Recovering 6D Object Pose: Multi-modal Analyses on Challenges

Caner Sahin
Imperial College London
c.sahin14@imperial.ac.uk

Tae-Kyun Kim
Imperial College London
tk.kim@imperial.ac.uk

Abstract

*A large number of studies analyse object detection and pose estimation at visual level in 2D, discussing the effects of challenges such as occlusion, clutter, texture, etc., on the performances of the methods, which work in the context of RGB modality. Interpreting the depth data, the study in this paper presents thorough multi-modal analyses. It discusses the above-mentioned challenges for full 6D object pose estimation in RGB-D images comparing the performances of several 6D detectors in order to answer the following questions: What is the current position of the computer vision community for maintaining "*automation*" in robotic manipulation? What next steps should the community take for improving "*autonomy*" in robotics while handling objects?*

Direct comparison of the detectors is difficult, since they are tested on multiple datasets with different characteristics and are evaluated using widely varying evaluation protocols. To deal with these issues, we follow a threefold strategy: five representative object datasets, mainly differing from the point of challenges that they involve, are collected. Then, two classes of detectors are tested on the collected datasets. Lastly, the baselines' performances are evaluated using two different evaluation metrics under uniform scoring criteria. Regarding the experiments conducted, we analyse our observations on the baselines along with the challenges involved in the interested datasets, and we suggest a number of insights for the next steps to be taken, for improving the autonomy in robotics.

1. Introduction

Object detection and pose estimation is an important problem in the realm of computer vision, for which a large number of solutions have been proposed. One line of the solutions is based on visual perception in RGB channel. Existing evaluation studies [18, 19] addressing this line of the solutions discuss the effects of challenges, such as occlusion, clutter, texture, *etc*, on the performances of the methods, which are mainly evaluated on large-scale datasets, *e.g.*, ImageNet [22], PASCAL [23]. These studies have made important inferences for generalized object detection, however, the discussions have been restricted to visual level in 2D, since the interested methods are designed to work in the context of RGB modality.

Figure 1: Benchmarks collected mainly differ from the point of challenges that they involve. Row-wise, the 1^{st} benchmark concerns texture-less objects at varying viewpoint with cluttered background, the 2^{nd} is interested in multi-instance, the 3^{rd} has scenes with severely occluded objects, the 4^{th} reflects the challenges found in bin-picking scenarios, and the 5^{th} is related to similar-looking distractors.

Increasing ubiquity of Kinect-like RGB-D sensors has prompted an interest in full 6D object pose estimation. Interpreting the depth data, state-of-the-art approaches for object detection and 6D pose estimation [1, 3, 11] report improved results tackling the aforesaid challenges in 6D. This improvement is of great importance to many higher level

tasks, *e.g.*, scene interpretation, augmented reality, and particularly, to robotic manipulation.

Robotic manipulators that pick and place the goods from conveyors, shelves, pallets, *etc.*, can facilitate several processes comprised within logistics systems, *e.g.*, warehousing, material handling, packaging. Amazon Picking Challenge (APC) [5] is an important example demonstrating the promising role of robotic manipulation for the facilitation of such processes. APC integrates many tasks, such as mapping, motion planning, grasping, object manipulation, *etc.*, with the goal of "*autonomously*" moving items by robotic systems from a warehouse shelf into a tote [4, 6]. Regarding the "*automated*" handling of items by robots, accurate object detection and 6D pose estimation is an important task that when successfully performed improves the autonomy of the manipulation. Within this context, we ask the following questions. What is the current position of the computer vision community for maintaining automation in robotic manipulation, with respect to the accuracy of the 6D detectors introduced? What next steps should the community take for improving the autonomy in robotics while handling objects? We aim at answering these questions performing multi-modal analyses for object detection and 6D pose estimation where we compare two baselines regarding the challenges involved in the interested datasets.

Direct comparison of the baselines is difficult, since they are tested on samples which are collected at non-identical scenarios by using RGB-D sensors with different characteristics. Additionally, different evaluation criteria are utilized for performance measure. In order to address such difficulties, we follow a threefold strategy: we firstly collect five representative object datasets [1, 3, 7, 11, 35] (see Fig. 1). Since the ground truth annotations are crucial for measuring the performance of a baseline, the uniformity of the datasets is provided by removing the test scenes where objects of interest are wrongly annotated. Then, we identify two baselines [1, 3]. Once we define the optimum parameters of each baseline employing control experiments, we test those on the collected datasets. Lastly, we evaluate the baselines' performance using two different evaluation metrics [1, 36] under uniform scoring criteria. Regarding the experiments conducted, we analyse our observations on the baselines and elaborate the analyses utilizing available results presented in the literature. We offer a number of insights for the next steps to be taken, for improving the autonomy in robotics. To summarize, our main contributions are as follows:

- This is the first time, the current position of the field is analysed regarding object detection and 6D pose estimation.
- We collect five representative publicly available datasets. The uniformity of the datasets is provided by removing the test scenes where objects of interest are wrongly annotated. In total, there are approximately 50 different object classes. We test two classes of 6D detectors on the collected datasets and evaluate the performances of the detectors using two different metrics.
- We discuss baselines' strength and weakness with respect to the challenges involved in the interested RGB-D datasets. We identify the next steps for improving the robustness of the detectors, and for improving the autonomy in robotic applications, consequently.

