2	48.3	61.5	56.2	53.4
V&K [20]	AVP ($\frac{\pi}{12}$)	58.2	48.6	17.8	69.3	50.5	23.7	23.1	51.8	40.4	55.1	40.4	43.5
	PEAP ($\frac{\pi}{12}$)	39	29.3	8.2	59.6	34.7	14.7	12	33.5	28.8	39.1	26.5	29.6

into account that when the GT BBs are used, many difficult and truncated or occluded objects, which might have not been detected, are considered.

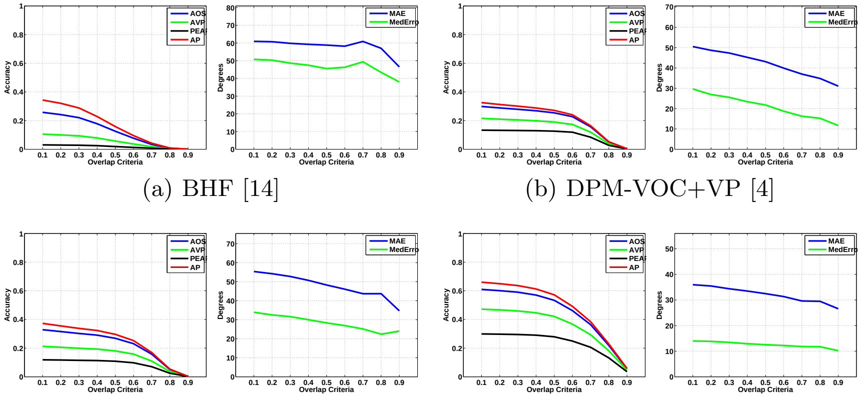
It is clearly the excellent performance of R-CNN [30] detector, which makes V&K the winner (observe the high AP for some categories). This suggests an intriguing question. Being the V&K model the only one that does not consider the localization and pose estimation jointly, is it adequate to decouple the detection and pose estimation tasks? We get back to this question in Section 4.

Note that the BHF performance for object detection is far from the state-of-the-art in the PASCAL 3D+ dataset. This conclusion is not new: already Gall *et al.* [31] manifested that Hough transform-based models struggle with the variation of the data that contains many truncated examples.

Does Table 2 offer the same conclusions we obtained before for the AOS metric and its bias towards the detection performance? First, it seems clear that AOS tends to be closer to AP than the rest of metrics. Second, while in terms of detection VDPM is better than DPM-VOC+VP, for the pose estimation task the DPM-VOC+VP is able to report a better performance, according to AVP and PEAP. Moreover, Figure 7 corroborates this fact too. However, in terms of AOS, the VDPM is better. This is a contradiction, which again reveals that AOS is *biased* towards the detection performance, while AVP and PEAP are more restrictive, penalizing harder the errors in the estimation of the poses.

Figure 6 shows an analysis of the influence of the overlap criterion in all the metrics. For this overlap criterion we follow the PASCAL VOC formulation: to be considered a true positive, the area of overlap between the predicted BB and GT BB must exceed a threshold. This figure also shows that the AOS metric is dominated by the detection, while AVP and PEAP are more independent.

This leads us to conclude that the AVP and PEAP metrics are more adequate to evaluate the performance of the models on pose estimation. We also observe that the overlap criterion can be relaxed, allowing less precise detections to be evaluated. This way we gain object hypotheses per method, and let the metrics choose which one estimates the viewpoints the best.



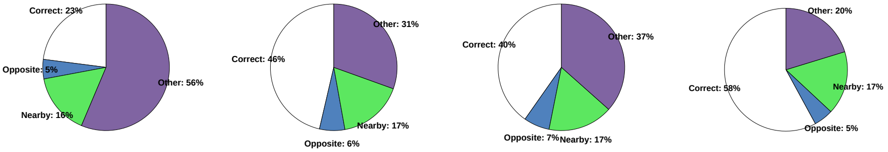
(a) BHF [14]

(b) DPM-VOC+VP [4]

(c) VDPM [19]

(d) V&K [20]

Fig. 6: Analysis of the influence of the overlap criterion in the different metrics.



