Recovering 6D Object Pose: A Review and Multi-modal Analysis

Caner Sahin(✉) and Tae-Kyun Kim

ICVL, Imperial College London, London, UK
c.sahin14@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? Our findings include: (i) reasonably accurate results are obtained on textured-objects at varying viewpoints with cluttered backgrounds. (ii) Heavy existence of occlusion and clutter severely affects the detectors, and similar-looking distractors is the biggest challenge in recovering instances' 6D. (iii) Template-based methods and random forest-based learning algorithms underlie object detection and 6D pose estimation. Recent paradigm is to learn deep discriminative feature representations and to adopt CNNs taking RGB images as input. (iv) 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.

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 [1,2] 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 [3], PASCAL [4]. 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.

L. Leal-Taixé and S. Roth (Eds.): ECCV 2018 Workshops, LNCS 11134, pp. 15–31, 2019.
https://doi.org/10.1007/978-3-030-11024-6_2

Fig. 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 [5–7] 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) [8] 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 [9,10]. 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 state-of-the-art 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 three-fold strategy: we firstly collect five representative object datasets [5–7,11,12] (see Fig. 1). Then, we investigate 10 state-of-the-art detectors [5–7,11,13–18] on the collected datasets under uniform scoring criteria of the Average Distance (AD) metric. We further extend our investigations comparing 2 of the detectors [5,6], which are our own implementations, using the Visible Surface Discrepancy (VSD) protocol. 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. In total, there are approximately 50 different object classes. We investigate ten classes of the state-of-the-art 6D detectors on the collected datasets under uniform scoring criteria.
- 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.

2 Related Work

Methods producing 2D bounding box hypotheses in color images [19–28] 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 [3,4], on which the methods have been tested. In [29], the effect of different context sources, such as geographic context, object spatial support, *etc.*, on object detection is examined. Hoiem et al. [1] 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. [2] use ImageNet in order to do meta-analysis, and to examine the influences

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. classes	Modality	# Total frame	Obj. Dist. [mm]
LINEMOD	VP + C + TL	15	RGB-D	15770	600–1200
MULT-I	VP + C + TL + O + MI	6	RGB-D	2067	600–1200
OCC	VP + C + TL + SO	8	RGB-D	9209	600–1200
BIN-P	VP + SC + SO + MI + BP	2	RGB-D	180	600–1200
T-LESS	VP + C + TL + O + MI + SLD	30	RGB-D	10080	600–1200

of color, texture, *etc.*, on the performances of object detectors. Torralba et al. [30] compares several datasets regarding the involved samples, cross-dataset generalization, and relative data bias, *etc.* Recently published retrospective evaluation [31] and benchmarking [32] 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.

3 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 [5] dataset involve *texture-less* objects at varying viewpoints 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° around the object, 0–90° degree tilt rotation, ∓45° 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 [6] 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, textureless 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 [7] 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 [5] dataset, various objects are present, but only ground truth poses for one object are given. Brachmann et al. [7] 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 [11] 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 [12] dataset particularly focuses on this problem. The RGB-D images of the objects located on a table are captured at different viewpoints covering 360° 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.

4 Baselines

State-of-the-art baselines for 6D object pose estimation address the challenges studied in Sect. 3, however, the architectures used differ between the baselines. In this section, we analyse 6D object pose estimators architecture-wise.

Template-Based. 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 [5], 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 [33]. Fast directional chamfer matching (FDCM) [34] is used in robotics applications.

Point-to-Point. Point-to-point techniques build point-pair features for sparse representations of the test and the model point sets. Drost et al. [13] propose create a global model description based on oriented point pair features and match that model locally using a fast voting scheme. Its further improved in [14] making the method more robust across clutter and sensor noise.

Conventional Learning-Based. These methods are in need of training sessions where training samples along with the ground truth annotations are learnt. Latent-class Hough forests [6,35], employing one-class learning, utilize surface normals and color gradients features in a part-based approach in order to provide robustness across occlusion. The random forest based method in [7] encodes contextual information of the objects with simple depth and RGB pixels, and improves the confidence of a pose hypothesis using a Ransac-like algorithm. An analysis-by-synthesis approach [36] and an uncertainty-driven methodology [15] are build upon random forests, using the architecture provided in [7]. The method based on random forests presented in [37] 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 architectures in [38,39] present iterative Hough forests that initially estimate coarse 6D pose of an object, and then iteratively refine the confidence of the estimation due to the extraction of more discriminative control point descriptors [40].

