pacman

Probabilistic and Compositional Representations for Object Manipulation

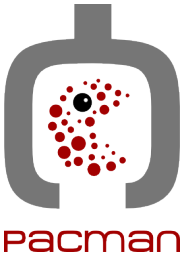
UoB Highly Occluded Object Challenge (UoB-HOOC)

1st International Workshop on Recovering 6D Object Pose

17 Dec 2015, In conjunction with ICCV 2015, Santiago, Chile

Challenge Organizers: Aleš Leonardis, Krzysztof Walas,
Hector Basevi, Jinpeng Wang

Outline

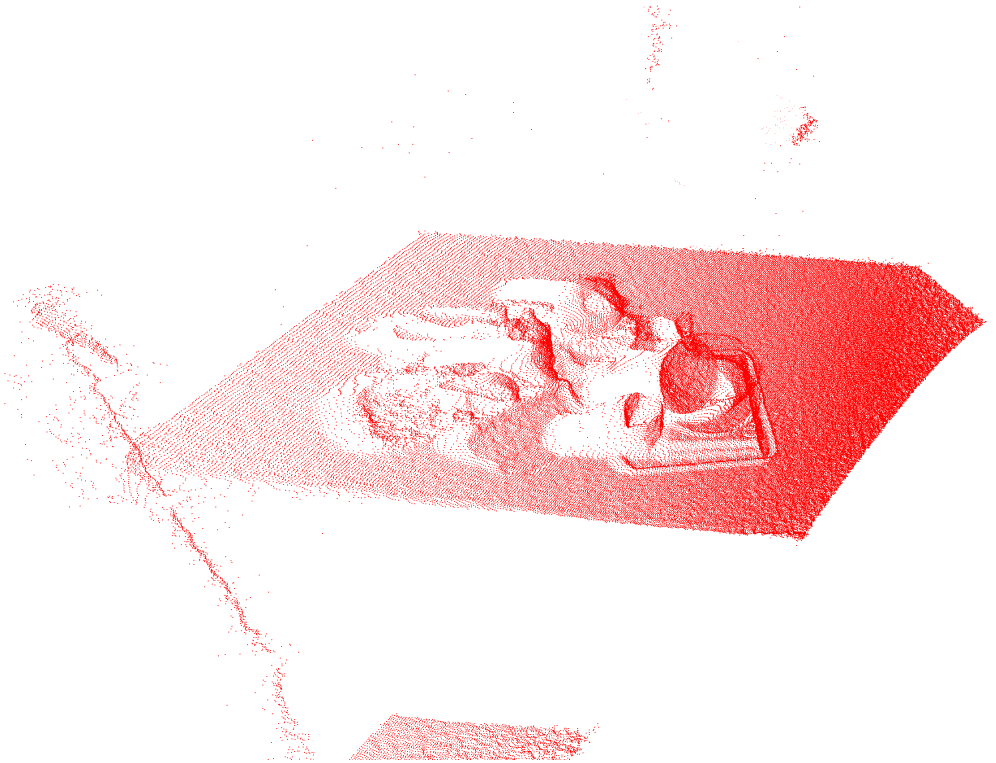


1. Dataset Description
    - capturing devices and scene views
    - scenes composition
    - multi-camera system calibration
    - annotations and scenes characterization
2. The Challenge
    - training data
    - expected results
    - performance evaluation
3. Results
    - base line method
    - submitted results
4. Summary and Future Work

Dataset Description

Scene views:

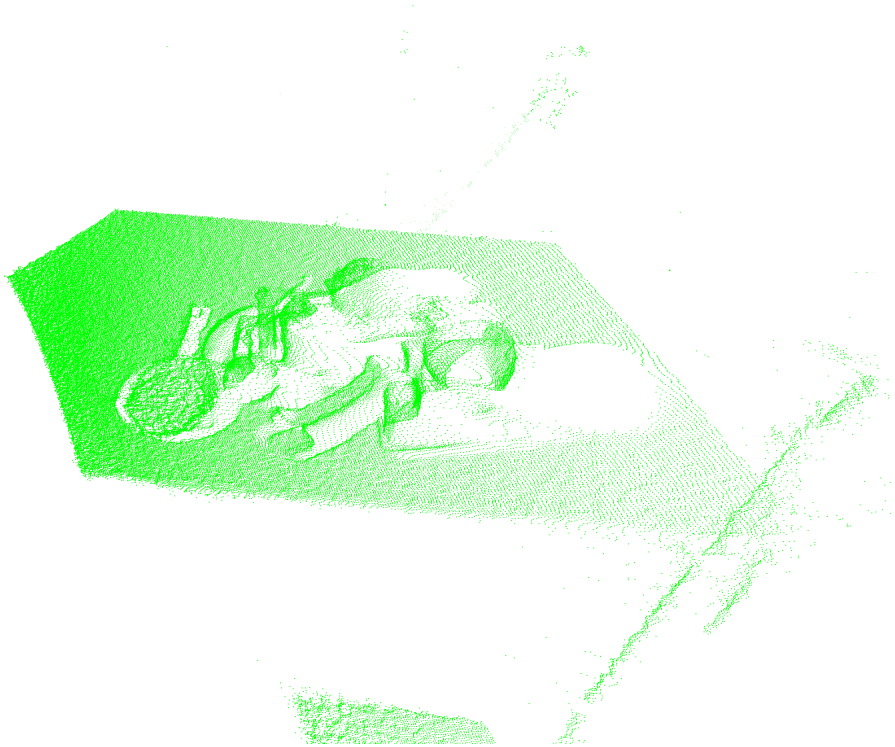
- data registered from three different Kinects
- all devices in fixed positions (mounted on tripods)



Dataset Description

Scene views:

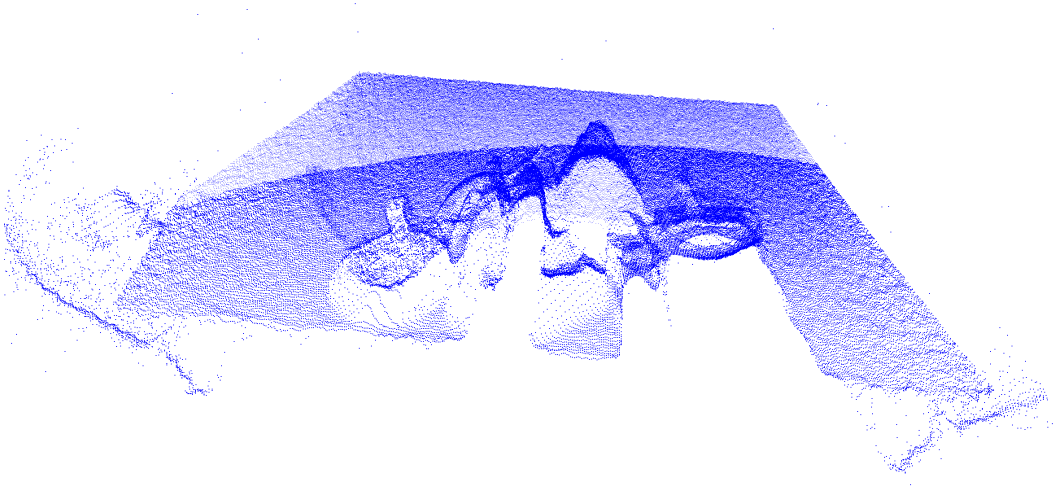
- data registered from three different Kinects
- all devices in fixed positions (mounted on tripods)



Dataset Description

Scene views:

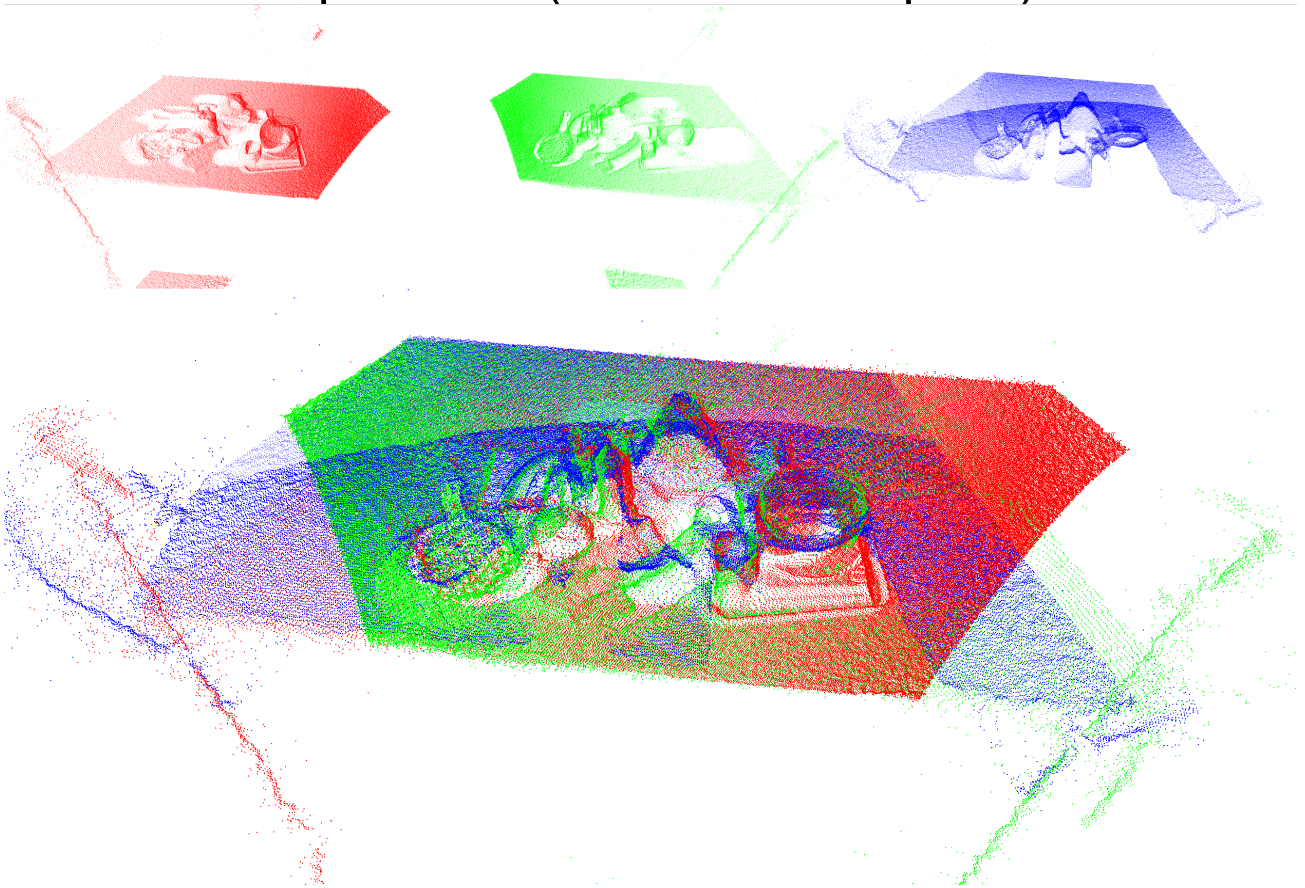
- data registered from three different Kinects
- all devices in fixed positions (mounted on tripods)