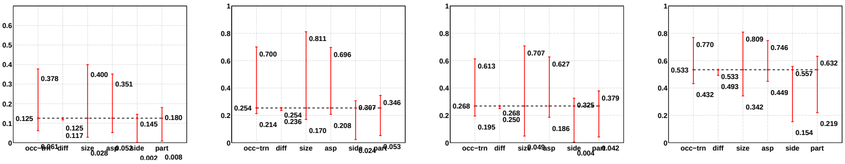
(a) BHF (b) DPMVOC+VP (c) VDPM (d) V&K

Fig. 7: False positive analysis on detection and pose estimation.

Observing the evolution of MAE and MedError, VDPM, V&K and BHF improve their performance with respect to the GT analysis. The detection seems to work as a filter stage letting pass only those candidates which are susceptible to be correctly estimated. Only V&K and BHF almost maintain these errors when the overlap criterion increases. This means that they are not sensitive to the detection accuracy. This is not the case for the DPM-based models, for which the more precise the BB localization, the better the pose estimation.

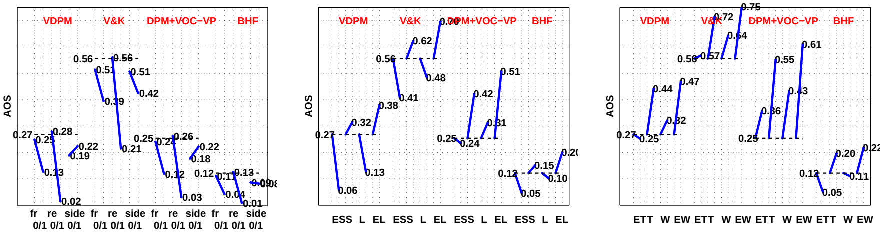
Types of false pose estimations Figure 7 shows the results corresponding to the type of false pose estimations for each method. We follow the same analysis detailed in Section 3.1. Remarkably, the detection stage has caused a decrease in the confusion with nearby viewpoints for V&K, improving the correct estimate percentage (from 47% with GT, to 58%). BHF is probably the most stable model, while VDPM is the most benefited by the detection: note that all the error percentages have been reduced. Like we said, overall, the detection stage seems to select those candidates for which the pose estimation is easy to estimate.

All methods exhibit a similar confusion with near and opposite poses. It seems that the classes with the highest opposite errors are boat, car and train. For the nearby poses, the most problematic classes are boat, motorbike and sofa.



(a) BHF (b) DPMVOC+VP (c) VDPM (d) V&K

Fig. 8: Summary of Sensitivity and Impact of Object Characteristics for simultaneous object detection and pose estimation performance.



(a) (b) (c)

Fig. 9: Effect of Object Characteristics on detection and pose estimation.

Impact of object characteristics In Figure 8 we show the impact of object characteristics for each method. Now, the size of the objects is the most influential characteristic, mainly affecting the performance of the detection (small objects are difficult to be detected – this is a common problem for all the approaches). Surprisingly, observing Figure 9(b), one can say that all methods improve their pose performances working with small or extra large objects (this does not happen in the GT analysis). This fact again reveals that the detection seems to work as a filter stage.

The second aspect which is worth analyzing is the effect of occluded/truncated objects. Now the impact of these characteristics is really considerable, compared with the numbers reported in the previous section with the GT. A conclusion is clear: if we jointly treat object localization and pose estimation tasks, more effort has to be done in order to tackle this problem.

All models are sensitive to the aspect ratio, but in this case the impact of this characteristic is reinforced by the detection. In contrast to the GT analysis, now, the models seem to prefer extra wide objects (see Fig. 9(c)). Interestingly, VDPM, V&K and BHF do not work well with tall objects when GT is used, but this seems to be solved by the detection stage. Remarkably, DPM-VOC+VP is the only one that works well with all unusual aspect ratios of the objects.

