Recognizing Objects In-the-wild: Where Do We Stand?

Mohammad Reza Loghmani[1], Barbara Caputo[2] and Markus Vincze[1]

***Abstract*— The ability to recognize objects is an essential skill for a robotic system acting in human-populated environments. Despite decades of effort from the robotic and vision research communities, robots are still missing good visual perceptual systems, preventing the use of autonomous agents for real-world applications. The progress is slowed down by the lack of a testbed able to accurately represent the world perceived by the robot in-the-wild. In order to fill this gap, we introduce a large-scale, multi-view object dataset collected with an RGB-D camera mounted on a mobile robot. The dataset embeds the challenges faced by a robot in a real-life application and provides a useful tool for validating object recognition algorithms. Besides describing the characteristics of the dataset, the paper evaluates the performance of a collection of well-established deep convolutional networks on the new dataset and analyzes the transferability of deep representations from Web images to robotic data. Despite the promising results obtained with such representations, the experiments demonstrate that object classification with real-life robotic data is far from being solved. Finally, we provide a comparative study to analyze and highlight the open challenges in robot vision, explaining the discrepancies in the performance.**

I. INTRODUCTION

Objects are ubiquitous in our everyday lives. Every common activity, such as cooking or cleaning, implies the capability of understanding and operating a set of objects to successfully complete a task. In order for a Service Robot (SR) to operate in human environments as well, the ability to recognize objects is a basic requirement. Object recognition is rarely a self-contained task, but it is rather a proxy for a large variety of high-level tasks, such as navigation, manipulation and user interaction, that heavily rely on an accurate description of the visual scene.

The advent of deep learning has had a huge impact on the object recognition task after decades of plateaued results. The progressive design of deeper and more sophisticated networks, starting from AlexNet [1], VGG [2], Inception [3] [4] [5] to ResNet [6] [7], has led to outstanding results in competitions such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [8]. Arguably the primary driving force of the deep learning revolution is the availability of large scale datasets. The majority of these datasets, such as the popular ImageNet [9], Pascal VOC [10], and Caltech-256 [11], are composed of images collected through Web search engines. However, the representation of the visual world provided by these datasets implies a bias from the observer (a human photographer) and the Web search engines [12] that are incoherent with the representation perceived by, for example, a SR. It is then legitimate to ask whether the features learned from Web-based datasets can generalize well to robotic data, despite the aforementioned bias.

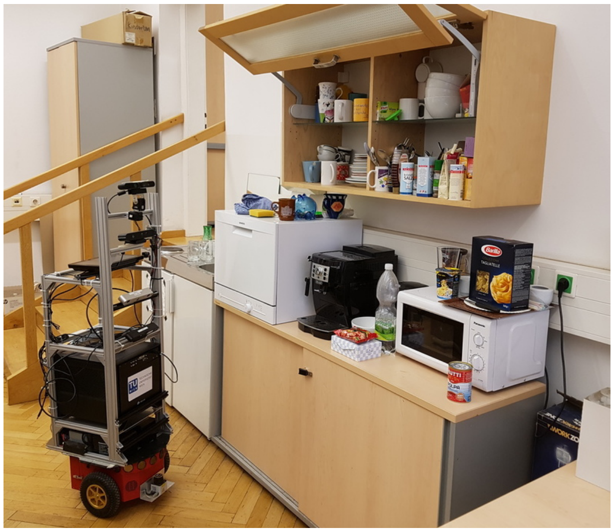

Fig. 1: Glimpse of the data collection process with the robotic platform (left) acquiring data of a cluttered scene populated with everyday objects.

In the last years, computer vision has progressed enormously due to the establishment of standard references and benchmarks, e.g. ImageNet, which have enabled consistent comparison and development of new methods. Unfortunately, the robot vision community has not experienced the same progress due to the lack of accurate testbeds for validating novel algorithms. In the past years, the RGB-D Object Dataset (ROD) [13] has become “de facto” standard in the robotic community for the object classification task [14] [15] [16]. Despite its well-deserved fame, this dataset has been acquired in a very constrained setting and does not present all the challenges that a robot faces in a real-life deployment. In order to fill the existing gap in the robot vision community between research benchmark and real-life application, we introduce a large-scale, multi-view object dataset collected with an RGB-D camera mounted on a mobile robot (see figure 1), called Autonomous Robot Indoor Dataset (ARID). The data are autonomously acquired by a robot patrolling in a defined human environment. The dataset presents 6,000+

[1]Mohammad Reza Loghmani and Markus Vincze are with the Vision4Robotics Group, Automation and Control Institute (ACIN), TU Wien, Vienna, Austria [loghmani, vincze]@acin.tuwien.ac.at

