ViDRILO: The Visual and Depth Robot Indoor Localization with Objects information dataset

International Journal of Robotics Research
000(00):1–6

DOI:doi number
http://mms.sagepub.com

Abstract

In this article we describe a semantic localization dataset for indoor environments named ViDRILO. The dataset provides five sequences of frames acquired with a mobile robot in two similar office buildings under different lighting conditions. Each frame consists of a point cloud representation of the scene and a perspective image. The frames in the dataset are annotated with the semantic category of the scene, but also with the presence or absence of a list of predefined objects appearing in the scene. In addition to the frames and annotations, the dataset is distributed with a set of tools for its use in both place classification and object recognition tasks. The large number of labeled frames in conjunction with the annotation scheme make this dataset different from existing ones. The ViDRILO dataset is released for use as a benchmark for different problems such as multimodal place classification and object recognition, 3D reconstruction or point cloud data compression.

1. Introduction

In robotics, the semantic localization problem consists in reporting the location of a mobile robot using semantic labels describing the scene, in contrast to the coordinates used in metric localization. The robot has to distinguish between different locations (e.g. kitchen, corridor, offices) given the information provided by its sensors. Moreover, it should recognize the objects present in its surroundings, which provides additional (but complementary) information. In this article we describe ViDRILO, the Visual and Depth Robot Indoor Localization with Objects information dataset, which was conceived as a challenging benchmark for multimodal semantic localization proposals. The dataset has been defined by taking into account issues such as object and scene variability, challenging lighting conditions, semantic-oriented labeling, and different environment configurations. ViDRILO (http://www.rovit.ua.es/dataset/vidrilo) contains five data sequences including temporal continuity that were acquired in two different buildings while a robot was moving. Each sequence consists of a list of frames composed of a perspective visual image and a depth image encoded as a 3D point cloud file. Each frame is annotated with: a) the semantic category of the room from a list of ten different labels (corridor, toilet, etc.); and b) the presence or absence of a list of fifteen predefined objects (trash, table, etc.).

The main novelty of this dataset is the object annotation scheme, that is, while other RGB-D datasets assign each 3D point to a semantic label, all the points from a ViDRILO frame share the same semantic annotations. The list of objects included in other datasets is defined without prior objectives (mainly availability) or based on spatial principles (tabletop objects). However, the 15 objects described in ViDRILO have been selected because of their intrinsic relationship with room categories. In other words, the presence/absence of the selected objects provides information that is useful to determine the room in which the robot is located.

The feasibility of this dataset as a benchmark for semantic localization has been proved in the RobotVision competition within the ImageCLEF lab (Caputo et al., 2013). More than 30 different research groups from around the world registered for, and participated in, the 2013 (Martínez-Gómez et al., 2013) and 2014 editions of the RobotVision challenge,

in which unreleased frames from the ViDRILO dataset were provided as training, validation and test sequences.

The rest of the article is organized as follows. Section 2 describes our motivation for creating ViDRILO and reviews some related datasets. In Section 3, the dataset and its main characteristics, including the ground truth information, are described. Section 4 is devoted to the acquisition procedure, while Section 5 presents an analysis of the dataset and its tools. Finally, some conclusions are drawn in Section 6.

2. Motivation

The main motivation for creating this dataset is the need for a challenging benchmark in the semantic localization problem within indoor environments, which is the objective of the RobotVision competition (Martínez-Gómez et al., 2013). This problem consists in answering the question "where am I?" from a semantic point of view. More specifically, we have to label the perceptions acquired by a robot in different indoor locations with semantic categories such as "kitchen" or "corridor". Humans perform this labeling on the basis of objects that are present in the scene and prior knowledge. That is, we are capable of performing this labeling even when we have never seen the scene before. For instance, if we see a fridge we can be almost certain that we are located in a kitchen.