Again, the visible sides aspect is the one that most adversely affects the performance of the models. Even for the winner method, *i.e.* V&K, the accuracy

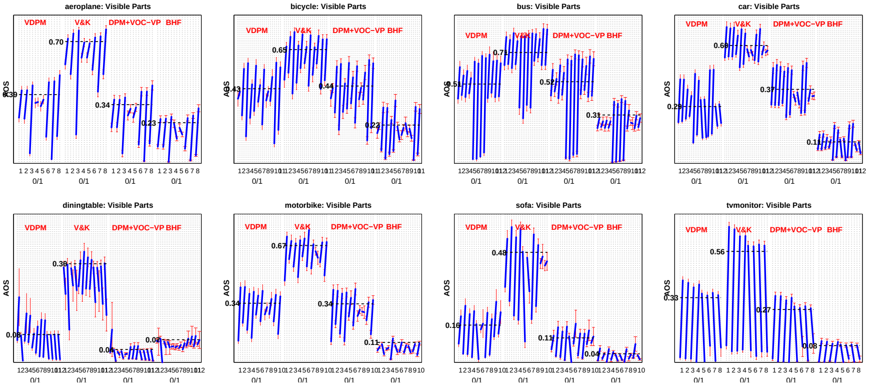


Fig. 10: Part Visibility Influence on detection and pose estimation.

dramatically drops from the average AOS of 0.533 to 0.154. From a careful inspection of the results, we conclude that the main problem for all approaches seems to be to obtain a simultaneous precise detection and quality pose estimation for the rear views of the objects (see Fig. 9(a)).

If we introduce the difficult objects in the evaluation, results show that these examples have a slight negative effect. V&K is the one which really suffers from this situation, although the AOS performance decreases just 0.04 points.

Visibility of Parts Influences Pose Estimation Performance What is the influence of the different parts for each of the models and categories? Figure 10 shows a detailed analysis for this question. The main difference with respect to the previous analysis is that now the parts play an important and different role for the detection. For instance, now, in car, wheels (the first 4 parts) are more influential than back trunks (the last 2 parts). In aeroplane, wings (*p2* and *p5*) are now not important for BHF. However, there are object parts that have the same effect as before: for instance, the models keep being very sensitive to the back part of the sofa (*p5*). The variability presented by the models towards train parts in the detection is similar to the one reported for the pose estimation analysis with GT. The class tvmonitor keeps being paradigmatic, affecting to all the approaches in a very sensitive way.

All the methods and classes exhibit a great variability, depending on the visible parts. But not all parts affect in the same way, as it has been discussed. Models vary their sensitivity according to whether they are detecting an object or estimating its pose. Therefore, we should seek parts that are very discriminative for both the object detection and pose estimation tasks.

4 Our prescription for success

We have performed a complete diagnosis for all the models. The following are the main problems identified, and our prescription for success.

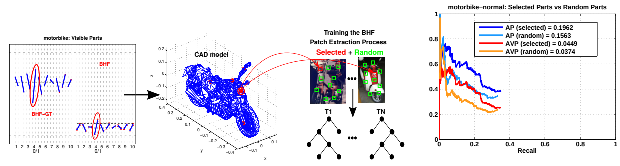
All the analyzed approaches are **biased towards the common modes of object appearance of the PASCAL 3D+ dataset**. For example, in aeroplane or tvmonitor classes, where the side-view and the frontal-view involve 62% and 63% of the objects, respectively, the models obtain for these views a pose estimation performance which is almost 10% better than the average. Furthermore, for categories that appear typically occluded in training, such as car (82% occluded/truncated) or chair (97% occluded/truncated), the models also do a good job with the slightly occluded or truncated test instances. That is, models are biased towards the data distribution used during training. To deal with the problem of scarcity of training data with viewpoint annotation, one solution could be to design models able to encode the shape and appearance of the categories with robustness to moderate changes in the viewpoint. The very recent work of [16] shows this is a promising direction, where a CNN-based architecture is trained using 3D CAD models.

