2017 26th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
Lisbon, Portugal, Aug 28 - Sept 1, 2017.

Impact of Embodied Training on Object Recognition

Priya Narayanan, Magdalena D. Bugajska, Wallace Lawson, and J. Gregory Trafton

Abstract— **The ability to perform robust, precise, real-time visual recognition is extremely critical for the use of robotic systems in real-world applications. This paper explores the use of Convolution Neural Networks (CNN) and human assisted training in teaching a robot to recognize novel objects.**

We investigated the impact of providing instructions to a human teacher during a training scenario for novel objects. Participants in the naïve condition were provided verbal instructions by the robot, and participants in the embodied condition were provided embodied demonstrations by the robot. The results showed that a vision system trained by participants with embodied instructions clearly outperformed a system trained by naïve participants. The latest computer vision techniques combined with human assisted teaching was found to provide excellent results for novel object recognition.

I. Introduction

The last several years have witnessed a huge surge in projects that transition robotic systems from research laboratories to natural environments for real-world applications [1,2,3]. Advanced autonomous robots are being designed to assist humans in a variety of settings ranging from domestic environments to workplace needs [4]. In spite of the advancements in autonomous robotic systems, efficient and accurate perception remains a significant impediment.

An important practical issue is the difficulty in gathering labeled images that can be used to train perception systems. Also, since these systems use two-dimensional images to learn and recognize three-dimensional objects, the availability of various poses or viewpoints in the training images is critical to the performance of learning algorithms [5]. In this study, we explore the use of a combination of deep learning techniques and human assisted teaching strategies for training novel objects on a robot quickly and accurately.

Deep learning using convolutional neural networks (CNN) provides an accurate and efficient way to recognize objects. Unfortunately, training a CNN requires a large amount of data, partially due to the number of free parameters that must be tuned. For example, the popular ImageNet dataset [6] contains 1.2 million images with 1000 categories. Existing datasets may not have objects that are of interest to the user. In such cases, *fine-tuning* can be an effective way to update an already trained CNN to recognize domain specific objects [7]. In fine-tuning, a CNN is adapted to new objects of interest by fixing most parameters, focusing on training of only the final layer(s) of the network. Although this requires substantially less data [7], the performance still depends on the availability and the quality of the data used for fine-tuning. The assistance of a human collaborator/teacher can be a very successful method for acquiring images to fine-tune the network.

We examine a scenario where the robot arrives at an unfamiliar environment with an already trained vision system (i.e., CNN). The robot then has to understand its new surroundings, and objects around it. In addition to the object classes it has been trained on, it will need to learn new objects in its new settings. A human teacher will assist the robot in collecting labeled data needed to learn new objects. We assume that a typical teacher is unfamiliar with the intricacies of robots and perceptual systems, but is familiar with different forms of social learning [8]. The human-robot interaction and teaching strategies are likely to vary between teachers and can vary greatly based on the instruction provided to them. In order to understand the impact of these instructions on the teaching approaches, we conducted a human subject study. We specifically investigated the teaching strategies adopted by a teacher who was provided embodied instructions versus a naive teacher. We evaluated how these strategies effect fine-tuning and the robot's vision system. In the human trials we collected real-time data with a humanoid robot in a setting that incorporates imperfections in real world conditions.

Thus, this article investigates the viability of using the latest machine learning algorithms combined with human assisted training for object recognition on a robotic platform.

II. Related Work

The past decade has seen extensive advances in the development of methodologies, algorithms and models to address the limitations in visual object recognition. Many of the earlier approaches use multiple handcrafted features extracted from two-dimensional RGB images and these features are then fed to classifiers to learn recognition models. Some of the popular hand-designed features are Scale Invariant Feature Transform (SIFT) [9] and Histogram of Oriented Gradients (HOG) [10]. In addition, several biologically inspired features have been developed that exhibit a good trade-off between generalization ability and discrimination ability for object recognition [11, 12]. With the advent of deep convolutional neural networks (CNN), these handcrafted features have been surpassed by efficient

Research supported by Naval Research Laboratory (NRL) and ONR.