[2]Barbara Caputo is with the VANDAL Laboratory, Department of Computer, Management and Control Engineering (DIAG), Sapienza Rome University, Rome, Italy caputo@dis.uniroma1.it

RGB-D scene images and 120,000+ 2D bounding boxes for 153 common everyday objects appearing in the scenes. Analogously to ROD, the object instances are organized into 51 categories, each containing three different object instances. In contrast, our dataset is designed to include real-world characteristics such as variation in lighting conditions, object scale and background as well as occlusion and clutter. To our knowledge, no other robotic dataset embedding all the challenges of real-life data can be found in the literature. Upon acceptance of the paper, all the collected data, together with the information needed to replicate the experiments, will be made publicly available.

In addition to introducing a new dataset, we inspect the effectiveness of features learned from the Web domain on robotic data and compare them with the features learned from the RGB-D domain. This comparison is made possible by collecting a second dataset containing the images downloaded from the Web representing the same categories as ARID. The acquisition of this Web-based dataset is performed by using query expansion strategies from [17] on different search engines followed by a manual cleaning to remove noisy images. Exhaustive experiments with different deep convolutional networks demonstrate that, despite the greater similarity between the RGB-D and the robotic domain, models learned from Web images are more effective. Finally, the best performing network, ResNet-50, is used to study the classification results on subsets of ARID representing three problematic characteristics of robotic data: small images, occlusion and clutter. The experiments point out small images as the main challenge of robotic data, indicating a path to follow for the resolution of the object classification problem for robotics.

In summary, our contributions are the following:

- a new RGB-D object dataset, collected in-the-wild with a mobile robot, that provides a "litmus test" for the validation of object recognition algorithms developed for robotic applications,
- a detailed analysis of publicly available RGB-D datasets from a robotic perspective,
- comprehensive experiments with several well-established deep convolutional networks, comparing the effectiveness of data coming from the Web and RGB-D domain in generating features for object classification in robotics, and
- a study of the main factors responsible for the difficulties faced by classifiers on robotic data.

The rest of the paper is organized as follows: the next section positions our approach compared to related work, section III introduces the proposed robotic dataset, section IV presents the experimental results and section V draws the conclusions.

II. Related Work

In the following, we first analyze the characteristics of existing RGB-D datasets from a robotic perspective. Then, we review related works on the transferability of learned features across different domains by focusing on the Web and RGB-D domains.

A. Datasets

During the last decade, a variety of datasets have been made publicly available for research. With the popularization of deep neural networks, which require a considerable amount of data for training, the race for large-scale datasets has become more intense. While Web images exist in abundance, robotic images are difficult to obtain because platforms are expensive and data acquisition is time consuming. Nevertheless, the robotic community has produced some interesting datasets. In particular, for indoor objects, the most relevant datasets are JHUIT-50, BigBIRD, iCubWorld Transformation, ROD, and the Active Vision Dataset.

ROD [13] contains 300 objects from 51 categories spanning from fruit and vegetables to tools and containers. Despite the availability of multiple views, each object is presented in isolation and variation in lighting condition, object scale and background are missing. The corresponding scene dataset, the RGB-D Scene Dataset [18], presents multiple objects in the same scene, but considers only five object categories. BigBIRD [19] contains 125 common human-made objects, with particular focus on boxes and bottles. This dataset is specifically designed for instance recognition and the selected objects belong to very few categories. In addition, occlusion, clutter, scale and light variation are not captured. A more recent dataset, the Active Vision Dataset [20], uses a subset of 33 objects from BigBIRD in densly acquired scenes. The data is directly acquired by a robot and it embeds most of the nuisances typical of real-life data. Nevertheless, the limited number of considered objects makes this dataset unsuitable for classification. JHUIT-50 [21] contains 50 industrial objects and hand tools used in mechanical operations. The objects are captured in isolation and from multiple viewpoints. Due to its limited scope, this dataset is more suitable for instance recognition rather that classification. In addition, nuisances such as occlusion, clutter, scale and light variation are not captured. The corresponding scene dataset, JHUScene-50 [22], includes occlusion and clutter, but limits the number of considered objects to 10. iCubWorld Transformation [23] contains 150 common indoor objects from 15 different categories. The data are collected directly with the iCub humanoid robot [24]. This dataset addresses specifically variance in the background as well as the variance in scale and rotation of the object. Nevertheless, each object is presented in isolation, avoiding problems caused by cluttered scenes.

