T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects

Tomáš Hodaň[1], Pavel Haluza[1], Štěpán Obdržálek[1], Jiří Matas[1], Manolis Lourakis[2], Xenophon Zabulis[2]

[1]Center for Machine Perception, Czech Technical University in Prague, Czech Republic
[2]Institute of Computer Science, Foundation for Research and Technology – Hellas, Heraklion, Greece

Abstract

We introduce T-LESS, a new public dataset for estimating the 6D pose, i.e. translation and rotation, of texture-less rigid objects. The dataset features thirty industry-relevant objects with no significant texture and no discriminative color or reflectance properties. The objects exhibit symmetries and mutual similarities in shape and/or size. Compared to other datasets, a unique property is that some of the objects are parts of others. The dataset includes training and test images that were captured with three synchronized sensors, specifically a structured-light and a time-of-flight RGB-D sensor and a high-resolution RGB camera. There are approximately 39K training and 10K test images from each sensor. Additionally, two types of 3D models are provided for each object, i.e. a manually created CAD model and a semi-automatically reconstructed one. Training images depict individual objects against a black background. Test images originate from twenty test scenes having varying complexity, which increases from simple scenes with several isolated objects to very challenging ones with multiple instances of several objects and with a high amount of clutter and occlusion. The images were captured from a systematically sampled view sphere around the object/scene, and are annotated with accurate ground truth 6D poses of all modeled objects. Initial evaluation results indicate that the state of the art in 6D object pose estimation has ample room for improvement, especially in difficult cases with significant occlusion. The T-LESS dataset is available online at `cmp.felk.cvut.cz/t-less`.

1. Introduction

Texture-less rigid objects are common in human environments and the need to learn, detect and accurately localize them from images arises in a variety of applications. The pose of a rigid object has six degrees of freedom, *i.e.* three in translation and three in rotation, and its full knowledge is often required. In robotics, for example, the 6D object pose facilitates spatial reasoning and allows an end-effector to act upon an object. In an augmented reality scenario, ob-

Figure 1. Examples of T-LESS test images (left) overlaid with colored 3D object models at the ground truth 6D poses (right). Instances of the same object have the same color. The goal is to find instances of the modeled objects and estimate their 6D poses.

ject pose can be used to enhance one's perception of reality by augmenting objects with extra information such as hints for assembly guidance.

The visual appearance of a texture-less object is dominated by its global shape, color, reflectance properties, and the configuration of light sources. The lack of texture implies that the object cannot be reliably recognized with traditional techniques relying on photometric local patch detectors and descriptors [9, 31]. Instead, recent approaches that can deal with texture-less objects have focused on local 3D feature description [33, 51, 19], and semi-global or

global description relying primarily on intensity edges and depth cues [20, 24, 54, 5, 14, 21, 27]. Therefore, RGB-D data consisting of aligned color and depth images, obtained with widely available Kinect-like sensors, have come to play an important role.

In this paper, we introduce a new public dataset for 6D pose estimation of texture-less rigid objects. An overview of the included objects and test scenes is provided in Fig. 2. The dataset features thirty commodity electrical parts which have no significant texture, discriminative color or distinctive reflectance properties, and often bear similarities in shape and/or size. Furthermore, a unique characteristic of the objects is that some of them are parts of others. For example, objects 7 and 8 are built up from object 6, object 9 is made of three copies of object 10 stacked next to each other, whilst the center part of objects 17 and 18 is nearly identical to object 13. Objects exhibiting similar properties are common in industrial environments.

The dataset includes training and test images captured with a triplet of sensors, *i.e.* a structured light RGB-D sensor Primesense Carmine 1.09, a time-of-flight RGB-D sensor Microsoft Kinect v2, and an RGB camera Canon IXUS 950 IS. The sensors were time-synchronized and had similar perspectives. All images were obtained with an automatic procedure that systematically sampled images from a view sphere, resulting in ~39K training and ~10K test images from each sensor. The training images depict objects in isolation with a black background, while the test images originate from twenty table-top scenes with arbitrarily arranged objects. Complexity of the test scenes varies from those with several isolated objects and a clean background to very challenging ones with multiple instances of several objects and with a high amount of occlusion and clutter. Additionally, the dataset contains two types of 3D mesh models for each object; one manually created in CAD software and one semi-automatically reconstructed from the training RGB-D images. All occurrences of the modeled objects in the training and test images are annotated with accurate ground truth 6D poses; see Fig. 1 for their qualitative and Sec. 4.1 for their quantitative evaluation.

The dataset is intended for evaluating various flavors of the 6D object pose estimation problem [23] and other related problems, such as 2D object detection [50, 22] and object segmentation [49, 17]. Since images from three sensors are available, one may also study the importance of different input modalities for a given problem. Another option is to use the training images for evaluating 3D object reconstruction methods [44], where the provided CAD models can serve as the ground truth.