HF and DPM-based models prefer balanced training sets. Other works [1, 3] have already shown how the performance of DPM-based solutions improves if the training data is sufficiently balanced and clean. One can try to artificially balance the dataset or control the number of training instances per viewpoint. Following these hints, we have completed an extra experiment. We have re-trained a BHF, but balancing the car training set. By doing so, we achieve a gain of 1% and 1.1%, for AP and AOS.

Size matters: for the detection and pose estimation tasks, all methods present **difficulties to work with (extra) small objects**. One solution could be to combine detectors at multiple resolutions (*e.g.* [32]). We also encourage to use contextual cues, which have been shown to be of great benefit for the related task of object detection. For instance, the Allocentric pose estimation model in [33] could be a good strategy to follow. This approach integrates the pose information of all the objects in a scene to improve the viewpoint estimations.

Should we decouple the pose estimation and the detection tasks? This is a fundamental question we wanted to answer in this study. In this diagnosis, only the V&K model decouples the two tasks, and it is the one systematically reporting the best performance. We could not find and analyze a deep learning approach where both problems were treated jointly. Hence we cannot provide a convincing answer to the question. However, our analysis reveals that the performance increases when the pose estimations are given over detections instead of GT bounding boxes. Figures 2 and 7 show that V&K has been able to *reduce* the large number of confusions with nearby views, simply thanks to the detection stage. This reveals that there is a correlation between easy-to-detect objects and easy-to-estimate-pose objects. Therefore we can affirm that the detection seems to help the pose estimation. Furthermore, knowing that when training and testing data belong to the same distribution, results are generally better, a good strategy could be to re-train the models, but on detected objects on the training set, *i.e.* not using GT BBs. We show the results of this extra experiment below.

What is the most convenient evaluation metric? After our diagnosis, the answer to this question is clear to us: PEAP and AVP are able to offer



(a) (b)

Fig. 11: Random Part vs. Selected Part Extractions for training a BHF.

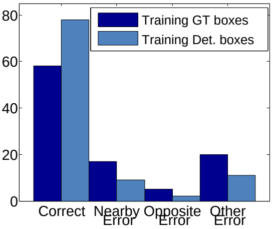


Fig. 12: Extra experiment for V&K.

more meaningful results and comparatives than AOS, which is greatly dominated by the detection performance. Both PEAP and AVP provide more information regarding the precision in pose estimations, while the localization precision is also considered.

How can we use this diagnostic tool to improve our models? The main objective of this work is that other researches can use it to improve the performance of their approaches. We provide here some examples. For instance, to improve the V&K performance, we proceed as follows. As it has been previously explained, our diagnosis indicates that the detection step improves the pose estimation performance. We propose to re-train the V&K model but with detections on the training set. First, we collect detected BBs using the R-CNN model [29] in the training images. Only those BBs satisfying the PASCAL VOC overlap criterion with respect to the annotations, with a threshold of 0.7, are selected (70% of the new training BBs). Following this strategy we achieve an improvement of 2%, 2.3% and 2.4% in terms of AOS, AVP and PEAP, respectively. Interestingly, nearby error is reduced by 8%, and correct estimations are increased by 20% (see Figure 12).

We can also improve the BHF performance. A careful inspection of the diagnosis for BHF shows that it exhibits the highest sensitivity to the visibility of object parts. For instance, for the class motorbike, see Fig. 11(a), there is a specific part (the headlight) that is very discriminative. BHF is normally trained performing a random extraction of image patches from the training images. If, instead of this random patch extraction, we check whether this part is visible and extract a patch centered at its annotated position, we get an increase of 4% for the AP, while the AVP increases from 0.037 to 0.045 (see Figure 11(b)).

Conclusion We hope that our work will inspire research that targets and evaluates reduction in specific error modes on object detection and pose estimation. Our tool is publicly available giving other researches the opportunity to perform similar analysis with other pose estimation methods working on the PASCAL 3D+ dataset.

Acknowledgements. This work is supported by projects DGT SPIP2015-01809, MINECO TEC2013-45183-R, FWO G069612N (Representations and algorithms for the captation, visualization and manipulation of moving 3D objects), Nissan (1188371-1-UDARQ).