Related Work. Methods producing 2D bounding box hypotheses in color images [25, 26, 29, 28, 31, 27, 32, 33, 34, 30] form one line of the solutions for object detection and pose estimation. Evaluation studies interested in this line of the solutions mainly analyse the performances of the methods regarding the challenges involved within the datasets [22, 23], on which the methods have been tested. In [17], the effect of different context sources, such as geographic context, object spatial support, *etc.*, on object detection is examined. Hoiem et al. [18] evaluate the performances of several baselines on PASCAL dataset particularly analysing the reasons why false positives are hypothesised. Since there are less number of object categories in PASCAL dataset, Russakovsky et al. [19] use ImageNet in order to do meta-analysis, and to examine the influences of color, texture, *etc.*, on the performances of object detectors. Torralba et al. [20] compares several datasets regarding the involved samples, cross-dataset generalization, and relative data bias, *etc*. Recently published retrospective evaluation [24] and benchmarking [21] studies perform the most comprehensive analyses on 2D object localization and category detection, by examining the PASCAL Visual Object Classes (VOC) Challenge, and the ImageNet Large Scale Visual Recognition Challenge, respectively. These studies introduce important implications for generalized object detection, however, the discussions are restricted to visual level in 2D, since the concerned methods are engineered for color images. In this study, we target to go beyond visual perception and extend the discussions on existing challenges to 6D, interpreting depth data.

2. Datasets

Every dataset used in this study is composed of several object classes, for each of which a set of RGB-D test images are provided with ground truth 6D object poses. The collected datasets mainly differ from the point of the challenges that they involve (see Table 1).

Viewpoint (VP) + Clutter (C). Every dataset involves the test scenes in which objects of interest are located at *varying viewpoints* and *cluttered backgrounds*.

VP + C + Texture-less (TL). Test scenes in the LINEMOD [1] dataset involve *texture-less* objects at varying viewpoints

Table 1: Datasets collected: each dataset shows different characteristics mainly from the challenge point of view (VP: viewpoint, O: occlusion, C: clutter, SO: severe occlusion, SC: severe clutter, MI: multiple instance, SLD: similar looking distractors, BP: bin picking).

Dataset	Challenge	# Obj. class	Modality	# Total Frame	Obj. dist. [mm]
LINEMOD	VP + C + TL	15	RGB-D	15770	600-1200
Multiple-Instance (MULT-I)	VP + C + TL + O + MI	6	RGB-D	2067	600-1200
Occlusion (OCC)	VP + C + TL + SO	8	RGB-D	9209	600-1200
Bin-Picking (BIN-P)	VP + SC + SO + MI + BP	2	RGB-D	180	600-1500
T-LESS	VP + C + TL + O + MI + SLD	30	RGB-D	10080	600-1500

with cluttered backgrounds. There are 15 objects, for each of which more than 1100 real images are recorded. The sequences provide views from 0 - 360 degree around the object, 0 - 90 degree tilt rotation, ∓ 45 degree in-plane rotation, and 650 mm - 1150 mm object distance.

VP + C + TL + Occlusion (O) + Multiple Instance (MI). Occlusion is one of the main challenges that makes the datasets more difficult for the task of object detection and 6D pose estimation. In addition to close and far range 2D and 3D clutter, testing sequences of the Multiple-Instance (MULT-I) dataset [3] contain *foreground occlusions* and *multiple object instances*. In total, there are approximately 2000 real images of 6 different objects, which are located at the range of 600 mm - 1200 mm. The testing images are sampled to produce sequences that are uniformly distributed in the pose space by $[0° - 360°]$, $[-80° - 80°]$, and $[-70° - 70°]$ in the yaw, roll, and pitch angles, respectively.

VP + C + TL + Severe Occlusion (SO). Occlusion, clutter, texture-less objects, and change in viewpoint are the most well-known challenges that could successfully be dealt with the state-of-the-art 6D object detectors. However, *heavy existence* of these challenges severely degrades the performance of 6D object detectors. Occlusion (OCC) dataset [11] is one of the most difficult datasets in which one can observe up to $70 - 80\%$ occluded objects. OCC includes the extended ground truth annotations of LINEMOD: in each test scene of the LINEMOD [1] dataset, various objects are present, but only ground truth poses for one object are given. Brachmann et al. [11] form OCC considering the images of one scene (benchvise) and annotating the poses of 8 additional objects.

VP + SC + SO + MI + Bin Picking (BP). In *bin-picking* scenarios, multiple instances of the objects of interest are arbitrarily stocked in a bin, and hence, the objects are inherently subjected to severe occlusion and severe clutter. Bin-Picking (BIN-P) dataset [7] is created to reflect such challenges found in industrial settings. It includes 183 test images of 2 textured objects under varying viewpoints.

VP + C + TL + O + MI + Similar Looking Distractors (SLD). *Similar-looking distractor(s)* along with similar looking object classes involved in the datasets strongly confuse recognition systems causing a lack of discriminative selection of shape features. Unlike the above-mentioned datasets and their corresponding challenges, the T-LESS [35] dataset particularly focuses on this problem. The RGB-D images of the objects located on a table are captured at different viewpoints covering 360 degrees rotation, and various object arrangements generate occlusion. Out-of-training objects, similar looking distractors (planar surfaces), and similar looking objects cause 6 DoF methods to produce many false positives, particularly affecting the depth modality features. T-LESS has 30 texture-less industry-relevant objects, and 20 different test scenes, each of which consists of 504 test images.

Efforts on the Benchmarks. Once the benchmarks are collected, wrongly annotated object samples are identified, and the test images involving these samples are removed. Since the reference coordinate frame of every benchmark is different from each other, the uniformity in between the benchmarks is obtained by applying the required transformations to the test images.