Our objectives in designing T-LESS were to provide a dataset of a substantial but manageable size, with a rigorous and complete ground truth annotation that is accurate to the level of sensor resolution, and with a significant variability

Figure 2. T-LESS includes training images and 3D models of 30 objects (top) and test images of 20 scenes (bottom) − shown overlaid with colored 3D object models at the ground truth poses. The images were captured from a systematically sampled view sphere around an object/scene and are annotated with accurate ground truth 6D poses of all modeled objects.

in complexity, so that it would provide different levels of difficulty and be reasonably future-proof, *i.e.* solvable, but not solved by the current state-of-the-art methods. The difficulty of the dataset for 6D object pose estimation is demonstrated by the relatively low performance of the method by Hodaň et al. [24]. This method otherwise achieves a performance close to the state of the art on the well-established dataset of Hinterstoisser et al. [20].

The remainder of the paper is organized as follows. Sec. 2 reviews related datasets, Sec. 3 describes technical details of the acquisition and post-processing of the T-LESS dataset, Sec. 4 assesses the accuracy of the ground truth poses and provides initial evaluation results, and Sec. 5 concludes the paper.

2. Related Datasets

First we review datasets for estimating the 6D pose of specific rigid objects, grouped by the type of provided images, then we mention a few datasets designed for similar problems. If not stated otherwise, these datasets supply ground truth annotations in the form of 6D object poses.

2.1. RGB-D Datasets

Only a few public RGB-D datasets, from over one hundred reported by Firman in [15], enable the evaluation of 6D object pose estimation methods. Most of the datasets reviewed in this section were captured with Microsoft Kinect v1 or Primesense Carmine 1.09, which represent the first generation of consumer-grade RGB-D sensors operating on the structured-light principle. The dataset introduced in [17] was captured with Microsoft Kinect v2, which is based on the time-of-flight principle.

For texture-less objects, the dataset of Hinterstoisser et al. [20] has become a standard benchmark used in most of the recent work, *e.g.* [38, 4, 47, 24, 54]. It contains 15 texture-less objects represented by a color 3D mesh model. Each object is associated with a test sequence consisting of ~1200 RGB-D images, each of which includes exactly one instance of the object. The test sequences feature significant 2D and 3D clutter, but only mild occlusion, and since the objects have discriminative color, shape and/or size, their recognition is relatively easy. In the 6D localization problem (where information about the number and identity of objects present in the images is provided beforehand [23]), state-of-the-art methods achieve recognition rates that exceed 95% for most of the objects. Brachmann et al. [4] provided additional ground truth poses for all modeled objects in one of the test sequences from [20]. This extended annotation introduces challenging test cases with various levels of occlusion and allows the evaluation of multiple object localization, with each object appearing in a single instance.

Tejani et al. [47] presented a dataset with 2 texture-less and 4 textured objects. For each object, a color 3D mesh model is provided together with a test sequence of over 700 RGB-D images. The images show several object instances with no to moderate occlusion, and with 2D and 3D clutter. Doumanoglou et al. [14] provide a dataset with 183 test images of 2 textured objects from [47] that appear in multiple instances in a challenging bin-picking scenario with heavy occlusion. Furthermore, they provide color 3D mesh

models of another 6 textured objects and 170 test images depicting the objects placed on a kitchen table.

The Challenge and Willow datasets [58], which were collected for the 2011 ICRA Solutions in Perception Challenge, share a set of 35 textured household objects. Training data for each object is given in the form of 37 RGB-D training images that show the object from different views, plus a color point cloud obtained by merging the training images. The Challenge and Willow datasets respectively contain 176 and 353 test RGB-D images of several objects in single instances placed on top of a turntable. The Willow dataset also features distractor objects and object occlusion. Similar is the TUW dataset [1] that presents 17 textured and texture-less objects appearing in 224 test RGB-D images. Instead of a turntable setup, images were obtained by moving a robot around a static cluttered environment with some objects appearing in multiple instances.

The Rutgers dataset [37] is focused on perception for robotic manipulation during pick-and-place tasks and comprises of images from a cluttered warehouse environment. It includes color 3D mesh models for 24 mostly textured objects from the Amazon Picking Challenge 2015 [11], that were captured in more than 10K test RGB-D images with various amounts of occlusion.

Aldoma et al. [2] provide 3D mesh models without color information of 35 household objects that are both textured and texture-less and are often symmetric and mutually similar in shape and size. There are 50 test RGB-D images of table-top scenes with multiple objects in single instances, with no clutter and various levels of occlusion.

The BigBIRD dataset [42] includes images of 125 mostly textured objects that were captured in isolation on a turntable with multiple calibrated RGB-D and DSLR sensors. For each object, the dataset provides 600 RGB-D point clouds, 600 high-resolution RGB images, and a color 3D mesh model reconstructed from the point clouds. Since BigBIRD was acquired under very controlled conditions, it is not concerned with occlusions, clutter, lighting changes or varying object-sensor distance. Georgakis et al. [17] provide 6735 test RGB-D images from kitchen scenes including a subset of the BigBIRD objects. Ground truth for objects in the test images is provided only in the form of 2D bounding boxes and 3D point labeling.

Lai et al. [29] created an extensive dataset with 300 common household objects captured on a turntable from three elevations. It contains 250K segmented RGB-D images and 22 annotated video sequences with a few hundred RGB-D frames in each. Ground truth is provided only in the form of approximate rotation angles for training images and in the form of 3D point labeling for test images.

