A Large-Scale 3D Object Recognition dataset

Anonymous 3DV submission

Paper ID 4

Abstract

This paper presents a new large scale dataset targeting evaluation of local shape descriptors and 3d object recognition algorithms. The dataset consists of point clouds and triangulated meshes from 292 physical scenes taken from 11 different views; a total of approximately 3204 views. Each of the physical scenes contain 10 occluded objects resulting in a dataset with 32040 unique object poses and 45 different object models. The 45 object models are full 360 degree models which are scanned with a high precision structured light scanner and a turntable. All the included objects belong to different geometric groups; concave, convex, cylindrical and flat 3D object models. The object models have varying amount of local geometric features to challenge existing local shape feature descriptors in terms of descriptiveness and robustness. The dataset is validated in a benchmark which evaluates the matching performance of 7 different state-of-the-art local shape descriptors. Further, we validate the dataset in a 3D object recognition pipeline. Our benchmark shows as expected that local shape feature descriptors without any global point relation across the surface have a poor matching performance with flat and cylindrical objects. It is our objective that this dataset contributes to the future development of next generation of 3D object recognition algorithms. The dataset will be made public available together with this paper.

1. Introduction

Object recognition from range images is a fundamental research area in computer vision with many different applications in different industries. With the continually introduction of new inexpensive 3D sensors for different applications, the ability to localize and recognize rigid and no rigid objects is an attractive and unavoidable technology. Applications areas such as robotic assembly, bin-picking, mobile robotic manipulation, biometric analysis, tracking and intelligent surveillance all benefits from 3D data for localizing objects. Mainly, because the third dimension is explicit given and not inferred as in 2D object pose estimation. In

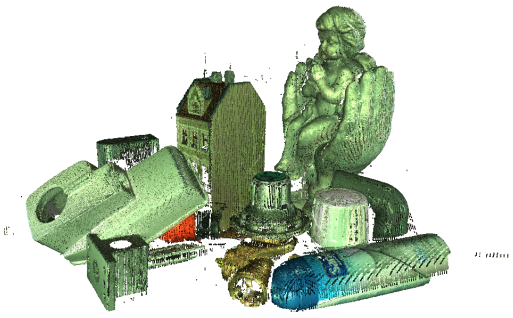


Figure 1: Example of a full registered scene included in the dataset

the last decades many contributions on 2D object recognition and classification have been published, where methods are evaluated on large-scale dataset like the PASCAL Visual Object Challenge (VOC) [8] and ImageNet [6]. The benefit from evaluating algorithms on a large-scale dataset has proven valuable in the continues improvement of recognition algorithms after the release of the PASCAL VOC and ImageNet datasets [8]. In 3D object recognition research, large-scale datasets that consist of a set of 3D querie models Q_n, 3D target scenes S_t and ground truth poses T_{gt} for each object in each scene are required in order to be able to evaluate existing and future 3D object recognition algorithms better. Until now, 3D object recognition algorithms are evaluated with eight smaller datasets [15], including in the UWA [25],[26], Queens [36],[36] and Bologna [33],[40] datasets. Other, smaller dataset like the Vienna Kinect[1], TUM [28], TUM-LineMod[17], BigBird[34] and RGB-D dataset version 1 & 2 [22],[21] are proposed. Common for all datasets is that the amount of scenes and/or models are limited.

In this paper we present a new large-scale dataset consisting of 45 objects and 3204 views. The new dataset is recorded systematically in an environment without ambient light and with controlled illumination. The system includes an industrial robot in a dark chamber, a high precision structured light scanner which records data of 292 different scenes from 11 different view points. Each scene

	M_q	S_t	Sensor	M_q in S_t	M_q normals	M_q mesh	S_t normals	S_t mesh	Full 6D pose	Occlusion	Clutter	[R]/[S]
UWA [25], [26]	5	50	LIDAR	(5)/(5)	✓	✓	✓	✓	✓	✓	%	R
Queens Lidar [37], [36]	5	80	LIDAR	(1-5)/(1-5)	✓	✓	%	✓	✓	%	%	R
Queens Stereo [37], [36]	5	100	Stereo	(3)/(3)	✓	✓	%	✓	✓	%	%	R
Bologna 1&2 [33], [40]	6	45	-	(3-5)/(3-5)	✓	✓	✓	✓	✓	%	%	S
Bologna 3 [33], [40]	8	15	Spacetime	(2)/(5-6)	✓	%	✓	✓	✓	%	%	R
Bologna 4 [33], [40]	8	16	Spacetime	(2)/(6)	✓	%	✓	✓	✓	%	%	R
Bologna 5 [33], [40]	6	16	Kinect V1	(2-4)/(5-9)	✓	%	✓	✓	✓	%	%	R
Vienna Kinect [1]	35	50	Kinect V1	(1-5)/(1-5)	✓	✓	%	%	✓	%	%	R
RGB-D Scenes V1 [22], [23]	5	8 videos	Kinect V1	(0)/(5)	%	%	%	%	%	%	%	R
RGB-D Scenes V2 [21]	9	14	Kinect V1	(0)/(9)	%	%	%	%	%	%	%	R
TÜM [28]	20	150	-	(3-5)/(3-5)	✓	✓	✓	✓	✓	✓	✓	S
Willow [41]	35	177	Kinect V1	*	✓	✓	%	%	%	%	%	R
ECCV12 [1]	35	50	Kinect V1	(3-7)/(3-7)	%	✓	%	%	✓	%	%	R
Alicante [12]	28	9	Kinect V2	*	%	✓	%	0	✓	%	%	R
Our dataset	**45**	**3204**	SL	(10)/(10)	✓	✓	✓	✓	✓	✓	✓	R