3. Baselines

State-of-the-art baselines for 6D object pose estimation address the challenges studied in Sect. 2, however, the architectures used differ between the baselines. In this section, we analyse 6D object pose estimators architecture-wise.

Template-based approaches: Template-based approaches, matching global descriptors of objects to the scene, are one of the most widely used approaches for object detection tasks, since they do not require time-consuming training effort. Linemod [1], being at the forefront of object detection research, estimates cluttered object's 6D pose using color gradients and surface normals. It is improved by discriminative learning in [13]. Fast directional chamfer matching (FDCM) [37] is used in robotics applications.

Learning-based methods: These methods are in need of training sessions where training samples along with the ground truth annotations are learnt. We subcategorize these methods regarding the features utilized:

Custom-designed features. Surface normals and color gradients, proposed in [1], are utilized in a part-based approach [3] in order to provide robustness across occlusion. In [11], contextual information of the objects is encoded with simple depth and RGB pixels, and the confidence of a pose hypothesis is improved using a Ransac-like algorithm. An analysis-by-synthesis approach [14] and an uncertainty-

driven methodology [9] are build upon the architecture provided in [11]. The method presented in [10] formulates the recognition problem globally and derives occlusion aware features computing a set of principal curvature ratios for all pixels in depth images. The depth-based architecture in [8, 41] initially estimates coarse 6D pose of an object, and then it iteratively refines the confidence of the estimation due to the extraction of more discriminative control point descriptors.

Learning deep features. Current paradigm in the community is to learn deep discriminative feature representations. Wohlhart et al. [16] utilize a CNN structure to learn discriminative descriptors and then pass the learnt descriptors to a Nearest Neighbor classifier in order to find the closest object pose. Although promising, this method has one main limitation, which is the requirement of background images during training along with the ones holistic foreground, thus making its performance dataset-specific. The studies in [7, 12] learn deep representation of parts in an unsupervised fashion using only foreground images. The features extracted in the course of the test are fed into a Hough forest in [7], and into a codebook of pre-computed synthetic local object patches in [12] in order to hypothesise object 6D pose.

According to the categorization presented, we identify the studies in [1, 3] as our baselines to test on the collected datasets.

4. Evaluation Metrics

Several evaluation metrics are proposed for measuring the performance of a 6D detector. Average Distance (AD) [1] outputs the score ω that calculates the distance between ground truth and estimated poses of a test object using its model. Hypotheses ensuring the following inequality is considered as correct:

$$\omega \leq z_{\omega}\Phi \tag{1}$$

where Φ is the diameter of the 3D model of the test object, and z_{ω} is a constant that determines the coarseness of an hypothesis which is assigned as correct. Translational and rotational error function [2], being independent from the models of objects, measures the correctness of an hypothesis according to the followings: i) $\mathcal{L}_2$ norm between the ground truth and estimated translations, ii) the angle computed from the axis-angle representation of ground truth and estimated rotation matrices.

Visible Surface Discrepancy (VSD) has recently been proposed to eliminate ambiguities arising from object symmetries and occlusions [36]. The model of an object of interest is rendered at both ground truth and estimated poses, and their depth maps are intersected with the test image itself in order to compute the visibility masks. Comparing the generated masks, the score normalized in $[0-1]$ determines whether an estimation is correct, according to the pre-defined thresholds.

In this study, we employ a twofold evaluation strategy for the 6D detectors using both AD and VSD metrics: i) Recall. The hypotheses on the test images of every object are ranked, and the hypothesis with the highest weight is selected as the estimated 6D pose. Recall value is calculated comparing the number of correctly estimated poses and the number of the test images of the interested object. ii) F1 scores. Unlike recall, all hypotheses are taken into account, and F1 score, the harmonic mean of precision and recall values, is presented.

Implementation Details. We test the baselines presented in [1] and [3] using in-house implemented versions. In our implementations, the color gradients and surface normal features, presented in [1], are computed using the built-in functions and classes provided by OpenCV. The features in Latent-Class Hough Forest (LCHF) [3] are the part-based version of the features introduced in [1]. Hence, we inherit the classes given by OpenCV in order to generate part-based features used in LCHF. We train each method for the objects of interest by ourselves, and using the learnt classifiers, we test those on all datasets. Note that, the methods use only foreground samples during training/template generation.

5. Multi-modal Analyses

In this section, we present a series of analyses for object detection and 6D pose estimation. Our analyses are based on the metrics used during the evaluation of the baselines. In this section, "LINEMOD" refers to the dataset, whilst "Linemod" is used to indicate the baseline itself.

5.1. Analyses Based on Average Distance

Utilizing the AD metric, we compare the chosen baselines along with the challenges, i) regarding the recall values that each baseline generates on every dataset, ii) regarding the F1 scores. In all experiments we conduct, the coefficient z_{ω} is set to the value of 0.15. In case we use different thresholds, we will specifically indicate in the related parts.

5.1.1 Recall-only Discussions

The bar charts in Fig. 2 show the superior performance of the LCHF algorithm [3] over the Linemod [1] detector. On the average, LCHF produces 80%, 72%, 53%, and 77% recall, whilst Linemod performs 63%, 45%, 26%, and 62%, on the LINEMOD, MULT-I, OCC, and BIN-P datasets, respectively.