Priya Narayanan was a NRC fellow at Naval Research Laboratory (NRL), Code 5515, Washington, DC 20375, USA. She is currently with the Army Research Laboratory (phone: 301-394-2376; fax: 301-394-5234; e-mail: Priya.narayanan.civ@mail.mil)

Magdalena D. Bugajska, Wallace Lawson, and J. Gregory Trafton are with the Naval Research Laboratory, Code 5515, Washington, DC 20375 USA.

supervised or semi-supervised algorithms that can directly learn and hierarchically extract features from the data to provide "end-to-end learning" [13, 14]. Further, it has been demonstrated in multiple studies [7, 15] that it is possible to fine-tune the pre-trained (on datasets such as ImageNet) models of these deep network architectures to new classes. This fine-tuning technique offers a powerful way to leverage existing datasets to train new tasks and has delivered very promising result in a wide range of applications from medical imaging [16] to remote sensing [17].

Most of these studies using deep neural networks are conducted on single-view and multiple instance databases [18, 19]. A study by Held et. al. [20] trains these networks with multi-view datasets consisting of multiple instances as well as multiple poses and orientations. This approach was shown to make the network more robust to viewpoint changes and could then be used for single-view instance recognition.

Robustness to viewpoint or poses is extremely critical during three-dimensional object learning in a robotic domain, since the robot is unlikely to encounter the same object from the identical viewpoints during the recognition phase without additional work. In such cases, the assistance of a human teacher can be used to gather multi-view dataset for training the network.

There are a few studies that investigate the challenges associated with learning novel objects in real-world environments with the assistance of human teachers. For example, Azagra et. al. [21] presented an incremental learning framework using human-robot assistance and interaction. The focus of the paper was to demonstrate the advantages of multimodal data using image and language features. The paper also showed an incremental approach using clustering ideas and nearest neighbor approach to learn object models. The system achieved results comparable to off line-trained system, while operating on a much more limited amount of stored data.

In another study, Pasquale et. al. [22] conducted a multi-day trial to test the incremental learning capabilities of CNN using an iCub platform with human assistance. The study focused on a single assistant (a domain expert) and showed preliminary evidence that the classification accuracy of the predictor trained on the mixed dataset from multiple days outperformed that trained on images acquired on a single day. While the focus of these papers is incremental learning and use of multimodal data, our goal in this paper is to explore the impact of providing instructions to the human assistants on training the perceptual system of a robot.

III. Methodology

This section describes the details of the hardware platform and the learning algorithm, experimental design and procedure, and data collection and analysis.

3.1 Hardware Platform

The robotic platform used in this study is Xitome's Mobile Dexterous Social (MDS) robot named Octavia. Octavia has two highly agile arms and a flexible humanoid torso (Fig. 1). Her face can be used to present a wide variety of expressions using gaze, eyelids, jaw and eyebrows. In this study, Octavia uses gaze (at the participant's face and objects in the world), facial expressions and hand gestures to keep the participant engaged in the training procedure. A Kinect 3D V2 sensor is mounted on her upper torso to obtain depth data and RGB Images.

3.1 Learning Algorithm

In this study, we use a Convolutional Neural Network (CNN) that contains eight learned layers. The Initial layers of the CNN architecture consist of convolutional, local response normalization, and max pooling layers (Fig. 2). The top layers of the network are three fully connected layers 'FC6', 'FC7', 'FC8'. The algorithm is implemented and trained using Caffe, a fully open-source software framework that offers clear access to deep architectures and is deployed on NVIDIA Titan X and 1070 GPUs. We refer the reader to the original papers for a more detailed description of the algorithm and framework [23]. We use the standard AlexNet deep CNN model available in the Caffe repository [24]. This model is trained on ILSVRC 2012 dataset with over a million images annotated with 1000 ImageNet classes. Since the numbers of training images are limited in a robotic domain, fine-tuning methodology [7] is used in this study. The AlexNet model is fine-tuned with experimental data by re-training the final fully connected layer using the Stochastic Gradient Descent (SGD) method with adaptive subgradient. The overall learning rate (base_lr) is set to 0.001 and the step-size to 20000. The lr_mult on the final layer is boosted so that the new layer learns faster than the rest of the model (10 for final layer and 1 for all the other layers). The architecture was trained for 100K iterations on mini-batches of size 50 samples.

Figure 1: Experimental environment for the human subject study. A participant teaching a handheld object to the humanoid robot Octavia.

