Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd

Supplementary Material

Andreas Doumanoglou[1,2], Rigas Kouskouridas[1], Sotiris Malassiotis[2], Tae-Kyun Kim[1]

[1]Imperial College London
[2]Center for Research & Technology Hellas (CERTH)

This supplementary material contains additional details that, because of space limitations, did not fit in our main submission. This material is not required for understanding the main paper, however it gives more details on our new dataset and shows more qualitative results (Section 1). Also in Section 4 we give the complete table of the results of the dataset proposed by [3].

1 Proposed Dataset Description

Our proposed dataset consists of two usage scenarios. One is related to domestic environments, where everyday objects are placed on a kitchen table. The second depicts a bin-picking scenario mainly found in industrial settings where a robot should pick objects successively from a bin, which contains many stacked objects of the same or different categories. In the following subsections we describe each usage scenario.

1.1 Usage Scenario 1

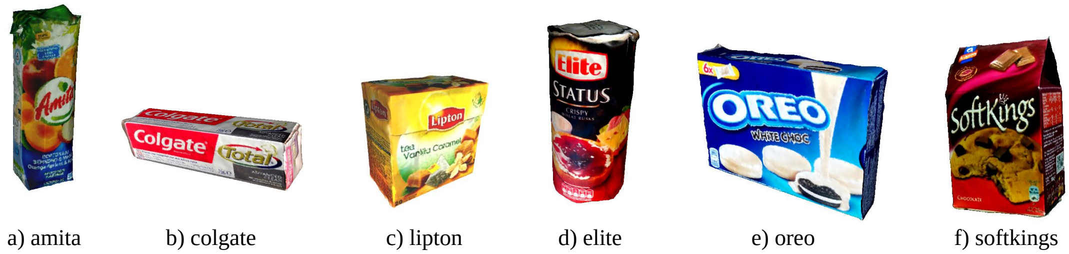
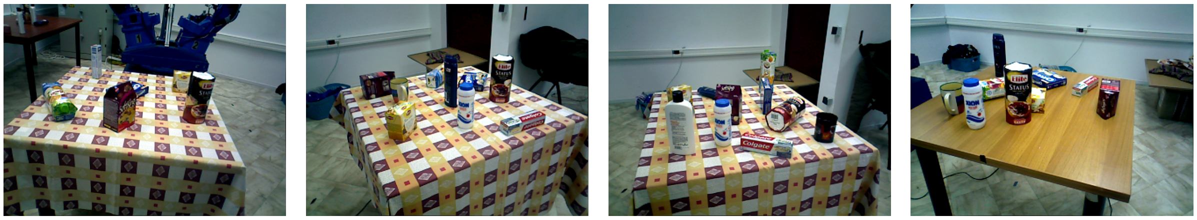

a) amita b) colgate c) lipton d) elite e) oreo f) softkings

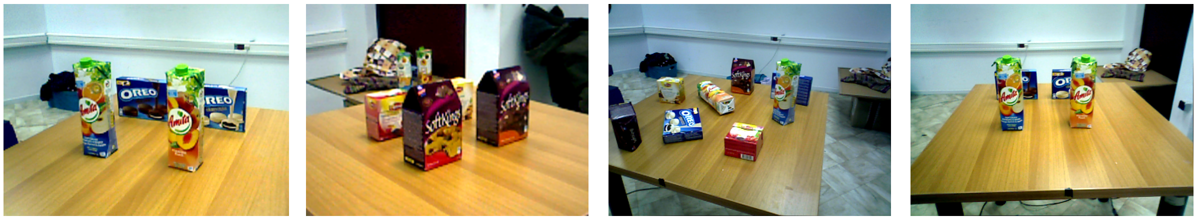
Figure 1: Dataset Objects. Images show renderings of the 3D models of the objects used in training.