Dataset Description

Scene views:

- data registered from three different Kinects
- all devices in fixed positions (mounted on tripods)



Dataset Description

Scene compositions:

- up to 20 objects selected from 25 categories no repetitions
- high level of occlusions
- objects stacked one on top of another
- many symmetric objects

Dataset Description

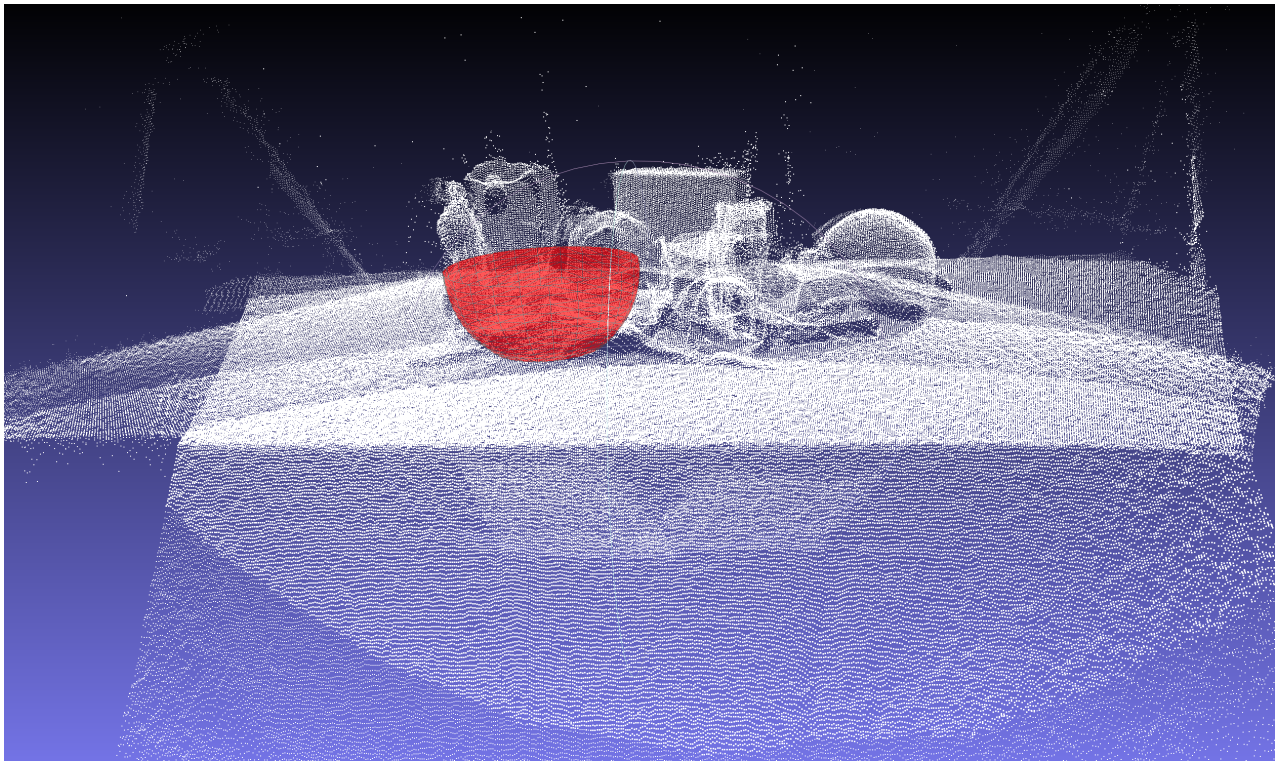
Calibration:

- infra-red image with long exposure (less noise)
- camera calibration toolbox for Matlab -- A2 chessboard
- intrinsic parameters, extrinsics computed for each scene separately
- all views transformed to the same reference frame associated with the chessboard lying on a table

Dataset Description

Annotations:

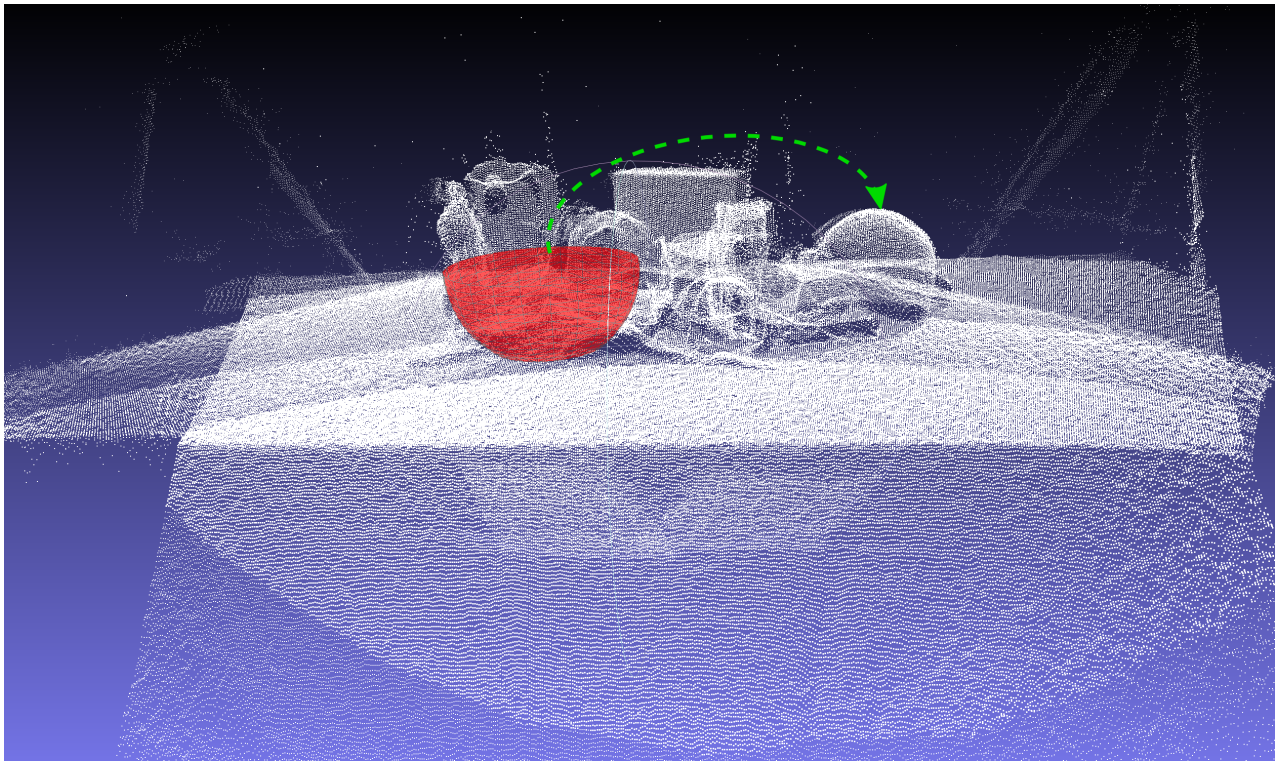
- using a set of 3D computer graphics models
- with 3 views of the scene loaded
- placing an object and establishing its scale



Dataset Description

Annotations:

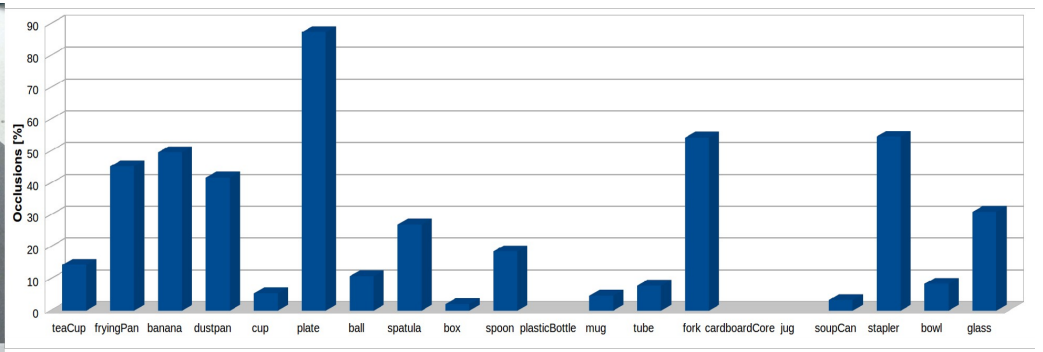
- using a set of 3D computer graphics models
- with 3 views of the scene loaded
- placing an object and establishing its scale



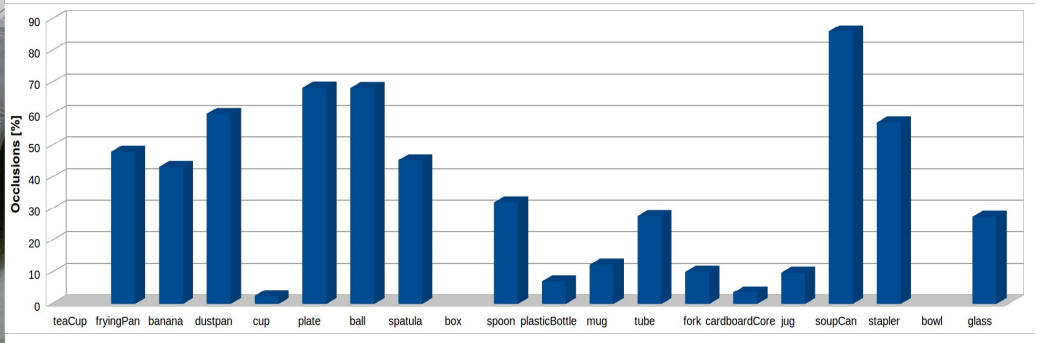
Dataset Description

Characterisations – object occlusion caused by other objects (points):