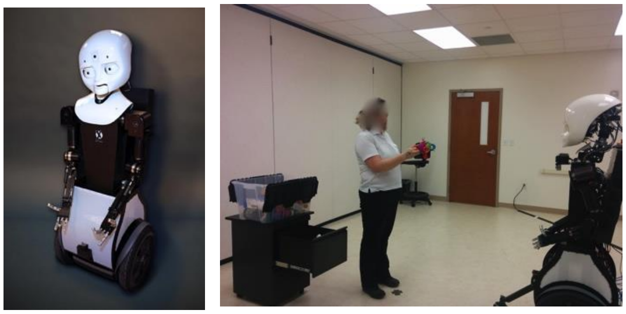
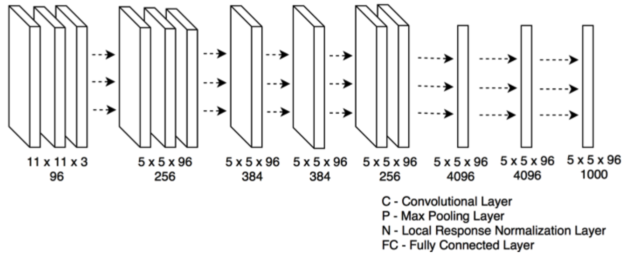


Figure 2: Schematic of the AlexNet Convolutional Network model.



3.3. Experimental Procedure

Forty English-speaking participants were recruited from Naval Research Laboratory (NRL) and randomly assigned to one of the two groups. The study was approved by the local Institutional Review Board (IRB) and written informed consent was obtained from each participant after a full explanation of the risks and benefits involved in the study. The participants were then introduced to Octavia.

The participants were first given a brief overview of the robotic platform, sensors and interfaces and the guidelines to be followed during the experiment (Fig. 1). Octavia then greeted the subjects and introduced herself. She provided instructions to the participants regarding the training procedure. Participants were asked to hold each of the 40 objects provided to them and train the robot, verbally naming (labeling) each object (see supplemental video). The participant then went through a practice session to familiarize themselves with the interactions with Octavia and the training procedure. Throughout the interaction, the participant initiated and ended the data collection procedure (i.e., the participant was able to control how much time was spent on each object). At the end of the experiment, a debriefing session provided the details and overall objective of the study.

3.4. Experimental Design

The primary goal of this study is to investigate how effectively a participant who was given embodied instructions, trains a robot when compared to a naïve participant who trains intuitively. We designed and conducted a between-group experiment where the participants are divided into two groups- *Embodied* and *Naive*.

For participants in the naïve condition, Octavia provided verbal directions on how she needs to be trained:

'You will pick up each object one at a time, tell me what it is, and show it to me.'

For participants in the embodied condition, Octavia provided instructions using her body as an example on how she needs to be trained:

'You will pick up each object one at a time, tell me what it is, and show it to me in various orientations [robot raises a fist and rotates it to couple poses].'

Participants were asked to select objects randomly from a bin that had a variety of objects (described below). Apart from the instructions, all participants underwent identical experimental procedures. Note that the differences in instructions were quite similar in wording and neither instruction set was emphasized. This was a between-subjects manipulation.

3.5 Data Collection.

Training data was simultaneously collected using a Kinect V2, GOPro mounted directly below Kinect and a Sony DCR-VX2100 HD Camera. The HD camera recorded the video data of the interaction between the subject and the robot during the entire duration of the training session.

Since one of the end goals of this project is to teach the robot to recognize objects on tabletops in an indoor environment, a testing dataset was created using a turntable at the end of all the training sessions. Each of the 40 objects was placed on a turntable and images were captured every 5 degrees using the Kinect V2. The objects were then isolated in the images by cropping manually.

3.6 Real-time Segmentation

In this study, we developed an automatic real-time object segmentation framework, which is an improved approach from prior studies employing manual object segmentation [22, 21]. The 3D data from the Kinect sensor was used for segmentation of the object using a series of steps to extract the object candidate of interest from the background. A combination of depth and color based method was used for segmentation and isolating the object from the background. First, the depth data from the Kinect sensor was used to detect the closest point to the sensor (closest-point criteria), which either should correspond to the subject's hand or the training object. A preliminary Region of Interest (ROI) of 600X600 was defined centering this point. Since in this study, the main source of occlusion was predicted to be the participant's hand, the skin color areas in the ROI were detected and eliminated. This was achieved by the histogram back projection method using the skin color on the facial region that was detected by Viola-Jones Haar feature-based cascade classifiers [25]. Further, pixels that had depth value greater than 100 were eliminated. The largest contour was then obtained and a bounding box drawn to isolate the object and the images were then saved. Real-time segmented images were obtained at approximately 14 images per second using an Intel NUC6i7KYK.

