Symmetry Aware Evaluation of 3D Object Detection and Pose Estimation in Scenes of Many Parts in Bulk

Romain Brégier[1,2], Frédéric Devernay[2], Laetitia Leyrit[1], James L. Crowley[2]

[1] Siléane, Saint-Étienne

[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LIG, F-38000 Grenoble

Supplementary material

1 Datasets details

1.1 Overall content description

The new datasets proposed in this paper consist in a total of 2601 independent scenes depicting various numbers of object instances in bulk. Synthetic datasets are made of scenes representing the inside of a bin, in which between 0 and $n \in \mathbb{N}^*$ object instances have been dropped randomly. n is chosen empirically as a compromise between simulation duration and scene complexity such as to produce plausible scenes of bulk (see table 1 for numerical values).

Real data consists similarly in scenes of various numbers of object instances (between 0 and 11) lying on a surface at various distances from the camera. Each instance was covered by 19 fiducial markers. We used two different background surfaces illustrated in figure 1: a planar one (*markers flat*, 308 scenes), representative of the typical bottom of a bin; and a bumpy surface (*markers bump*, 325 scenes), increasing the variability of poses and producing a pose distribution more consistent with the scenario of many instances piled up. An additional dataset of 46 scenes (*markers clutter*) targets the problem of object detection and pose estimation in a cluttered environment.

The symmetry class considered for a rigid object depends on what static configurations of the object we wish to distinguish, and this choice is not necessarily obvious. For example, the *gear* object could be considered as an object with a cyclic symmetry of order 2 – *i.e.* an invariance under rotation of 1/2 turn about a given axis – , a cyclic symmetry of order its number of teeth, or a revolution symmetry depending on the level of details considered. We considered this latter option in our experiments, and table 1 synthesize the choices of symmetry classes we made.

1.2 Scene data description

We provide for each scene the following content, illustrated on figure 2:

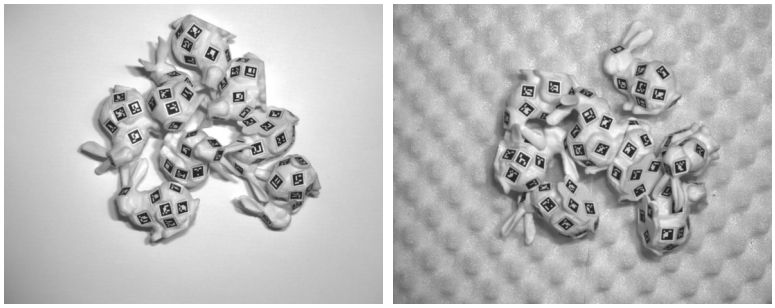


Figure 1: Different backgrounds used for real data acquisition

Table 1: Main properties of the datasets used.

	Dataset	Class of symmetry considered	Number of scenes	Number of instances per scene
Real data	markers bump	no	325	0 - 11
	markers clutter	no	46	0 - 4
	markers flat	no	308	0 - 11
	juice [1]	no	60	5
	coffee cup [1]	revolution without rotoreflection invariance	56	19
Synthetic data	markers flat	no	308	0 - 11
	tless 22	no	202	0 - 100
	bunny	no	324	0-80
	tless 20	cyclic 2	200	0 - 99
	tless 29	cyclic 2	160	0 - 79
	brick	cyclic 2	302	0 - 150
	gear	revolution without rotoreflection invariance	122	0 - 60
	candlestick	revolution without rotoreflection invariance	122	0 - 60
	pepper	revolution without rotoreflection invariance	182	0 - 90