Despite the high-quality that characterizes each of these datasets, their constrained setting makes them incoherent with real-life data. In addition, only the Active Vision Dataset and the iCubWorld Transformation present data collected directly from a robot. Table I presents a summary of the characteristics of the datasets discussed above and highlights that, differently from other datasets, ARID embeds all these characteristics.

TABLE I: Summary of the characteristics of different RGB-D datasets with focus on variation in lighting condition, variation in scale, multiple views, occlusion, clutter, variation in background and whether or not the data are collected directly from a robot. *Not Available* (NA) indicates that the dataset focuses on object instances rather than categories and the number of categories is unknown.

Dataset		**Characteristic**						
Name	**# classes**	**light var.**	**scale var.**	**multiview**	**occlusion**	**clutter**	**bkg var.**	**robot**
RGB-D Object Dataset [13]	51			✓				
RGB-D Scene Dataset [18]	5		✓	✓	✓	✓	✓	
BigBIRD [19]	NA			✓				
Active Vision Dataset [20]	NA	✓	✓	✓	✓		✓	✓
JHUIT-50 [21]	NA			✓				
JHUScene-50 [22]	NA			✓	✓	✓		
iCubWorld Transf. [23]	15	✓	✓	✓			✓	✓
Autonomous Robot Indoor Dataset (ARID)	51	✓	✓	✓	✓	✓	✓	✓

B. Transfer Learning

Deep convolutional networks are currently dominating several computer vision tasks. One of the key factors contributing to their success is the transferability of the produced deep representation for a variety of visual recognition tasks. The deep representations, also called features, of these networks have been empirically proven to be superior to traditional hand-crafted features, e.g. [25] [26] [27]. In order to take advantage of the generalization power of deep models, the networks need a large amount of training data. For this reason, large-scale datasets, such as ImageNet, with millions of samples, have been extensively used across different domains. It is common practice to further adapt the deep representation learned from a large dataset to the specific domain of interest through fine-tuning [28] [29], i.e., by refining the representation using annotated data from the novel task in a subsequent training stage.

The effectiveness of using features learned from the Web domain in the robotic domain has been previously studied [17] [30]. In Massouh et al. [17], features learned from the Web domain are tested on the RGB-D Object Dataset, while in Pasquale et al. [30], features learned from ImageNet are used to train a classifier on the iCubWorld28 [30], a former version of the iCubWorld Transformation. Although both works exhibit interesting results, we claim that, due to the intrinsic constraints discussed in section II-A, the utilized datasets cannot be considered as reliable representatives of real-life robotic data. In addition, only AlexNet and Inception are used to produce the analyzed features. Our work exhaustively benchmarks deep models obtained with five different networks against a robotic dataset collected in-the-wild.

III. AUTONOMOUS ROBOT INDOOR DATASET

In the following, we describe the characteristics of the proposed robotic dataset by highlighting its most significant peculiarities. In addition, we unveil the protocol used for the autonomous data collection and the details of the provided annotation.

A. Scope and Motivation

The Autonomous Robot Indoor Dataset contains RGB and depth images of daily life objects belonging to 51 categories. Each object category contains three instances, for a total of 153 physical objects, and it coincides with one of the 51 WordNet leaf nodes used to determine the categories of a very well-known dataset, the RGB-D Object Dataset. In other words, there is a complete overlap between the categories represented in the two datasets, ARID and ROD. Figure 2 gives a concrete idea of the dataset's content by showing one sample object per category.

Since we are mostly interested in autonomous assistive robots operating in indoor environments, the object classes considered in ROD are a valid representative. These objects consist of a large variety of food items, such as fruit, vegetables and packed goods, and human-made objects common to homes and offices. Nevertheless, our goal is not to extend and contribute to ROD, but rather fill the gap between research-oriented datasets and real-life data by introducing a robotic dataset collected in-the-wild. While ROD contains images collected in a constrained setting (fixed camera-object distance, static background, invariant light conditions), our dataset includes all the nuisances of robotic data by acquiring it directly with a mobile robot navigating autonomously in an indoor environment. More precisely, the following challenges are taken into account:

- variation of lighting conditions,
- object scale variation,
- significant changes in the viewpoint,
- partial view and occlusion,
- clutter, and
- background variation.

We hope that this work provides the robot vision community with a tool to advance the visual capabilities of robots in order to accelerate their integration in our lives.