Schlette et al. [40] synthesized RGB-D images from simulated object manipulation scenarios involving 4 texture-less objects from the Cranfield assembly benchmark [10].

Several small datasets that were used for evaluation of the SHOT descriptor are provided by Salti et al. [39]. These datasets include synthetic data as well as data acquired with a spacetime-stereo method and an RGB-D sensor.

2.2. Depth-only and RGB-only Datasets

The depth-only dataset of Mian et al. [34] includes 3D mesh models of 5 objects and 50 test depth images acquired with an industrial range scanner. The test scenes contain only the modeled objects that occlude each other. A similar dataset is provided by Taati et al. [46]. The Desk3D dataset [3] comprises of 3D mesh models for 6 objects which are captured in over 850 test depth images with occlusion, clutter and similarly looking distractor objects. The dataset was obtained with an RGB-D sensor, however only the depth images are publicly available.

The IKEA dataset by Lim et al. [30] provides RGB images with objects being aligned with their exactly matched 3D models. Crivellaro et al. [12] supply 3D CAD models and annotated RGB sequences with 3 highly occluded and texture-less objects. Muñoz et al. [36] provide RGB sequences of 6 texture-less objects that are each imaged in isolation against a clean background and without occlusion. Further to the above, there exist RGB datasets such as [13, 50, 38, 25], for which the ground truth is provided only in the form of 2D bounding boxes.

2.3. Datasets for Similar Problems

The RGB-D dataset of Michel et al. [35] is focused on articulated objects, where the goal is to estimate the 6D pose of each object part, subject to the constraints introduced by their joints. There are also datasets for categorical pose estimation. For example, the 3DNet [55] and the UoB-HOOC [53] contain generic 3D models and RGB-D images annotated with 6D object poses. The UBC VRS [32], the RMRC (a subset of NYU Depth v2 [41] with annotations derived from [18]), the B3DO [26], and the SUN RGB-D [43] provide no 3D models and ground truth only in the form of bounding boxes. The PASCAL3D+ [57] and the ObjectNet3D [56] provide generic 3D models and ground truth 6D poses, but only RGB images.

3. The T-LESS Dataset

Compared to the reviewed datasets, T-LESS is unique in its combination of the following characteristics. It contains 1) a larger number of industry-relevant objects, 2) training images captured under controlled conditions, 3) test images with large viewpoint changes, objects in multiple instances, affected by clutter and occlusion; including test cases that are challenging even for the state-of-the-art methods, 4) images captured with a synchronized and calibrated triplet of sensors, 5) accurate ground truth 6D poses for all modeled objects, and 6) two types of 3D models for each object.
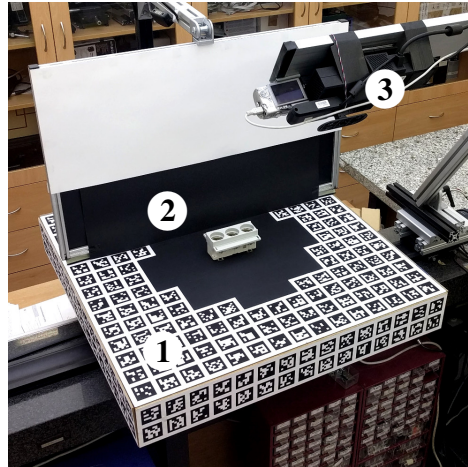

Figure 3. Acquisition setup: 1) turntable with marker field, 2) screen ensuring a black background for training images, removed when capturing test images, 3) triplet of sensors attached to a jig with adjustable tilt.
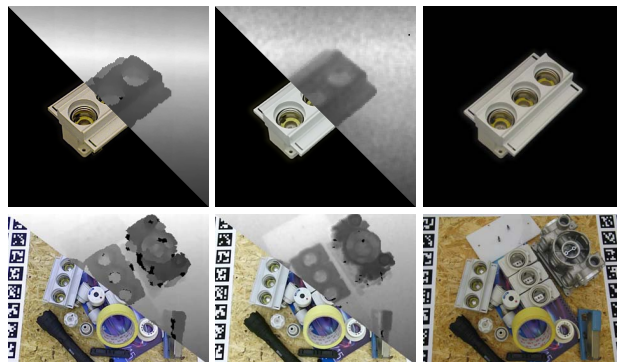

Figure 4. Sample training (top) and test (bottom) images. Left: RGB-D images from Primesense Carmine 1.09. Middle: RGB-D images from Microsoft Kinect v2. Right: High-resolution RGB images from Canon IXUS 950 IS. For the RGB-D images, bottom-left halves show the RGB components whereas the top-right show the depth components.

The rest of the section describes the process of dataset preparation, which includes image acquisition, camera calibration, depth correction, 3D object model generation and the ground truth pose annotation.