Table 1: Comparison of existing datasets for 3D object recognition with the presented datasets. M_q is the amount of different models in the dataset and S_t is the amount of scenes. M_q in S_t shows how many models annotated in each scene and how many objects there are in each scene (No. annotated in scene/No. of object in scene).[R]/[S] indicates whether the dataset is synthetic or acquired in the real world. *(* dataset unavailable online)*

consists of 10 occluded objects, automatically taken with a structured light sensor, which results in a dataset with 32040 unique object poses. With 11 different views of the same scene our dataset is especially suited for studying the effect of view point changes in 3D object recognition. The objects are configured as a classic table-top scenario where different objects are depicted form the side. The dataset is not meant as an evaluation platform for bin-picking and top-view scenarios where many instances of the same object are present in the scene. The table-top scenario is selected because it encapsulates many of the problems and challenges in 3D object recognition in favour of a larger research community. Our 45 object models are scanned in a similar dark chamber with a high precision structured light scanner and a rotation table. With this setup we are able to scan objects with an average point resolution of 275 microns. We validate our dataset by evaluating state of the art local shape features in a 3D object recognition pipeline.

Why does the 3D Object recognition community need another dataset? Previous proposed datasets are limited in the total number of objects, object per. scene and total number of scenes. Additional, the included objects in the different datasets are mostly geometrical ideal objects. With geometrical ideal objects we refer to objects with concave water tight and closed surfaces that are rich of local descriptive geometrical features, like the UWA-chef model [25],[26], the bologna-Armadillo [33],[40] and the Queens-BigBird [36],[36]. In our dataset we include not only ideal concave objects, but flat, cylindrical, feature-rich, simple concave and convex objects. It is expected that some local shape features will have a poor performance on our dataset because the lack of local geometric features. However, our main goal for proposing this dataset is to highlight the challenges in 3D object recognition research and strengthen the data foundation in future algorithm development. This work is initiated by our previous experiences on pose estimation of geometric simple objects in an industrial robotic context. We hope that new novel methods for 3D object recognition will emerge in the future, as a side-effect of this dataset. Not only local shape descriptors but to a great extend features that use point relations across the surface e.g. Point Pair Features [2], semi-global features and template matching. We have selected the different objects based on previous experience, as a combination of lab-objects and industrial objects. The industrial objects are provided by companies that need to pick the objects from boxes or bins with a robot. Thus, the dataset is a mix of real industrial objects that we know from experience is difficult to detect and objects from our lab which are more suited for a 3D object recognition pipeline with local shape features.

This paper is structured as follows: In Section 2 related datasets and local feature descriptors are presented. Section

3 outlines our experimental design followed by Section 4 which presents our benchmarking methodology. In Section 5 the results are given followed by a discussion and conclusion in Section 6.

2. Related work

We will now relate our work to state of the art. First, we review existing 3D object recognition datasets followed by a concise review of significant local feature descriptors.

3D Object Recognition datasets:
The last 10 years a few 3D object recognition datasets have been published. Mainly in conjunction with algorithms for local feature description [25],[36],[37],[40],[33], correspondence matching and rejection [28], pose hypothesis verification [1] and 3D keypoint detection [26]. Common for all existing dataset are the limited number of 3D object models and scenes. A comparison of the different datasets are presented in Table 1, from which it is seen that our dataset is magnitudes larger. The most common used datasets are the UWA [25], [26], Queens [36],[37] and the Bologna [33], [40]. These datasets are widely used in performance evaluations of local feature descriptors [15], [3], keypoint detectors [32] and surveys [14], [9]. The main problems of all the datasets are; size, variety of objects, few objects per. scene and missing occlusion/clutter estimates. In this work we are not considering 2.5D recognition methods where either a full object model or sampled templates is used for recognizing object in a RGB-D image. However, some datasets exist for this problem e.g. the Line-Mod dataset [17] or the recent published RGB-D dataset for warehouse pick-and-place tasks [27].

Local feature descriptors:
During the last three decades, a vast number of different 3D local feature descriptors have been proposed including SPLASH [35], Spin Image (SI) [18], 3D Shape Context (3DSC) [11], LSP [4], 3D Tensors [25], THRIFT [10], MESH-HOG [42], ISS [43], Unique shape context (USC) [39], Point Feature Histogram (PFH) [30], Fast Point Feature Histograms (FPFH) [29], SHOT [40], ROPS [16], ECSAD [19] and Tri-Spin-Image (TriSI)[15]. The descriptors find usages in applications such as 3D object categorization, recognition, retrieval, analysis, registration and reconstruction among others. Designing descriptors which are distinctive and robust toward occlusion and noise is still an ongoing research topic.