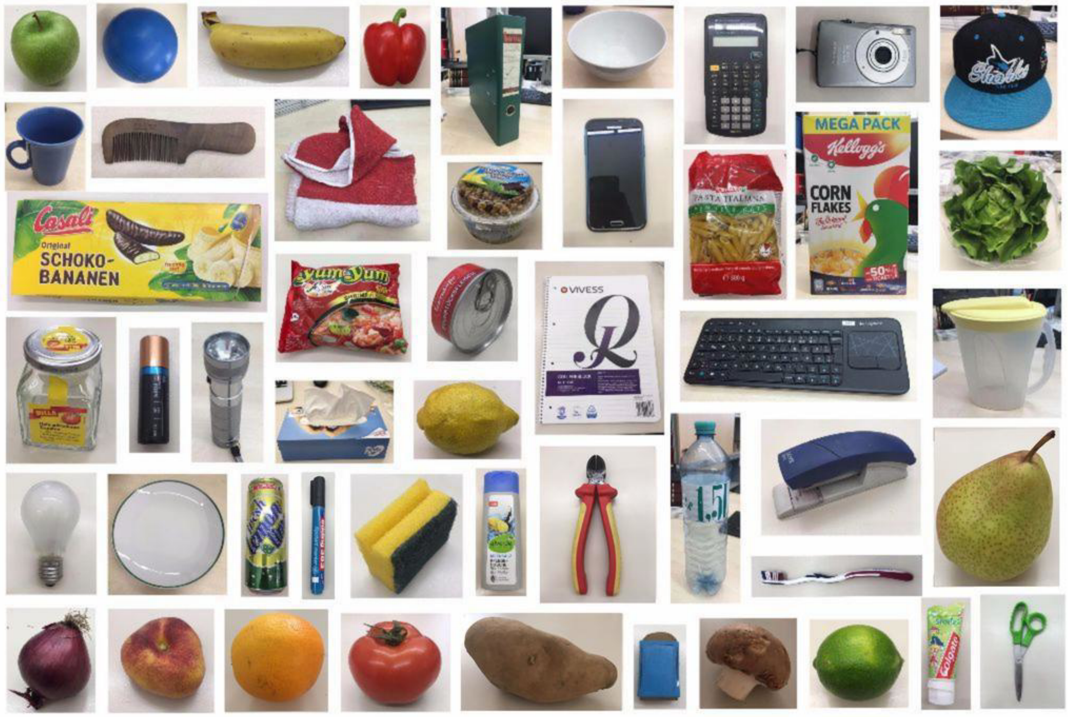

Fig. 2: Sample of objects used in Autonomous Robot Indoor Dataset. Each object shown belongs to a different category.

B. Data Acquisition Protocol

In order to avoid a human bias in data acquisition and to observe the objects from the robot's perspective, a mobile robot with an RGB-D camera is used. In particular, the mobile robotic platform is powered by a Pioneer P3-DX with a customized structure that supports an Asus Xtion Pro camera mounted on a pan/tilt unit (see figure 1).

The data collection is performed in 10 different sessions conducted during different days and at different times of the day: this allows a natural variation of the lighting conditions among the data. At each run, 30-31 objects are spread in the environment where the mobile robot patrols predefined waypoints. When a waypoint is reached, the camera scans the scene with a horizontal movement of the pan/tilt unit and acquires RGB and depth data, both with a resolution of 640x480 pixels and a frame rate of 30 Hz. The RGB and the depth frames are later synchronized based on their acquisition time and unmatched frames are discarded. Each session lasts for approximately one hour in which the robot continuously loops over four distinct waypoints. In order to guarantee the appropriate variability in terms of camera-object distance and object view, the objects are randomly moved in between two patrolling loops.

C. Annotation

In order to discard similar frames, every fifth frame is chosen for annotation for a total of over 6,000 frames. For each frame, a bounding box annotation indicates the location and the label (at instance level) of every visible object for a total of over 120,000 2D bounding boxes for the whole dataset. A modified version of Sloth annotation tool [31] is used for this purpose. In case of occlusion or partial view, if the object is still distinguishable, a bounding box is drawn around the visible part of the object. Figure 3 shows a sample frame, together with its bounding box annotation. Since the objects are captured in a realistic scenario rather than in isolation, the dataset is also suitable for object detection. In addition, the availability of object labels at instance level allows the dataset to be used for object classification as well as object identification (also referred to as instance recognition).