3.1. Acquisition Setup

The training and test images were captured with the aid of the setup shown in Fig. 3. It consists of a turntable, where the imaged objects were placed, and a jig with adjustable tilt, to which the sensors were attached. A marker field used for camera pose estimation was affixed to the turntable. The field was extended vertically to the sides of the turntable to

facilitate pose estimation at lower elevations. To capture training images, the objects were placed in the middle of the turntable and in the front of a black screen, which ensured a uniform background at all elevations. To introduce a non-uniform background in the test images, a sheet of plywood with markers at its edges was placed on the top of the turntable. In some scenes, the objects were placed on the top of other objects (*e.g.* books) to give them different elevations and thus invalidate a ground plane assumption that might be made by an evaluated method. The depth of object surfaces in the training and test images is in the range $0.53 - 0.92\,m$, which is within the sensing ranges of the used RGB-D sensors that are $0.35 - 1.4\,m$ for Carmine and $0.5 - 4.5\,m$ for Kinect.

3.2. Calibration of Sensors

Intrinsic and distortion parameters of the sensors were estimated with the standard checkerboard-based procedure using OpenCV [6]. The root mean square re-projection error calculated at corners of the calibration checkerboard squares is $0.51\,px$ for Carmine, $0.35\,px$ for Kinect, and $0.43\,px$ for Canon. For the RGB-D sensors, the calibration was performed with the RGB images. The depth images were aligned to the RGB images using the factory depth-to-color registration available through manufacturer's SDKs (OpenNI 2.2 and Kinect for Windows SDK 2.0). The color and aligned depth images, which are included in the dataset, are already processed to remove radial distortion. The intrinsic parameters can be found at the dataset website.

All sensors were synchronized and extrinsically calibrated with respect to the turntable, making it possible to register any pair of images. Synchronization was essential since the images were taken while the turntable was spinning. The extrinsic calibration was achieved using fiducial BCH-code markers from ARToolKitPlus [52]. Specifically, the detection of particular markers in an image combined with the knowledge of their physical location on the turntable provided a set of 2D-3D correspondences. These were used to estimate the camera pose in the turntable coordinate system by robustly solving the PnP problem and then refining the estimated 6D pose by non-linearly minimizing the cumulative re-projection error with the posest library from [31]. The root mean square re-projection error, which was calculated at marker corners in all test images, is $1.27\,px$ for Carmine, $1.37\,px$ for Kinect, and $1.50\,px$ for Canon. This measure combines errors in sensor calibration, marker field detection and sensor pose estimation and is therefore larger than the aforementioned error in sensor intrinsic calibration.

3.3. Training and Test Images

A common strategy for dealing with poorly textured objects is to adopt a template-based approach trained on object images that are acquired with a dense sampling of viewpoints, *e.g.* [13, 20, 38, 24]. To support such approaches, T-LESS offers training images of every object in isolation from a full view sphere. These images were obtained with a systematic acquisition procedure which uniformly sampled elevation from $85°$ to $-85°$ with a $10°$ step and the complete azimuth range with a $5°$ step. Views from the upper and lower hemispheres were captured separately, turning the object upside down in between. In total, there are $18 \times 72 = 1296$ training images per object from each sensor. Exceptions are objects 19 and 20, for which only views from the upper hemisphere were captured, specifically 648 images from elevation $85°$ to $5°$. These objects are horizontally symmetric at the pose in which they were placed on the turntable, thus the views from the upper hemisphere are sufficient to capture their appearance. Test scenes were captured from a view hemisphere with a $10°$ step in elevation (ranging from $75°$ to $15°$) and a $5°$ step in azimuth. A total of $7 \times 72 = 504$ test images were captured per scene by each sensor.

To remove irrelevant parts of the scene in the images periphery, the provided images are cropped versions of the captured ones. Resolution of the provided images is as follows: $400 \times 400\,px$ for training RGB-D images from Carmine and Kinect, $1900 \times 1900\,px$ for training RGB images from Canon, $720 \times 540\,px$ for test RGB-D images from Carmine and Kinect and $2560 \times 1920\,px$ for test RGB images from Canon. Sample images are shown in Fig. 4.

Parts of the marker field were visible in some of the training images, especially at lower elevations. These were masked to ensure a black background everywhere around the objects. To achieve this, we identified an object mask in an image by back-projecting its CAD model and gradually darkened the image moving from the mask perimeter towards the image border.

3.4. Depth Correction

Similarly to [16, 45], we observed that the depths measured by the RGB-D sensors exhibit a systematic error. To remove it, we collected depth measurements d at projections of the marker corners and computed their expected depth values d_e from the known marker coordinates. The measurements were collected from the depth range $0.53 - 0.92\,m$ in which the objects appear in the training and test images. We found the following linear correction models by least squares fitting: $d_c = 1.0247 \cdot d - 5.19$ for Carmine, and $d_c = 1.0266 \cdot d - 26.88$ for Kinect (depth measured in mm). In [45], only scaling is used for the depth correction. According to Foix et al. [16], a 3-degree polynomial function suffices to correct depth in the $1 - 2\,m$ range. In our case, a narrower range is used and we found a simple linear polynomial to adequately account for the error: the correction reduced the mean absolute difference from the

expected depth d_e from $12.4\,$mm to $2.8\,$mm for Carmine and from $7.0\,$mm to $3.6\,$mm for Kinect. The estimated correction was applied to all depth images, requiring no further action from the dataset user.