Local feature descriptors aim at computing a distinctive and robust N-dimensional feature vector around a point, by considering the points in an Euclidean neighbourhood. The support radius determining the size of the neighbourhood is often one of the critical parameters in the pipeline. Local feature descriptors are often split into two different categories, spatial and geometrical histograms [33]. Recent studies have shown that the state of the art features are not generalizing well over many types of geometry classes, [15],[3]. The studies proved one of the main issues in 3D object recognition today, that there exist no local shape feature which describes the geometry well over many different object classes, e.g. flat, rotational symmetrical and geometrically feature rich objects. What feature(s) to use are still dependent on the object geometry. However, the experimental results in [3] showed the advantage of fusing several local feature descriptors. The studies from [15],[3] leave a relevant research question to be answered; are the local feature descriptors not descriptive enough to generalize different object classes or are the data used for evaluation too limited? The evaluation datasets used for the evaluation in general and in [15],[3] consist mainly of objects with many descriptive and distinctive geometrical features.

3. Experimental design

We have constructed a dataset by achieving highly accurate 3D model of each of the 45 individual objects, as described in Section 3.1. These are then used to compose 292 individual scenes where 10 objects are placed in different configurations. A robot moves our scanner to 11 fixed positions in order to create 11 independent view points of the same object configuration, as described in Section 3.2. Hence, we have 3204 individual observations of 10 objects included in the dataset. We argue that these observations are independent because of the large view point change at 36 degree horizontal and 45 degree vertical. Thus, the objects surface are very different in each view. We achieve very accurate ground truth poses by annotating the entire dataset with our high resolution object models in full resolution, as described in Section 3.3.

3.1. Object model scanning

Our object models are scanned with a high precision structured light setup consisting of two industrial cameras (Point Grey Research GS3-U3-91S6C-C) and a high resolution DLP projector (LG PF80G) mounted on a rigid aluminium frame. In addition, a high precision turntable (New-

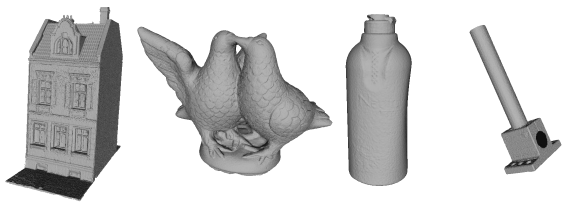


Figure 2: A sample of the scanned object models

mark Systems RT-5) is used in order to provide automatic rotation of the object. Each of the 45 objects are incremental scanned with a rotation of 20 degrees. All individual scan views are reconstructed by the Line shifting algorithm [13], which results in accurate and dense point clouds of the objects. Eleven temporal binary gray code patterns are projected followed by eight line shifting patterns. The point resolution of the scanned objects are in average 275 microns and consist of around 908000 vertices and 1.8 million faces in average. Once a single view of the model is scanned, noisy outliers of the measurement are manual removed and surface normals are estimated to ensure consistent normals. All 18 views are registered with Iterative Closest Point (ICP) and a new object frame is computed with principal component analysis on the point set. All models are sampled with a Poisson disk sampling algorithm [5] and triangulated with the poisson reconstruction algorithm [20] with an octree depth=14, solver divide = 8 and iso divide = 5. We use the PCL implementation [31]. All models are provided as coloured point clouds and triangular meshes, all in the *.ply* format, see Figure 2. Note that the modelling setup is not radiometrical calibrated.

3.2. Scene scanning

For the data collection we use a 6-axis ABB IRB 1600 industrial robot to provide a precise and highly repeatable camera pose. The robot is equipped with two PointGray Grasshopper3 GS3-U391S6C-C USB3 color cameras with resolution of 9.1 Mp and a Wintech Pro4500 projector with a resolution of 1140x912 pixels. The sensor cluster with the structured light sensor (SL) is mounted in the robot tool. The Robot setup is constructed as a radiometric "dead" which gives a scene representation with zero ambient light. All scene illumination is controlled in the recording process. The sensor cluster is calibrated with an automatic calibration procedure which includes a stereo calibration of the SL sensor using `OpenCV`[1].An automatic hand eye calibration [7] is conducted in order to align the structured light scans in the world frame. The world frame is placed in the robot base frame. Each physical scene is scanned from 11 different views with the structured line scanner (SL). During the structured light scanning process the chamber is completely dark in order to increase the signal-to-noise ration of the scan. The structured light scans are reconstructed with the Line shifting algorithm [13] with ten temporal pattern levels. The views are distributed equally on a quarter sphere around the scene, such that each scene is depicted from -90 to 90 degrees horizontally and 0 to 45 degree vertically. The distance between the sensor views and the center of the scene is 0.8 meter. In order to be able to reproduce the results we store all images in raw Portable Network Graphics in full resolution (9.1 MP). De-bayering, white balance, image rectification, pattern decoding and point cloud reconstruction are a off line process. Local shape features like

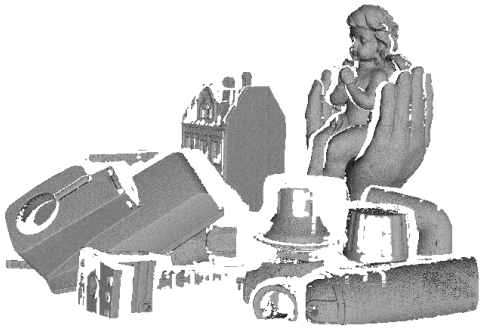


Figure 3: Example on one of the 3204 triangulated scene

3D Tensor [25], ROPS [16] and FPFH [29] apply the underlying mesh surface during feature computation. In order for the dataset to support these algorithms all scenes are triangulated to a polygonal mesh. For each scene a set of corresponding triangles are computed with a 2D delaynay triangulation algorithm in VTK[2]. Each (x, y) point coordinates of the scene point cloud are normalized with its z component in order to project the 3D point cloud into 2D and triangulated. The 3D mesh structures are created by reassigning all 3D point with the triangles of the same index. Long edges in the mesh are then removed form the mesh in order to avoid triangles between step edges. An example of a triangulated scene is shown in Figure 3.