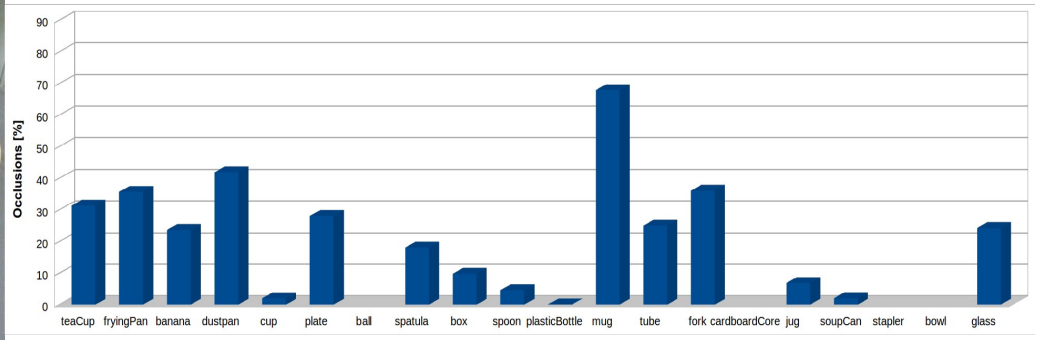
Camera 1



Camera 2

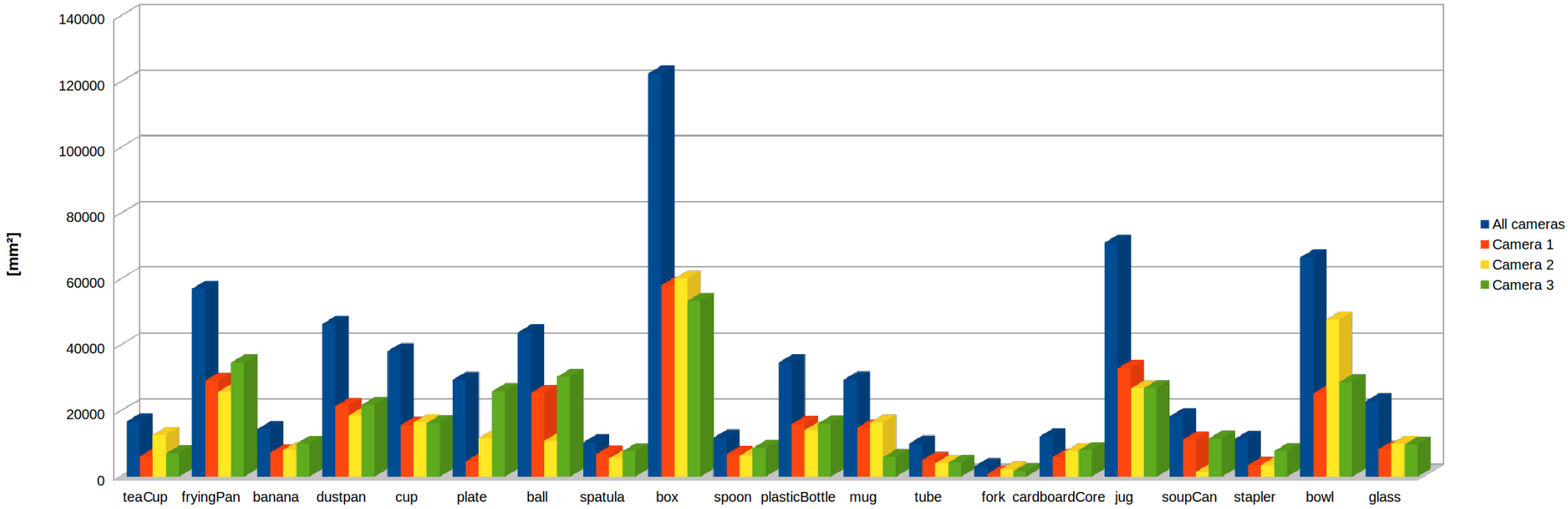


Camera 3



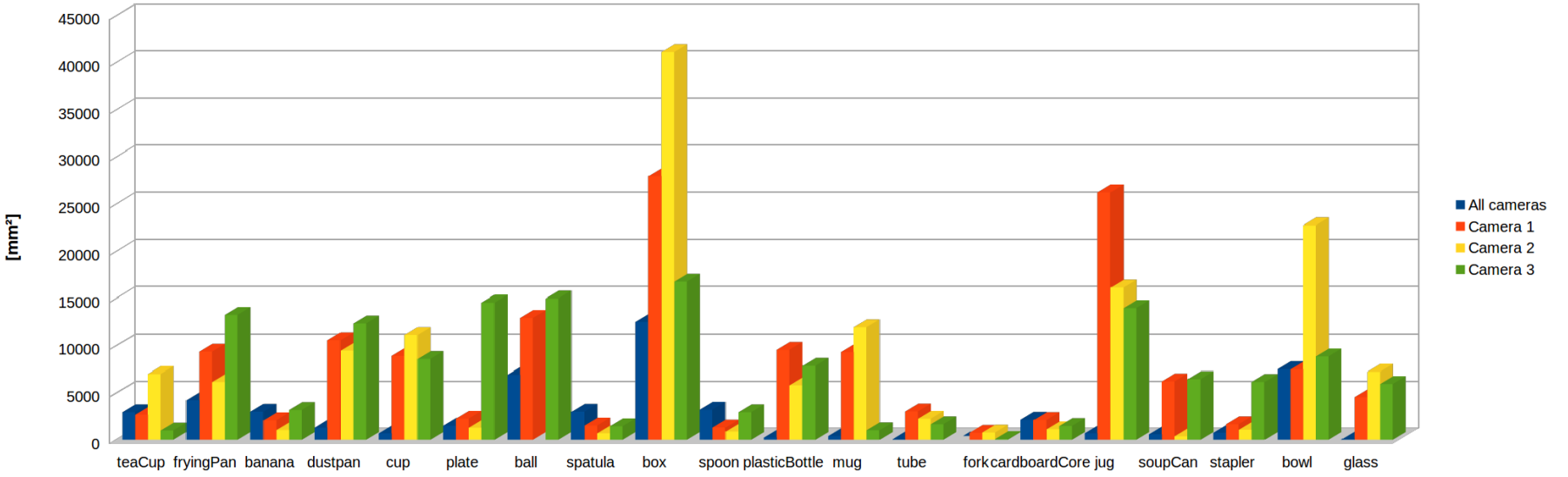
Dataset Description

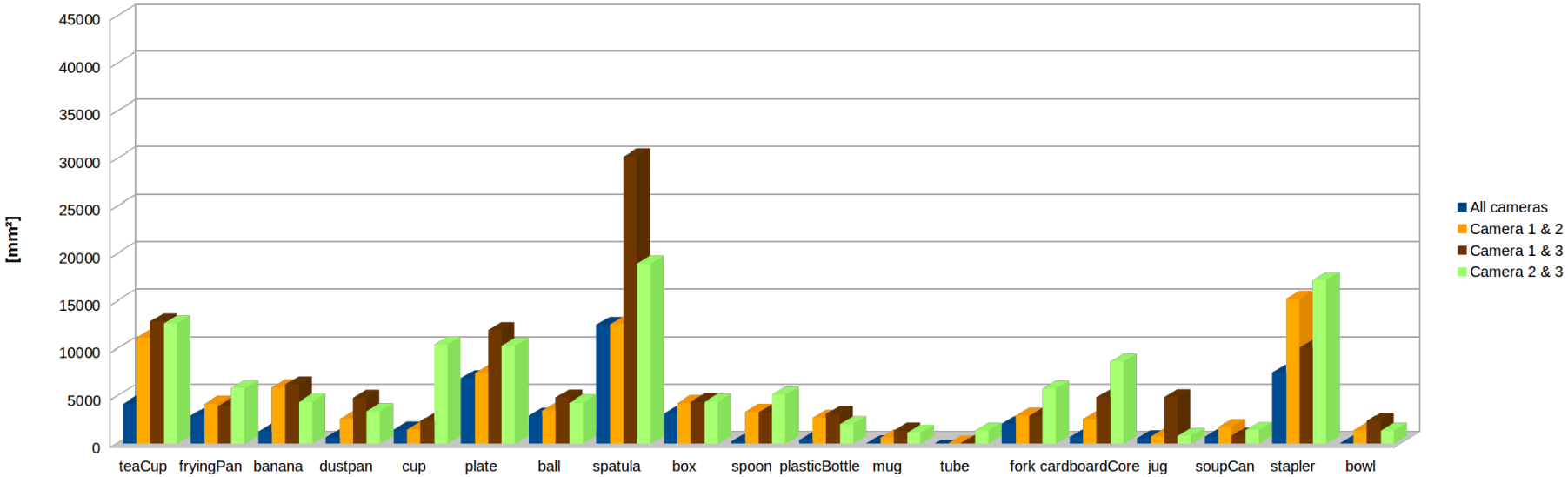
Characterisations – object visibilities (area):



Dataset Description

Characterisations – object visibilities overlap (area):





Training Data

PACMAN

Scene may contain up to 20 objects from 25 different categories:

Dataset Description

Content of the repository:

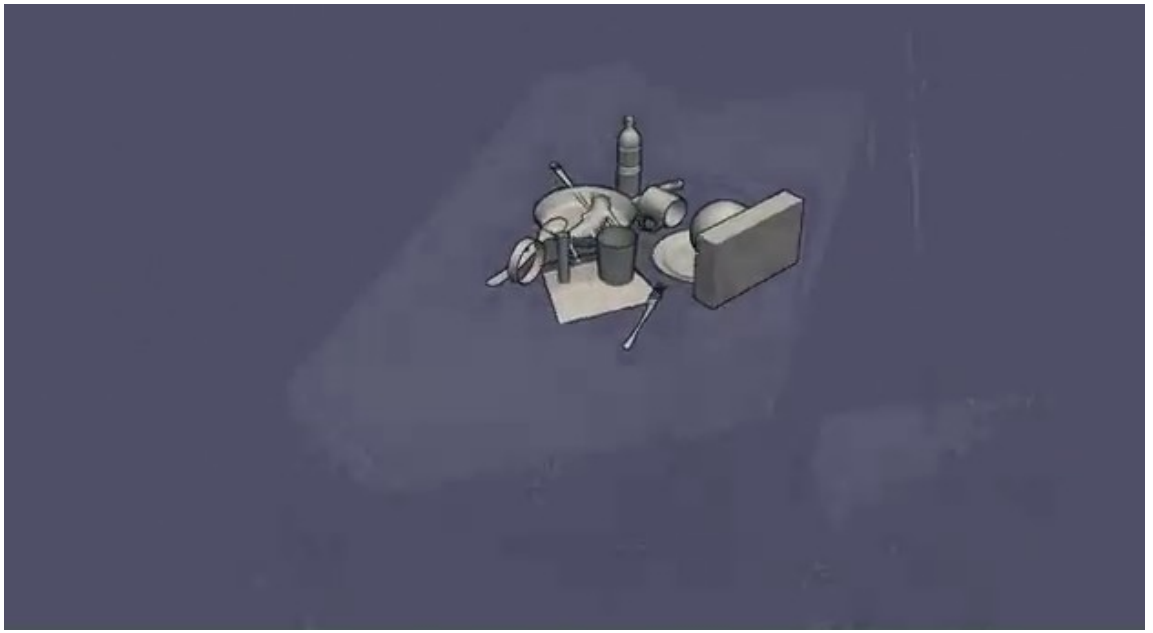
- pointclouds in camera reference frame, ordered pointclouds
- pointclouds in a common reference frame placed on a table top
- RGB images of the scenes (calibration data relating RGB and D image is not provided, RGB image should be used for illustration purposes only)
- depth data together with intrinsic and extrinsic parameters of the camera
- set of generic objects which has to be fit into test sceneData registered from three different Kinects
- scripts for downloading training data

Expected result

Given a set of computer graphics models of generic representatives:

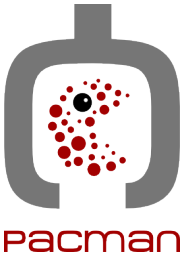
- establish category of each object in a registered pointcloud
- find its 6D pose and appropriate scale

In other words, the generic exemplar has to be fitted as accurately as possible into the object in the registered scene .



https://www.youtube.com/watch?v=P-7rUhYeYBE

Performance measures



Common statistical measures for multi-label classification[1]:

$$Accuracy(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Y_i \bigcup Z_i|} \tag{1}$$

$$Precission(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Z_i|} \tag{2}$$

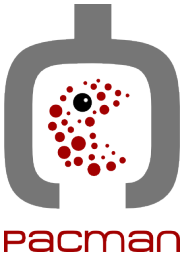
$$Recall(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap Z_i|}{|Y_i|} \tag{3}$$

$$F_1(H, D) = 2 \times \frac{Precission \times Recall}{Precission + Recall} \tag{4}$$

where:
H – multi-label classifier
D – multi-label dataset consisting of $|D|$ multi-label examples (x_i, Y_i); $i = 1 \ldots |D|$
$Z_i = H(x_i)$ – a set of labels predicted by H for example x_i;

[1] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. Int J Data Warehousing and Mining, 2007:1–13, 2007.