Deep Learning. Current paradigm in the community is to learn deep discriminative feature representations. Wohlhart et al. [41] 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 [11,16] learn deep representation of parts in an unsupervised fashion only from foreground images using auto-encoder architectures. The features extracted in the course of the test are fed into a Hough forest in [11], and into a codebook of pre-computed synthetic local object patches in [16] in order to hypothesise object 6D pose. While [41] focuses on learning feature embeddings based on metric learning with triplet comparisons, Balntas et al. [42] further examine the effects of using object poses as guidance to learning robust features for 3D object pose estimation in order to handle symmetry issue.

More recent methods adopt CNNs for 6D pose estimation, taking RGB images as inputs [17]. BB8 [43] and Tekin et al. [44] perform corner-point regression followed by PnP for 6D pose estimation. Typically employed is a computationally expensive post processing step such as iterative closest point (ICP) or a verification network [18].

5 Evaluation Metrics

Several evaluation metrics are proposed for measuring the performance of a 6D detector. Average Distance (AD) [5] 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 [45], 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 [46]. 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.

6 Multi-modal Analyses

We analyse ten baselines on the datasets with respect to both challenges and the architectures. Two of the baselines [5,6] are our own implementations. The color gradients and surface normal features, presented in [5], are computed using the built-in functions and classes provided by OpenCV. The features in Latent-Class Hough Forest (LCHF) [6] are the part-based version of the features introduced in [5]. 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. In this section, "LINEMOD" refers to the dataset, whilst "Linemod" is used to indicate the baseline itself.

6.1 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. The coefficient z_{ω} is 0.10, and in case we use different thresholds, we will specifically indicate in the related parts.

Table 2. Methods' performance are depicted object-wise based on recall values computed using the Average Distance (AD) evaluation protocol.

Method	ch.	ape	bvise	cam	can	cat	dril	duck	box	glue	hpunch	iron	lamp	phone	AVER
Kehl et al [16]	RGB-D	96.9	94.1	97.7	95.2	97.4	96.2	97.3	99.9	78.6	96.8	98.7	96.2	92.8	95.2
LCHF [6]	RGB-D	84	95	72	74	91	92	91	48	55	89	72	90	69	78.6
Linemod [5]	RGB-D	95.8	98.7	97.5	95.4	99.3	93.6	95.9	99.8	91.8	95.9	97.5	97.7	93.3	96.3
Drost et al [13]	D	86.5	70.7	78.6	80.2	85.4	87.3	46	97	57.2	77.4	84.9	93.3	80.7	78.9
Kehl et al [18]	RGB	65	80	78	86	70	73	66	100	100	49	78	73	79	76.7

(a) LINEMOD dataset

Method	ch.	camera	cup	joystick	juice	milk	shampoo	AVER
LCHF [6]	RGB-D	52.5	99.8	98.3	99.3	92.7	97.2	90
Linemod [5]	RGB-D	18.3	99.2	85	51.6	72.2	53.1	63.2

(b) MULT-I dataset

Method	ch.	ape	can	cat	dril	duck	box	glue	hpunch	AVER
Xiang et al. [17]	RGB-D	76.2	87.4	52.2	90.3	77.7	72.2	76.7	91.4	78
LCHF [6]	RGB-D	48.0	79.0	38.0	83.0	64.0	11.0	32.0	69.0	53
Hinters et al. [14]	RGB-D	81.4	94.7	55.2	86.0	79.7	65.5	52.1	95.5	76.3
Linemod [5]	RGB-D	21.0	31.0	14.0	37.0	42.0	21.0	5.0	35.0	25.8
Xiang et al. [17]	RGB	9.6	45.2	0.93	41.4	19.6	22.0	38.5	22.1	25

(c) OCC dataset

Method	ch.	cup	juice	AVER
LCHF [6]	RGB-D	90.0	89.0	90
Brach et al. [7]	RGB-D	89.4	87.6	89
Linemod [5]	RGB-D	88.0	40.0	64

(d) BIN-P dataset

Recall-Only Discussions. Recall-only discussions are based on the numbers provided in Table 2, and Fig. 2.