3.5. 3D Object Models

For each object, a manually created CAD model and a semi-automatically reconstructed model are available (Fig. 5). Both models are provided in the form of 3D meshes with surface normals at model vertices. Surface color is included only for the reconstructed models. The normals were calculated using MeshLab [7] as the angle-weighted sum of face normals incident to a vertex [48].

The reconstructed models were created using fastfusion, a volumetric 3D mapping system by Steinbrücker et al. [44]. The input to fastfusion were the RGB-D training images from Carmine and the associated camera poses estimated using the fiducial markers (see Sec. 3.2). For each object, two partial models were first reconstructed, one for the upper and another for the lower view hemisphere. The partial models were then aligned using the iterative closest point (ICP) algorithm applied to their vertices. This was followed by manual refinement that ensured correct registration of surface details that are visible only in color. The resulting alignment was applied to the camera poses to transform them into a common reference frame, and the updated poses were used to reconstruct the full object model from all images. These models contained some minor artifacts, *e.g.* small spikes, which were removed manually. It is noted that some of the objects contain small shiny metal parts whose depth is not reliably captured by the current depth sensors; in general, any glossy or translucent surface is problematic. Hence, some of these parts, such as the plug poles, were not reconstructed.

The reconstructed models were aligned to the CAD models using the ICP algorithm and the alignment was further refined manually. Models of both types are therefore defined in the same coordinate system and the provided ground truth poses are valid for both of them. The origin of the model coordinate system coincides with the center of the bounding box of the CAD model.

The geometrical similarity of the two model types was assessed by calculating the average distance from vertices of the reconstructed models to the closest surface points of the corresponding CAD models. The average distance over all object models was found to be $1.01\,$mm, which is very low compared to the size of objects that ranges from $58.13\,$mm for object 13 to $217.16\,$mm for object 8. Distances in the opposite direction, *i.e.* from the CAD models to the reconstructed models, are not informative since some CAD models contain inner parts that are not represented in the reconstructed models. The Metro software by Cignoni et al. [8] was used to measure the model differences.
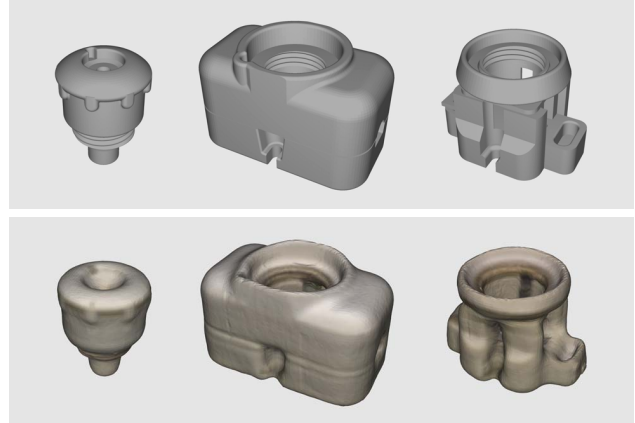

Figure 5. Examples of 3D object models. Top: Manually created CAD models. Bottom: Semi-automatically reconstructed models which include also surface color. Surface normals at model vertices are included in both model types.

3.6. Ground Truth Poses

To obtain ground truth 6D object poses for images of a test scene, a dense 3D model of the scene was first reconstructed with the system of Steinbrücker et al. [44]. This was accomplished using all 504 RGB-D images of the scene along with the sensor poses estimated using the turntable markers. The CAD object models were then manually aligned to the scene model. To increase accuracy, the object models were rendered into several selected high-resolution scene images from Canon, misalignments were identified and the poses were manually refined accordingly. This process was repeated until a satisfactory alignment of the renderings with the scene images was achieved. The final poses were distributed to all test images with the aid of the known camera-to-turntable coordinate transformations. The transformed poses are provided as the ground truth poses with each test image.

4. Design Validation and Experiments

This section presents an accuracy assessment of the ground truth poses and examines the difficulty of T-LESS with a recent 6D localization method.

4.1. Accuracy of the Ground Truth Poses

Aiming to evaluate the accuracy of the ground truth poses, we compared the captured depth images, after the correction described in Sec. 3.4, with depth images obtained by graphically rendering the 3D object models at the ground truth poses. At each pixel with a valid depth value in both images, we calculated the difference $\delta = d_c - d_r$, where d_c is the captured and d_r is the rendered depth. Table 1 presents statistics of these differences, aggregated over all

training and test depth images. Differences exceeding $5\,\mathrm{cm}$ and amounting to around $2.5\,\%$ of the measurements were considered to be outliers and were pruned before calculating the statistics. The outlying differences may be caused by erroneous depth measurements, or by occlusion induced by distractor objects in the case of test images.