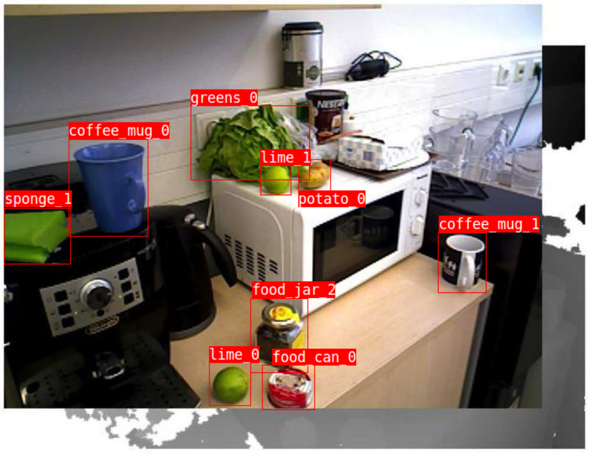

Fig. 3: Example of an RGB-D frame from the Autonomous Robot Indoor Dataset with 2D bounding box annotation.

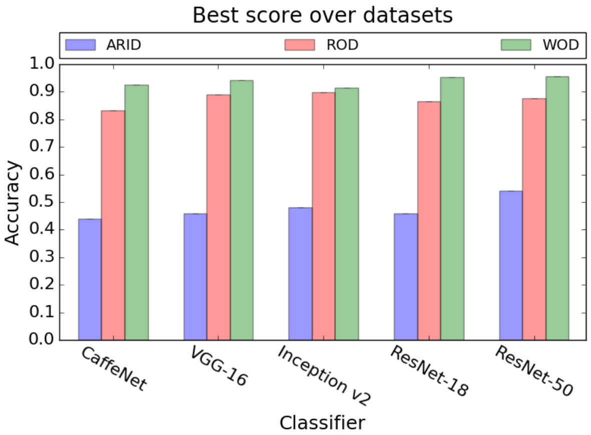

Fig. 4: Accuracy of different deep convolutional networks on three datasets: Autonomous Robot Indoor Dataset (ARID), RGB-D Object Dataset (ROD) [13] and Web Object Dataset (WOD). The results are obtained by training and testing on different splits of the same dataset.

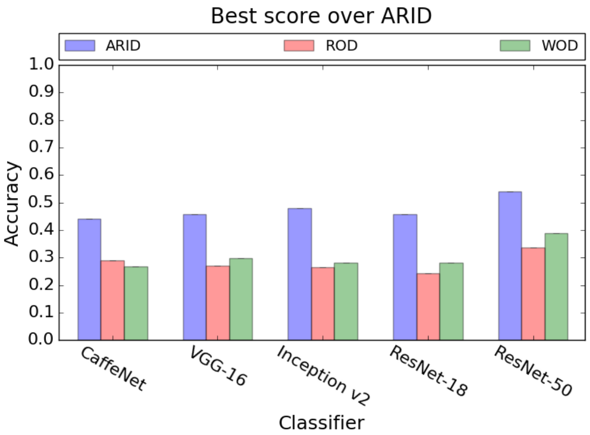

Fig. 5: Accuracy of different deep convolutional networks on Autonomous Robot Indoor Dataset (ARID). The results are obtained by training independently on ARID, RGB-D Object Dataset (ROD) [13] and Web Object Dataset (WOD) and testing on ARID.

IV. Experiments

We take advantage of the availability of ARID to disclose the characteristics of robotic data. In particular, we want to (i) analyze the transferability of features from the Web domain to the robotic domain (Section IV-A) and (ii) study the characteristics of robotic data to identify the main source(s) of complication for classifying objects (Section IV-B). In order to accomplish these goals, another dataset, called Web Object Dataset (WOD), is collected. WOD is composed of images downloaded from the Web representing objects from the same categories as ARID. The images are downloaded from multiple search engines (Google, Yahoo, Bing and Flickr) using the method proposed by Massouh et al. [17]. This method uses a concept expansion strategy by leveraging visual and natural language processing information to minimize the noise while maximizing the visual variability. The remaining noise is then manually removed, leaving a total of 50,547 samples.

A. *Baseline and Features Transferability*

The limited availability of robotic data raises the question of whether data coming from a more accessible domain, the Web domain, can be effectively used instead of data from the RGB-D domain to learn features that are transferable to the robotic data. In particular, we want to compare the performance of well-known deep convolutional networks on robotic data (ARID), when trained on Web data (WOD) and on RGB-D data (ROD). In order to allow a fair evaluation, a subset of 40,000 samples from ARID dataset is selected, such that all the involved datasets are approximately the same size. It is worth noticing that, since WOD does not contain depth information, only RGB data are considered for all datasets. For this benchmark, we employ some of the most utilized network architectures in the literature, CaffeNet[1], VGG-16, Inception v2, ResNet-18 and ResNet-50. All networks are pre-trained on ImageNet and then fine-tuned on the desired dataset, according to the guidelines provided in [28] [29].

