Human Perceptions of Task Load and Trust when Interactively Teaching a Continual Learning Robot

Ali Ayub[1] Zachary De Francesco[1] Patrick Holthaus[2] Chrystopher L. Nehaniv[1] Kerstin Dautenhahn[1]
[1]University of Waterloo [2]University of Hertfordshire
[1]{a9ayub, zdefrancesco, cnehaniv, kdautenh}@uwaterloo.ca [2]p.holthaus@herts.ac.uk

Abstract

Although machine learning models for continual learning (CL) can mitigate forgetting on static, systematically collected datasets, it is unclear how human users might perceive a robot that continually learns over multiple interactions with them. In this paper, we developed a system that integrates CL models for object recognition with a mobile manipulator robot and allows humans to directly teach and test the robot over multiple sessions. We conducted an in-person between-subject study with two CL models and 40 participants that interacted with our system in 200 sessions (5 sessions per participant). Our results indicate that state-of-the-art CL models might perform unreliably when applied on robots interacting with human participants. Our results also suggest that participants' trust in a continual learning robot significantly decreases over multiple sessions if the robot forgets previously learned objects. However, the perceived task load on participants for teaching and testing the robot remains low for all sessions, indicating the feasibility of continual learning robots in the real world. Our code is available at https://github.com/aliayub7/cl_hri.

1. Introduction

To operate in daily environments, a general task for a robot is to learn and understand the objects in its environment [1, 5, 8, 30, 31]. Various machine learning models have been developed in the last decade for achieving remarkable performance on object recognition tasks [12,26]. However, one of the main challenges faced by robots using ML models to continually learn objects is *catastrophic forgetting* [9, 19]. Catastrophic forgetting occurs when a continual learning (CL) agent forgets the previously learned knowledge when learning new information [22] (note that the degree of forgetting may turn out to be far from what users find "catastrophic".). In recent years, various research directions (some inspired by neuroscience [14, 15]) have been taken in the field of continual learning to mitigate the catastrophic forgetting problem [6, 11, 17, 20, 21, 27]. While SOTA CL models alleviate catastrophic forgetting, they still suffer from some forgetting when learning over a large number of repeated sessions [2, 15, 16].

Another challenge faced by continual learning robots is that their users might not provide a sufficiently large amount of data to train an ML model. In the past few years, robotics researchers developed CL models that can learn continually from only a few training examples per object, while also mitigating catastrophic forgetting [4, 29]. This problem is known as Few-Shot Incremental Learning (FSIL) [4, 28, 29, 34]. Although FSIL approaches have produced promising results on systematically collected "non-social" datasets by the experimenters, it is unknown how these systems might perform when learning from human participants. It is also unknown how people might perceive robots that continually learn through interaction with their users. To the best of our knowledge, we know of no other work on testing CL or FSIL models deployed on robots that learn from human users (unlike experts who are familiar with programming the particular robot and have in-depth knowledge of the CL models) over multiple interactions.

In this paper, we consider a system for socially guided continual learning (SGCL) and conduct an in-person user study to explore how people perceive a robot that continually learns common household objects over multiple interactions. Our system integrates a graphical user interface (GUI) with a CL model deployed on the Fetch mobile manipulator robot [33]. In this system, we focused solely on the continual learning of objects and avoided adding any extra social cues to the robot that might affect human perceptions of the robot. We performed a long-term between-subject user study (N=40) where participants interacted for 5 sessions with a fully autonomous Fetch robot that used two different CL models: one that suffers from catastrophic forgetting on static datasets, and another state-of-the-art (SOTA) approach for FSIL that mitigates forgetting. We used two questionnaires in the study to answer the following research questions:

RQ1 How does forgetting affect humans' trust in a continual learning robot?

RQ2 How does humans' trust change when interacting with a continual learning robot over multiple sessions?

RQ3 Do people consider a continual learning robot easy to use?

2. Socially Guided Continual Learning

We studied human perceptions of a continual learning robot in the context of an object recognition task. In this setup, the robot learns household objects from the user (in multiple sessions) on a table-top environment, and then finds and points to the requested object on the table after learning them from the user. Figure 1 (left) shows the table-top experimental setup for this study. The simplicity of the setup and the task makes it clear what the user should do to teach the robot different objects and what the robot should do to find the learned objects during the testing phase.