3.7. Data Analysis

Analysis was primarily performed on a subset of the data that was created after visual manual inspection of the segmented images where the images with unacceptable segmentation results were removed. If the object in the image was occluded (typically by the hand) or out of the frame, the image was excluded; all other images were included in for analysis. Analysis was also conducted on the original dataset to determine the difference in the recognition results with and without the bias introduced by imperfect segmentation. This original dataset contained all segmented images except that of one participant with an incomplete dataset due to a hardware malfunction. All images were manually labeled in post-processing to account for the language variability between participants.

Two datasets were created for testing / evaluation. The turntable dataset had a controlled range of poses for each object and provides a standard testing set that can be used for each participant's model. The individual-based dataset used a leave-out-one-subject approach where each person's fine-tuned model was tested on training data from the remaining subjects. The classification accuracy was computed as the ratio of the total number of images correctly recognized to the total number of images used in testing.

The individual based dataset allowed us to examine how well a single individual's training mapped to other participant's data. Note that the individual-based dataset is especially difficult because of both individual variability (e.g., some participants spent longer or shorter on each object) as well as condition differences (e.g., participants in the naïve condition may have emphasized the canonical form of the object more than participants in the embodied condition).

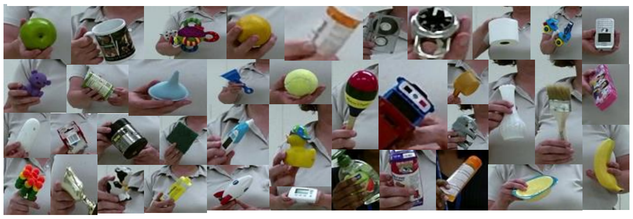
The AlexNet model was fine-tuned for each participant. Each fine-tuned model was then used to classify images in both testing datasets. The statistical significance of the results was computed using the Welch two sample t-test.

IV. Results

A. Segmentation Results

For most participants, the segmentation approach was successful in isolating the objects and produced good segmentation results (Fig. 3). Best performance was achieved for spherical objects such as apple, ball and orange. In these cases, the regions of interests (the object) were fully detected and occupied a large portion of the cropped image. However, the segmentation results were not always ideal for smaller objects such as the combination lock, and skin-colored objects such as wooden spoon, measuring cup etc. In addition, the skin tone of the participants also had an effect on accurate segmentation results of these skin-colored objects. In both these cases, the vision system detected incomplete or partial interest regions due to imperfect segmentation. In certain cases, the sensor did not detect the object or the hand as the closest points (closest- point criteria) leading to imperfect segmentation. These failures highlight the difficulties of purely automatic object training and recognition by non-experts.

Figure 3: Examples of segmented images obtained from the human subject study.



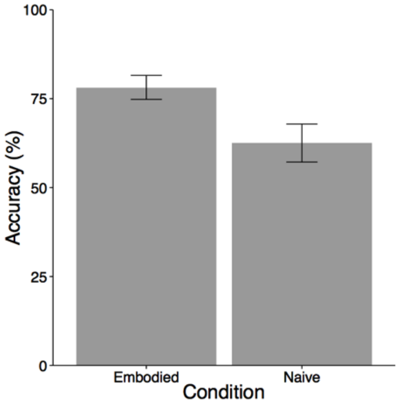
These images with imperfect segmentations were eliminated in the hand-tuned dataset. It was observed that the majority of excluded images had segmentation failures due to the closest point criteria or presence of skin-color objects. Dataset of five participants (*Embodied* - 4, *Naive* - 1) were removed completely since they had more than 20% images with imperfect segmentations. In the remaining 35 participants *Embodied*-16, *Naive* – 19) the amount of images removed ranged from 0.01%-16.3 %.