3.3. Ground Truth 6D pose

For each of the 3204 scenes the 6D ground truth pose, occlusion and clutter estimates for each object are provided. We estimate the occlusion from Equation 1 and clutter from Equation 2 by counting the number of points at the object model in which the squared euclidean distance to a scene point is less than two times the scene resolution. Both the object model and the scene are sampled to ensure equal point distance.

$$Occlusion = 1 - \frac{visible\ object\ points}{total\ object\ points} \quad (1)$$

$$Clutter = 1 - \frac{visible\ object\ points}{total\ scene\ points} \quad (2)$$

Accurate ground true poses are ensured by manual annotation of each object in the scenes. First, one point cloud that covers 180 degree of the scene is stitched together form each of the 11 views and registered with ICP. This full scene point cloud is applied in the annotation process, see Figure 1. The full scene point cloud covers the geometric structures of each model in the scene with more points compared to the

[1] `http://opencv.org/`

[2] `http://www.vtk.org/`

individual views. Thus, it is possible to get a more accurate ICP registration of the object model in the scenes. All 290 combined scenes are manually annotated by selecting four identical points for each model and the scene, followed by an estimation of the rigid transform. An iterative ICP, which incremental decreases the allowed correspondence distance ensures a very accurate final ground truth pose of each object in the scene. The final ICP iterations are accomplished with the full resolution model to get the best fit of the model points to the scene points. The individual ground truth poses in each sensor view T_{sensor_n} is computed by transforming the ground truth poses from the world frame T_{world} to each of the individual sensor frames T_{sensor_n}. Equation 3 shows the final transformation.

$$T^{world}_{sensor_n} = T^{-1}_{robot} \cdot T^{-1}_{HandEye} \cdot T^{-1}_{icp} \cdot T_{gt} \qquad (3)$$

where T_{robot} is the known robot pose for each view point, $T_{HandEye}$ is the calibrated hand eye transform, T_{icp} is the alignment transform which align view n to view 0 in the world frame and T_{gt} is the annotated ground truth pose in the full scene point cloud with the world frame as reference. This methodology guarantees accurate pose in $Tsensor_n$ independent of the amount of occlusion. Even in views with limited number of scene points of an object the ground truth pose is accurate because the ground truth pose is computed in the full scene point cloud. For each ground truth pose in each sensor view, we compute the RMS error between the model and the scene to guarantee the overall accuracy of all ground truth poses. On average the RMS error of all ground truth poses is within 0.15 mm.

4. Benchmark

This section outlines the experimental protocol defined to validate the dataset. The protocol is inspired by Salti *et al*. [33]. The evaluation is divide into two parts; feature matching accuracy and object recognition rate. The selected local feature descriptors for our evaluation include; Spin Image (SI) [18], PFH [30], FPFH [29], USC [39], SHOT [40], ROPS [16], ECSAD [19] and NDHIST [3]. The features are selected based on implementation availability and results from previous studies on feature descriptor benchmarking [15],[3].

4.1. Feature Matching

The descriptiveness and accuracy of a feature descriptor are measured with Precision-Recall and presented as 1-Precision vs. Recall Cruves (PRC). First we sample both the query models and the target scenes with a voxel grid sampling [31] which results in equal point distance. The voxel size is tuned to give approximately 1000 seed points per. object in both the query and target. The target seed points are found by transforming the query seeds into the target by applying the object ground truth pose. The target seeds are selected in a nearest neighbour search with a distance threshold. A feature descriptor for each seed point in the query and target mesh is computed. For a fair comparison individually tuned support radii for each descriptors are used. We use the following feature resolution multiplier: SI(20), 3DSC(22.5), FPFH(17.5), USC(25), SHOT(17.5), ROPS(20), ECSAD(20), NDHIST(31). Hence, the radius for each feature is a function of the average model or scene resolution. Upon feature computation the underlying scene and object meshes are utilized in 0.25 and 0.05 decimated version; respectively. The level of decimation are empirical determined to the level with best overall matching results for all features. During decimation the normal orientation is re-computed for each vertex by the area weighted mean of the mesh triangle [38] and normalized. In order to resolve the exact number of correct feature matches, a brute-force linear kd-tree search is used with a L_2 distance function. Other distance metrics such as L_1, L_∞, are tested in previous work but the best results are achieved with the L_2 distance metric. Thus, we only present results for the L_2 metric. During matching we are computing the ratio of the nearest and the second-nearest matching distances. This matching strategy is adapted from multiple previous studies e.g. Lowe *et al*. [24] that proved a performance enhancement compared to a native matching strategy where only the nearest neighbour is considered. Once all matches for all queries are computed they are ranked and sorted according to the L_2 distance in one array. The correct matches are found by traversing the array of matches and count the number of matches that are spatial close, determined by a distance threshold. The PRC curves are presented in Section 5 where precision refers to the number of correct matches compared to the total number of matches. Recall refers to the number of correct matches compared to the total amount of possible matches (i.e. feature seed points found in the target). In addition to PRC curves, we compute the area under the PRC curve (AUC) as a single quantitative measure of the overall accuracy. The AUC is computed as the numerical integration over all (P,R) per feature.