The object of this scenario are shown in Fig. 1. We collected six objects usually bought from a supermarket, and captured their 3D models using 123D Catch from Autodesk [2] (with the Android application on a smartphone). Fig. 1 shows real renderings of the 3D models used for training. We can see that the quality of the models created by this Structure From Motion solution is much better compared to Kinect Fusion [5] (Fig. 4d,4e) for textured, non-glossy objects. Furthermode, these models capture the complete object including the bottom part so that they can be used to detect objects lying on the table in any possible orientation.



(a) (b) (c) (d)

Figure 2: Examples of test images



(a) (b) (c) (d)

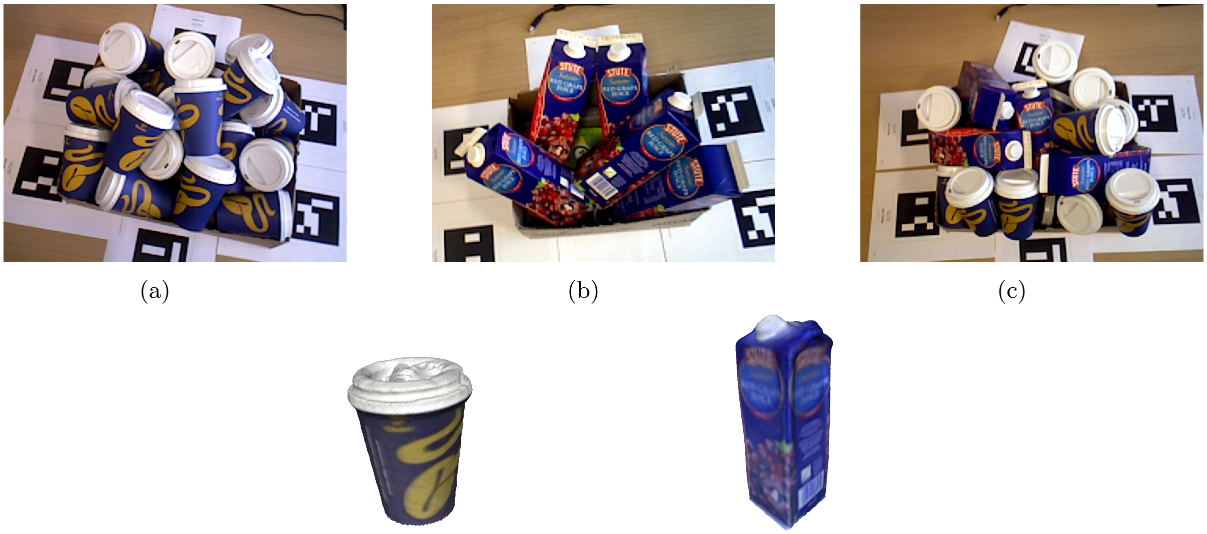
Figure 3: Examples of test images for evaluation active vision methods

Fig. 2 shows examples of the test images of our dataset. We created a variety of different scenes, with and without a table top, including sometimes objects not present in the training set. We capture RGB-D images covering 360 degrees around the table from two different heights for various object arrangements. These simple arrangements often create occlusions, which combined with the table top and the out-of-training objects make 6 DoF methods produce many false positives. In addition, we provide full annotation of the test images for the objects in the training set using our own semi-automatic algorithms for camera and object registration, to avoid placing markers that makes the scenes look artificial. This scenario contains 6 different scenes with 170 test images in total.

For evaluating active vision methods, we created some additional scenes shown in Fig. 3. We added objects that differ with the existing ones only in some parts of the object (for example, oreo with white and dark chocolate). Thus, the next-best-view of the camera should ideally focus on these distinctive parts and one can qualitatively evaluate an active vision method. Also, we have arranged the objects in such a way so that objects are occluded in viewpoints where the distinctive parts should have been observed (Fig. 3d). Such situations can be resolved by the refinement step of our active method described in the main paper. We created 6 additional scenes with a total of 181 test images.