B. Object Recognition Result

AlexNet was fine-tuned individually for each of the participants. The number of training images ranged from 752 to 15586 images per participant with a total of 164952 images from 35 participants.

The fine-tuned models were first tested on the turntable dataset (Figure 4). This dataset consisted of 2890 images with an average of 72 images per object. Models of participants in the *Embodied* condition (μ = 78.1 %, σ = 7.7 %) were more accurate than the models of participants in the *Naïve* condition (μ = 62.6 %, σ = 13.1 %), t(29.7) = 4.3, p < 0.01. Our findings were consistent with the hypothesis that naïve instructions are likely to be insufficient to meet system needs and lead to less robust models.

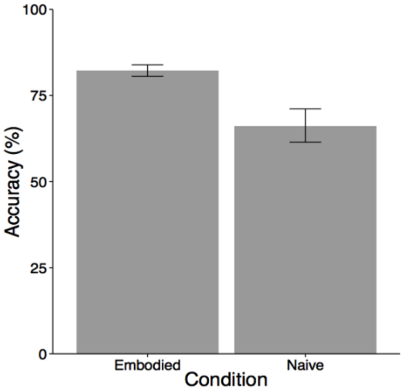
Figure 4: Classification results for the turntable dataset. The error bars show the 95% confidence interval (CI).



Next, the fine-tuned models were tested on the individual dataset (Figure 5). Models of participants in the *Embodied* condition (μ = 82.2 %, σ = 3.7 %) were more accurate than the models of participants in the *Naïve* condition (μ = 66.1 %, σ = 11.3 %), t(22.5) = 5.7, p < 0.01. Our findings were consistent with the hypothesis that naïve instructions are likely to be insufficient to meet system needs and would produce less robust models.

We also examined performance on the original dataset (no hand removal of images). Models of participants in the *Embodied* condition (μ = 73.6 %, σ = 4.3 %) were more accurate than the models of participants in the *Naïve* condition (μ = 58.8 %, σ = 10.3 %), t(25.6) = 5.9, p < 0.01. Hence, while hand coding the data to remove images with imperfect segmentation was helpful in improving the accuracy, results from both datasets showed that embodied instruction resulted in better models than naïve instruction.

Figure 5: Classification results for the individual dataset. The error bars show the 95% confidence interval (CI).



V. DISCUSSION

In this paper, we investigate the impact of embodied training on object recognition through a human subject study conducted using a humanoid robot. The participants were broadly classified into two groups: *Embodied* and *Naïve*. Each participant was either given embodied examples or verbal instructions for teaching novel objects to the robot. The collected image data was then used for training a Convolution Neural Networks. An integrated software and hardware framework was created for data acquisition and real-time object segmentation, which represents system advancement over previous studies. The results of the study indicate that embodied training has a significant impact on the classification performance of the learning algorithm. Generally, the network trained by participants with embodied instructions outperformed a system trained by naïve participants.

Our results clearly show that embodied training is an excellent way to provide instructions to teachers during a training scenario to improve vision performance of robotic systems. There are several possible reasons why the models created by data from participants in the embodied training condition performed better than models created by data from participants in the naïve condition. Our primary hypothesis is that participants who were given naive instructions showed the canonical view of objects, while participants who received embodied training showed a variety of different viewpoints, including the canonical one. Future work will examine the type of training that actually occurred and how it impacted network learning as well as improve the object segmentation algorithm, and analyze the language variability of object labeling to automate the labeling process.

The study also shows that the fine-tuning methodology is a strong candidate for object recognition in a robotic domain. The standard AlexNet model combined with real-time segmentation provided impressive classification accuracy in the initial results. There is great potential for improving the results by various data augmentation techniques and improving and customizing the deep learning models.

Our findings confirm the fact that the current learning algorithms are reliant on the detailed availability of information about the objects being learned. As a result, the specificity of instructions has a large influence in the learning performance. Thus, the use of the latest machine learning algorithms combined with good embodied training provides an efficient object recognition approach in a robotic domain in real-world environments.

REFERENCES

1. S. Najarian, M. Fallahnezhad, and E. Afshari, “Advances in medical robotic systems with specific applications in surgery - a review,” *Journal of Medical Engineering & Technology*, vol. 35, no. 1, pp. 19– 33, 2011.