4.2. Pose estimation

In this section the experimental protocol for the pose estimation experiments is presented. The sampling and seed point selection are identical with the feature matching benchmark presented, except that we cannot use the ground truth pose for selecting target seed points. Instead, the target resolution is doubled, thus quadrupling the number of feature descriptors which increase the chance for describing the same feature. To increase the efficiency during matching we apply approximate nearest neighbour search to determine correspondences hypotheses instead of exact matching. Again, the ratio of the nearest and second nearest

neighbour feature distances are used. A multiple randomized kd-trees with a bound of 512 checks and 4 trees are used as a good trade-off between accuracy and efficiency. Correspondences are ranked by the L_2 distance which inputs potential feature correspondences to a hypothesis and test RANSAC algorithm. During random sampling, three correspondences are sampled which is sufficient to generate a hypothesis pose. The hypothesis pose is tested by transforming the query points and counting the number of query points close to the target feature up to a tolerance given by the inlier threshold. The algorithm filters out false positives by setting a lower minimum of the number of inliers required to accept a pose hypothesis to 1%. The pose with the highest number of inliers is returned as the object pose. Our RANSAC implementation deviates from classic RANSAC, which treats all data points uniformly. Instead we sample correspondences according to their quality score. The quality scores are given by the negative normalized L_2 distance ratio. The efficiency of the algorithm is further increased by only considering the top 10% of the best correspondences with the highest quality score before running the RANSAC algorithm. Upon RANSAC completion the final pose is refined by 150 ICP iterations on the query/target seed points. We accept a pose estimate as valid by computing the euclidean and geodesic distances between the computed pose and the ground true pose from the annotation process. If the euclidean and geodesic distances are less than the threshold the object is correctly recognized. The euclidean and geodesic distance metrics are computed in accordance to Equation 4 and 5.

$$arccos\left(\frac{trace(\mathbf{R}^T\hat{\mathbf{R}})-1}{2}\right) \leq 7.5^\circ \tag{4}$$

$$||\mathbf{t}-\hat{\mathbf{t}}|| \leq 5mm \tag{5}$$

The recognition rate is computed as the ratio of true positive poses compared to all detected poses as a quantitative measure of the overall recognition performance.

5. Evaluation

In this section we present the results of our evaluation benchmark. All experiments are based on the proposed dataset where 3204 views, 45 objects and 5675 ground truth poses are included. As a first evaluation we benchmark the matching accuracy of the seven feature descriptors with the parameters outlined in Section 4.1 and all 45 object models included. The PRC curve of the overall matching accuracy is presented in Figure 4. As expected the total matching accuracy is much lower compared to previous studies e.g. Guo *et al.* [15] and Buch *et al.* [3], where both are running a similar benchmark but on previous proposed datasets. The result indicates that our objective of the proposed dataset has been