1.2 Usage Scenario 2

The second scenario in our dataset, named as **bin-picking scenario**, is shown in Fig. 4. We used two objects, a coffee cup and a juice, and created three different scenes, two containing each object separately (Fig. 4a-4b) and one containing both objects (4c). This is a very challenging scenario with much occlusion, while stacking similar objects makes it hard to combine features extracted from different locations and estimate an object pose. To compare with the state of the art in this challenging problem, we chose the objects of [6] that showed the best performance on their dataset and also used the same 3D model given by the authors captured with Kinect Fusion [5] (Fig. 4d-4e). We provide full annotation of the objects that are visible in each scene. The three bin-picking scenes contain 183 test images in total.



(a) (b) (c)

(d) (e)

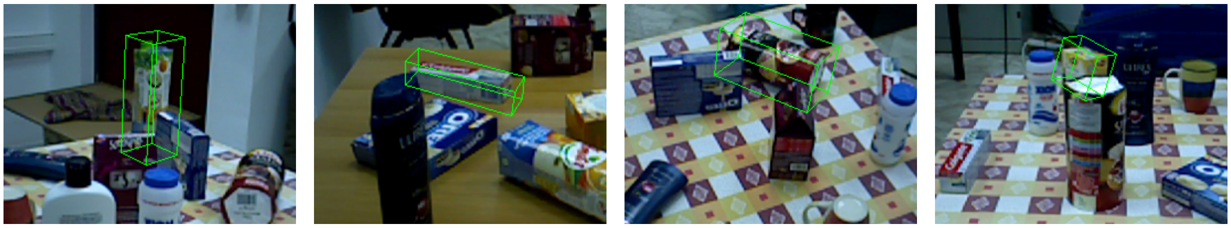
Figure 4: a) test image with a bin of coffee cups, b) test image with a bin of juices, c) test image with both objects, d) coffeecup 3D model, e) juice 3D model

2 Further Analysis on Pose Estimation

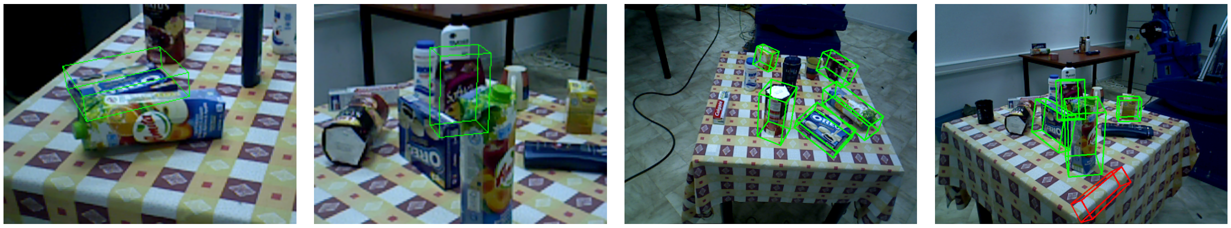
The pose estimation of an object is achieved by passing the patches through the forest and accumulating the votes of the leaf nodes that were reached in a Hough space. The modes of the Hough space represent the object location and pose hypotheses. Each object has its own Hough space for voting. The dimension of the Hough space is 6, 3 for object location and 3 for pose. Extracting the modes in such a high dimensional space is not efficient due to the high sparsity of the data and the computational complexity. We have experimented with three different implementations of the voting space: a) direct 6D voting, b) 3D voting in $\{x, y, z\}$, with each extracted mode subsequently voting in the 3D space $\{yaw, pitch, roll\}$, and c) voting in 2D space $\{x, y\}$ with the extracted modes subsequently voting in $\{z\}$, and the same process repeated in $\{yaw, pitch\}$ and then in $\{roll\}$. The most efficient in terms of time complexity and accuracy was the latter, which was used in our experiments. An explanation of why such approximation works is that the modes in 6D space, are also modes in any 2D projected space. We chose to vote first in $\{x, y\}$ because they provide more information than z axis. However, the order of 2D projection for the pose estimation (i.e. first voting in $\{yaw, pitch\}$ and then in $\{roll\}$) was proven experimentally that does not affect the results. Finally, the quantization of the Hough sapce is 5mm for X, Y and Z, and 1 degree for yaw, pitch and roll.