2. J. Bohren, R. B. Rusu, E. G. Jones, E. Marder-Eppstein, C. Pantofaru, M. Wise, L. Mosenlechner, W. Meeussen, and S. Holzer, “Towards autonomous robotic butlers: Lessons learned with the PR2,” in *IEEE International Conference on Robotics and Automation, ICRA*, pp. 5568-5575. May 2011.

3. H. Iwata and S. Sugano, “Design of human symbiotic robot TWENDY-ONE,” in *IEEE International Conference on Robotics and Automation, ICRA*, pp. 580-586. May 2009.

4. B. Gates, “A robot in every home,” *Scientific American*, *296*(1), pp.58-65. 2007.

5. Y. Xiang, R. Mottaghi and S. Savarese, “Beyond pascal: A benchmark for 3d object detection in the wild,” in *IEEE Winter Conference on Applications of Computer Vision, WACV,* pp. 75-82. March 2014.

6. 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 (* = equal contribution), “ ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision, IJCV.* 2015.

7. S. Karayev, M. Trentacoste, H. Han, A. Agarwala, T. Darrell, A. Hertzmann and H. Winnemoeller, “ Recognizing image style,” *arXiv preprint arXiv:1311.3715*. 2013

8. C. Breazeal and A. L. Thomaz, “Learning from human teachers with socially guided exploration,” in *IEEE International Conference on Robotics and Automation, ICRA,* pp. 3539-3544, May 2008.

9. D. Lowe. “Object Recognition from Local Scale-Invariant Features,” in the proceedings of the seventh *International Conference on Computer Vision (ICCV),* Vol. 2, pp. 1150-1157, 1999.

10. N. Dalal and B. Triggs. “Histogram of Oriented Gradients for Human Detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 1, pp. 886-893, 2005.

11. C. Gao, N. Sang, J. Gao, L.Zou and Q. Tang. “Generic Object Recognition with Biologically-inspired Features,” Fourth *International Conference on Bio-Inspired Computing*, pp. 1-7, 2009.

12. R.C. O’Reilly, and Y. Munakata. “ Computational Exploration in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain,” The MIT Press, Massachusetts. 2000.

13. A. Krizhevsky, I. Sutskever, and G. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks,” in *Advances in neural information processing systems*, pp. 1097-1105, 2012.

14. M. Rastegar, V. Ordonez, J. Redmon and A. Farhadi, “Xnor-net: ImageNet classification using binary convolutional neural networks,” in *European Conference on Computer Vision*, pp. 525-542, October 2016.

15. R. Girshick, J. Donahue, T. Darrell and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR,* (pp. 580-587). 2014.

16. N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway and J. Liang, “ Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?” *IEEE Transactions on Medical Imaging*, *35*(5), pp.1299-1312. 2016

17. L. Zhang, L. Zhang and B. Du, “Deep learning for remote sensing data: A technical tutorial on the state of the art”, *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), pp.22-40. 2016

18. R. Girshick, "Fast r-cnn," *Proceedings of the IEEE International Conference on Computer Vision*, ICCV, pp. 1440-1448, 2015.

19. A. S. 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, pp. 806-813, 2014.

20. D. Held, S. Thrun, and S. Savarese, “Deep Learning for Single-View Instance Recognition,” *arXiv preprint arXiv:1507.08286*, 2015.

21. P. Azagra, Y. Mollard, F. Golemo, A. Murillo, M. Lopes, and J. Civera. "A Multimodal Dataset for Interactive and Incremental Learning of Object Models," in *IEEE International Conference on Robotics and Automation, ICRA*, pp. 5568-5575, 2017.

22. G. Pasquale, C. Ciliberto, F. Odone, L. Rosasco, L. Natale, and I. dei Sistemi, “Teaching iCub to recognize objects using deep Convolutional Neural Networks,” proceedings of the 4th *Workshop on Machine Learning for Interactive Systems*, pp. 21-25, 2015

23. A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” Advances in *Neural Information Processing Systems*, NIPS, pp. 1097-1105, 2012.

24. Y. Jia, E. Shelhamer, J. Donahue, S. Karayey, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional Architecture for Fast Feature Embedding,” Proceedings of the *ACM International Conference on Multimedia,* ACM, pp. 675-678, 2014.

25. P. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," Proceedings of the 2001 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR, pp.511–518, 2001.