Architectures: Linemod formulates the detection problem globally representing the windows extracted from RGB and depth images by the surface normals and color gradients features. Distortions along the object borders arising from occlusion and clutter, that is, the distortions of the color gradient and surface normal information in the test

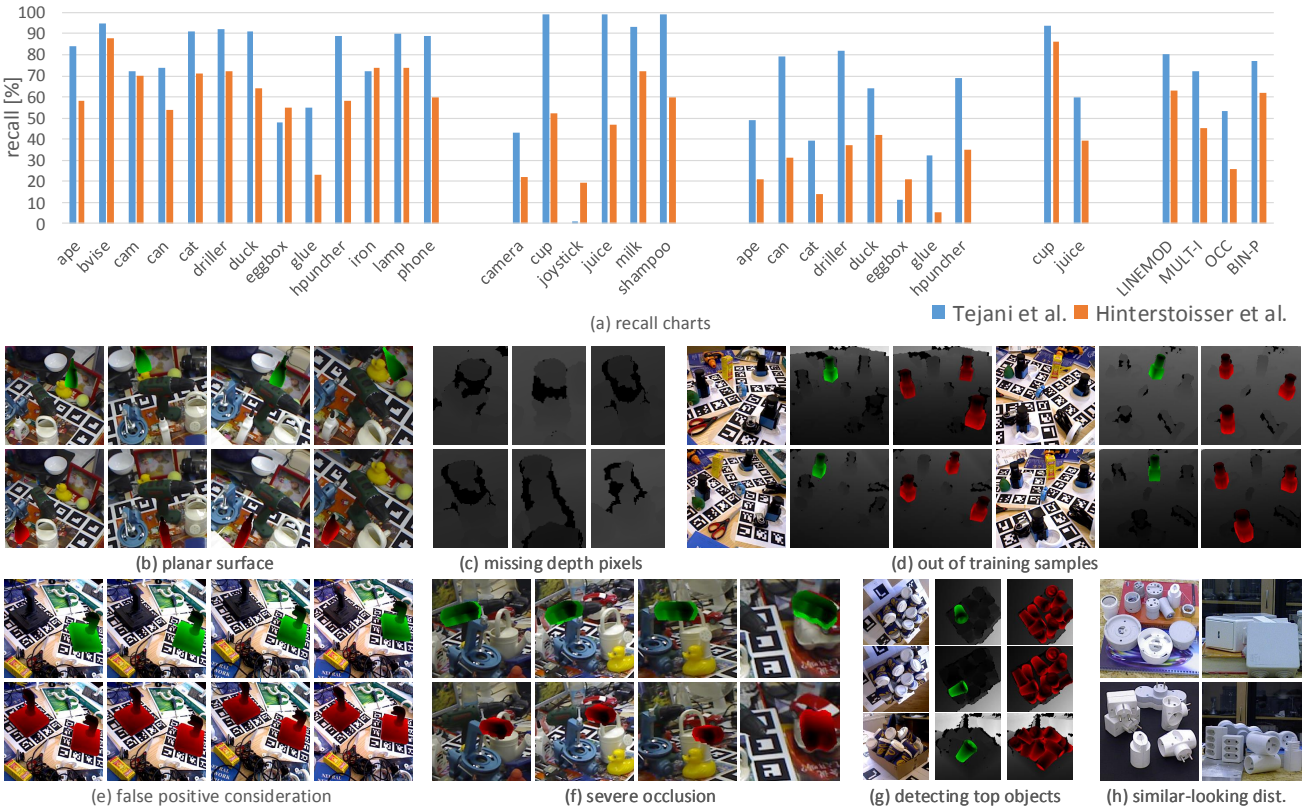

Figure 2: Baselines are compared using Average Distance (AD) with respect to recall values. (a) The performances of the baselines are depicted object-wise: from left to right, LINEMOD, MULT-I, OCC, and BIN-P. The rightmost bar chart demonstrates the success of each baseline on every dataset, averaging the recall values of individual objects. (b)-(h) challenges encountered during test are exemplified (green renderings are hypotheses, and the red ones are ground truths).

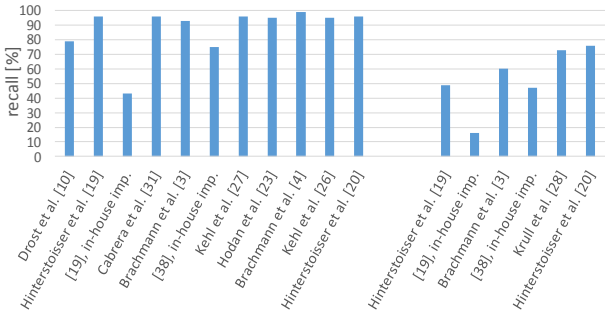

Figure 3: Performance of the state-of-the-art on the "LINEMOD" (left) and the "OCC" (right) datasets are compared with respect to the recall values, which are evaluated using AD when $z_\omega = 0.10$.

processes, mainly degrade the performance of this detector. Several imperfections of depth sensors, such as missing depth values, noisy measurements, *etc.*, impair the surface normal representations at depth discontinuities, causing extra degradation in the detector's performance. Despite the fact that LCHF uses the same kinds of features as in Linemod, LCHF detects objects extracting parts, thus making the method more robust to occlusion and clutter. However, it still suffers from missing depth pixels as in Linemod. Linemod matches the features extracted during test to a set of templates, and hence, it cannot easily be generalized well to unseen ground truth annotations, that is, the translation and rotation parameters in our case. Unlike Linemod, LCHF is based on random forests. Randomisation embedded in LCHF provides good generalisation performance on new unseen samples.

Challenges: We analyse the baselines along with the challenges involved in the interested datasets.