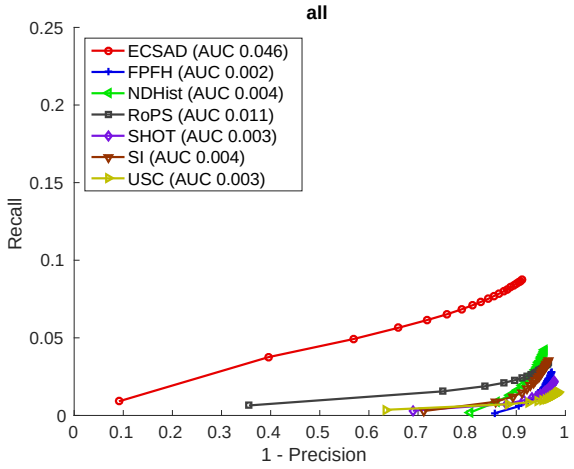


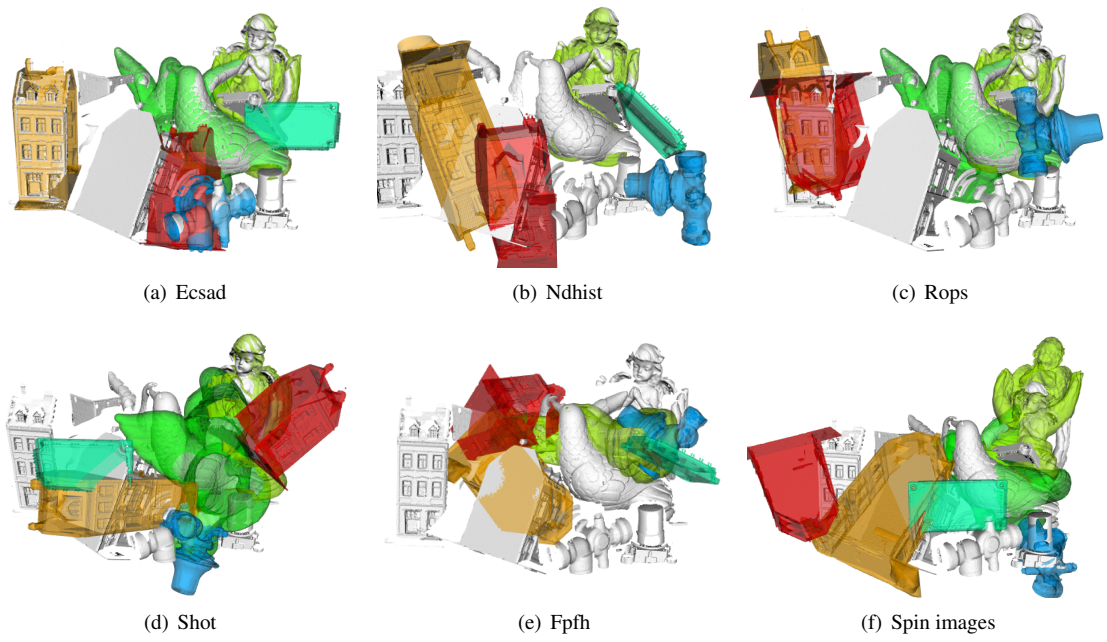
Figure 4: Overall PRC curve

fulfilled. Moreover, the evaluation shows that the ECSAD feature has a best performance with a AUC = 0.046 followed by ROPS with AUC = 0.011. In order to investigate the performance further we run the matching benchmark for each object model and compute PRC curves for each of the 45 models. We have categorize the object into three different groups and present 3 PRC curves for each group in Figure 7-14 as a sample. The three groups are a) Geometric complex objects, b) Cylindrical objects and c) Flat or box shaped objects. The objects that corresponds to the PRC curves in Figure 7-14 are presented in Figure 6. The geometric complex objects are the Angel, the Birds and the Rabbit in Figure 6(a)-(c), the cylindrical objects are Neutral, Pringles and Hand soap in Figure 6(d)-(f), and the flat or box shaped objects are the button, the brake disc and the Psu in Figure 6(g)-(i). The results show that a recent precision/recall is achieved with the Angel,Birds and Rabbit, but it is much lower than previous studies e.g. Buch *et al.* [3] who obtain very matching accuracy for some datasets. These low numbers for our geometric ideal objects indicate that the our dataset has the desired level of complexity. Regarding, the cylindrical objects which features the Neutral, Pringles and Hand Soap objects, it is clear that current local shape descriptors are very little descriptive for these uniform shaped objects. Again, ECSAD performs in general best which might results from ECSAD' ability to capture edges. For the flat and boxed shaped models we can conclude that current state of the art feature descriptors is not suitable as the only detection method.

The results for the pose estimation experiments are presented in Table 2 and Figure 5. Our object recognition pipeline in these experiment is in accordance to the presented pipeline in Section 4.2. In the first recognition experiment we run the pipeline with all 45 object model

Feature	Overall	Angel	Birds	Rabbit	Neutral	Pringels	Hand soap	Button	Brake Disc	Psu	*Mean (λ)*
ECSAD	0.21	**0.59**	**0.86**	**0.62**	0.00	0.06	**0.03**	0.00	0.00	0.00	**0.24**
FPFH	0.01	0.07	0.01	0.04	0.00	0.00	0.02	0.00	0.00	0.00	0.02
NDHIST	0.02	0.13	0.05	0.14	**0.02**	**0.06**	0.02	0.00	0.00	0.00	0.05
ROPS	0.13	0.45	0.70	0.29	0.01	0.00	0.00	0.00	0.00	0.00	0.16
SHOT	0.04	0.25	0.09	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.05
SI	0.05	0.20	0.16	0.10	0.01	0.00	0.00	0.00	0.00	0.00	0.05
USC	0.00	0.01	0.04	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.02

Table 2: Overall recognition rates. **Column 1:** Features descriptors. **Column 2:** Overall recognition rate. **Column 3-11:** Recognition rate for each sample object in Figure 6. **Column 12:** Mean recognition rate for the 9 sample object per. feature.



(a) Ecsad (b) Ndhist (c) Rops

(d) Shot (e) Fpfh (f) Spin images

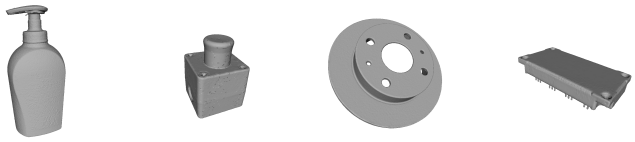
Figure 5: Qualitative recognition results for view 94

included and present the recognition rate in Table 2, second column. In the second experiment we run the recognition benchmark for each individual object, which is a more complex problem because we in the benchmark remove the scene points corresponding to each detected object. Hence, in case of more objects to recognize, the number of scene points are reduced each time the algorithm detect an object. Again, as expected ECSAD and ROPS perform best on average. However, the recognition rate is lower than seen in other datasets which is expected since some views in our dataset are heavy occluded. In Figure 5 qualitative results are presented for the recognition result in view 94. Again, is is clear that ECSAD and ROPS perform best with 4 and 2 correct recognize objects, respectively.

6. Conclusion

This paper has introduce a new large scale dataset for 3D object recognition and a evaluation benchmark to validate the dataset. Our benchmark results show as expected a general low matching score due to the level of complexity of the dataset. Especially, repetitive, symmetric, flat and thin-edge objects without many local features demonstrate as expected a very low precision/recall. Our matching results shows that the ECSAD feature performs best followed by the ROPS feature. This result is in accordance to some of the experiment in a previous benchmark by Buch *et al.* [3]. Our object recognition results reflected as expected the low matching accuracy from the matching experiments. From the evaluation we conclude that the dataset full-fill our requirement in terms of difficulty. Furthermore, the objects and scenes included in the dataset represents the real world problems, which 3D object recognition systems need to handle in the future. Our main goal of this work has been to challenge existing 3D object recognition algorithms and create a dataset which contains real world object from the robot and automation industry.