In order to provide a reference for the upcoming evaluations, we assess the performances of all considered networks for each of the three datasets (ARID, ROD, WOD) when training and test set come from the same dataset. For each dataset, multiple training/test splits are considered and the results are averaged to obtain the final classification accuracy. In particular, for ARID, each split uses one different object instance per class in the test set, for ROD, the first three splits indicated by the authors is used and, for WOD, each split uses 25% of the data in the test set. From the results displayed in figure 4, it can be noticed that the different networks consistently obtain a higher accuracy on WOD. Unsurprisingly, ARID appears to be the most challenging dataset and all the networks achieve an accuracy much lower (on average, ~ 0.4 lower) on ARID than on the other two datasets.

The networks fine-tuned on ROD and WOD are then tested on ARID to evaluate the transferability of the learned features to the robotic data. From the results displayed in figure 5, it can be noticed that, as expected, all the networks undergo a performance drop when the training and test set belong to different datasets with respect to the case in which both sets belong to the same dataset (see figure 4). The domain shift responsible for this negative inflection of the classification results occurs because the data composing training and test set are drawn from different distributions [32]. However, features learned from Web data (WOD) consistently allow a higher classification accuracy (with improvements up to 0.05) on robotic data (ARID) than features learned from

[1] A slightly modified version of AlexNet in which the normalization is performed after the pooling.

TABLE II: Accuracy of multiple deep convolutional networks on different training/test combination of three datasets: Autonomous Robot Indoor Dataset (ARID), RGB-D Object Dataset (ROD) [13] and Web Object Dataset (WOD). For each training/test set combination, the mean and maximum accuracy among the considered networks is shown.

Dataset		**Network**					**Statistics**	
Train on	**Test on**	**CaffeNet**	**VGG-16**	**Inception-v2**	**ResNet-18**	**ResNet-50**	**Mean**	**Max**
ROD	ROD	0.832	0.889	0.897	0.864	0.876	0.872	0.897
ROD	ARID	0.291	0.270	0.266	0.243	0.337	0.281	0.337
WOD	WOD	0.924	0.942	0.914	0.953	0.956	0.938	0.956
WOD	ARID	0.268	0.297	0.282	0.282	0.388	0.303	0.388
ARID	ARID	0.441	0.458	0.481	0.458	0.540	0.476	0.540

RGB-D data (ROD) on all networks, with the exception of CaffeNet. The key factor to interpret this phenomenon is the greater variability of Web images: while ROD contains a limited number of instances per class, with some classes containing only 3 instances, in WOD each sample potentially represents a different object instance. Very deep networks, like ResNet-50, with high capacity and generalization power, take advantage of this richness in information to generate better models. This is further highlighted by the difference between the accuracy of ResNet-50 and the mean accuracy of all tested networks when training with WOD (see table II). The results of this experiment have a twofold implication: (i) despite the greater visual affinity between the RGB-D and the robotic domain, data from the Web domain generate more effective models for object classification in robotic applications, and (ii) the currently well-established deep convolutional network, when used in their plain stand-alone form and without any prior, do not perform satisfactorily for object classification in robotics.

B. Robotic Challenges

In order to better understand which characteristics of robotic data negatively influence the results of the object classification task, we independently analyze three key variables: image dimension, occlusion and clutter[2]. Image dimension is a variable related to the camera-object distance: when the camera is not near enough to clearly capture the object, the object occupies only few pixels in the whole frame, making the classification task more challenging. For obvious reasons, this problem is emphasized when dealing with small and/or elongated objects, such as dry batteries or glue sticks. Occlusion occurs when a portion of an object is hidden by another object or when only part of the object enters the field of view. Since distinctive characteristics of the object might be hidden, occlusion makes the classification task considerably more challenging. Clutter refers to the presence of other objects in the vicinity of the considered object. The simultaneous presence of multiple objects may interfere with the classification task.

TABLE III: Accuracy of ResNet-50, trained on Web Object Dataset and on its augmented version (++), for three subsets of Autonomous Robot Indoor Dataset containing small images, occluded objects and clutters. The model is also tested on the whole dataset to show the overall impact of data augmentation.