The rendered depths align well with the depths captured by Carmine, as indicated by the mean difference μ_δ being close to zero. In the case of Kinect, we observed that the RGB and depth images are slightly misregistered, which is the cause of the positive bias in μ_δ. The average absolute difference $\mu_{|\delta|}$ is less than $5\,\mathrm{mm}$ for Carmine and $9\,\mathrm{mm}$ for Kinect, which is near the accuracy of the sensors [28] and is relatively small compared to the size of objects. The error statistics are slightly favorable for the reconstructed models (as opposed to the CAD models), as they were obtained from the captured depth images and therefore exhibit similar characteristics and artifacts. For example, the plug poles are invisible to the RGB-D sensors and are missing in the reconstructed models, but are present in the CAD models.

4.2. 6D Localization

The recent template-based method of Hodaň et al. [24] was evaluated on the 6D localization problem. The input is comprised of a test image together with the identities of object instances that are present in the image, and the goal is to estimate the 6D poses of these instances [23]. The method was evaluated on all test RGB-D images from the Carmine sensor. The parameters were set as described in [24], the templates were generated from the training images from Carmine, and the CAD models were employed in the pose refinement stage as detailed in [59]. Pose estimates were evaluated as in [20], using the average distance error for objects with indistinguishable views. This error measures the misalignment between the surface of model $\mathcal{M}$ at the ground truth pose $(\bar{\mathbf{R}}, \bar{\mathbf{t}})$ and at the estimated pose $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$, and is defined as:

$$e = \underset{\mathbf{x}_1 \in \mathcal{M}}{\mathrm{avg}} \, \underset{\mathbf{x}_2 \in \mathcal{M}}{\min} \left\| \left(\bar{\mathbf{R}}\mathbf{x}_1 + \bar{\mathbf{t}}\right) - \left(\hat{\mathbf{R}}\mathbf{x}_2 + \hat{\mathbf{t}}\right) \right\|_2 .$$

Pose estimate $(\hat{\mathbf{R}}, \hat{\mathbf{t}})$ is considered correct if $e \le k \cdot d$, where $k = 0.1$ and d is the largest distance between any pair of model vertices, *i.e.* the object diameter. Only the ground truth poses at which at least 10% of the object surface is visible were considered for the evaluation. The visibility was estimated as in [23].

The performance is measured by recall, *i.e.* the percentage of the ground truth poses for which a correct pose was estimated. Fig. 6 presents achieved recall per object (top) and recall per scene (middle). The objects with the lowest recall are those that are similar to other objects. For example, object 1 is often confused with object 2, as are objects 20, 21 and 22. Likewise, test scenes containing similar objects are harder, with the hardest one being scene 20 that

Sensor, model type	μ_δ	σ_δ	$\mu_{|\delta|}$	$med_{|\delta|}$
Carmine, CAD	-0.60	8.12	4.53	2.57
Carmine, reconst.	-0.79	7.72	4.28	2.46
Kinect, CAD	4.46	11.76	8.76	5.67
Kinect, reconst.	4.08	11.36	8.40	5.45

Table 1. Statistics of differences between the depth of object models at the ground truth poses and the captured depth (in mm). μ_δ and σ_δ is the mean and the standard deviation of the differences, $\mu_{|\delta|}$ and $med_{|\delta|}$ is the mean and the median of the absolute differences.
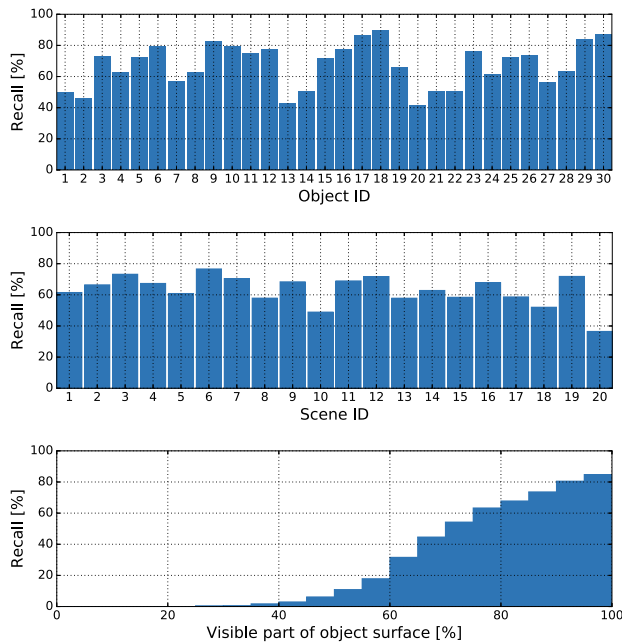

Figure 6. Performance of the method by Hodaň et al. [24] on the 6D localization problem. Shown are the recall per object (top), recall per scene (middle), and recall w.r.t. the percentage of the visible object surface (bottom).

contains many similar objects and severe occlusions. Bottom of Fig. 6 plots the recall, accumulated over all objects, as a function of the fraction of their image projection that is unoccluded. The recall increases proportionally with this fraction, illustrating that occlusion is one of the main challenges in T-LESS.

The achieved mean recall over all objects is 67.2%, which suggests a significant margin for improvement. We note that the same method achieved a mean recall of 95.4% on the dataset of Hinterstoisser et al. [20], which is close to the state of the art: [20] reports 96.6% and [5] reports 99.0% on this dataset. The latter is not directly comparable since it was calculated only over 13 out of 15 objects included in the dataset.