For this setup, we consider a socially guided continual learning (SGCL) system for the object recognition task, which integrates CL models with the robot for interactive and transparent learning from human users. Figure 1 (right) shows the SGCL system for the object recognition task. In this system, in each session (or increment) t the user interacts with the robot through a graphical user interface (GUI) to teach the robot L_t number of objects. The robot captures images of the L_t objects and pre-process them, and gets the labels of the processed object images from the user to generate a dataset $D^t = \{x_i^t, y_i^t\}_{i=1}^{|D^t|}$, where x_i^t is the ith image in the dataset with the class label y_i^t. The CL model $\mathcal{M}$ then trains on the dataset D^t. Note that unlike static CL setups (such as FSIL [4]), the number of objects per object class in a session is not fixed as it is dependent on the number of times the user teaches an object to the robot. Further, there can be an overlap in the object classes taught in different sessions depending on how the user labels the objects.

In the testing phase, the robot receives the request from the user through the GUI to find an object. The robot passes the pre-processed images to the CL model to get the predicted object labels. If the object is found, the robot finds the 3D location of the object on the table and points to the object using its arm. Note that the user has flexibility in terms of the total number of objects to be tested in an increment, as well as which objects to test (old or new objects).

2.1. Continual Learning Models

The main goal of our study is to do an in-depth analysis of how users perceive CL models over repeated, long-term interactions. To do such an analysis, it is important to choose a meaningful baseline. The naive finetuning (FT) approach [22] has been used extensively in CL literature as a baseline on static datasets. Therefore, we chose to test FT as our study's baseline model. FT approach uses a convolutional neural network (CNN) [12] that is trained on the image data of the object classes in each increment. The model does not train on any of the objects learned in the previous increments (sessions) and therefore it catastrophically forgets the previously learned objects. More details about this model can be found in [22].

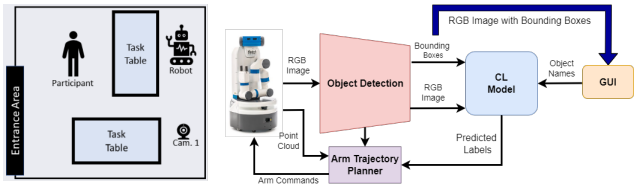


Figure 1. (Left) Experimental layout for the SGCL setup with the participant and the robot. (Right) Our complete SGCL system. Processed RGB images from Fetch's camera are sent to the GUI for transparency and also passed on to the CL Model. The user sends object names to the CL model either for training the CL model or finding an object. The arm trajectory planner takes point cloud data, processed RGB data, and predicted object labels from the CL model as input and sends the arm trajectory for the Fetch robot to point to the object.

For the second model, we consider a SOTA CL approach designed for robotics applications [4, 5]. This approach, termed centroid-based concept learning (CBCL) [3], uses a CNN pre-trained on the ImageNet dataset [25] as a feature extractor for object images. In each increment t, CBCL clusters the feature vectors of all the object classes in the increment and generates a set of centroids $C^y = \{c_1^y, ..., c_{n_y}^y\}$ for each object class separately, where n_y is the total number of centroids for class y. CBCL avoids forgetting by generating separate centroids for each class in a new increment t. For the classification of a new object, CBCL uses a weighted voting scheme to find the most common class (prediction for the test object) among the closest centroids to the test feature vector. More details about CBCL can be found in [4]. CBCL has been shown to produce promising results when learning from systematically collected object datasets by experts. However, it was never trained or tested in real-time with human participants. In this paper, we integrate both CBCL and FT in a fully-autonomous system that allows users to interact with these models in real-time through the Fetch mobile manipulator robot.

3. Method

To answer the three research questions, we tested the following hypotheses, related to those research questions, in a repeated measures study where users interacted over five sessions with the system (Section 2):

H1.1 A forgetful robot is perceived as less trustworthy than a

robot that remembers most previously learned objects.

H1.2 Task load for teaching and testing a forgetful robot is less than a robot that remembers most previous objects.

H2.1 Users' trust decreases in the robot over multiple sessions regardless of the CL model.

H2.2 The task load for teaching the robot increases over multiple sessions.

H3.1 The overall task load for teaching the robot is minimal.

Where, H$n.m$ is the mth hypothesis related to the research question n, e.g. H1.2 is the second hypothesis for RQ1.