Clutter, Viewpoint, Texture-Less Objects. Highest recall values are obtained on the LINEMOD dataset (see Fig. 2(a)), meaning that the state-of-the-art methods for 6D object pose estimation can successfully handle the challenges, clutter, varying viewpoint, and texture-less objects. LCHF, detecting more than half of the objects with over 80% accuracy, worst performs on "box" and "glue" (see Table 2a), 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 feature representations in color and depth channels. LCHF performs better on this dataset than Linemod. Since LCHF 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,

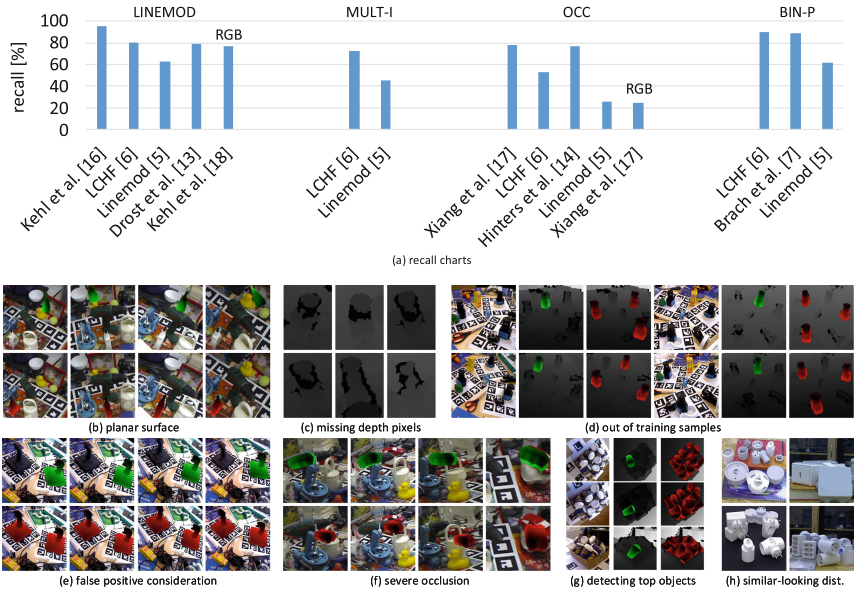

Fig. 2. (a) Success of each baseline on every dataset is shown, recall values are computed using the Average Distance (AD) metric. (b)–(h) challenges encountered during test are exemplified (green renderings are hypotheses, and the red ones are ground truths). (Color figure online)

"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 0.15, we observe 89% accuracy.

Severe Occlusion. OCC involves challenging test images where the objects of interest are cluttered and severely occluded. The best performance on this dataset is caught by Xiang et al. [17], and there is still room for improvement in order to fully handle this challenge. 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, methods produce relatively higher recall values (*e.g.* "can, driller, duck, holepuncher", Table 2c). On the other hand, when the target object has additionally other challenges such as planar surfaces, methods' performance (LCHF and Linemod) decreases (*e.g.* "box", 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, where objects are arbitrarily stacked in a pile. According to the recall values presented in Table 2d, LCHF and Brachmann et al. [7] perform 25% better than Linemod. 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)).

Part-Based vs. Holistic Approaches. Holistic methods [5,13,14,17,18] formulate the detection problem globally. Linemod [5] represents 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 processes, mainly degrade the performance of this detector. Part-based methods [6,7,11,15,16] extract parts in the given image. 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.

Template-Based vs. Random Forest-Based. Template-based methods, *i.e.*, Linemod, match the features extracted during test to a set of templates, and hence, they cannot easily be generalized well to unseen ground truth annotations, that is, the translation and rotation parameters in object pose estimation. Methods based on random forests [6,7,11,15] efficiently benefit the randomisation embedded in this learning tool, consequently providing good generalisation performance on new unseen samples.

RGB-D vs. Depth. Methods utilizing both RGB and depth channels demonstrate higher recall values than methods that are of using only depth, since RGB provides extra clues to ease the detection. This is depicted in Table 2a where learning- and template-based methods of RGB-D perform much better than point-to-point technique [13] of depth channel.

Table 3. Methods' performance are depicted object-wise based on F1 scores computed using the Average Distance (AD) evaluation protocol.