(a) Angel (b) Birds (c) Rabbit (d) Neutral (e) Pringels



(f) Hand soap (g) Button (h) Brake disc (i) Psu

Figure 6: (a)-(c) Geometric complex objects used for dataset verification. (d)-(e) Cylinderical objects. (f)-(h) Flat and box shaped objects

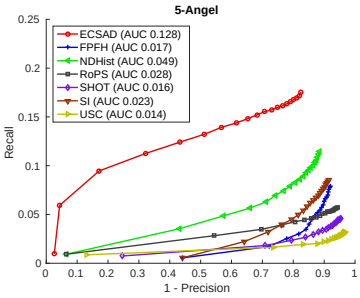


Figure 7: PRC curve: Angel

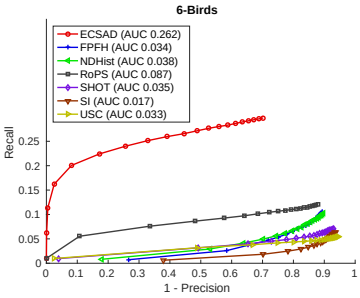


Figure 8: PRC curve: Birds

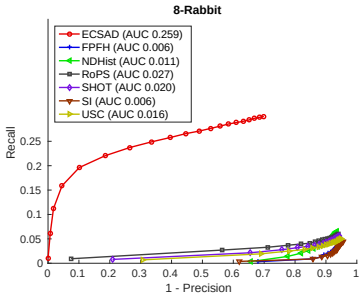


Figure 9: PRC curve: Rabbit

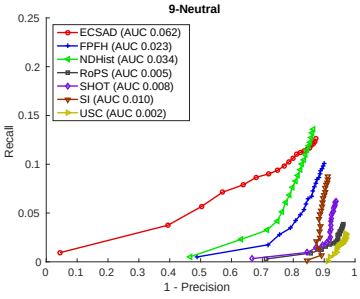


Figure 10: PRC curve: Neutral

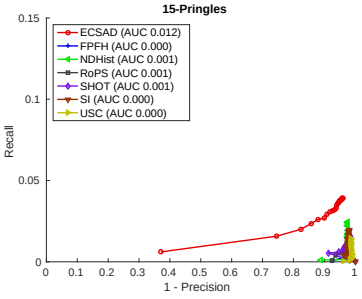


Figure 11: PRC curve: Pringles

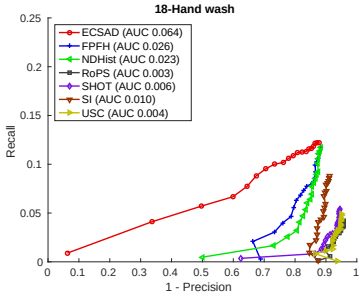


Figure 12: PRC curve: Hand soap

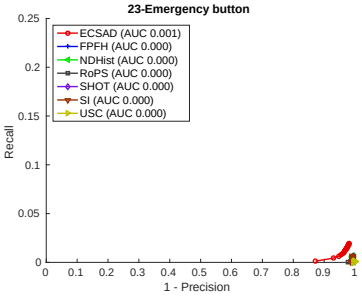


Figure 13: PRC curve: Button

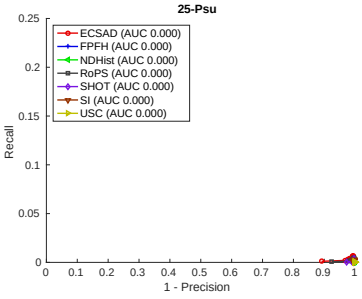


Figure 14: PRC curve: Psu

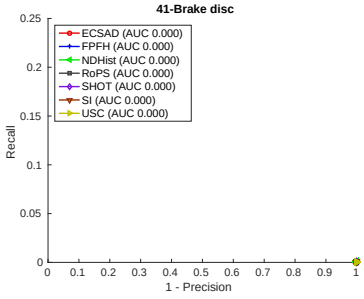


Figure 15: PRC curve: Brake disc

References

[1] A. Aldoma, F. Tombari, L. Di Stefano, and M. Vincze. A global hypotheses verification method for 3d object recognition. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part III*, ECCV'12, pages 511–524, Berlin, Heidelberg, 2012. Springer-Verlag.

[2] T. Birdal and S. Ilic. Point pair features based object detection and pose estimation revisited. In *2015 International Conference on 3D Vision, 3DV 2015, Lyon, France, October 19-22, 2015*, pages 527–535, 2015.

[3] A. G. Buch, H. G. Petersen, and N. Krüger. Local shape feature fusion for improved matching, pose estimation and 3d object recognition. *SpringerPlus*, 5(1):1–33, 2016.

[4] H. Chen and B. Bhanu. 3d free-form object recognition in range images using local surface patches. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 136–139 Vol.3, Aug 2004.

[5] M. Corsini, P. Cignoni, and R. Scopigno. Efficient and flexible sampling with blue noise properties of triangular meshes. *IEEE Transactions on Visualization and Computer Graphics*, 18(6):914–924, June 2012.

[6] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255, June 2009.

[7] F. Dornaika and R. Horaud. Simultaneous robot-world and hand-eye calibration. *IEEE Transactions on Robotics and Automation*, 14(4):617–622, Aug 1998.