References

1. Ghodrati, A., Pedersoli, M., Tuytelaars, T.: Is 2D information enough for viewpoint estimation? In: BMVC. (2014)
2. Glasner, D., Galun, M., Alpert, S., Basri, R., Shakhnarovich, G.: Viewpoint-aware object detection and continuous pose estimation. Image and Vision Computing **30** (2012) 923–933
3. Lopez-Sastre, R.J., Tuytelaars, T., Savarese, S.: Deformable part models revisited: A performance evaluation for object category pose estimation. In: ICCV 2011, 1st IEEE Workshop on Challenges and Opportunities in Robot Perception. (2011)
4. Pepik, B., Stark, M., Gehler, P., Schiele, B.: Teaching 3D geometry to deformable part models. In: CVPR. (2012)
5. Pepik, B., Stark, M., Gehler, P., Ritschel, T., Schiele, B.: 3D object class detection in the wild. In: Workshop on 3D from a Single Image (3DSI) (in conjunction with CVPR'15). (2015)
6. Sun, M., Su, H., Savarese, S., Fei-Fei, L.: A multi-view probabilistic model for 3D object classes. In: CVPR. (2009)
7. Fanelli, G., Gall, J., Van Gool, L.: Real time head pose estimation with random regression forests. In: CVPR. (2011)
8. Fenzi, M., Ostermann, J.: Embedding geometry in generative models for pose estimation of object categories. In: BMVC. (2014)
9. Gu, C., Ren, X.: Discriminative mixture-of-templates for viewpoint classification. In: ECCV. (2010)
10. Liebelt, J., Schmid, C.: Multi-view object class detection with a 3D geometric model. In: CVPR. (2010)
11. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Schiele, B., Van Gool, L.: Towards multi-view object class detection. In: CVPR. Volume 2. (2006) 1589–1596
12. Savarese, S., Fei-Fei, L.: 3D generic object categorization, localization and pose estimation. In: ICCV. (2007) 1 – 8
13. Ozuysal, M., Lepetit, V., Fua, P.: Pose estimation for category specific multiview object localization. In: CVPR. (2009)
14. Redondo-Cabrera, C., Lopez-Sastre, R.J.: Because better detections are still possible: Multi-aspect object detection with boosted hough forest. In: BMVC. (2015)
15. Roman, J., Adam, H., Markata, D., Pavel, Z.: Real-time pose estimation piggybacked on object detection. In: ICCV. (2015)
16. Su, H., Qi, C.R., Li, Y., Guibas, L.J.: Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3d model views. In: ICCV. (December 2015)
17. Zia, Z., Stark, M., Schiele, B., Schindler, K.: Detailed 3D representations for object recognition and modeling. PAMI **35**(11) (2013) 2608–2623
18. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: ECCV. (2012)
19. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3D object detection in the wild. In: IEEE Winter Conference on Applications of Computer Vision. (2014)
20. Tulsiani, S., Malik, J.: Viewpoints and keypoints. In: CVPR. (2015)
21. Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A., Hebert, M.: An empirical study of context in object detection. In: CVPR. (2009)
22. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC)challenge. IJCV **88**(2) (2010) 303–338

23. Pepik, B., Benenson, R., Ritschel, T., Schiele, B.: What is holding back convnets for detection? In: GCPR. (2015)
24. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: CVPR. (2012)
25. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html
26. Deng, J., Dong, W., Socher, R., Li, L.J., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
27. Maji, s., Malik, J.: Object detection using a max-margin hough transform. In: CVPR. (2009)
28. Razavi, N., Gall, J., Van Gool, L.: Scalable multi-class object detection. In: CVPR. (2011)
29. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NIPS. (2015)
30. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR. (2014)
31. Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V.: Hough forests for object detection, tracking, and action recognition. PAMI **33** (2011) 2188–2202
32. Park, D., Ramanan, D., Fowlkes, C.: Multiresolution models for object detection. In: ECCV. (2010)
33. Oramas-Mogrovejo, J., De Raedt, L., Tuytelaars, T.: Allocentric pose estimation. In: ICCV. (2013)