Method	ch.	ape	bvise	cam	can	cat	dril	duck	box	glue	hpunch	iron	lamp	phone	AVER
Kehl et al. [16]	RGB-D	0.98	0.95	0.93	0.83	0.98	0.97	0.98	1	0.74	0.98	0.91	0.98	0.85	0.93
LCHF [6]	RGB-D	0.86	0.96	0.72	0.71	0.89	0.91	0.91	0.74	0.68	0.88	0.74	0.92	0.73	0.82
Linemod [5]	RGB-D	0.53	0.85	0.64	0.51	0.66	0.69	0.58	0.86	0.44	0.52	0.68	0.68	0.56	0.63
Kehl et al. [18]	RGB	0.76	0.97	0.92	0.93	0.89	0.97	0.80	0.94	0.76	0.72	0.98	0.93	0.92	0.88

(a) LINEMOD dataset

Method	ch.	camera	cup	joystick	juice	milk	shampoo	AVER
Kehl et al. [16]	RGB-D	0.38	0.97	0.89	0.87	0.46	0.91	0.75
LCHF [6]	RGB-D	0.39	0.89	0.55	0.88	0.40	0.79	0.65
Drost et al. [13]	D	0.41	0.87	0.28	0.60	0.26	0.65	0.51
Linemod [5]	RGB-D	0.37	0.58	0.15	0.44	0.49	0.55	0.43
Kehl et al. [18]	RGB	0.74	0.98	0.99	0.92	0.78	0.89	0.88

(b) MULT-I dataset

Method	ch.	ape	can	cat	dril	duck	box	glue	hpunch	AVER
LCHF [6]	RGB-D	0.51	0.77	0.44	0.82	0.66	0.13	0.25	0.64	0.53
Linemod [5]	RGB-D	0.23	0.31	0.17	0.37	0.43	0.19	0.05	0.30	0.26
Brach et al. [15]	RGB	-	-	-	-	-	-	-	-	0.51
Kehl et al. [18]	RGB	-	-	-	-	-	-	-	-	0.38

(c) OCC dataset

Method	ch.	cup	juice	AVER
LCHF [6]	RGB-D	0.48	0.29	0.39
Doumanoglou et al. [11]	RGB-D	0.36	0.29	0.33
Linemod [5]	RGB-D	0.48	0.20	0.34

(d) BIN-P dataset

RGB-D vs. RGB (CNN Structures). More recent paradigm is to adopt CNNs to solve 6D object pose estimation problem taking RGB images as inputs [17,18]. Methods working in the RGB channel in Table 2 are based on CNN structure. According to the numbers presented in Table 2, RGB-based SSD-6D [43] and RGB-D-based LCHF achieve similar performance. These recall values show the promising performance of CNN architectures across random forest-based learning methods.

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 cluttered.

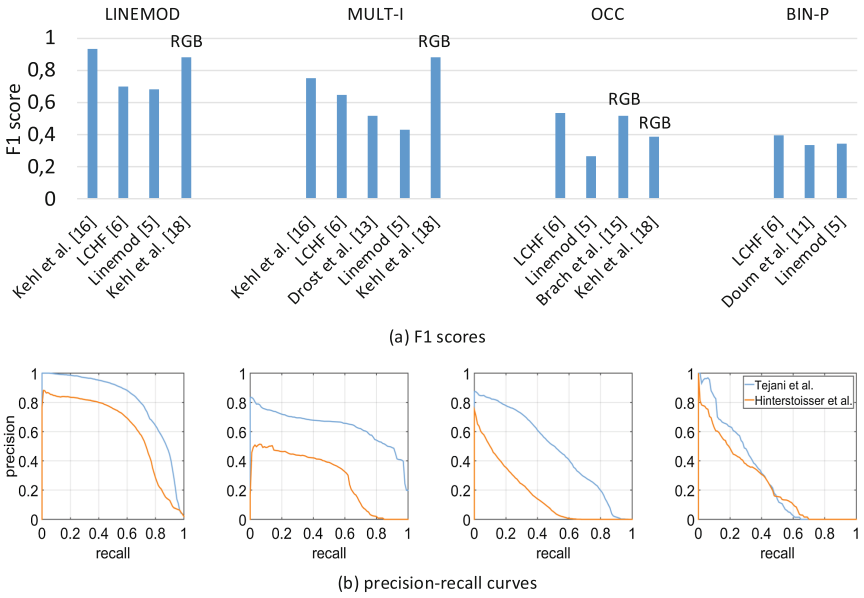

Fig. 3. (a) Success of each baseline on every dataset is shown, F1 scores are computed using the Average Distance (AD) metric. (b) Precision-recall curves of averaged F1 scores for Tejani et al. [6] and Hinterstoisser et al. [5] are shown: from left to right, LINEMOD, MULT-I, OCC, BIN-P.

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

Precision-Recall Discussions. Our precision-recall discussions are based on the F1 scores provided in Table 3, and Fig. 3(a).