Challenge	**Accuracy**	
	Top-1	**Top-5**
Small image	0.230	0.511
Occlusion	0.273	0.508
Clutter	0.558	0.777
Small image ++	0.240	0.513
Occlusion ++	0.318	0.577
Clutter ++	0.543	0.802
All ++	0.441	0.702

Table III shows the classification results of the best performing model of section IV-A (ResNet-50 trained on WOD) on three subsets of ARID, each containing samples with the characteristics discussed above. The set of small images is obtained by taking half of ARID containing images with the smallest area, while the occlusion and clutter set have been manually selected. It is worth noticing that the three sets are mutually exclusive in order to avoid interference between the analyzed variables. The occlusion and, especially, the small images set exhibit low accuracy, thus negatively affecting the classification score of the whole dataset. It is possible to improve the classification by performing problem-specific data augmentation during the training phase. In particular, we augmented WOD by resizing the original samples to different scales and by randomly adding rectangular noise patches to simulate occlusion. These two strategies are commonly used to encourage the network to learn scale-/occlusion-invariant models [3] [16] [33]. Table III also shows the performances of ResNet-50 trained with this augmented WOD on the three subsets and on the whole ARID dataset. Even though the occlusion set benefits from this strategy (and so does the whole dataset) the classification of small images does not

[2] Since ARID is collected in-the-wild, by definition, the data acquisition is performed in an unconstrained manner. For this reason, rigorously isolating other characteristics of the data, such as light variation, background variation and different object view is prohibitive.

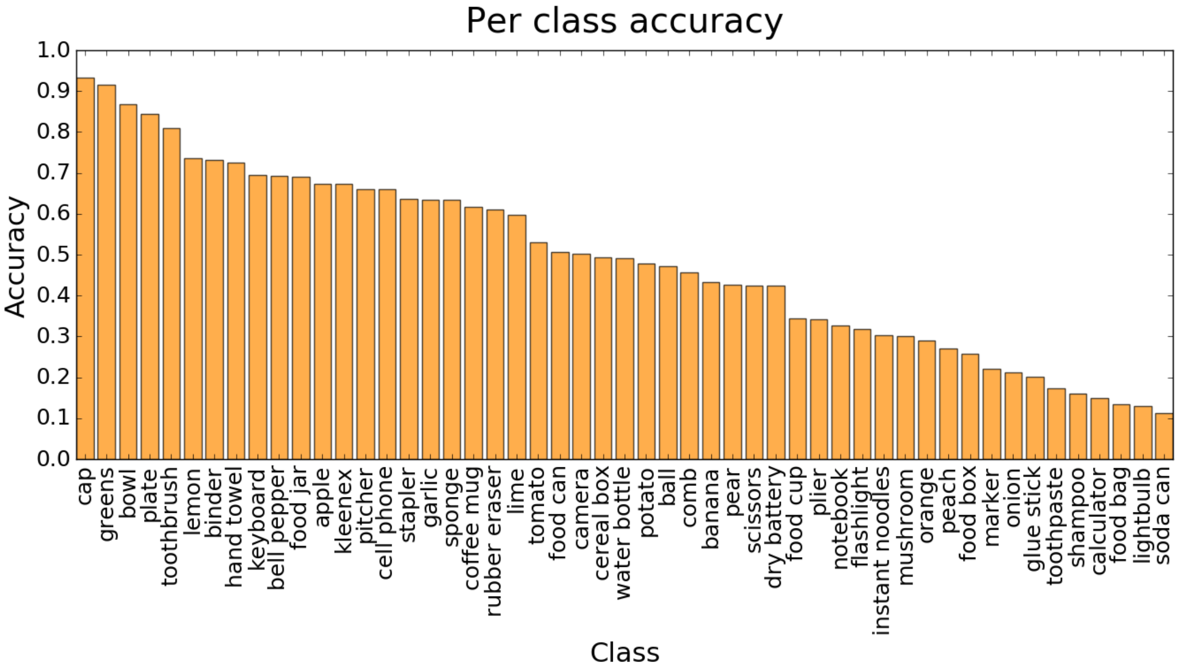

Fig. 6: Accuracy of each of the 51 classes of the Autonomous Robot Indoor Dataset obtained with a ResNet-50 trained on the augmented Web Object Dataset.

exhibit significant improvement. The difficulty of classifying small images is further confirmed by the results in figure 6, where classes representing small or elongated objects have the lowest accuracy.