Clutter, Viewpoint, Texture-less objects. Both methods perform well on the LINEMOD dataset across the challenges, texture-less objects, varying viewpoint, and clutter. LCHF detects more than half of the objects with over 80% accuracy, whilst this ratio is approximately 60% for the Linemod detector. LCHF worst performs on "eggbox" and "glue", since these objects have planar surfaces, which confuses the features extracted in depth channel (example images are given in Fig. 2 (b)).

Occlusion. In addition to the challenges involved in LINEMOD, occlusion is introduced in MULT-I. Linemod's performance decreases, since occlusion affects holistic fea-

ture representations in color and depth channels. LCHF performs better on this dataset. Since the algorithm is trained using the parts coming from positive training images, it can easily handle occlusion, using the information acquired from occlusion-free parts of the target objects. However, LCHF degrades on "camera". In comparison with the other objects in the dataset, "camera" has relatively smaller dimensions. In most of the test images, there are non-negligible amount of missing depth pixels (Fig. 2 (c)) along the borders of this object, and thus confusing the features extracted in depth channel. In such cases, LCHF is liable to detect similar-looking out of training objects and generate many false positives (see Fig. 2 (d)). The hypotheses produced by LCHF for "joystick" are all considered as false positive (Fig. 2 (e)). When we re-evaluate the recall that LCHF produces on the "joystick" object setting z_ω to the value of 0.20, we observe 89% accuracy.

Severe Occlusion. OCC involves challenging test images where the objects of interest are cluttered and severely occluded. This benchmark clearly demonstrates the superiority of part-based approaches over holistic ones. Linemod has 26% accuracy on the average, whilst LCHF is approximately 50% accurate. It is worth discussing the accuracy produced by LCHF. Despite the fact that the distinctive feature of this benchmark is the existence of "severe occlusion", there are occlusion-free target objects in several test images. In case the test images of a target object include unoccluded and/or naively occluded samples (with the occlusion ratio up to $40\% - 50\%$ of the object dimensions) in addition to severely occluded samples, LCHF can produce recall values over 60% (*e.g.* "can, driller, duck, holepuncher"). On the other hand, when the target object has additionally other challenges such as planar surfaces, LCHF's performance decreases (*e.g.* "eggbox", Fig. 2 (f)).

Severe Clutter. In addition to the challenges discussed above, BIN-P inherently involves severe clutter, since it is designed for bin-picking scenarios. According to the recall values presented in Fig. 2, LCHF performs 15% better than Linemod, and both methods demonstrate better recognition rates on the "cup" object than "juice". Despite having severely occluded target objects in this dataset, there are unoccluded/relatively less occluded objects at the top of the bin. Since our current analyses are based on the top hypothesis of each method, the produced success rates show that the methods can recognize the objects located on top of the bin with reasonable accuracy (Fig. 2 (g)).

Similar-Looking Distractors: We test both Linemod and LCHF on the T-LESS dataset. Since most of the time the algorithms fail, we do not report quantitative analyses, instead we discuss our observations from the experiments. The dataset involves various object classes with strong shape and color similarities. When the background color is different than that of the objects of interest, color gradient features are successfully extracted. However, the scenes involve multiple instances, multiple objects similar in shape and color, and hence, the features queried exist in the scene at multiple locations. The features extracted in depth channel are also severely affected from the lack of discriminative selection of shape information. When the objects of interest have planar surfaces, the detectors cannot easily discriminate foreground and background in depth channel, since these objects in the dataset are relatively smaller in dimension (see Fig. 2 (h)).

We next elaborate our discussions on the challenges utilizing the results available in the literature. In order to provide the uniformity, we re-evaluate the results of the in-house implemented algorithms setting the threshold z_ω to the value of 0.10, since the methods in the literature reported the recall values using this threshold. Figure 3 compares the 6D detectors introduced in [1, 3, 13, 11, 15, 9, 12, 38, 39, 40] on the LINEMOD, and the 6D methods presented in [1, 3, 11, 14, 40] on the OCC datasets. According to the bar chart, satisfactorily accurate results are acquired when the target objects are at varying viewpoints with cluttered background. It is shown that template-based methods [1, 13, 15, 39] and random forest based learning algorithms [3, 11, 9] underlie the 6D problem. Recent trend in the community is to learn deep discriminative feature representations. Kehl et al. [12], using deep features, report the state-of-the-art results tackling the core challenges of the problem. Whilst these architectures are designed to utilize information both from RGB and depth modalities, point-to-point techniques, relying on depth cameras, build point-pair features for sparse representations of the test and model point sets. The approach presented by Drost et al. [38] has recently been improved in [40] proposing a novel voting scheme and making the algorithm more robust across sensor noise and background clutter. Although promising, the introduction of heavy occlusion degrades the performance of the methods. The average success rate of all methods on the OCC dataset is approximately 40%. One important point to be discussed is that, most of the state-of-the-art baselines addressing severe occlusion are part-based learning algorithms.