[8] M. Everingham, S. M. A. Eslami, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2014.

[9] S. Filipe and L. A. Alexandre. A comparative evaluation of 3d keypoint detectors in a RGB-D object dataset. In *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 1, Lisbon, Portugal, 5-8 January, 2014*, pages 476–483, 2014.

[10] A. Flint, A. Dick, and A. v. d. Hengel. Thrift: Local 3d structure recognition. In *Digital Image Computing Techniques and Applications, 9th Biennial Conference of the Australian Pattern Recognition Society on*, pages 182–188, Dec 2007.

[11] A. Frome, D. Huber, R. Kolluri, T. Bülow, and J. Malik. *Computer Vision - ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part III*, chapter Recognizing Objects in Range Data Using Regional Point Descriptors, pages 224–237. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[12] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, J. Garcia-Rodriguez, J. Azorin-Lopez, M. Saval-Calvo, and M. Cazorla. Multi-sensor 3d object dataset for object recognition with full pose estimation. *Neural Computing and Applications*, pages 1–12, 2016.

[13] J. Guehring. Dense 3d surface acquisition by structured light using off-the-shelf components. volume 4309, pages 220–231, 2000.

[14] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, and J. Wan. 3d object recognition in cluttered scenes with local surface features: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2270–2287, Nov 2014.

[15] Y. Guo, M. Bennamoun, F. Sohel, M. Lu, J. Wan, and N. M. Kwok. A comprehensive performance evaluation of 3d local feature descriptors. *International Journal of Computer Vision*, 116(1):66–89, 2015.

[16] Y. Guo, F. Sohel, M. Bennamoun, M. Lu, and J. Wan. Rotational projection statistics for 3d local surface description and object recognition. *International Journal of Computer Vision*, 105(1):63–86, 2013.

[17] S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real-time detection of textureless objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):876–888, May 2012.

[18] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):433–449, May 1999.

[19] T. B. Jørgensen, A. G. Buch, and D. Kraft. *Geometric Edge Description and Classification in Point Cloud Data with Application to 3D Object Recognition*. 2015.

[20] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, SGP '06, pages 61–70, Aire-la-Ville, Switzerland, Switzerland, 2006. Eurographics Association.

[21] K. Lai, L. Bo, and D. Fox. Unsupervised feature learning for 3d scene labeling. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 3050–3057, May 2014.

[22] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824, May 2011.

[23] K. Lai, L. Bo, X. Ren, and D. Fox. Detection-based object labeling in 3d scenes. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 1330–1337, May 2012.

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004.

[25] A. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1584–1601, Oct 2006.

[26] A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision*, 89(2-3):348–361, 2010.

[27] 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. *CoRR*, abs/1509.01277, 2015.

[28] E. Rodol, A. Albarelli, F. Bergamasco, and A. Torsello. A scale independent selection process for 3d object recognition in cluttered scenes. *International Journal of Computer Vision*, 102(1-3):129–145, 2013.

[29] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 3212–3217, May 2009.

[30] R. B. Rusu, N. Blodow, Z. C. Marton, and M. Beetz. Aligning point cloud views using persistent feature histograms. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3384–3391, Sept 2008.

[31] R. B. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4, May 2011.

[32] S. Salti, F. Tombari, and L. D. Stefano. A performance evaluation of 3d keypoint detectors. In *2011 International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pages 236–243, May 2011.

[33] S. Salti, F. Tombari, and L. D. Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125(0):251 – 264, 2014.

[34] A. Singh, J. Sha, K. Narayan, T. Achim, and P. Abbeel. Bigbird: A large-scale 3d database of object instances. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 509–516, May 2014.

[35] F. Stein and G. Medioni. Structural indexing: efficient 3-d object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):125–145, Feb 1992.

[36] B. Taati, M. Bondy, P. Jasiobedzki, and M. Greenspan. Variable dimensional local shape descriptors for object recognition in range data. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007.

[37] B. Taati and M. Greenspan. Local shape descriptor selection for object recognition in range data. *Computer Vision and Image Understanding*, 115(5):681 – 694, 2011. Special issue on 3D Imaging and Modelling.

[38] G. Thürmer and C. A. Wüthrich. Computing vertex normals from polygonal facets. *J. Graph. Tools*, 3(1):43–46, Mar. 1998.

[39] F. Tombari, S. Salti, and L. Di Stefano. Unique shape context for 3d data description. In *Proceedings of the ACM Workshop on 3D Object Retrieval*, 3DOR '10, pages 57–62, New York, NY, USA, 2010. ACM.

[40] F. Tombari, S. Salti, and L. Di Stefano. Unique signatures of histograms for local surface description. In *Proceedings of the 11th European Conference on Computer Vision Conference on Computer Vision: Part III*, ECCV'10, pages 356–369, Berlin, Heidelberg, 2010. Springer-Verlag.

[41] W. G. . T. U. Wien. The willow garage object recognition challenge, 2015. [Online; accessed 5-May-2015].

[42] A. Zaharescu, E. Boyer, K. Varanasi, and R. P. Horaud. Surface feature detection and description with applications to mesh matching. In *International Conference on Computer Vision and Pattern Recognition, CVPR'09, June, 2009*, pages 373–380, Miami, Etats-Unis, June 2009. IEEE.

[43] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 689–696, Sept 2009.