Early editions of the RobotVision competition took advantage of two datasets that were specifically created for the semantic localization problem: COLD (Pronobis et al., 2006) and KTH-IDOL2 (Luo et al., 2006). These datasets provide enough illumination and distribution variability to serve as a benchmark in the RobotVision competition. ViDRILO was created as a natural evolution of these datasets but including two new characteristics: the use of 3D information and the annotation of the objects in the scene. Indeed, it has been used for this competition since 2013. The 3D information can be useful to classify scenes under challenging lighting conditions (even in darkness), and the presence/absence of objects provides a higher-level of reasoning that is closer to the human way of thinking.

Since the first acquisitions of ViDRILO (September of 2012), new datasets with similar characteristics have been released [1]. One dataset which includes 3D scene labeling and object annotation is NYU Depth V2 (Silberman et al., 2012). However, we can find the following differences with respect to ViDRILO: a) ViDRILO provides a higher number of annotated images (22454 against 1449) split into training, validation and test sequences; b) the sequences in ViDRILO were recorded in temporal continuity fashion, which is desirable for many robotic applications; c) NYUD provides information about the exact 3D position of 894 different objects (useful for object recognition and scene reconstruction problems), while ViDRILO includes the presence/absence of 15 objects, appropriate for the semantic localization problem. Another dataset containing room and object information is the RGB-D Scenes Dataset v2, proposed in Lai et al. (2014), in which every 3D point in the scene is assigned to an object label. The main difference with respect to ViDRILO is the size of the dataset (24 vs. 22454 scenes), which makes it inappropriate for training semantic localization classifiers. Moreover, the set of objects is limited to small tabletop objects (cap, bowl, cereal box, coffee mug, and soda can), which are not so discriminative for distinguishing between different room categories. Other related datasets are the NewCollegeData (Smith et al., 2009), which contains outdoor images, but not object annotations, and the KIT object database (Kasper et al., 2012), which on the contrary does not include scene information. There are even some proposals that generate datasets from information retrieved from the Internet through interactive search (Thomee & Lew, 2012), but there are still technological limitations to relying on this information.

3. Dataset

The ViDRILO dataset consists of five different sequences of frames acquired over a span of 12 months. These sequences differ in the path followed by the robot and the building used for the acquisition. All the frames were acquired using a Powerbot robot with an onboard Microsoft Kinect device (Fig. 1 left) in two different buildings (Fig. 1 right). Table 1 shows the overall characteristics of the dataset, while the specific details of the acquisition procedure are given in Section 4. The two buildings used for the acquisition (buildings

[1] Please visit `http://www0.cs.ucl.ac.uk/staff/M.Firman/RGBDdatasets/` for an updated list of RGB-D datasets

Table 1. Global sequences distribution.

Sequence	#Frames	Floors	Dark Rooms	Time Span
Seq.1	2389	1st,2nd	0/18	0 months
Seq.2	4579	1st,2nd	0/18	0 months
Seq.3	2248	2nd	4/13	3 months
Seq.4	4826	1st,2nd	6/18	6 months
Seq.5	8412	1st,2nd	0/20	12 months

A and B) present a similar structure, but considerable variations in object and room layout. Both buildings are poorly illuminated by external light and need the continuous use of artificial lighting systems. This fact reduces dependency on the external lighting conditions, and allowed us to introduce lighting variations by turning on/off the artificial illumination.

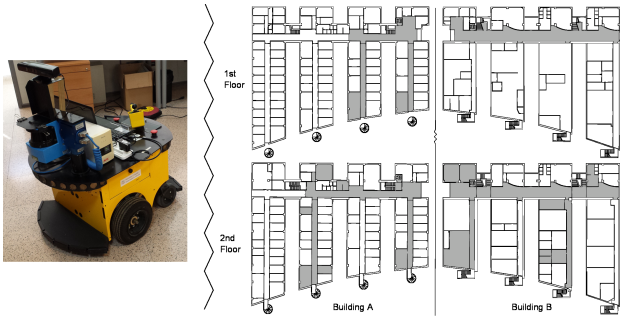

Figure 1. Robotic platform (left) and buildings used for the acquisition (right).