5. Conclusion

This paper has presented T-LESS, a new dataset for evaluating 6D pose estimation of texture-less objects that can facilitate systematic comparison of pertinent methods. The dataset features industry-relevant objects and is characterized by a large number of training and test images, accurate 6D ground truth poses, multiple sensing modalities, test scenes with multiple object instances and with increasing difficulty due to occlusion and clutter. Initial evaluation results using the dataset indicate that the state of the art in 6D object pose estimation has ample room for improvement.

The T-LESS dataset is available online at:
`cmp.felk.cvut.cz/t-less`

Acknowledgements

References

[1] A. Aldoma, T. Fäulhammer, and M. Vincze. Automation of "ground truth" annotation for multi-view RGB-D object instance recognition datasets. In *IROS*, 2014. `repo.acin.tuwien.ac.at/tmp/permanent/dataset_index.php`.

[2] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze. A global hypotheses verification method for 3D object recognition. In *ECCV*, 2012. `users.acin.tuwien.ac.at/aaldoma/datasets/ECCV.zip`.

[3] U. Bonde, V. Badrinarayanan, and R. Cipolla. Robust instance recognition in presence of occlusion and clutter. In *ECCV*, 2014. `sites.google.com/site/ujwalbonde/publications/downloads`.

[4] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6D object pose estimation using 3D object coordinates. In *ECCV*, 2014. `cvlab-dresden.de/iccv2015-occlusion-challenge`.

[5] 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. 2016.

[6] G. Bradski and A. Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc., 2008.

[7] P. Cignoni, M. Callieri, M. Corsini, M. Dellepiane, F. Ganovelli, and G. Ranzuglia. MeshLab: an open-source mesh processing tool. In *Eurographics Italian Chapter Conf.*, 2008.

[8] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, volume 17, pages 167–174. Wiley Online Library, 1998.

[9] A. Collet, M. Martinez, and S. S. Srinivasa. The MOPED framework: Object recognition and pose estimation for manipulation. *IJRR*, 2011.

[10] K. Collins, A. Palmer, and K. Rathmill. The development of a European benchmark for the comparison of assembly robot programming systems. In *Robot technology and applications*. 1985.

[11] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman. Lessons from the Amazon picking challenge. *arXiv preprint arXiv:1601.05484*, 2016.

[12] A. Crivellaro, M. Rad, Y. Verdie, K. M. Yi, P. Fua, and V. Lepetit. A novel representation of parts for accurate 3D object detection and tracking in monocular images. In *ICCV*, 2015. `cvlab.epfl.ch/data/3d_object_tracking`.

[13] D. Damen, P. Bunnun, A. Calway, and W. Mayol-Cuevas. Real-time learning and detection of 3D texture-less objects: A scalable approach. In *BMVC*, 2012.

[14] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim. Recovering 6D object pose and predicting next-best-view in the crowd. In *CVPR*, 2016. `www.iis.ee.ic.ac.uk/rkouskou/research/6D_NBV.html`.

[15] M. Firman. RGBD datasets: Past, present and future. *arXiv:1604.00999*, 2016.

[16] S. Foix, G. Alenya, and C. Torras. Lock-in time-of-flight (ToF) cameras: a survey. *Sensors Journal*, 2011.

[17] G. Georgakis, M. A. Reza, A. Mousavian, P.-H. Le, and J. Kosecka. Multiview RGB-D dataset for object instance detection. *arXiv preprint arXiv:1609.07826*, 2016. `cs.gmu.edu/~robot/gmu-kitchens.html`.

[18] R. Guo and D. Hoiem. Support surface prediction in indoor scenes. In *ICCV*, 2013. `ttic.uchicago.edu/~rurtasun/rmrc/indoor.php`.

[19] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok. A comprehensive performance evaluation of 3D local feature descriptors. *IJCV*, 2016.

[20] 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. In *ACCV*, 2012. `campar.in.tum.de/Main/StefanHinterstoisser`.

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

[22] T. Hodaň, D. Damen, W. Mayol-Cuevas, and J. Matas. Efficient texture-less object detection for augmented reality guidance. In *ISMARW*, 2015.

[23] T. Hodaň, J. Matas, and Š. Obdržálek. On evaluation of 6D object pose estimation. In *ECCV Workshop on Recovering 6D Object Pose*, 2016.

[24] T. Hodaň, X. Zabulis, M. Lourakis, Š. Obdržálek, and J. Matas. Detection and fine 3D pose estimation of texture-less objects in RGB-D images. In *IROS*, 2015.

[25] E. Hsiao and M. Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. *TPAMI*, 2014. `www.cs.cmu.edu/~./hebert/occarbview.html`.

[26] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell. A category-level 3D object dataset: Putting the Kinect to work. In *Consumer Depth Cameras for Computer Vision*. 2013. `kinectdata.com`.

[27] 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. In *ECCV*, 2016.

[28] K. Khoshelham and S. Elberink. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*, 2012.

[29] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view RGB-D object dataset. In *ICRA*, 2011. `rgbd-dataset.cs.washington.edu`.