Robotic manipulators that pick and place the items from conveyors, shelves, pallets, *etc.*, need to know the pose of one item per RGB-D image, even though there might be multiple items in its workspace. Hence our recall-only analyses mainly target to solve the problems that could be encountered in such cases. Based upon the analyses currently made, one can make important implications, particularly from the point of the performances of the detectors. On the other hand, recall-based analyses are not enough to illustrate which dataset is more challenging than the others. This is especially true in crowded scenarios where multiple instances of target objects are severely occluded and

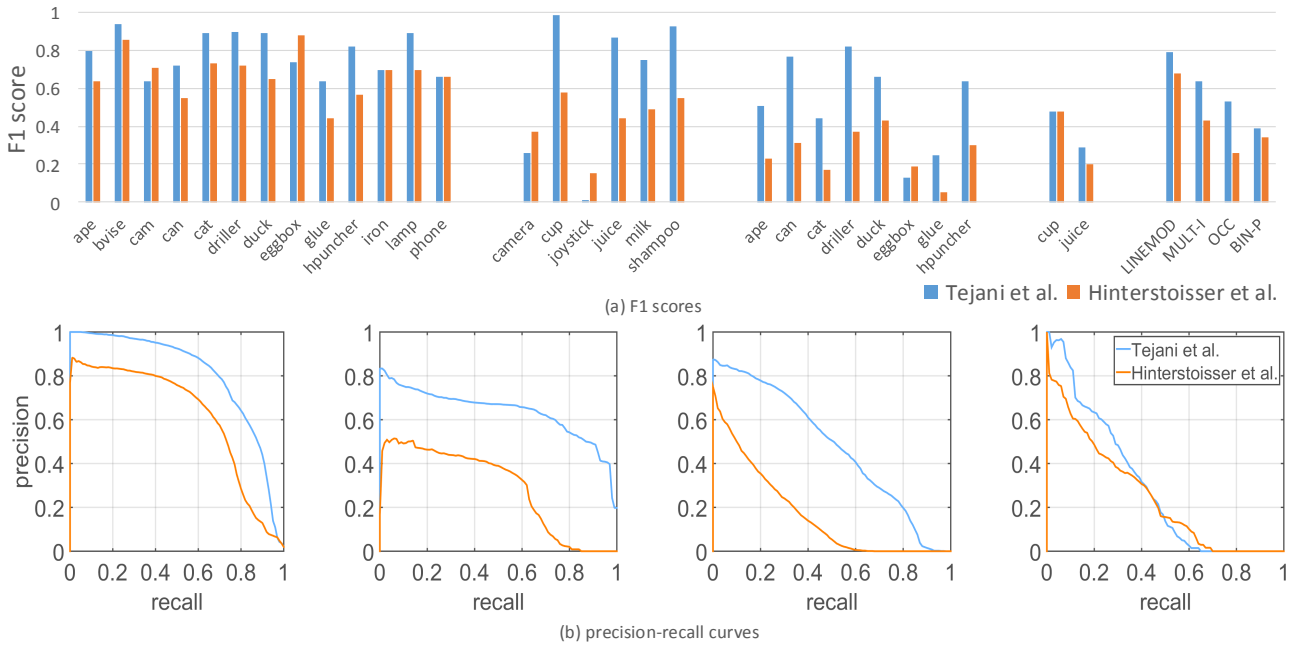

Figure 4: We compare baselines regarding F1 scores calculated using Average Distance: (a) first four bar charts illustrate the performances of the 6D detectors object-wise, on LINEMOD, MULT-I, OCC, and BIN-P, respectively. The rightmost bar graph shows the average. (b) The corresponding precision-recall curves of averaged F1 scores are shown: from left to right, LINEMOD, MULT-I, OCC, BIN-P.

cluttered. Therefore, in the next part, we discuss the performances of the baselines from another aspect, regarding precision-recall curves and F1 scores, where we evaluate the 6D detectors sorting all detection scores across all images.

5.1.2 Precision-Recall Discussions

We compute the F1 scores for both baselines and report in Fig. 4. On the average, LCHF produces 0.79, 0.64, 0.53, and 0.39, whilst Linemod perform 0.68, 0.43, 0.26, and 0.34 on LINEMOD, MULT-I, OCC, and BIN-P, respectively. In our precision-recall analyses, we compare the computed F1 scores with the success rate of the baselines determined with respect to the recall values.

The comparison between Fig. 2 (a) and Fig. 4 (a) reveals that the results produced by both methods have approximately the same characteristics on three datasets, LINEMOD, MULT-I, and OCC. They perform better on the LINEMOD dataset than OCC with respect to the recall values and F1 scores. However, LCHF slightly degrades on MULT-I when we evaluate its performance using F1 score. This is due to the further introduction of occlusion arising from the consideration of multiple instances of the objects of interest. The most important difference is observed on the BIN-P dataset. While the success rates of the detectors on this dataset are higher than 60% with respect to the recall values, according to the presented F1 scores, their performance are less than 40%. When we take into account all hypotheses and the challenges particular to this dataset, which are severe occlusion and severe clutter, we observe strong degradation in the accuracy of the detectors.

In Fig. 4 (b), we lastly report precision-recall curves that correspond to the bar chart on the top-right of the figure. Regarding these curves, one can observe that as the datasets are getting more difficult, from the point of challenges involved, the methods produce less accurate results.

Effect of refinement. The numbers depicted in Fig. 2 and Fig. 4 regarding MULT-I are based on the refined version of this dataset. Figure 5 compares the baselines' accuracy before and after the refinement with respect to the F1 scores. According to the areas under the curves, LCHF demonstrates approximately the same results, whilst Linemod degrades on the average after the refinement. Based on this outcome, one can discuss the Linemod baseline: it can localize the objects of interest, however, from the fine pose estimation point of view, it produces less accurate results. When the hypotheses of this baseline are evaluated by another metric, *e.g.*, 3D bounding box error, higher F1 scores could be reached.