Sequences 1 to 4 were acquired in building A over a time span of 6 months. Sequences 1 and 2 were captured on two consecutive days and the robot followed the same path, but in the opposite direction. These two sequences are proposed for use as training. Sequence 3 was acquired 3 months later, and it was created for validation purposes with the inclusion of dark rooms (imaged without artificial lights). The robot followed a path that was much shorter than for the rest of the sequences, and only the second floor of building A was imaged. Sequence 4 also contains dark rooms and it is proposed for testing systems that need to cope with drastic lighting variations. Finally, Sequence 5 was acquired in building B, 12 months after the first acquisition. This sequence does not contain dark rooms, and is proposed for semantic localization problems with domain adaptation.

3.1. Ground Truth Data

Each frame in the dataset consists of a visual image and a point cloud file representing the same scene. Both are labeled with the semantic category of the room where they were acquired, as well as the list of 15 predefined objects appearing in the images. Fig. 2 shows a pair of images for each one of the 10 room categories using the following codes: CR (Corridor), HA (Hall), PO (Professor Office), SO (Student Office), TR (Technical Room), TO (Toilet), SE (Secretary Office), VC (Video Conference Room), WH (Warehouse), and EA (Elevator Area).

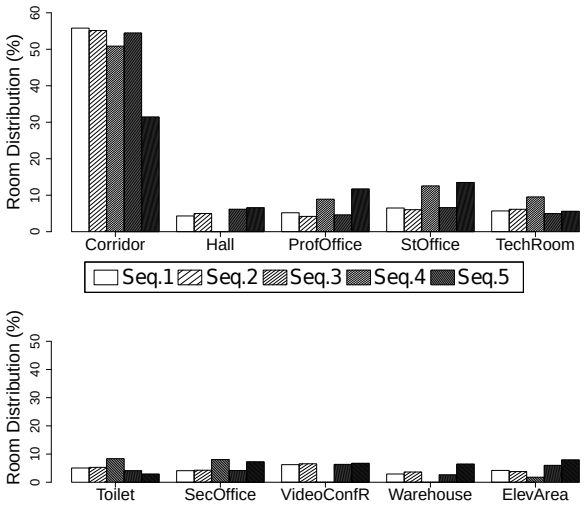

Figure 3. Room category distribution for the 5 sequences.

A histogram of the room category distribution for the 5 sequences is shown in Fig. 3. As expected for office buildings with large corridors and several rooms, the Corridor category is the predominant class in all the sequences, with an appearance ratio higher than 50% in most of the cases. All the room categories present significant variations in the distribution between sequences, especially for sequences 3 (acquired just on the 2nd floor of building A) and 5 (acquired in Building B).

An example image of each one of the 15 objects is shown in Fig. 4. These objects were selected because their occurrence (or lack of) within a scene provides useful cues for detecting the semantic category of the corresponding room. Fig. 5 shows the conditional distribution of each object given the room category. This figure presents some statistics (1st and 3rd quartiles, maximum, minimum, and average conditional probabilities) computed from the whole dataset. We

Figure 2. Exemplar visual images for all room categories in buildings A(top) and B(bottom).

Figure 4. Exemplar visual images for all the objects in the dataset.

can observe how some objects and rooms are highly related, such as the Hand-drier and the Toilet. However, there exist significant differences in the object distribution between sequences, as they were acquired with variations in the time span and environment.