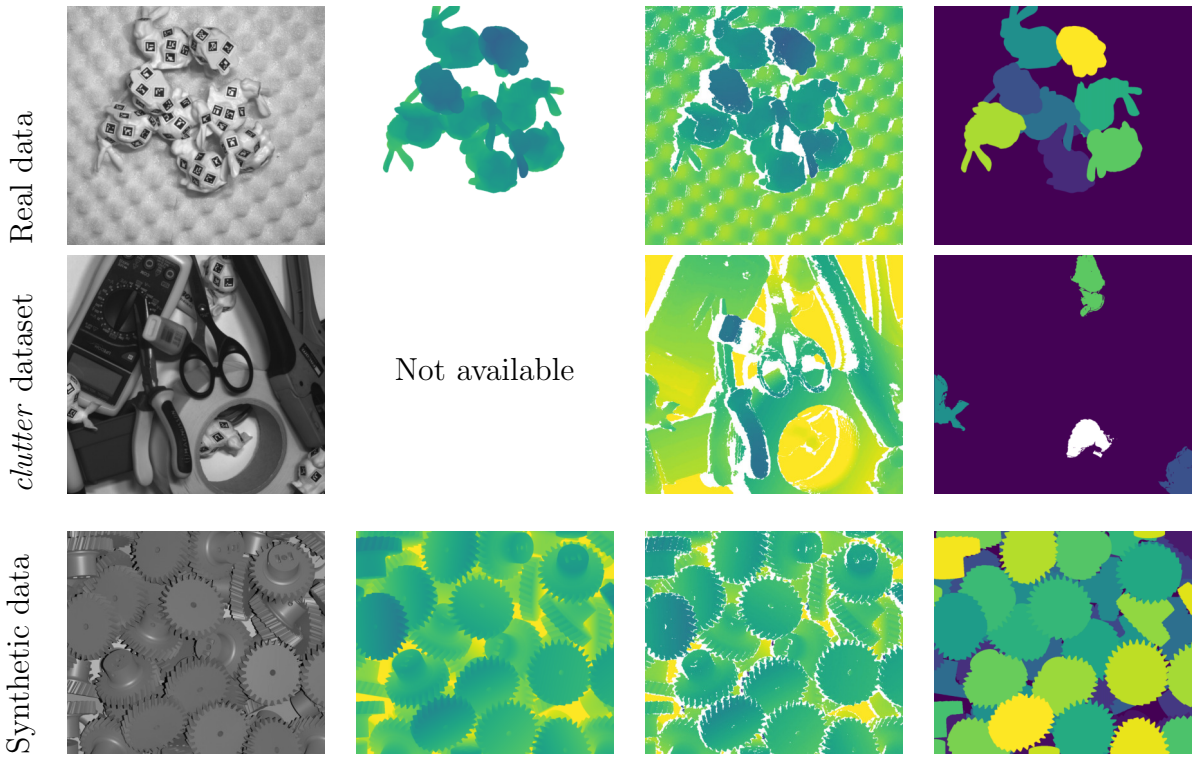
- An RGB or intensity image (figure 2a).
- Depth images, from the same viewpoint as the RGB data. In addition to the depth data affected by stereo reconstruction noise considered in the paper (figure 2c), we propose an ideal depth image of the scene, synthetically generated (figure 2b). Because we did not model the bumpy background used in the case of the *bunny* dataset, it is not present in the ideal data. We do not provide ideal data for the *clutter* dataset for similar reasons.
- A segmentation image of the different object instances present in the scene (figure 2d). In the case of the *clutter* dataset, the segmentation label is only provided for pixels with a defined depth, since occlusion due to clutter is difficult to assess automatically where depth information is missing.
- Camera parameters, enabling to recover 3D coordinates from the RGBD data.
- Ground truth annotation, consisting in the pose, occlusion rate and segmentation label of each object instance present in the scene.

2 Implementation details

2.1 PPF

We used the following parameters for the PPF method (see the original paper [2] for more information):

- Sampling rate of $\tau_d = 0.05$.
- Every sampled point of the scene is used as reference point.
- Angles are discretized into $n_{angle} = 30$ bins.



(a) RGB or intensity image (b) Ideal range image (c) Noisy range image (d) Instances segmentation

Figure 2: Typical content of our dataset

2.2 LINEMOD+

For each object, LINEMOD templates are generated based on depth renderings of the 3D object model, considering a virtual depth camera of identical properties to those of the depth sensor considered in the dataset (focal length, width, height, principal point). As stated in the paper, we ignore the potential symmetries of the object for templates generation, and assimilate therefore poses to rigid transformations. The position of the templated poses is chosen along the optical axis at a given distance from the virtual camera, which corresponds to the typical distance of an object instance from the camera in the dataset (estimated as the average distance between the near and far clipping planes of our depth images). The orientation of those poses is sampled uniformly with a 10° step based on a Tait-Bryan angles parametrization, leading to the generation of 22104 templates. We did not perform multiscale detection, because of prohibitive computation times with the LINE3D implementation used.

During test, local maxima of 2D templates response maps are converted into pose hypotheses, by matching the average depth of a template with the one of its projected silhouette in the range image.

2.3 Generating pose hypotheses

The modes finding step performed for the PPF method described in section 4.2 of the paper consists in an adaptation of the Mean Shift algorithm for the proposed distance and is described in [3]. The radius of the Mean Shift kernel used corresponds to half the typical thickness of the object, and is estimated as $\min(\lambda_1, \lambda_2, \lambda_3)$, where $\lambda_1, \lambda_2, \lambda_3$ are the standard deviations of the object's surface along its principal axes. The same value is used as minimum distance between two hypotheses for them to be considered as duplicates, in the duplicates filtering step performed for both PPF and LINEMOD+ methods.

2.4 Postprocessing

The post-processing step (PP) evoked in the paper is performed as follows. For each pose hypothesis, we compare the depth values of a rendering of the object at the hypothesized pose with the actual depth values of the data, and the contours of its silhouettes with contours detected in the RGBD data. Hypotheses are sorted by decreasing consistency with the input data, and pose hypotheses intersecting one another are filtered in a *winner-takes-all* approach.

References

[1] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6d Object Pose and Predicting Next-Best-View in the Crowd," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on.* IEEE, 2010, p. 998–1005.

[3] R. Brégier, F. Devernay, L. Leyrit, and J. Crowley, "Defining the pose of any 3D rigid object and an associated distance," 2016, manuscript submitted for publication. [Online]. Available: https://hal.inria.fr/hal-01415027