V. Discussion and Conclusion

In this paper, we have presented ARID: a large-scale, multi-view, RGB-D object dataset collected with a mobile robot in-the-wild. This dataset is designed to capture the challenges a robot faces when deployed in an indoor environment and fills the current gap in the robot vision community between research oriented datasets and real-life data. Furthermore, with an extensive comparative study, we have shown that it is possible to overcome the complication of collecting a large amount of robotic data for training data-craving deep convolutional networks by using images downloaded from the Web. We have found that, despite being relatively easy to obtain, Web-based data allow the generation of more effective deep models than the RGB-D counterpart for the classification of robotic images. Nevertheless, object classification remains a challenging task in robotics and current algorithms present results that are insufficient for a successful integration of robotic systems in our homes. In order to shed light on the difficulties of this task, we have analyzed the effects of specific factors, such as object dimension, occlusion and clutter, on the performance. Results indicate that clutter is rather a secondary problem: occlusions and, especially, small objects more seriously degrade the classification accuracy. These observations suggest a research path in which visual tasks for robotic applications are tackled through methods designed to cope with domain-specific challenges. ARID is a valuable resource to pursue this goal and provides an important testbed for the robot vision community. In addition, the dataset may also be used to explore other aspects of robotic data, such as the integration of RGB and depth information.

ACKNOWLEDGMENT

This work has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676157, project ACROSSING, by the ERC grant 637076 - RoboExNovo (B.C.), and the CHIST-ERA project ALOOF (B.C.). The authors would like to thank Victor León, Mirco Planamente and Silvia Bucci for their help during the data collection and annotation process. The authors are also grateful to Tim Patten for the support in finalizing the work.

References

[1] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25 (NIPS)*, 2012, pp. 1097–1105.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[4] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[5] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017, pp. 4278–4284.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[7] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2016, pp. 87.1–87.12.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[9] J. Deng, W. D. R. Socher, L. Li, K. Li, and F. Li, "Imagenet: A large-scale hierarchical image database," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[10] M. Everingham, L. V. Gool, C. K. I. W. J., Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision (IJCV)*, vol. 88, no. 2, pp. 303–338, 2010.

[11] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," 2007, technical Report 7694, California Institute of Technology.

[12] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1521–1528.

[13] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1817–1824.

[14] L. Bo, K. Lai, X. Ren, and D. Fox, "Object recognition with hierarchical kernel descriptors," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1729–1736.

[15] R. Socher, B. Huval, B.Bhat, C. Manning, and A. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS)*, 2012, pp. 656–664.

[16] F. Carlucci, P. Russo, and B. Caputo, "A deep representation for depth images from synthetic data," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2017, in press.

[17] N. Massouh, F. Babiloni, T. Tommasi, J. Young, N. Hawes, and B. Caputo, "Learning deep visual object models from noisy web data: How to make it work," *CoRR*, vol. abs/1702.08513, 2017.

[18] K. Lai, L. Bo, and D. Fox, "Unsupervised feature learning for 3D scene labeling," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 3050–3057.

[19] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3D database of object instances," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 509–516.

[20] P. Ammirato, P. Poirson, E. Park, and A. Berg, "A dataset for developing and benchmarking active vision," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2017, in press.

[21] C. Li, A. Reiter, and G. D. Hager, "Beyond spatial pooling: Fine-grained representation learning in multiple domains," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4913–4922.

[22] C. Li, J. Bohren, E. Carlson, and G. D. Hager, "Hierarchical semantic parsing for object pose estimation in densely cluttered scenes," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 5068–5075.

[23] G. Pasquale, C. Ciliberto, L. Rosasco, and L. Natale, "Object identification from few examples by improving the invariance of a deep convolutional neural network," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4904–4911.

[24] G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, C. von Hofsten, K. Rosander, M. Lopes, J. Santos-Victor, A. Bernardino, and L. Montesano, "The icub humanoid robot: An open-systems platform for research in cognitive development," *Neural Networks*, vol. 23, no. 8, pp. 1125 – 1134, 2010.

[25] A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR)*, 2014, pp. 512–519.

[26] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014: 13th European Conference, Proceedings, Part I*, 2014, pp. 818–833.

[27] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML) - Volume 32*, 2014, pp. I–647–I–655.

[28] B. Chu, V. Madhavan, O. Beijbom, J. Hoffman, and T. Darrell, "Best practices for fine-tuning visual classifiers to new domains," in *Computer Vision – ECCV 2016 Workshops, Proceedings, Part III*, 2016, pp. 435–442.

[29] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014, pp. 3320–3328.

[30] G. Pasquale, C. Ciliberto, F. Odone, L. Rosasco, and L. Natale, "Teaching icub to recognize objects using deep convolutional neural networks," in *Proceedings of The 4th Workshop on Machine Learning for Interactive Systems at ICML 2015*, vol. 43, 2015, pp. 21–25.

[31] "Sloth," https://github.com/cvhciKIT/sloth, accessed: 2017-09-05.

[32] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *Computer Vision – ECCV 2010: 11th European Conference on Computer Vision (ECCV), Part IV*, 2010, pp. 213–226.

[33] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, in press.