4. Acquisition Procedure

The whole dataset was acquired using a Powerbot robot with a Kinect device installed on top of it at a total height of 90 cm. (see left part of Figure 1). The robot was manually driven using a joystick at an approximate mean linear velocity of 0.3 m/s, and the pose (metric localization) of the robot was not stored. Frames were acquired, processed and saved using the OpenNI tools[2] and the Point Cloud Library (PCL) (Rusu & Cousins, 2011). The acquisition process provided color images a with resolution of 640×480 pixels and point clouds with 307200 3D colored points. The point cloud was stored using the binary Point Cloud Data format (PCD_V7). All frames were manually labeled with the corresponding room category and the presence or absence of the list of 15 predefined objects. Sequences 1 and 4 were generated following the path shown in Fig. 6: the robot started on the 2nd floor, visited 13 rooms, and then moved to the 1st floor using the elevator, where it visited 4 rooms and finished the path. Sequence 2 was acquired following the same but inverse path of Sequences 1 and 4. In other words, the robot moved from the 1st floor to the 2nd one making counterclockwise turns. Sequence 3 was acquired following the same path as for Sequence 2, but starting on the 2nd floor. Finally, Sequence 5 was generated in building B (Fig. 7), where the robot started on the 2nd floor and was also moved to the 1st floor using the elevator.

5. Analysis of the dataset

The ViDRILO dataset is released with a MATLAB toolbox that provides several functionalities. These functionalities are extensively described in the guide of use [3], and they include point cloud representation, statistics generation and frame (visual and depth information) visualization. Moreover, the toolbox includes all the basic steps in both visual place classification and object recognition problems: feature extraction, learning stage, classification, and evaluation of the results.

Once the toolbox and the dataset have been correctly downloaded, a complete semantic localization system is