3.1. Participants

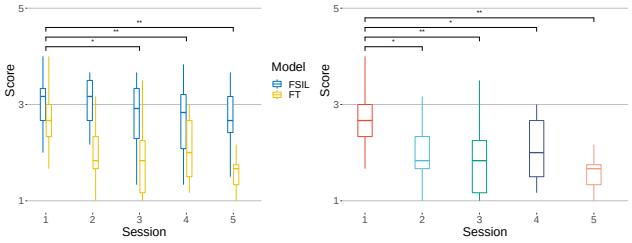
We recruited 40 participants (19 female (F); 21 male (M), all students) from the University of Waterloo, between the ages of 18 and 37 years ($M = 23.48$, $SD = 4.49$). 20 participants (ages: $M = 24.15$, $SD = 4.21$, 10 F, 10 M) were randomly assigned to the *FT* condition, and the other 20 (ages: $M = 22.78$, $SD = 4.68$, 9 F, 11 M) were randomly assigned to the *FSIL* condition. The participants had diverse backgrounds in terms of their majors, but most of them were engineering and computer science students. Based on their self-assessments in a pre-experiment survey, 40% of the participants reported that they were familiar with robot programming, 55% reported that they had previously interacted with a robot, 5% were familiar with the Fetch robot, and 10% had previously participated in an HRI study. All procedures were approved by the University of Waterloo Human Research Ethics Board.

3.2. Measures

To verify the hypotheses and thus evaluate the different learning models, we applied the following measures in both experimental conditions.

Subjective Measures. After each trial, we asked participants to fill in the following questionnaire scales as subjective measurements aimed to test the hypotheses. We measured people's trust in the robot using the cognition-based trust subscale of Madsen's *Human-Computer Trust (HCT)* questionnaire [18] to address **H1.1** and **H2.1**. The scale contains six individual questions that can be rated on a 5-point Likert scale, ranging from "Strongly disagree" to "Strongly agree". Additionally, we used the *Nasa-Task Load Index (NASA-TLX)* [10] to estimate participants' mental workload to gain insights about **H1.2**, **H2.2**, and **H3.1**. *TLX* is comprised of six questions that participants rate on a 21-point scale, ranging from "Very low" to "Very high", resulting in a single factor.

Objective Measures. We also used an objective measure to analyze the performance of the two CL approaches. Classification accuracy per session (increment) has been



(a) Overall scores by session (b) Scores by session in *FT*

Figure 2. Boxplots for *cognition based trust* scores on the *HCT* scale, ranging from 1 to 5. Significance levels ($* := p < .05$; $** := p < 0.01$) are indicated on bars between the columns.

commonly used in the CL literature [22, 29] for quantifying the performance of CL models for object recognition tasks. Therefore, for each session, during the testing phase, we recorded the total number of objects tested by the participant and the total number of objects that were correctly found by the robot. Using this data, we calculated the accuracy $\mathcal{A}$ of the robot in each session as: $\mathcal{A} = \frac{number\ of\ objects\ correctly\ found}{number\ of\ objects\ tested}$

4. Results

We evaluated all questionnaire scales using a Wilcoxon rank sum test [32] comparing the scores between the two models and five sessions, respectively. For the remainder of the paper, we term the finetuning model as *FT*, and the FSIL model CBCL, as *FSIL*.

4.1. Cognition based trust

Scores for *cognition-based trust* on *HCT* are calculated as mean values of six individual items with a minimum value of 1 and a maximum value of 5, resulting in an overall value of $\mu = 2.37, \sigma = 0.92$. Figure 2a details how this score differs between the subsequent experimental sessions. In particular, as displayed in Figure 2b, trust decreases significantly only in the *FT* condition when comparing the first session with any of the subsequent sessionsWhen only considering the *FSIL* condition, no significant differences in scores can be observed between any of the sessions.

Moreover, *cognition based trust* scores are significantly different between the *FSIL* condition ($\mu = 2.83, \sigma = 0.639$) and *FT* condition ($\mu = 1.865, \sigma = 0.91$) when looking at all sessions combined, ($p < 0.0001, W = 1698$), and consistently across all five sessions (Figure 2a).

4.2. Task load index

Simplified scores for NASA task load index (TLX) are calculated as average values of six individual items (21-point scale, which is then translated into a score that ranges from 0 to 100). No significant differences were seen between the two conditions overall (*FSIL* condition: $\mu = 26.38, \sigma =$

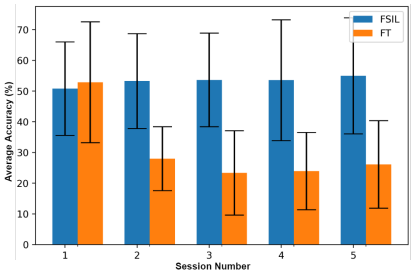


Figure 3. Average classification accuracy of the two CL models over 5 sessions.