[30] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing IKEA Objects: Fine Pose Estimation. In *ICCV*, 2013. `ikea.csail.mit.edu`.

[31] M. Lourakis and X. Zabulis. Model-based pose estimation for rigid objects. In *Computer Vision Systems*. 2013. `www.ics.forth.gr/~lourakis/posest`.

[32] D. Meger and J. J. Little. Mobile 3D object detection in clutter. In *IROS*, 2011. `www.cs.ubc.ca/labs/lci/vrs`.

[33] A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes. *IJCV*, 2010.

[34] A. S. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *TPAMI*, 2006. `staffhome.ecm.uwa.edu.au/~00053650/recognition.html`.

[35] F. Michel, A. Krull, E. Brachmann, M. Y. Yang, S. Gumhold, and C. Rother. Pose estimation of kinematic chain instances via object coordinate regression. In *BMVC*, 2015. `cvlab-dresden.de/iccv2015-articulation-challenge`.

[36] E. Muñoz, Y. Konishi, V. Murino, and A. D. Bue. Fast 6D pose estimation for texture-less objects from a single RGB image. In *ICRA*, 2016. `www.iit.it/datasets/vgm-6d-pose-of-texture-less-objects-dataset`.

[37] C. Rennie, R. Shome, K. E. Bekris, and A. F. D. Souza. A dataset for improved RGBD-based object detection and pose estimation for warehouse pick-and-place. *RA-L*, 2016. `www.pracsyslab.org/rutgers_apc_rgbd_dataset`.

[38] R. Rios-Cabrera and T. Tuytelaars. Discriminatively trained templates for 3D object detection: A real time scalable approach. In *ICCV*, 2013.

[39] S. Salti, F. Tombari, and L. D. Stefano. SHOT: Unique signatures of histograms for surface and texture description. *CVIU*. `www.vision.deis.unibo.it/research/80-shot`.

[40] C. Schlette et al. A new benchmark for pose estimation with ground truth from virtual reality. *Production Engineering*, 2014. `www.mmi.rwth-aachen.de/exchange/data/pesi2014/benchmark.htm`.

[41] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV*, 2012. `cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html`.

[42] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel. BigBIRD: A large-scale 3D database of object instances. In *ICRA*, 2014. `rll.berkeley.edu/bigbird`.

[43] S. Song, S. P. Lichtenberg, and J. Xiao. Sun RGB-D: A RGB-D scene understanding benchmark suite. In *CVPR*, 2015. `rgbd.cs.princeton.edu`.

[44] F. Steinbrücker, J. Sturm, and D. Cremers. Volumetric 3D mapping in real-time on a CPU. In *ICRA*, 2014. `github.com/tum-vision/fastfusion`.

[45] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *IROS*, 2012.

[46] B. Taati, M. Bondy, P. Jasiobedzki, and M. Greenspan. Variable dimensional local shape descriptors for object recognition in range data. In *ICCV*, 2007. `rcvlab.ece.queensu.ca/~qridb/lsdPage.html`.

[47] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim. Latent-class hough forests for 3D object detection and pose estimation. In *ECCV*, 2014. `www.iis.ee.ic.ac.uk/rkouskou/research/LCHF.html`.

[48] G. Thürrner and C. A. Wüthrich. Computing vertex normals from polygonal facets. *Journal of Graphics Tools*, 1998.

[49] F. Tombari, L. Di Stefano, and S. Giardino. Online learning for automatic segmentation of 3D data. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4857–4864. IEEE, 2011.

[50] F. Tombari, A. Franchi, and L. Di Stefano. BOLD features to detect texture-less objects. In *ICCV*, 2013.

[51] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *ECCV*, 2010.

[52] D. Wagner and D. Schmalstieg. ARToolKitPlus for pose tracking on mobile devices. In *CVWW*, 2007.

[53] K. Walas and A. Leonardis. UoB highly occluded object challenge II, 2016. `www.cs.bham.ac.uk/research/projects/uob-hooc`.

[54] P. Wohlhart and V. Lepetit. Learning descriptors for object recognition and 3D pose estimation. In *CVPR*, 2015.

[55] W. Wohlkinger, A. Aldoma, R. B. Rusu, and M. Vincze. 3DNet: Large-scale object class recognition from CAD models. In *ICRA*, 2012. `repo.acin.tuwien.ac.at/tmp/permanent/3d-net.org`.

[56] Y. Xiang et al. ObjectNet3D: A large scale database for 3D object recognition. In *ECCV*, 2016. `cvgl.stanford.edu/projects/objectnet3d`.

[57] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond PASCAL: A benchmark for 3D object detection in the wild. In *Winter Conference on Applications of Computer Vision*, 2014.

[58] Z. Xie, A. Singh, J. Uang, K. S. Narayan, and P. Abbeel. Multimodal blending for high-accuracy instance recognition. In *IROS*, 2013. `rll.berkeley.edu/2013_IROS_ODP`.

[59] X. Zabulis, M. Lourakis, and P. Koutlemanis. 3D object pose refinement in range images. In *ICVS*, 2015.