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

[3] `http://www.rovit.ua.es/dataset/vidrilo/downloads.html`

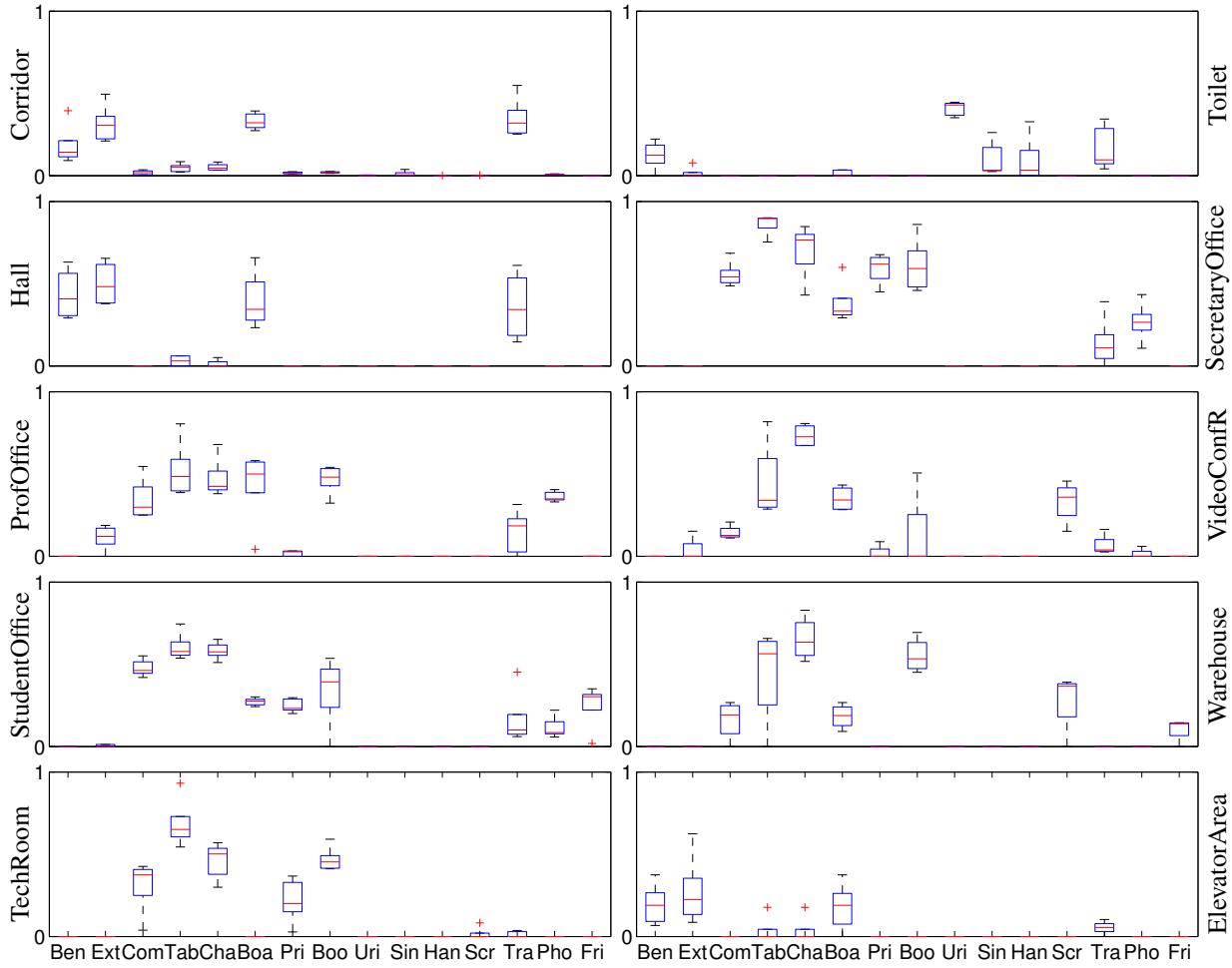

Figure 5. Conditional distribution of each object given the room category for the whole dataset.

trained and evaluated by simply executing the following MATLAB command: `runVidriloClassifier()`. It takes as arguments the training (1-5) and test sequences (1-5), the type of input data (visual=0, depth=1, both=2), the classification model (SVM=0, k-NN=1, RF=2), the visual input descriptor (GIST=0, PHOW=1, HIST=2), and the depth input descriptor (ESF=0, HIST=1).

In order to analyze the dataset, we carried out an extensive experimentation by using: a) five (visual and depth) descriptors, b) three classifiers, and c) every training/test combination of the five sequences. The combination of them resulted into 375 semantic localization systems to be evaluated. Each semantic localization system consists of a single multiclass classifier (room classification), and 15 binary classifiers (object recognition). We computed three visual descriptors: PHOG (Bosch et al., 2007), GIST (Oliva & Torralba, 2001), and a basic grayscale histogram; and two 3D descriptors: ESF (Wohlkinger & Vincze, 2011), and a basic depth histogram. Regarding the classification models, we used: SVM with an exponential chi-squared kernel, a k-NN classifier (with $k = 7$), and a Random Forest (with 50 decision trees). A more detailed description of all the experiments carried out, the descriptors, the classification models, and the results obtained are presented on the ViDRILO web page[4].

In table 2, a selection of these results is shown. Concretely we show the best room classification results, which were obtained by using GIST and ESF as visual and depth descriptors, and a SVM as classification model in both cases. The results are shown for each training/test sequence combination. According to these results, certain points are worth