13.32, *FT* condition: $\mu = 28.14, \sigma = 13.38$) or between any of the five sessions.

4.3. Classification Accuracy

Figure 3 shows the classification accuracy of the two models averaged over all the participants per model. The classification accuracy of both models is similar in the first two sessions (μ =∼50%). However, for the second session, FT's accuracy significantly decreased (μ =∼ 27%) and stayed consistent for the four sessions. FSIL's accuracy remained similar in all five sessions. Huge variations were seen in classification accuracy for both models in all five sessions. This variation was because of the differences in the accuracy of the models for different participants.

5. Discussion

Results obtained in the repeated measures experiment with the interactive system allow us to validate the hypotheses introduced in Section 3 and conclusions to be drawn with regards to the research questions in Section 1.

In comparison to other studies [24], overall *cognition-based trust* is rated at mediocre levels only. Such a result is within our expectations since in CL approaches (including FSIL) forgetting plays an important role and the cognitive function of the system is therefore not reliably identifiable by the user. Moreover, the imperfect nature of object teaching might have influenced the user's impression of the system, because even the FSIL approach achieved only $\sim 50\%$ classification accuracy in all sessions. Considering the two conditions, trust towards the system drops in the *FT* condition as opposed to the *FSIL* condition, where it remains on similar levels. This indicates that people, over time, lose trust in a model that forgets learned objects but they keep a similar amount of trust if it remembers previous objects. As a consequence, we can accept **H1.1** but we only find evidence for **H2.1** in the *FT* condition. Hence **H2.1** can only be confirmed partially. This result is supported by the experiment's objective measures since trust seems to correlate with the classification performance of both models. The classification accuracy for *FT* condition decreased because of forgetting and so did the trust. For *FSIL* condition, both the trust and the accuracy stayed similar. Further, note that although it might be expected that trust in an imperfect robot (FT condition) would drop, Chi et al. [7] showed that trust towards an imperfect robot evolves over multiple interactive sessions, if participants are involved in directly teaching the robot. Their study, however, was not conducted with an embodied agent or with continual learning models. Therefore, it was imperative for us to conduct the study with both CL models integrated with a fully autonomous robot, to understand participants' trust towards continual learning robots that might forget previous objects.

With an overall low *task load index*, **H3.1** can be approved firmly. Both models had similarly low task load ratings, which is expected for *FT* condition as the model is simple and it continues to forget previous objects. However, even for more complex models that mitigate forgetting, participants' workload did not increase. Neither the accuracy of the model nor any subsequent iterations have an effect on the task load and hence **H2.2** has to be rejected. Similarly, **H1.2** has to be rejected since we cannot find evidence that would support any difference between the conditions with regard to task load. There is no correlation between the task load and the model's performance. However, task load seems to be linked with the total number of images shown per object, as participants for both models showed only a few images per object. These results are quite promising as they indicate the feasibility of personalized continual learning robots that learn from the users. The results also suggest that researchers might need to focus more on the task (and task load) than the choice of the model alone when developing continual learning robots.

6. Conclusions

In this work, we designed a novel user study to understand human perceptions of a continual learning robot while teaching and testing the robot over five sessions. We conducted a between-subject study with two CL models and asked participants about their perceptions of the robot in terms of trust and task load of the system, after directly teaching and testing the robot over five sessions. Our results indicate that users' perceptions of trust are negatively affected by forgetting of the CL models. Our results also indicate that the performance of even the SOTA CL models is unreliable (only ∼50% accuracy) when learning from users instead of learning on static datasets. Therefore, with the current SOTA CL models, continual learning robots are not perceived to be very trustworthy by their users. However, the task load for teaching and testing the robot stayed low and was not affected by the choice of the CL model. Our results indicate that future CL research should also focus on the task load and the needs and tendencies of the users when designing CL models that learn through human interactions.