Performance measures



Dataset characterization:

$$LD(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \tag{5}$$

where:
L – a set of disjoint labels;
$|L|$ – cardinality of the set of labels;
$Y \subseteq L$ – set of labels for current example (scene);
D – multi-label dataset consisting of $|D|$ multi-label examples (x_i, Y_i); $i = 1 \ldots |D|$

Hamming Loss:

$$HammingLoss(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta Z_i|}{|L|} \tag{6}$$

where:
Δ – symmetric difference of two sets;
$Z_i = H(x_i)$ – a set of labels predicted by H for example x_i;
H – multi-label classifier.

Performance measures

Detection measure:

$$DetectionError(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{3} \cdot \left(\frac{trans_{dist}}{0.2} + \frac{rot_{dist}}{\pi} + scale_{dist} \right) \cdot \left(1 - \frac{|Y_i|}{|L|} \right) \tag{7}$$

$$T = \begin{bmatrix} s_x r_{11} & s_y r_{12} & s_z r_{13} & t_x \\ s_x r_{21} & s_y r_{22} & s_z r_{23} & t_y \\ s_x r_{31} & s_y r_{32} & s_z r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{8}$$

$$T = Trans \cdot Rot \cdot Scale \tag{9}$$

$$trans_{dist} = \sqrt{(t_{x1} - t_{x2})^2 + (t_{y1} - t_{y2})^2 + (t_{z1} - t_{z2})^2}. \tag{10}$$

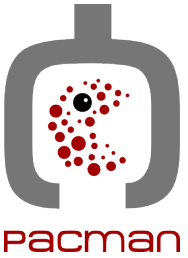
$$rot_{dist} = \Phi(Q_1, Q_2) = arccos(\|Q_1 \cdot Q_2\|), \tag{11}$$

where:
$Q_1 \cdot Q_2 = w_1 w_2 + x_1 x_2 + y_1 y_2 + z_1 z_2$

$$Q = \begin{bmatrix} \frac{1}{2}\sqrt{1 + r_{11} + r_{22} + r_{33}} = w \\ \frac{r_{32} - r_{23}}{4w} = x \\ \frac{r_{13} - r_{31}}{4w} = y \\ \frac{r_{21} - r_{12}}{4w} = z \end{bmatrix} \tag{12}$$

$$scale_{dist} = \left| \frac{s_1^3 - s_2^3}{s_1^3} \right| \tag{13}$$

Performance measures



Average Distance (AD) measure[2]:

$$m = \underset{\mathbf{x} \in \mathcal{M}}{\text{avg}} \|(\mathbf{R}\mathbf{x} + \mathbf{T}) - (\widetilde{\mathbf{R}}\mathbf{x} + \widetilde{\mathbf{T}})\| \tag{14}$$

where:
m – a matching score,
$\mathcal{M}$ – a 3D model,
$\mathbf{x}$ – vertex of a 3D model $\mathcal{M}$,
$\mathbf{R}$ – a ground truth rotation matrix,
$\mathbf{T}$ – a ground truth translation vector,
$\widetilde{\mathbf{R}}$ – an estimated rotation matrix,
$\widetilde{\mathbf{T}}$ – an estimated translation vector,

$$m = \underset{\mathbf{x}_1 \in \mathcal{M}}{\text{avg}} \min_{\mathbf{x}_2 \in \mathcal{M}} \|(\mathbf{R}\mathbf{x}_1 + \mathbf{T}) - (\widetilde{\mathbf{R}}\mathbf{x}_2 + \widetilde{\mathbf{T}})\| \tag{15}$$

where:
$\mathbf{x}_1$ – vertex of a 3D model $\mathcal{M}$ in a ground truth pose,
$\mathbf{x}_2$ – vertex of a 3D model $\mathcal{M}$ in an estimated pose,

[2] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary R. Bradski, Kurt Konolige, Nassir Navab, Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes. ACCV 2012

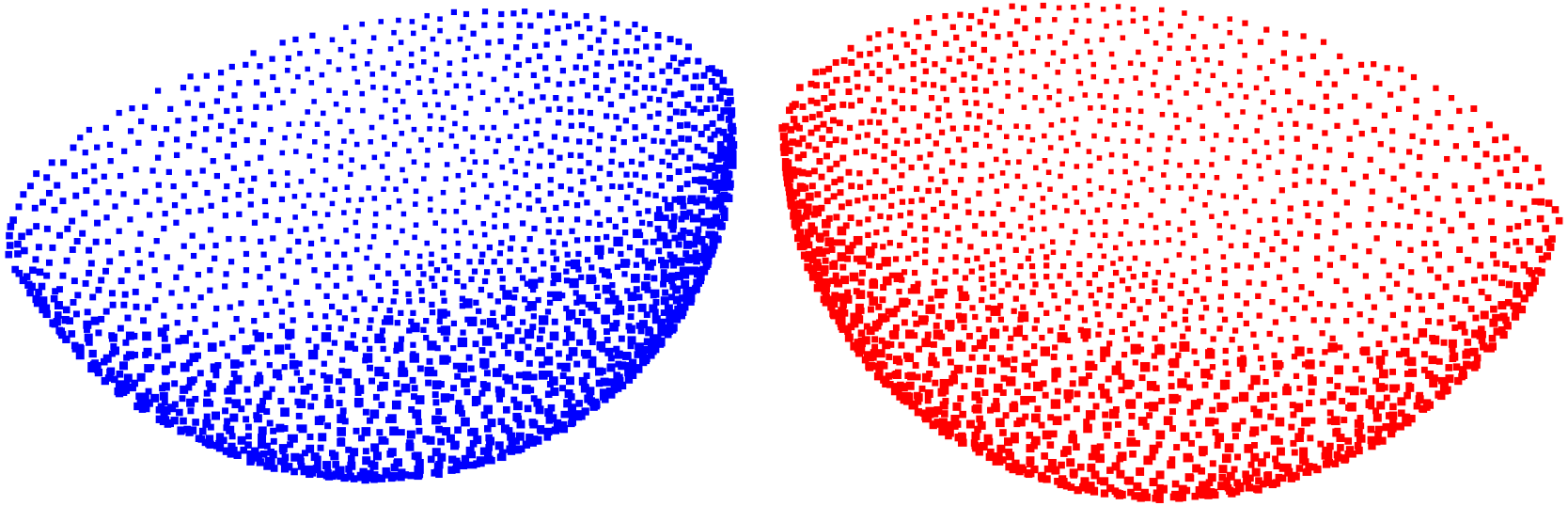
Performance measures

Improved Average Distance (IAD) measure:

$$m = \underset{\mathbf{x}_1 \in \mathcal{M}}{\text{avg}} \min_{\mathbf{x}_2 \in \mathcal{M}} \|(\mathbf{R}\mathbf{x}_1) - (\widetilde{\mathbf{R}}\mathbf{x}_2)\| + \|\mathbf{T} - \widetilde{\mathbf{T}}\| \tag{16}$$

Actual object pose

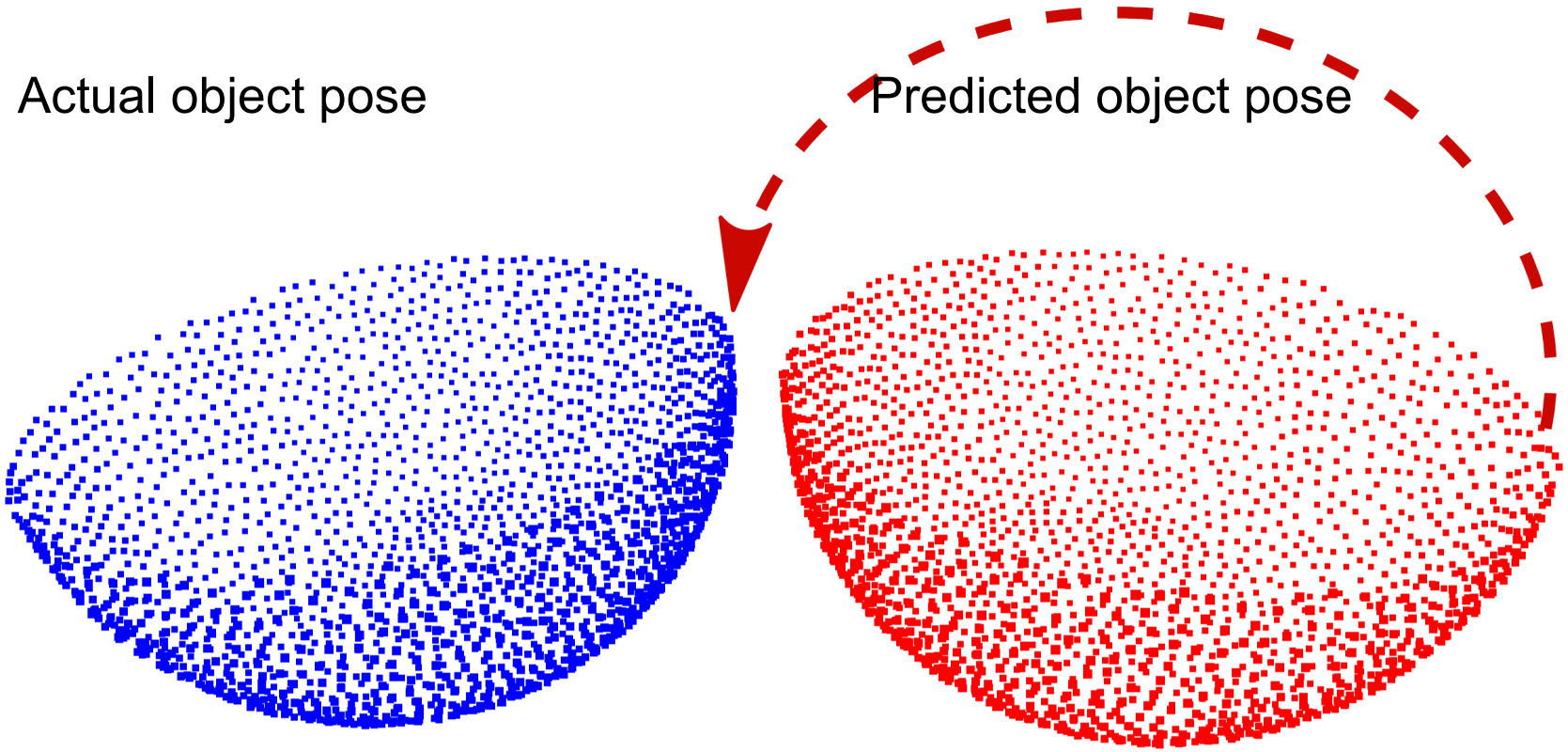
Predicted object pose



Performance measures

Improved Average Distance (IAD) measure:

$$m = \underset{\mathbf{x}_1 \in \mathcal{M}}{\text{avg}} \min_{\mathbf{x}_2 \in \mathcal{M}} \|(\mathbf{R}\mathbf{x}_1) - (\widetilde{\mathbf{R}}\mathbf{x}_2)\| + \|\mathbf{T} - \widetilde{\mathbf{T}}\| \tag{16}$$