3 Qualitative Results

Below we provide some qualitative results of our method. Fig. 5a - 5f show results of single object detection in scenario 1. Fig. 5g and 5h show examples of joint object registration using global optimization. Results on scenario 2 are shown in Fig. 6 where again joint registration is used. Finally, in Fig. **??** there are two examples of next-best-view estimation results. These scenes help in qualitatively evaluating a next-best-view strategy. However, the results of the paper regarding active vision were measured using all images in both scenarios in our dataset. More results and comparisons side-by-side with state of the art methods can be found on the video attached to the supplementary material.

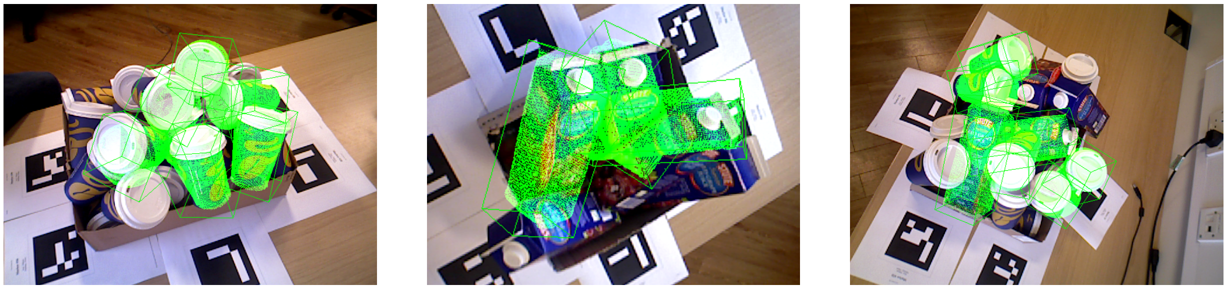


(a) Amita (b) Colgate (c) Elite (d) Lipton



(e) Oreo (f) Softkings (g) Joint Registration (h) Joint Registration

Figure 5: Our dataset - Scenario 1

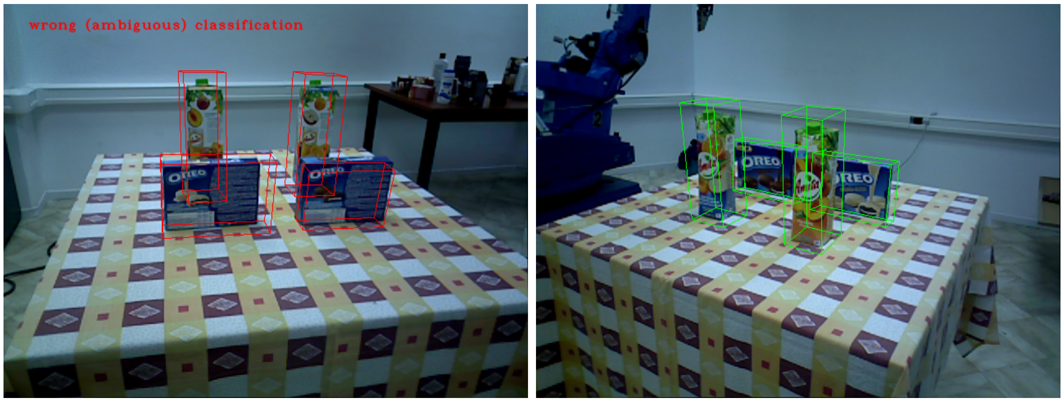


(a) Coffee cups (b) Juice (c) Both objects

Figure 6: Our dataset - Scenario 2 (Bin picking). Results taken using joint registration.