References

[1] Ali Ayub and Carter Fendley. Few-shot continual active learning by a robot. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1

[2] Ali Ayub and Alan Wagner. Eec: Learning to encode and regenerate images for continual learning. In *International Conference on Learning Representations*, 2021. 1

[3] Ali Ayub and Alan R. Wagner. Centroid based concept learning for RGB-D indoor scene classification. *British Machine Vision Conference (BMVC)*, 2020. 2

[4] Ali Ayub and Alan R. Wagner. Cognitively-inspired model for incremental learning using a few examples. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020. 1, 2

[5] Ali Ayub and Alan R. Wagner. Tell me what this is: Few-shot incremental object learning by a robot. *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020. 1, 2

[6] Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *The European Conference on Computer Vision (ECCV)*, September 2018. 1

[7] Vivienne Bihe Chi and Bertram F. Malle. People dynamically update trust when interactively teaching robots. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, page 554–564, New York, NY, USA, 2023. Association for Computing Machinery. 4

[8] M. Dehghan, Z. Zhang, M. Siam, J. Jin, L. Petrich, and M. Jagersand. Online object and task learning via human robot interaction. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2132–2138, May 2019. 1

[9] Robert M. French. Dynamically constraining connectionist networks to produce distributed, orthogonal representations to reduce catastrophic interference. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, pages 335–340, 2019. 1

[10] Sandra G. Hart. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, Oct. 2006. 3

[11] Tyler L. Hayes, Kushal Kafle, Robik Shrestha, Manoj Acharya, and Christopher Kanan. Remind your neural network to prevent catastrophic forgetting. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 466–483, Cham, 2020. Springer International Publishing. 1

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1, 2

[13] Jan Hosang, Rodrigo Benenson, and Bernt Schiele. Learning non-maximum suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[14] Ronald Kemker and Christopher Kanan. Fearnet: Brain-inspired model for incremental learning. In *International Conference on Learning Representations*, 2018. 1

[15] James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114(13):3521–3526, 2017. 1

[16] Z. Li and D. Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, Dec 2018. 1

[17] Vincenzo Lomonaco and Davide Maltoni. Core50: a new dataset and benchmark for continuous object recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, volume 78, pages 17–26, 2017. 1

[18] Maria Madsen and Shirley Gregor. Measuring human-computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*, pages 6–8, 2000. 3

[19] James L. Mcclelland, Bruce L. Mcnaughton, and Randall C. Oreilly. Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457, 1995. 1

[20] Martin Mundt, Yong Won Hong, Iuliia Pliushch, and Visvanathan Ramesh. A wholistic view of continual learning with deep neural networks: Forgotten lessons and the bridge to active and open world learning. *CoRR*, abs/2009.01797, 2020. 1

[21] Martin Mundt, Steven Lang, Quentin Delfosse, and Kristian Kersting. CLEVA-compass: A continual learning evaluation assessment compass to promote research transparency and comparability. In *International Conference on Learning Representations*, 2022. 1

[22] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 1, 2, 3

[23] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[24] Paul Robinette, Ayanna M. Howard, and Alan R. Wagner. Effect of robot performance on human–robot trust in time-critical situations. *IEEE Transactions on Human-Machine Systems*, 47(4):425–436, 2017. 4

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252, Dec. 2015. 2

[26] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural In-*

formation Processing Systems - Volume 1, NIPS'14, pages 568–576, Cambridge, MA, USA, 2014. MIT Press. 1

[27] James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9374–9384, October 2021. 1

[28] Xiaoyu Tao, Xinyuan Chang, Xiaopeng Hong, Xing Wei, and Yihong Gong. Topology-preserving class-incremental learning. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX*, page 254–270. Springer-Verlag, 2020. 1

[29] Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[30] Andrea L. Thomaz and Maya Cakmak. Learning about objects with human teachers. In *2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 15–22, 2009. 1

[31] Sepehr Valipour, Camilo Perez Quintero, and Martin Jägersand. Incremental learning for robot perception through hri. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2772–2777, 2017. 1

[32] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. 3

[33] Melonee Wise, Michael Ferguson, Derek King, Eric Diehr, and David Dymesich. Fetch and freight: Standard platforms for service robot applications. In *IJCAI, Workshop on Autonomous Mobile Service Robots*, 2016. 1, 7

[34] Chi Zhang, Nan Song, Guosheng Lin, Yun Zheng, Pan Pan, and Yinghui Xu. Few-shot incremental learning with continually evolved classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12455–12464, June 2021. 1