[4] `http://www.rovit.ua.es/dataset/vidrilo/experimentation.html`

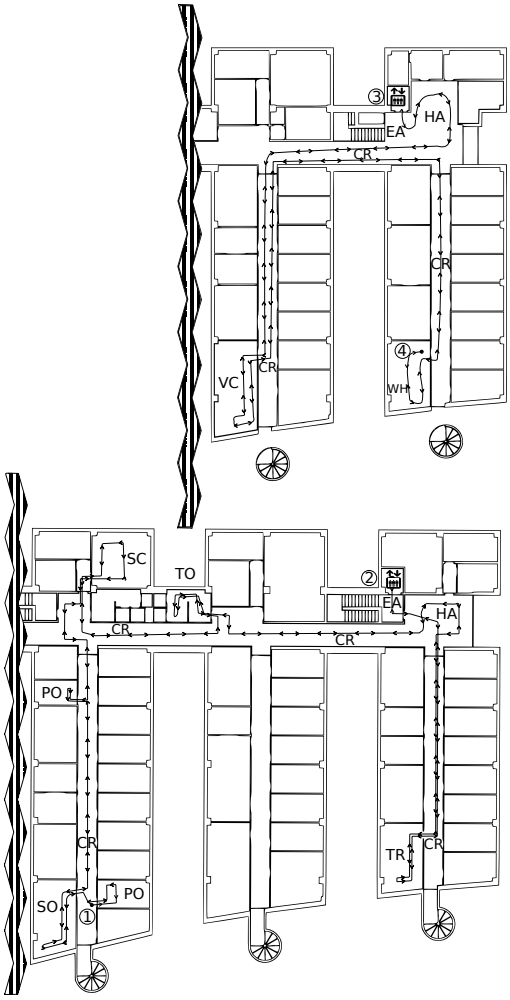

Figure 6. Building A with the paths followed by the robot during the acquisition of Sequences 1 to 4.

highlighting. Firstly, the accuracies present a lower variance when using ESF features due to the effect of the viewpoint and lighting variations on the GIST descriptor, which relies on visual information. As expected, the poorest results with both descriptors were obtained when using Sequence 5 for testing. This reveals the environment generalization as an open problem that cannot be properly managed with generic proposals.

For object recognition, we computed the average F1-score measure for each combination of training/test sequences and the same selection of descriptors and classification model. This is shown in Table 3, where it can be seen how the use of the depth information (ESF) outperforms the visual information (GIST). A deeper analysis of the results shows that ESF descriptors increased a 18% the average object recall with respects to GIST, while the precision decreased

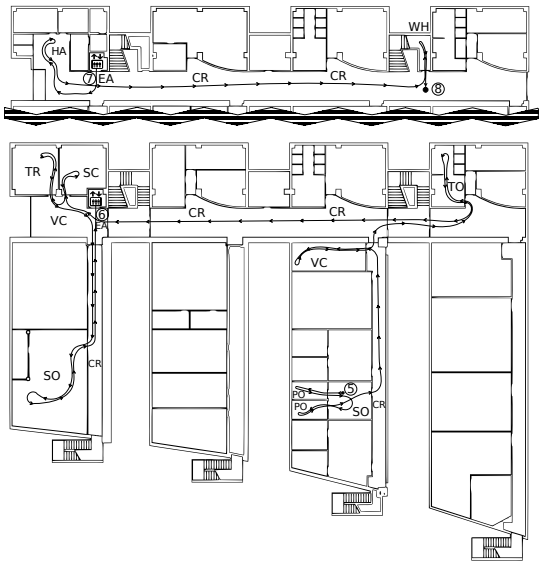

Figure 7. Building B with the paths followed by the robot during the acquisition of Sequence 5.

a 11% on average. The small recall values can be explained by the high number of object absences in the whole dataset.

6. Conclusions

We have presented ViDRILO, a dataset that consists of five sequences captured with a ground mobile robot in two different buildings. Each acquisition records the following information: a perspective image, a 3D point cloud file, the semantic category of the scene, and the presence/absence of a list of 15 predefined objects. The main novelties with respect to similar existing datasets are a large number of labeled frames and the annotation scheme. ViDRILO can be mainly applied to semantic place classification problems, but could also be used in other related ones such as object presence, 3D reconstruction or dense image compression/transmission. The dataset is released in conjunction with a complete toolbox for processing and visualization. This toolbox has been used to assess the dataset in a set of baseline experiments. The use of unreleased ViDRILO sequences in previous editions of the RobotVision competition has proven the suitability of this dataset as a benchmark for semantic localization tasks.

Acknowledgements

This work was supported by grant DPI2013-40534-R of the Ministerio de Economia y Competitividad of the Spanish

Table 2. Room classification accuracy for GIST (top) and ESF (bottom) descriptors.