Beyond localizing objects, accurate Euler angles of the target objects are the required inputs for robotic systems, more particularly for the robotic arms that have hand-like graspers as end-effectors. Since such graspers have high degrees of freedom, during the hand-object interaction under uncontrolled conditions, fine Euler parameters help the robotic systems to plan more robust paths, and mapping. On

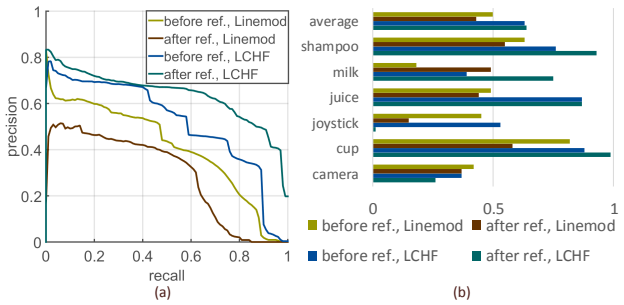

Figure 5: Effect of refinement on the MULT-I dataset is shown. (a) Precision-Recall curves before and after refinement. (b) Object-wise F1 scores.

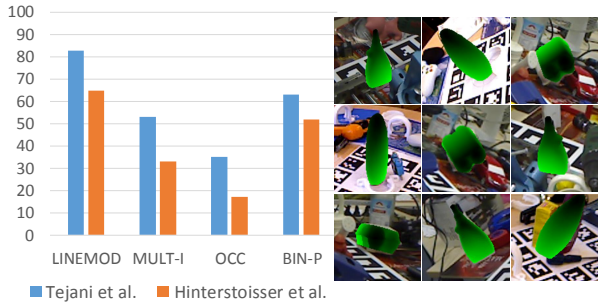

Figure 6: Methods are evaluated based on Visible Surface Discrepancy. Samples on the right are considered as false positive with respect to Average Distance, whilst VSD deems correct.

the other hand, objects' shape can relax this requirement up to a certain degree in cases the objects are symmetric.

5.2. Analyses Based on Visible Surface Discrepancy

The analyses presented so far have been employed using the AD metric. We continue our discussions computing the recall values using the VSD metric, which is inherently proposed for tackling the pose-ambiguities arising from symmetry. We set δ, τ, and t, the thresholds defined in [36], to the values of 20 mm, 100 mm, and 0.5 respectively. Figure 6 shows the accuracy of each baseline on the LINEMOD, MULT-I, OCC, BIN-P datasets, respectively. Comparing the numbers in this chart with the ones depicted in Fig. 2, one can observe that the generated results are relatively lower than that are of the AD metric. This arises mainly from the chosen parameters. However, the characteristics of both charts are the same, that is, both methods, according to AD and VSD, perform best on the LINEMOD dataset, whilst worst on OCC. On the other hand, the main advantage of the proposed metric is that it features ambiguity-invariance: Since it is designed to evaluate the baselines over the visible parts of the objects, it gives more robust measurements across symmetric objects. Sample images in Fig. 6 shows the hypotheses of symmetric objects which are considered as false positive according to the AD metric, whilst VSD accepts those as correct. Due to space limitation, we provide object-wise results and in depth discussions regarding the thresholds of VSD in the supplementary material.

6. Discussions and Conclusions

We outline our key observations that provide guidance for future research.

From the challenges aspect, reasonably accurate results have been obtained on textured-objects at varying viewpoints with cluttered backgrounds. In case occlusion is introduced in the test scenes, depending on the architecture of the baseline, good performance demonstrated. Part-based solutions can handle the occlusion problem better than the ones global, using the information acquired from occlusion-free parts of the target objects. However, heavy existence of occlusion and clutter severely affects the detectors. It is possible that modelling occlusion during training can improve the performance of a detector across severe occlusion. But when occlusion is modelled, the baseline could be data-dependent. In order to maintain the generalization capability of the baseline contextual information can additionally be utilized during the modelling. Currently, similar looking distractors along with similar looking object classes seem the biggest challenge in recovering instances' 6D, since the lack of discriminative selection of shape features strongly confuse recognition systems. One possible solution could be considering the instances that have strong similarity in shape in a same category. In such a case, detectors trained using the data coming from the instances involved in the same category might report better detection results.

Architecture-wise, template-based methods, matching model features to the scene, and random forest based learning algorithms, along with their good generalization performance across unseen samples, underlie object detection and 6D pose estimation. Recent paradigm in the community is to learn deep discriminative feature representations. Despite the fact that several methods addressed 6D pose estimation utilizing deep features [7, 12], end-to-end neural network-based solutions for 6D object pose recovery are still not widespread. Depending on the availability of large-scale 6D annotated depth datasets, feature representations can be learnt on these datasets, and then the learnt representations can be customized for the 6D problem.

These implications are related to automation in robotic systems. The implications can provide guidance for robotic manipulators that pick and place the items from conveyors, shelves, pallets, *etc*. Accurately detecting objects and estimating their fine pose under uncontrolled conditions improves the grasping capability of the manipulators. Beyond accuracy, the baselines are expected to show real-time performance. Although the detectors we have tested cannot perform real-time, their run-time can be improved by utilizing APIs like OpenMP.