A. Fetch Mobile Manipulator Robot

Manipulator robots with an RGB-D camera are well-suited for recognizing and manipulating objects. In our setup, we use the Fetch mobile manipulator robot [33]. Fetch consists of a mobile base and a 7 DOF arm. The robot also contains an RGB camera, a depth sensor, and a Lidar sensor. These sensors can be used for 3D perception, slam mapping, and obstacle detection in the robot's environment. In our setup, we do not ask the robot to manipulate objects or move its base, allowing us to solely focus on continual learning which is principally about learning and recognizing objects. We mainly use the RGB-D camera to recognize objects and the 7 DOF arm to point to objects. We use ROS packages available with the Fetch robot for moving the torso, and the arm of the robot. We did a safety analysis of the robot (approved by our University's ethics review board) and also adopted several mitigating strategies. Therefore, the robot was considered safe to be used with human participants in our study.

As there can be multiple objects on the table in front of the robot's camera, we process the RGB images further by passing them through a generic object detector [23]. The object detector finds regions in the image that are likely to contain objects (Figure 4). The detected regions are filtered using non-max suppression [13] to remove any overlaps. We also filter out the detected objects that are not on the table (background objects, participant interacting with the robot as seen in Figure 4) using the depth perception of the objects. The resulting regions are cropped into separate images for objects detected on the table and then forwarded to the CL model.

B. Graphical User Interface

For users to be able to interact and teach the robot different objects in an open-ended manner, we created a simple graphical user interface and deployed it on an Android tablet. Figure 4 shows a screenshot of the GUI. The top left side of the GUI shows the pre-processed camera output of the robot which contains bounding boxes for detected objects. The camera output was used as a transparency device so that the participants could clearly understand what the robot was seeing on the table. On the bottom left of the GUI, there is a toggle button that can be used to start a teaching session with the robot. Once the button is pressed, it turns green indicating that the system was in the teaching phase. After starting the teaching phase, participants can type the name (class label) of the objects in the space below the toggle button. Participants can save an image of the object using the save button next to the empty space. The bottom right of the GUI contains another toggle button that can be used by the participants to start the testing phase. The button turns green once pressed. During the testing phase,

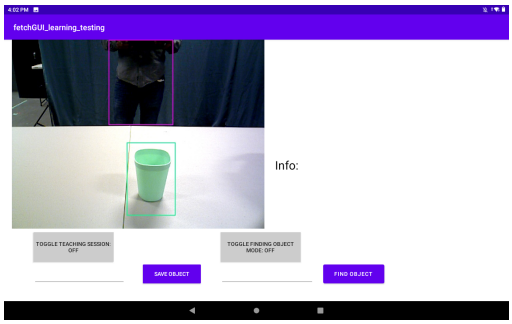


Figure 4. The graphical user interface (GUI) used to interact with the robot. The RGB camera output with bounding boxes is on the top left. The buttons at the bottom can be used to teach objects to the robot and ask it to find objects in the testing phase. The top right of the GUI shows information sent by the robot to the user.

participants can type the name (class label) of the object to be found on the table in the space below the testing toggle button. Participants can then press the Find Object button next to the empty space to ask the robot to find the requested object on the table. Finally, the top right section of the GUI shows the messages communicated by the robot to the user during the session. The robot also spoke these messages using a text-to-speech module available in ROS.

C. Overall Results of the Study

Tables 1 shows the overall statistics of the study across all 40 participants.

Table 1. Detailed results for the two questionnaires in the two conditions. NS stands for not significant.

Trust						
	FT		FSIL			
Session Value	μ	σ	μ	σ	p	W
1	2.49	0.97	3.02	0.50	0.0353	122
2	1.86	0.77	3.05	0.47	3.2×10^{-5}	34.5
3	1.61	0.84	2.79	0.67	0.0001	50.5
4	1.74	0.93	2.64	0.70	0.0054	72.5
5	1.52	0.71	2.64	0.73	0.0003	49.5
all	1.86	0.91	2.83	0.64	5.3×10^{-13}	1698

TLX						
	FT		FSIL			
Session Value	μ	σ	μ	σ	p	W
1	28.5	9.98	24.4	8.16	NS	NS
2	29.7	11.9	25.6	13.5	NS	NS
3	27.6	15.2	25.2	11.9	NS	NS
4	27.1	14.9	27.8	14.6	NS	NS
5	27.3	16.3	28.9	17.6	NS	NS
all	28.1	13.4	26.4	13.3	NS	NS