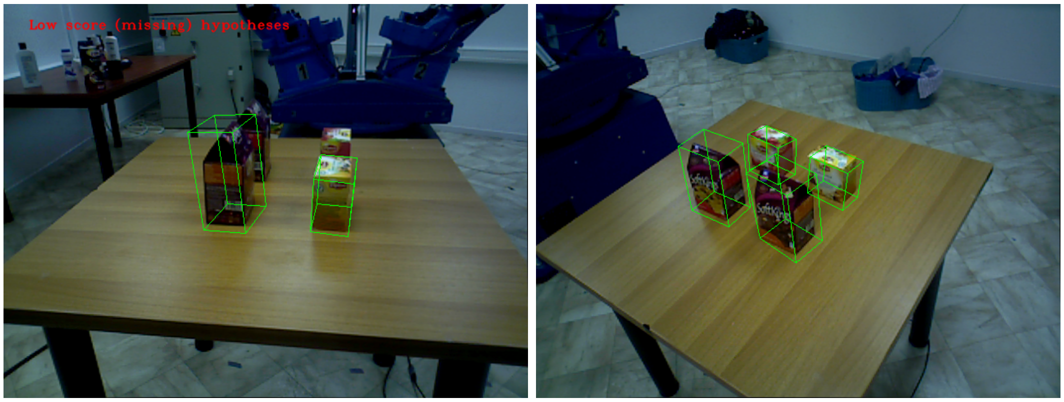
4 Brahmann et al. [3] Dataset Results - Evaluation

Below is the complete table of the results of our method evaluated on the dataset of [3]. The only result of [4] is borrowed from the authors of [3]. One difficulty we faced when evaluating this dataset was that the annotations provided in some cases, were not accurate enough. Therefore, some better estimations from the ground truth exceeded the metric threshold of accepting a correct pose. We manually corrected only few of such cases. Furthermore, the spot light used in testing, made the depth sensor in some cases unable to capture most of the depth information of the object resulting in few corrupted test images. Last, objects like the Pump are symmetric in a wide range of viewpoints, making an accurate estimation not possible, whereas many correct estimations in this case reveal some overfitting on the object context, which does not differ in training and testing.



(a) Classes cannot be disambiguated from this viewpoint

(b) Next best view of (a)



(c) Hypotheses of hidden objects have low score

(d) Next best view of (c)

Figure 7: Active Next Best View estimation using occlusion refinement in order to disambiguate the four objects.

References

[1] Euler angles. https://en.wikipedia.org/wiki/Euler_angles.

[2] Autodesk. *123D Catch (http://www.123dapp.com/catch)*, 2015.

[3] 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.

[4] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *ICCV*, 2011.

[5] S. Izadi, D. Kim, O. Hilliges, Molyneaux, et al. Kinectfusion: Real-time 3d reconstruction and interaction using a moving depth camera. In *ACM*, 2011.

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

Table 1: Results on the dataset of [3]

Object	[4] (%)	[3] (%)	**Our** (%)
Audiobox		**75.4**	71.5
Carry Case		**95.9**	90.7
Dish Soap		**100**	**100**
Helmet		**77.6**	74.5
Hole Puncher		**98.1**	94.3
Pump		**69.3**	67.4
Toolbox		99.5	**100**
Toy (Battle Cat)	70.2	91.8	**92.4**
Toy (Panthor)		**96.9**	94.2
Toy (Stridor)		94	**94.3**
Stuffed Cat		**98.3**	94
Duck		81.6	**87.7**
Dwarf		**67.6**	65.6
Mouse		89.1	**90.1**
Owl		60.5	**90.27**
Elephant		94.7	**96.13**
Samurai		98.5	**99.6**
Sculpture 1		82.7	**89.5**
Sculpture 2		**100**	**100**
Avg.		88.2	**89.1**
Med.		**93.0**	92.4