References

[1] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige and N. Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes", *ACCV*, 2012. 1, 2, 3, 4, 6

[2] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in RGB-D images", *CVPR*, 2013. 4

[3] A. Tejani, D. Tang, R. Kouskouridas and T-K. Kim, "Latent-class hough forests for 3D object detection and pose estimation", *ECCV*, 2014. 1, 2, 3, 4, 6

[4] R. Jonschkowski, C. Eppner, S. Hofer, R. Martin-Martin and O. Brock, "Probabilistic multi-class segmentation for the amazon picking challenge", *IROS*, 2016. 2

[5] C. Eppner, S. Hofer, R. Jonschkowski, R. Martin-Martin, A. Sieverling, V. Wall and O. Brock, "Lessons from the amazon picking challenge: Four aspects of building robotic systems", *Proceedings of Robotics: Science and Systems*, 2016. 2

[6] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano and P. R. Wurman, "Analysis and Observations From the First Amazon Picking Challenge", *IEEE Transactions on Automation Science and Engineering*, 2016. 2

[7] A. Doumanoglou, R. Kouskouridas, S. Malassiotis and T-K. Kim, "Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd", *CVPR*, 2016. 2, 3, 4, 8

[8] C. Sahin, R. Kouskouridas and T-K. Kim, "Iterative Hough Forest with Histogram of Control Points for 6 DoF Object Registration from Depth Images", *IROS*, 2016. 4

[9] E. Brachmann, F. Michel, A. Krull, M.Y. Yang, S. Gumhold and C. Rother, "Uncertainty-Driven 6D Pose Estimation of Objects and Scenes from a Single RGB Image", *CVPR*, 2016. 4, 6

[10] U. Bonde, V. Badrinarayanan and R. Cipolla, "Robust instance recognition in presence of occlusion and clutter", *ECCV*, 2014. 4

[11] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton and C. Rother, "Learning 6D object pose estimation using 3D object coordinates", *ECCV*, 2014. 1, 2, 3, 4, 6

[12] W. Kehl, F. Milletari, F. Tombari, S. Ilic and N. Navab, "Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation", *ECCV*, 2016. 4, 6, 8

[13] R. Rios-Cabrera and T. Tuytelaars, "Discriminatively trained templates for 3d object detection: A real time scalable approach", *ICCV*, 2013. 3, 6

[14] A. Krull, E. Brachmann, F. Michel, M. Y. Yang, S. Gumhold and C. Rother, "Learning analysis-by-synthesis for 6d pose estimation in rgb-d images", *ICCV*, 2015. 3, 6

[15] T. Hodan, X. Zabulis, M. Lourakis, S. Obdrzalek and J. Matas, "Detection and fine 3D pose estimation of texture-less objects in RGB-D images", *IROS*, 2015. 6

[16] P. Wohlhart and V. Lepetit, "Learning descriptors for object recognition and 3d pose estimation", *CVPR*, 2015. 4

[17] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros and M. Hebert, "An empirical study of context in object detection", *CVPR*, 2009. 2

[18] D. Hoiem, Y. Chodpathumwan and Q. Dai, "Diagnosing error in object detectors", *ECCV*, 2012. 1, 2

[19] O. Russakovsky, J. Deng, Z. Huang, A. C. Berg and L. Fei-Fei, "Detecting avocados to zucchinis: what have we done, and where are we going?", *ICCV*, 2013. 1, 2

[20] A. Torralba and A. A. Efros, "Unbiased look at dataset bias", *CVPR*, 2011. 2

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "Imagenet large scale visual recognition challenge", *IJCV*, 2015. 2

[22] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database", *CVPR*, 2009. 1, 2

[23] M. Everingham, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, "The pascal visual object classes (voc) challenge", *IJCV*, 2010. 1, 2

[24] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn and A. Zisserman, "The pascal visual object classes challenge: A retrospective", *IJCV*, 2015. 2

[25] P.F. Felzenszwalb, R.B. Girshick, D. McAllester and D. Ramanan, "Object detection with discriminatively trained part-based models", *TPAMI*, 2010. 2

[26] H. Azizpour and I. Laptev, "Object detection using strongly-supervised deformable part models", *ECCV*, 2012. 2

[27] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation", *CVPR*, 2014. 2

[28] A. Shrivastava and A. Gupta, "Building part-based object detectors via 3d geometry", *ICCV*, 2013. 2

[29] B. Pepik, M. Stark, P. Gehler and B. Schiele, "Teaching 3d geometry to deformable part models", *CVPR*, 2012. 2

[30] R. Girshick, F. Iandola, T. Darrell and J. Malik, "Deformable part models are convolutional neural networks", *CVPR*, 2015. 2

[31] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng and T. Darrell, "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", *ICML*, 2014. 2

[32] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus and Y. LeCun, "Overfeat: Integrated recognition, localization and detection using convolutional networks", *ICLR*, 2014. 2

[33] K. He, X. Zhang, S. Ren and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition", *PAMI*, 2015. 2

[34] R. Girshick, "Fast R-CNN", *ICCV*, 2015. 2

[35] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis and X. Zabulis, "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects", *WACV*, 2017. 2, 3

[36] T. Hodan, J. Matas and S. Obdrzalek, "On Evaluation of 6D Object Pose Estimation", *ECCVW*, 2016. 2, 4, 8

[37] M. Y. Liu, O. Tuzel, A. Veeraraghavan, Y. Taguchi, T. K. Marks and R. Chellappa, "Fast object localization and pose estimation in heavy clutter for robotic bin picking", *IJRR*, 2012. 3

[38] B. Drost, M. Ulrich, N. Navab and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition", *CVPR*, 2010. 6

[39] W. Kehl, F. Tombari, N. Navab, S. Ilic and V. Lepetit, "Hashmod: A Hashing Method for Scalable 3D Object Detection", *BMVC*, 2015. 6

[40] S. Hinterstoisser, V. Lepetit, N. Rajkumar and K. Konolige, "Going further with point pair features", *ECCV*, 2016. 6

[41] C. Sahin, R. Kouskouridas and T-K. Kim, "A learning-based variable size part extraction architecture for 6D object pose recovery in depth images", *Image and Vision Computing (IVC)*, 2017. 4