GIST		Test				
		Seq.1	Seq.2	Seq.3	Seq.4	Seq.5
Training	Seq.1	98.70	64.38	60.99	67.80	35.02
	Seq.2	67.10	99.59	61.12	72.90	33.27
	Seq.3	56.05	59.79	99.47	55.30	32.14
	Seq.4	69.28	71.50	56.32	99.65	33.31
	Seq.5	53.24	46.87	43.51	46.37	99.73

ESF		Test				
		Seq.1	Seq.2	Seq.3	Seq.4	Seq.5
Training	Seq.1	68.94	61.87	57.03	59.90	35.21
	Seq.2	64.49	78.23	61.39	63.34	34.15
	Seq.3	59.48	64.69	80.96	62.29	34.27
	Seq.4	63.54	65.30	65.52	80.11	36.71
	Seq.5	49.56	47.04	46.57	46.75	84.18

Table 3. Object recognition F1 score for GIST (top) and ESF (bottom) descriptors.

GIST		Test				
		Seq.1	Seq.2	Seq.3	Seq.4	Seq.5
Training	Seq.1	1.00	0.45	0.36	0.55	0.37
	Seq.2	0.40	1.00	0.42	0.53	0.14
	Seq.3	0.28	0.40	1.00	0.34	0.13
	Seq.4	0.57	0.57	0.43	1.00	0.26
	Seq.5	0.44	0.26	0.22	0.32	1.00

ESF		Test				
		Seq.1	Seq.2	Seq.3	Seq.4	Seq.5
Training	Seq.1	1.00	0.47	0.55	0.43	0.40
	Seq.2	0.53	1.00	0.65	0.54	0.39
	Seq.3	0.49	0.53	1.00	0.55	0.34
	Seq.4	0.61	0.54	0.67	1.00	0.36
	Seq.5	0.38	0.36	0.42	0.35	1.00

Government, and by Consejería de Educación, Cultura y Deportes of the JCCM regional government through project PPII-2014-015-P. Jesus Martínez-Gómez is also funded by the JCCM grant POST2014/8171.

References

Bosch A, Zisserman A and Munoz X (2007) Representing Shape with a Spatial Pyramid Kernel. In: *Proceedings of the 6th ACM international conference on image and video retrieval.*

Caputo B, Muller H, Thomee B, Villegas M, Paredes R, Zellhofer D, Goeau H, Joly A, Bonnet P, Martínez-Gómez J, García-Varea I and Cazorla M (2013) ImageCLEF 2013: the vision, the data and the open challenges. In: *Information access evaluation. Multilinguality, multimodality, and visualization*.

Kasper A, Xue Z and Dillmann R (2012) The KIT object models database: An object model database for object recognition, localization and manipulation in service robotics. *International Journal of Robotics Research.* 31(8): 927–934.

Lai K, Bo L and Fox D (2014) Unsupervised Feature Learning for 3D Scene Labeling. In: *International conference on robotics and automation.*

Luo J, Pronobis A, Caputo B and Jensfelt P (2006) (2006) The kth-idol2 database. *KTH, CAS/CVAP, technical report* 304.

Martínez-Gómez J, García-Varea I, Cazorla M and Caputo B (2013) Overview of the ImageCLEF 2013 Robot Vision Task. In: *CLEF 2013 evaluation labs and workshop, Online working notes*.

Oliva A and Torralba A (2001) Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision.* 42(3): 145–175.

Pronobis A and Caputo B (2009) COLD: COsy Localization Database. *International Journal of Robotics Research.* 28(5): 588–594.

Rusu RB and Cousins S (2011) 3D is here: Point Cloud Library (PCL). In: *International conference on robotics and automation (ICRA)*.

Silberman N, Hoiem D, Kohli P and Fergus R (2012) Indoor Segmentation and Support Inference from RGBD Images. In: *European conference on computer vision (ECCV).*

Smith M, Baldwin I, Churchill W, Paul R and Newman P (2009) The New College Vision and Laser Data Set. *International Journal of Robotics Research.* 28(5): 595–599.

Thomee B and Lew MS (2012) Interactive search in image retrieval: a survey. *International Journal of Multimedia Information Retrieval.* 1(2): 71–86.

Wohlkinger W and Vincze M (2011) Ensemble of shape functions for 3D object classification. In: *International conference on robotics and biomimetics (ROBIO).*