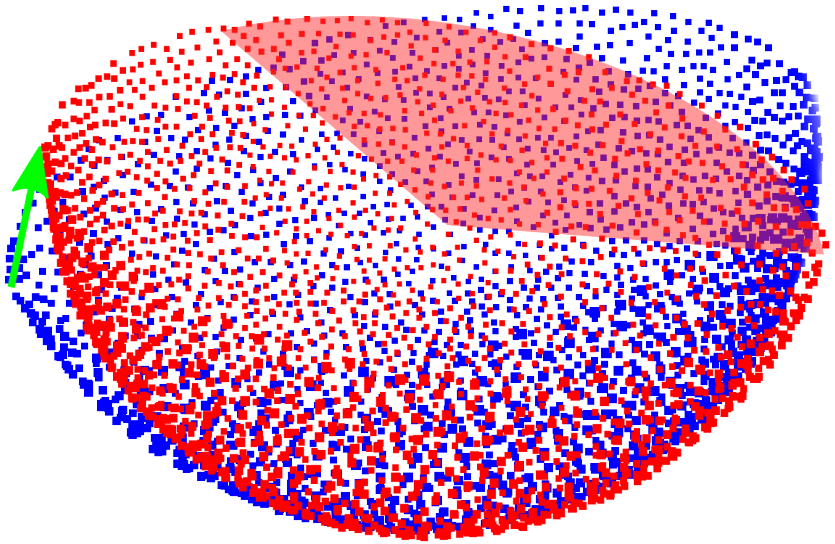


Performance measures

Improved Average Distance (IAD) measure:

$$m = \underset{\mathbf{x}_1 \in \mathcal{M}}{\text{avg}} \min_{\mathbf{x}_2 \in \mathcal{M}} \|(\mathbf{R}\mathbf{x}_1) - (\widetilde{\mathbf{R}}\mathbf{x}_2)\| + \|\mathbf{T} - \widetilde{\mathbf{T}}\| \quad (16)$$

Actual object rotation/Predicted object rotation



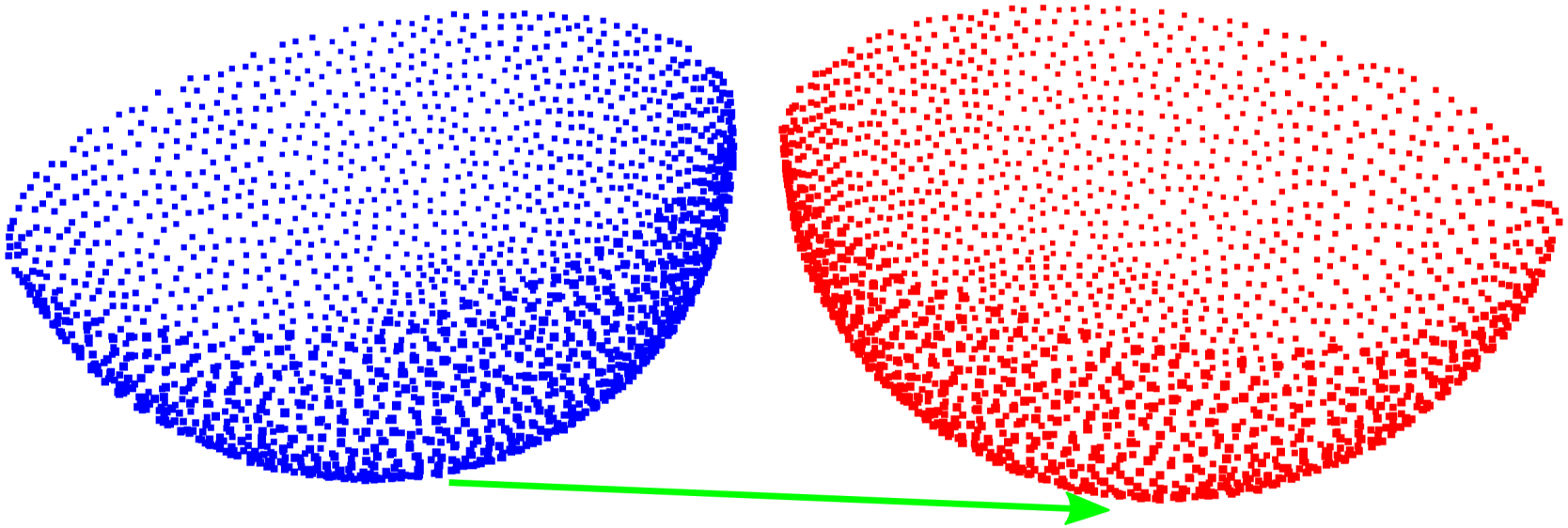
Performance measures

Improved Average Distance (IAD) measure:

$$m = \underset{\mathbf{x}_1 \in \mathcal{M}}{\text{avg}} \min_{\mathbf{x}_2 \in \mathcal{M}} \|(\mathbf{R}\mathbf{x}_1) - (\widetilde{\mathbf{R}}\mathbf{x}_2)\| + \|\mathbf{T} - \widetilde{\mathbf{T}}\| \quad (16)$$

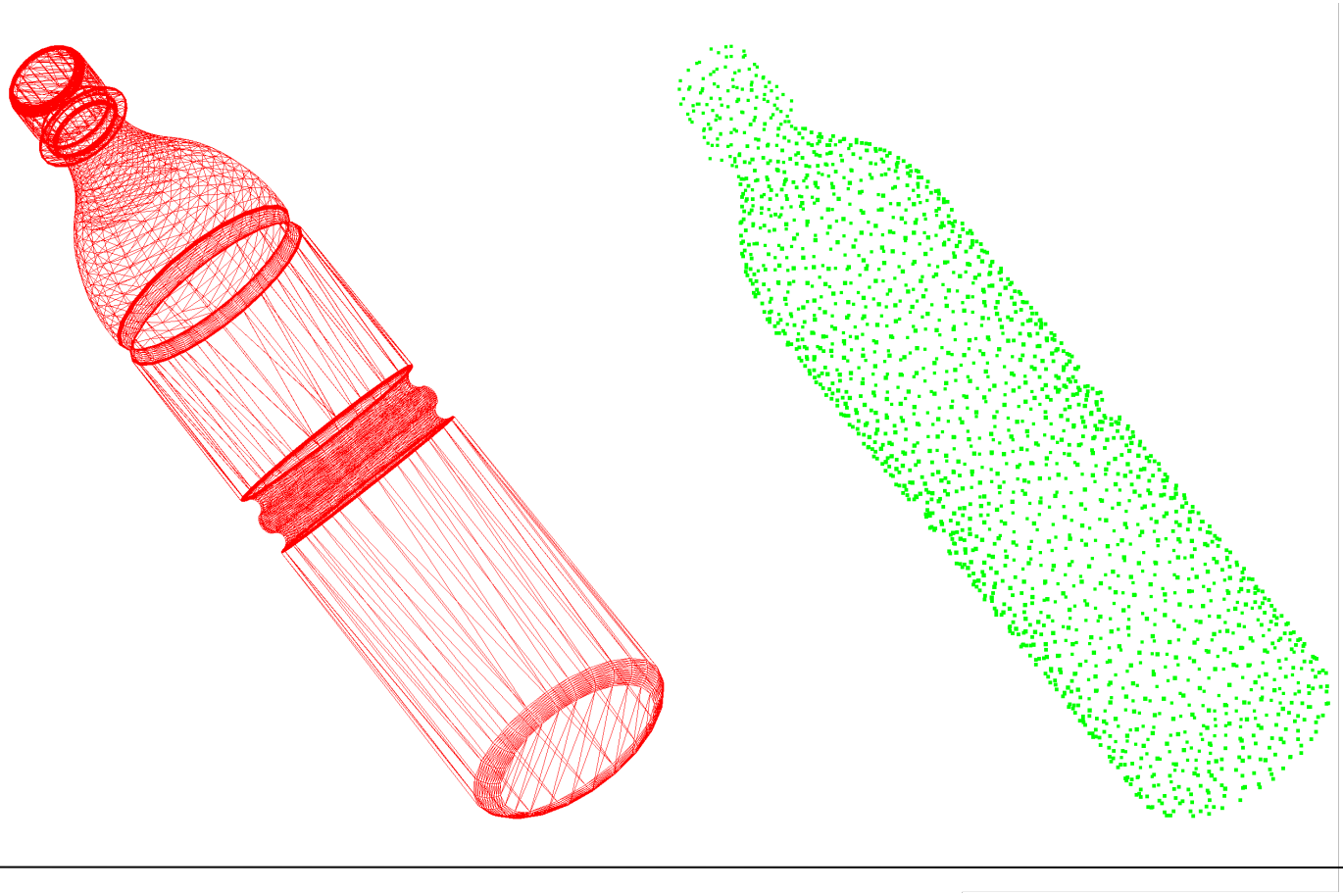
Actual object translation

Predicted object translation



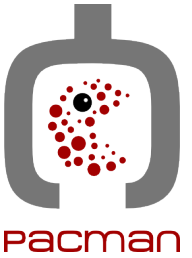
Performance measures

Non-unifrom discretisation of object for meshes – resampling[3]:



[3] M.; Cignoni, P.; Scopigno, R., "Efficient and Flexible Sampling with Blue Noise Properties of Triangular Meshes," in Visualization and Computer Graphics, IEEE Transactions on , vol.18, no.6, pp.914-924, June 2012

Results



Baseline method based on Brachmann et al., Learning 6D Object Pose Estimation using 3D Object Coordinates, ECCV 2014 :

- decision forests for calculating class probabilities and estimated object coordinates for each pixel in a RGB-D image
- decisions via binary linear classifiers on RGB or D local gradients
- forests trained to separate pixels by object class and discretised object coordinates
- estimated object coordinates used to perform rough pose estimation
- RANSAC-based algorithm and energy function for pose refinement and final pose selection

Results

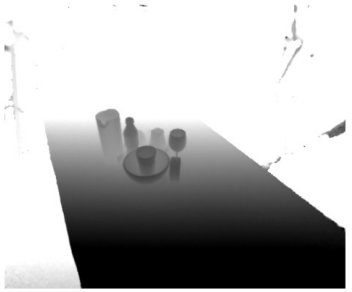
Baseline method – modifications to Brachmann et al. method:

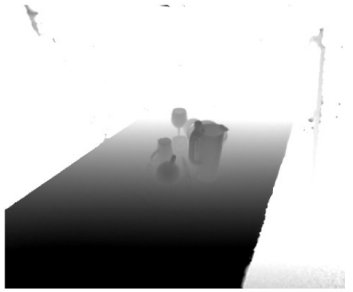
- Data pre-processing:
  - UoB-HOOC data transformed using calibration data to correct for lens spherical aberration
- Training:
  - on synthetic images of generic objects in randomised poses
  - no RGB data, decisions on depth features alone
- Decision forests:
  - different forests trained for each camera
  - data integrated to perform pose estimation using all views
- Pose estimation:
  - rough pose estimation and final pose selection process modified for compatibility with symmetric objects

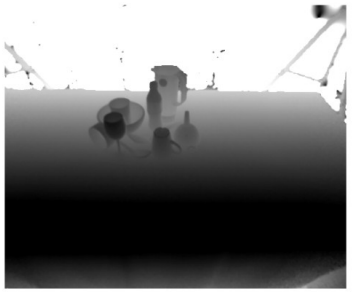
Results

Baseline – workflow on selected scene:

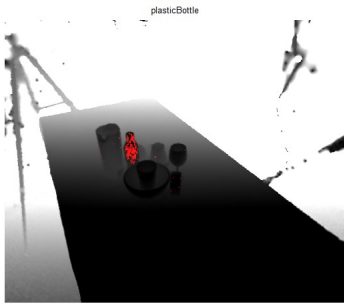
corrected depth maps

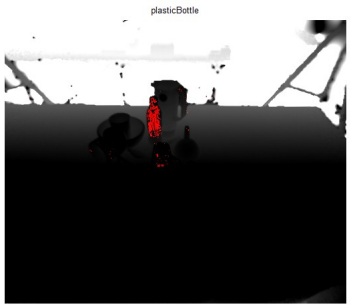




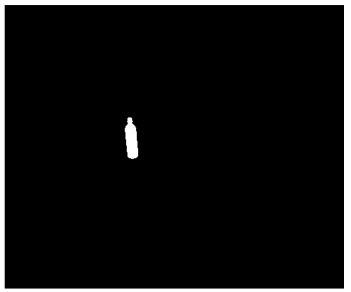


probability maps from decision forests



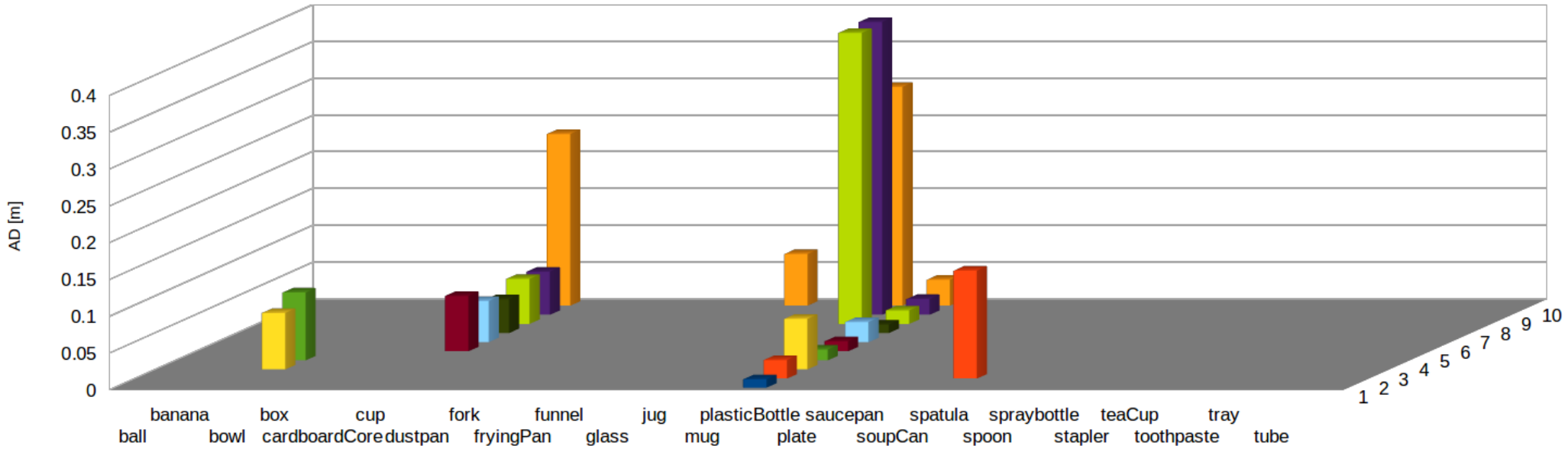


estimated depth from probable depth data



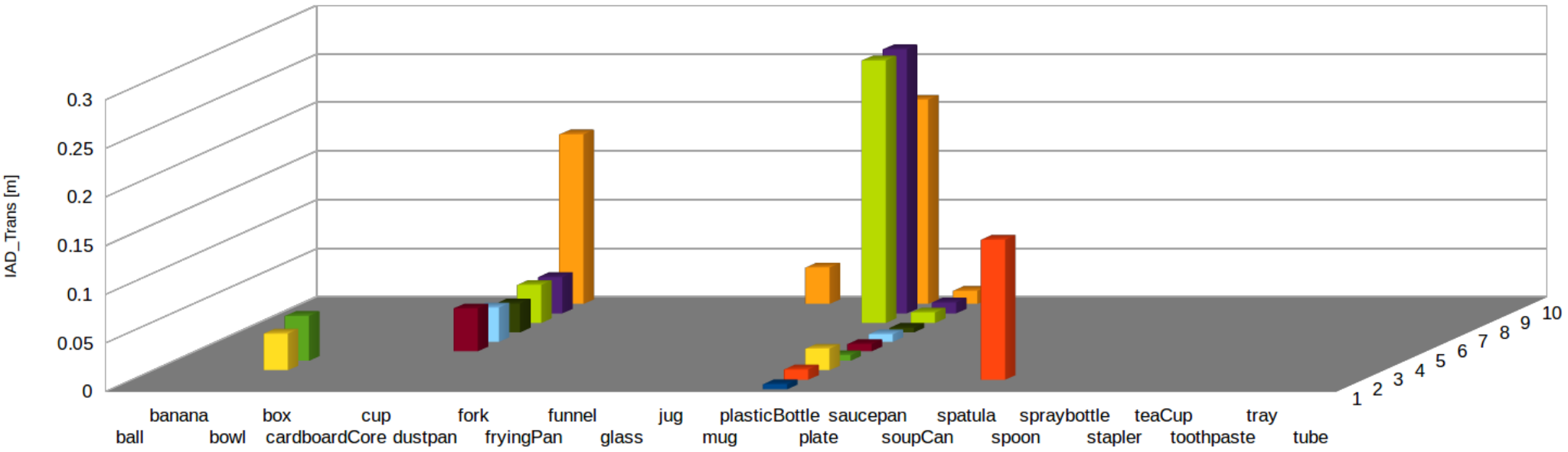
Preliminary Results

Up to 10 objects – Birmingham:



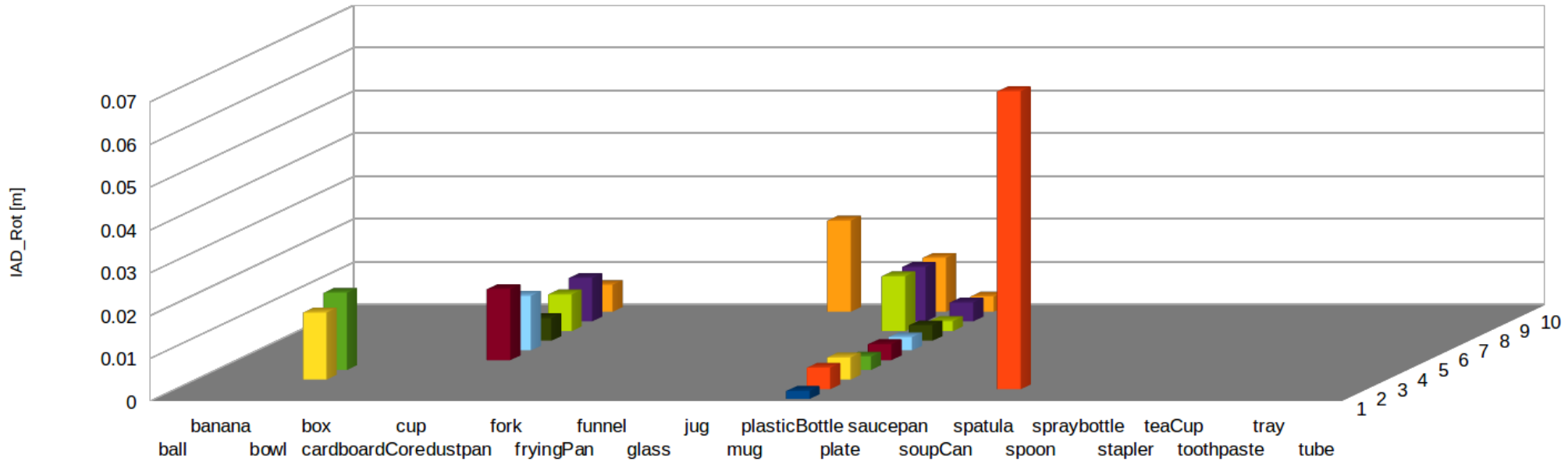
Preliminary Results

Up to 10 objects – Birmingham:



Preliminary Results

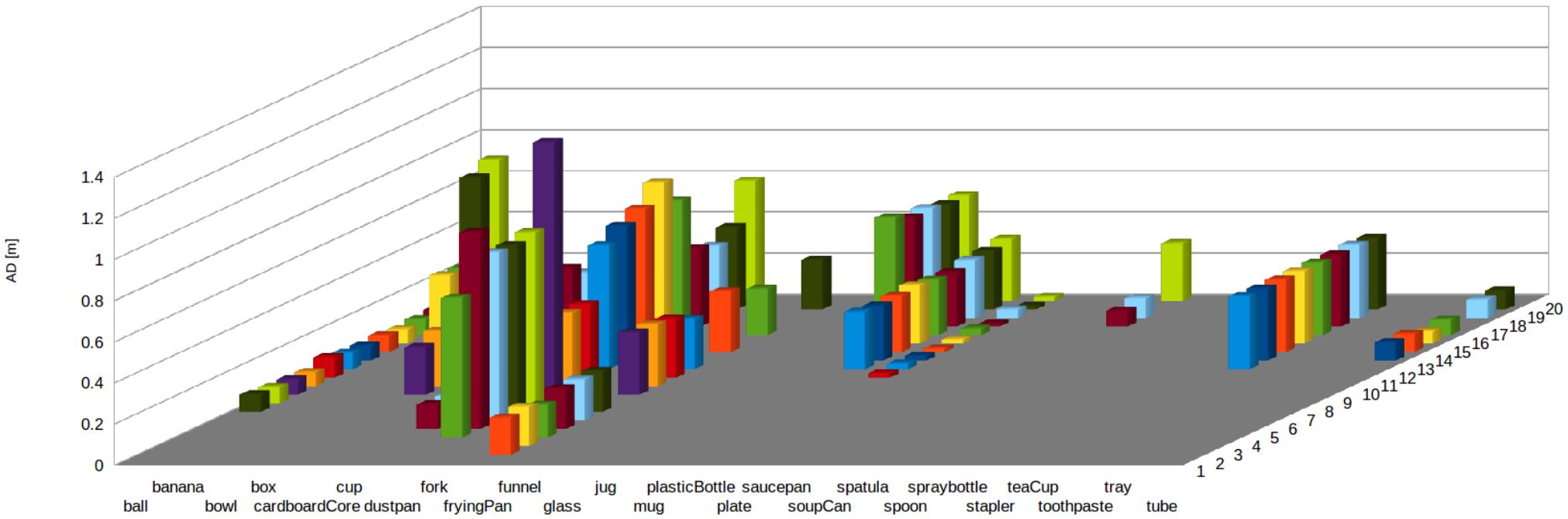
Up to 10 objects – Birmingham:



Preliminary Results

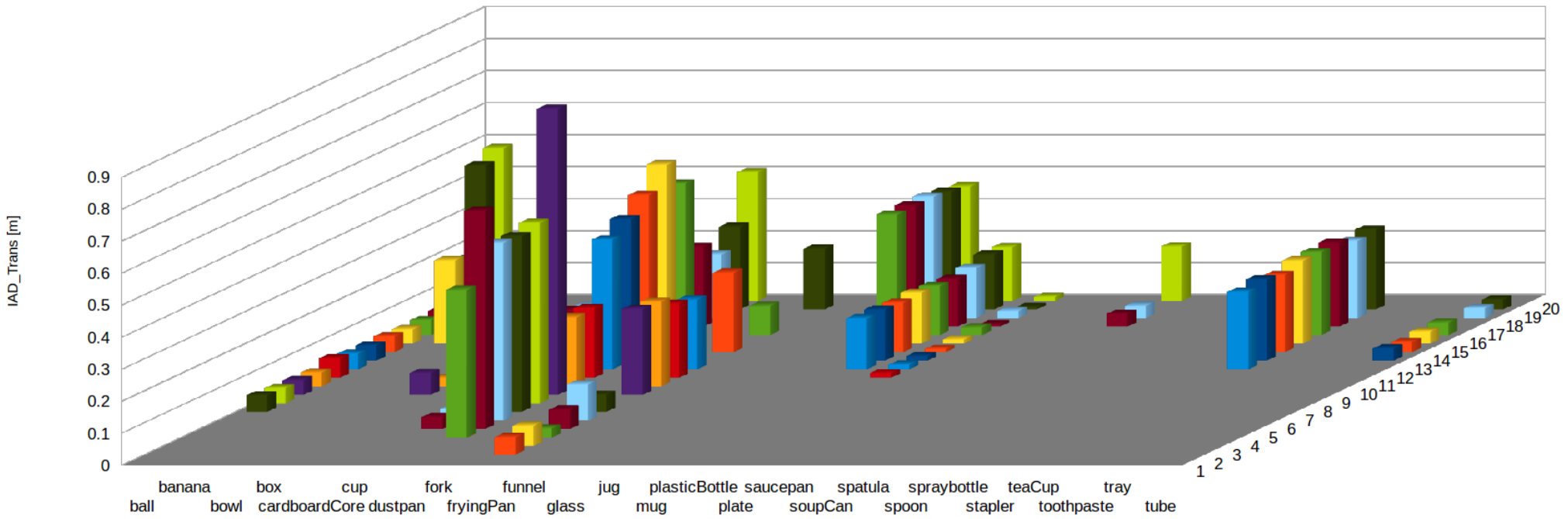
Pacman

Up to 20 objects – Birmingham:



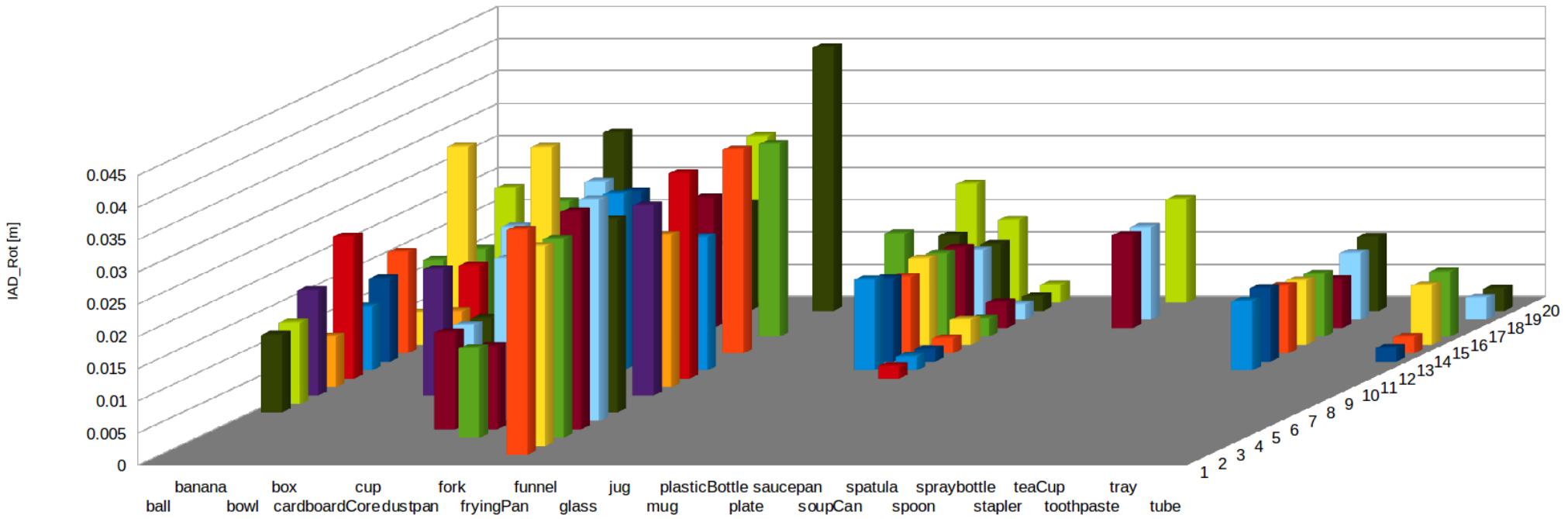
Preliminary Results

Up to 20 objects – Birmingham:



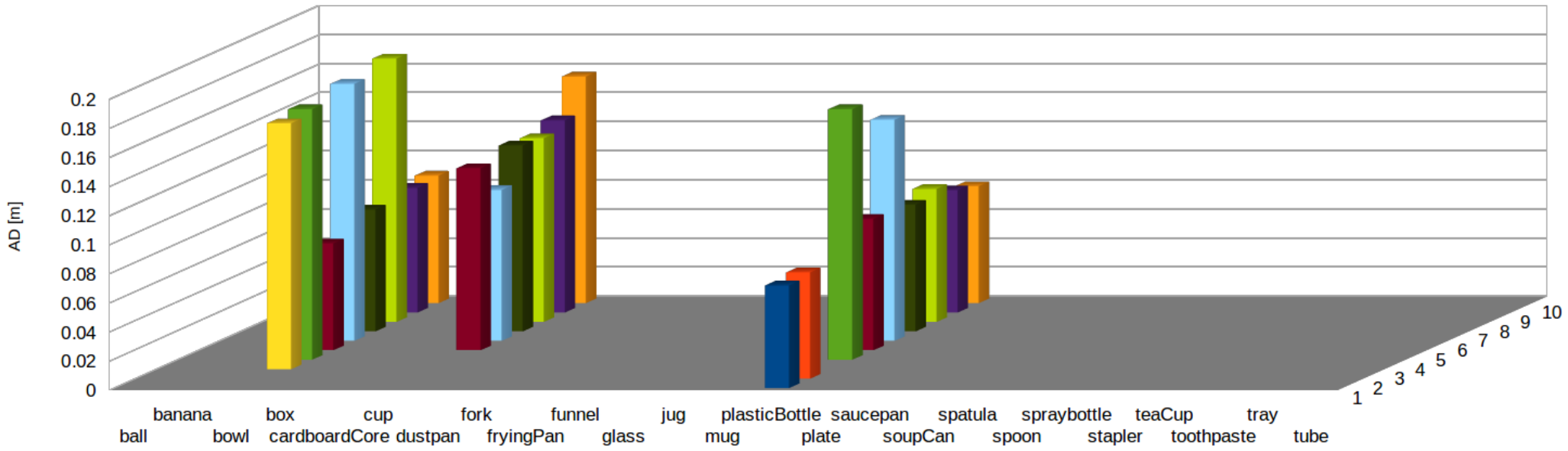
Preliminary Results

Up to 20 objects – Birmingham:



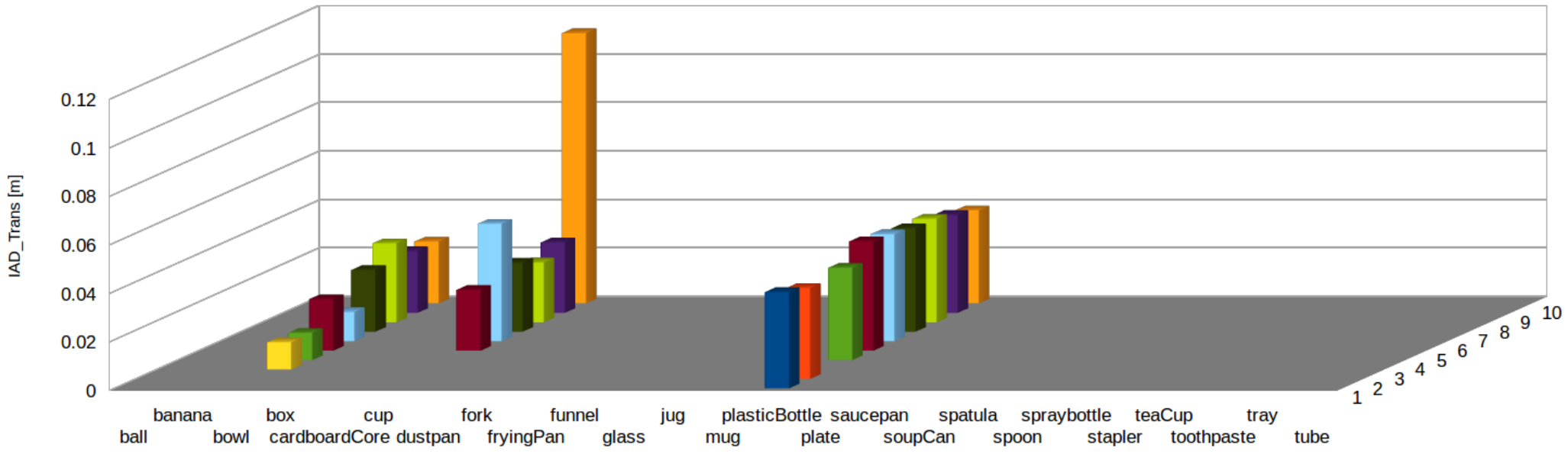
Preliminary Results

Up to 10 objects – Innsbruck:



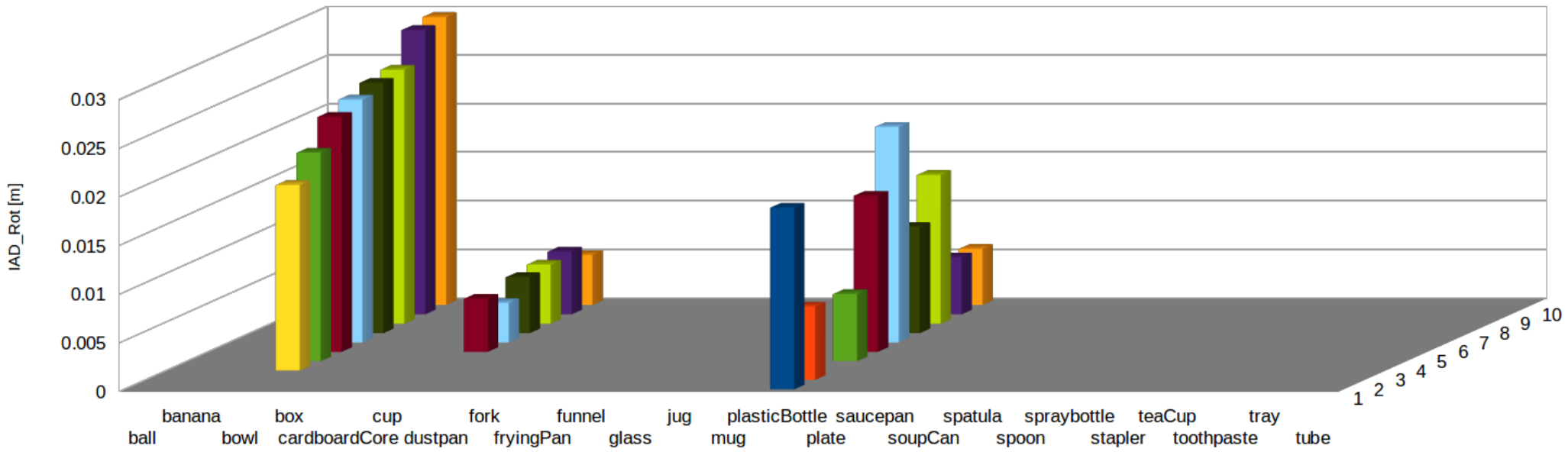
Preliminary Results

Up to 10 objects – Innsbruck:



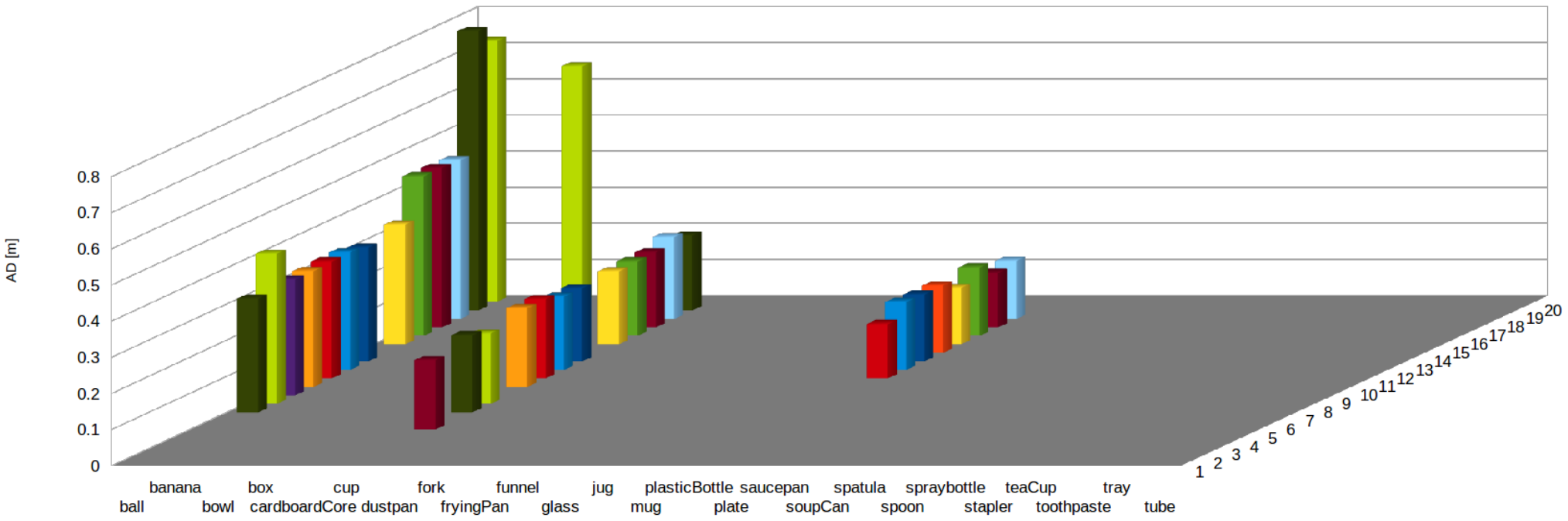
Preliminary Results

Up to 10 objects – Innsbruck:



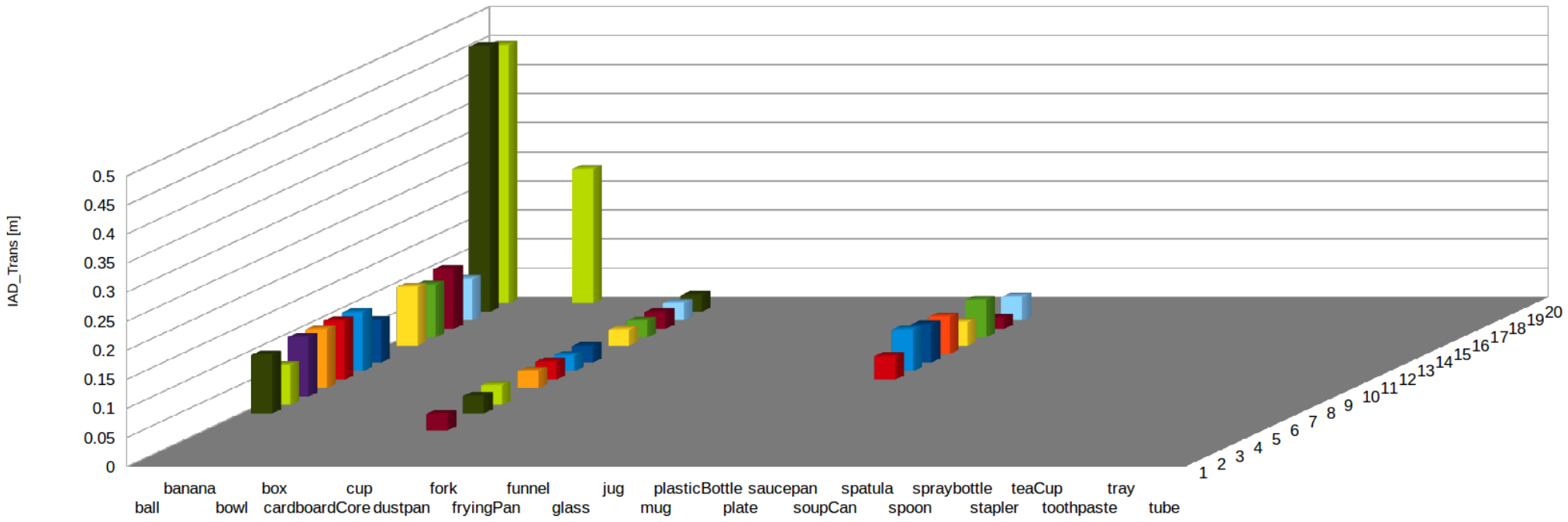
Preliminary Results

Up to 20 objects – Innsbruck:



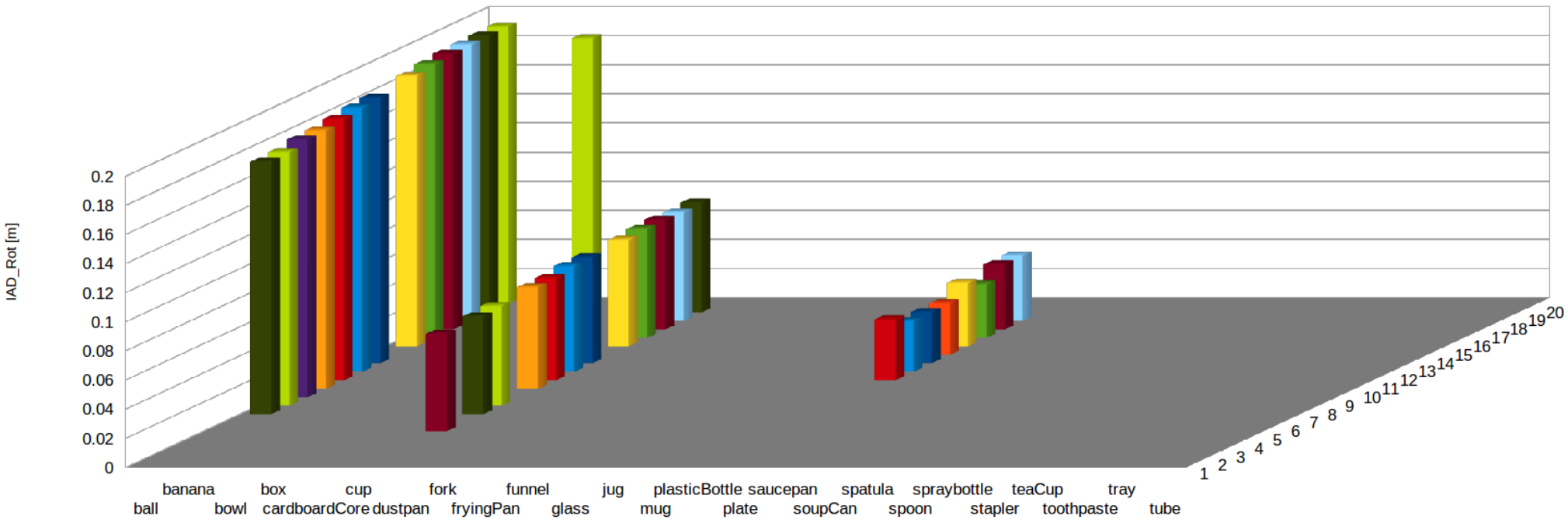
Preliminary Results

Up to 20 objects – Innsbruck:

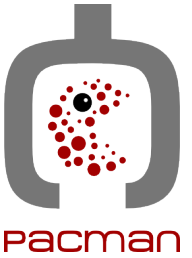


Preliminary Results

Up to 20 objects – Innsbruck:



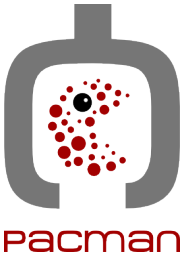
Summary



UoB HOOC :

- high level of occlusions
- symmetric objects
- learning on synthetic data
- testing on real data
- different evaluation measures

Future work



UoB HOOC :

- more attributes
- different object types
- standardized measures