We first analyse the performance of the methods [5,6,18,47] on the LINEMOD dataset. On the average, Kehl et al. [47] outperforms other methods proving the superiority of learning deep features. Despite estimating 6D in RGB images, SSD-6D [18] exhibits the advantages of using CNN structures for 6D object pose estimation. LCHF and Linemod demonstrate lower performance, since the features used by these methods are manually-crafted. The comparison between Figs. 2(a) and 3(a) reveals that the results produced by the methods have approximately the same characteristics on the LINEMOD dataset, with respect to recall and F1 scores.

The methods tested on the MULT-I dataset [5,6,13,47] utilize the geometry information inherently provided by depth images. Despite this fact, SSD-6D [18], estimating 6D pose only from RGB images, outperforms other methods clearly proving the superiority of using CNNs for the 6D problem over other structures.

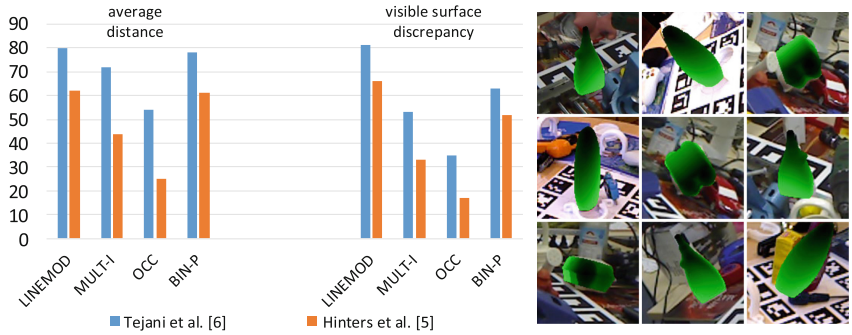

Fig. 4. Methods are evaluated based on Visible Surface Discrepancy. Samples on the right are considered as false positive with respect to Average Distance, whilst VSD deems correct.

LCHF [5] and Brachmann et al. [15] best perform on OCC with respect to F1 scores. As this dataset involves test images where highly occluded objects are located, the reported results depict the importance of designing part-based solutions.

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 (see Fig. 2(a)), 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. 3(b), we lastly report precision-recall curves of LCHF and Linemod. 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.

6.2 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 [46], to the values of 20 mm, 100 mm, and 0.5 respectively. Figure 4 shows the accuracy of each baseline on the LINEMOD, MULT-I, OCC, BIN-P datasets, respectively. Comparing the numbers in this chart, one can observe that the results from VSD 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. 4 show the hypotheses of symmetric objects which are considered as false positive according to the AD metric, whilst VSD accepts those as correct.

7 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 can 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 [11,16], 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. Hoiem, D., Chodpathumwan, Y., Dai, Q.: Diagnosing error in object detectors. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7574, pp. 340–353. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_25
2. Russakovsky, O., Deng, J., Huang, Z., Berg, A.C., Fei-Fei, L.: Detecting avocados to zucchinis: what have we done, and where are we going? In: ICCV (2013)
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: CVPR (2009)
4. Everingham, M., Gool, L.V., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. IJCV **88**, 303–338 (2010)
5. Hinterstoisser, S., et al.: Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012. LNCS, vol. 7724, pp. 548–562. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-37331-2_42
6. Tejani, A., Tang, D., Kouskouridas, R., Kim, T.-K.: Latent-class hough forests for 3D object detection and pose estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 462–477. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_30
7. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D object pose estimation using 3D object coordinates. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 536–551. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_35
8. Eppner, C., et al.: Lessons from the Amazon picking challenge: four aspects of building robotic systems. In: Proceedings of Robotics: Science and Systems (2016)
9. Jonschkowski, R., Eppner, C., Hofer, S., Martin-Martin, R., Brock, O.: Probabilistic multi-class segmentation for the Amazon picking challenge. In: IROS (2016)
10. Correll, N., et al.: Analysis and observations from the first Amazon picking challenge. IEEE Trans. Autom. Sci. Eng. **15**, 172–188 (2016)
11. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: Recovering 6D object pose and predicting next-best-view in the crowd. In: CVPR (2016)
12. Hodan, T., Haluza, P., Obdrzalek, S., Matas, J., Lourakis, M., Zabulis, X.: T-less: an RGB-D dataset for 6D pose estimation of texture-less objects. In: WACV (2017)
13. Drost, B., Ulrich, M., Navab, N., Ilic, S.: Model globally, match locally: efficient and robust 3D object recognition. In: CVPR (2010)
14. Hinterstoisser, S., Lepetit, V., Rajkumar, N., Konolige, K.: Going further with point pair features. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 834–848. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_51
15. Brachmann, E., Michel, F., Krull, A., Yang, M., Gumhold, S., Rother, C.: Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image. In: CVPR (2016)
16. Kehl, W., Milletari, F., Tombari, F., Ilic, S., Navab, N.: Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 205–220. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_13
17. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. arxiv (2017)

18. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: SSD-6D: making RGB-based 3D detection and 6D pose estimation great again. In: CVPR (2017)
19. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. In: TPAMI (2010)
20. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 836–849. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33718-5_60
21. Pepik, B., Stark, M., Gehler, P., Schiele, B.: Teaching 3D geometry to deformable part models. In: CVPR (2012)
22. Shrivastava, A., Gupta, A.: Building part-based object detectors via 3D geometry. In: ICCV (2013)
23. Donahue, J., et al.: DeCAF: a deep convolutional activation feature for generic visual recognition. ICML (2014)
24. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR (2014)
25. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: OverFeat: integrated recognition, localization and detection using convolutional networks. In: ICLR (2014)
26. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: PAMI (2015)
27. Girshick, R.: Fast R-CNN. In: ICCV (2015)
28. Girshick, R., Iandola, F., Darrell, T., Malik, J.: Deformable part models are convolutional neural networks. In: CVPR (2015)
29. Divvala, S.K., Hoiem, D., Hays, J.H., Efros, A.A., Hebert, M.: An empirical study of context in object detection. In: CVPR (2009)
30. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
31. Everingham, M., Eslami, S.A., Gool, L.V., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. IJCV **111**, 98–136 (2015)
32. Russakovsky, O., et al.: Imagenet large scale visual recognition challenge. IJCV **115**, 211–252 (2015)
33. Rios-Cabrera, R., Tuytelaars, T.: Discriminatively trained templates for 3D object detection: a real time scalable approach. In: ICCV (2013)
34. Liu, M.Y., Tuzel, O., Veeraraghavan, A., Taguchi, Y., Marks, T.K., Chellappa, R.: Fast object localization and pose estimation in heavy clutter for robotic bin picking. IJRR **31**, 951–973 (2012)
35. Sock, J., Kasaei, S.H., Lopes, L.S., Kim, T.K.: Multi-view 6D object pose estimation and camera motion planning using RGBD images. In: 3rd International Workshop on Recovering 6D Object Pose (2017)
36. Krull, A., Brachmann, E., Michel, F., Yang, M.Y., Gumhold, S., Rother, C.: Learning analysis-by-synthesis for 6D pose estimation in RGB-D images. In: ICCV (2015)
37. Bonde, U., Badrinarayanan, V., Cipolla, R.: Robust instance recognition in presence of occlusion and clutter. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8690, pp. 520–535. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10605-2_34
38. Sahin, C., Kouskouridas, R., Kim, T.K.: Iterative hough forest with histogram of control points for 6 DoF object registration from depth images. In: IROS (2016)
39. Sahin, C., Kouskouridas, R., Kim, T.K.: A learning-based variable size part extraction architecture for 6D object pose recovery in depth images. Image Vis. Comput. (IVC) **63**, 38–50 (2017)

40. Michel, F., et al.: Global hypothesis generation for 6D object pose estimation. In: CVPR (2017)
41. Wohlhart, P., Lepetit, V.: Learning descriptors for object recognition and 3D pose estimation. In: CVPR (2015)
42. Balntas, V., Doumanoglou, A., Sahin, C., Sock, J., Kouskouridas, R., Kim, T.K.: Pose guided RGBD feature learning for 3D object pose estimation. In: ICCV (2017)
43. Rad, M., Lepetit, V.: BB8: a scalable, accurate, robust to partial occlusion method for predicting the 3D poses of challenging objects without using depth. In: ICCV (2017)
44. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6D object pose prediction. arxiv (2017)
45. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: CVPR (2013)
46. Hodaň, T., Matas, J., Obdržálek, Š.: On evaluation of 6D object pose estimation. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 606–619. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_52
47. Kehl, W., Tombari, F., Navab, N., Ilic, S., Lepetit, V.: Hashmod: a hashing method for scalable 3D object detection. In: BMVC (2015)