2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

A Social Referencing Disambiguation Framework for Domestic Service Robots

Kevin Fan[1*], Melanie Jouaiti[1], Ali Noormohammadi-Asl[2], Chrystopher L. Nehaniv[2,1], Kerstin Dautenhahn[1,2]

***Abstract*— The successful integration of domestic service robots into home environments can bring significant services and convenience to the general population and possibly mitigate important societal issues, such as care provision for older adults. However, home environments are complex, dynamic and object-rich. It is, thus, very probable that service robots will encounter ambiguity while interacting with household items. To enable service robots to be more adaptive, we proposed a learning social referencing computational framework and experimentally evaluated the framework on a mobile manipulator robot, Fetch, in object selection scenarios. The framework allows the robot to (1) detect and analyze the ambiguity level based on the robot's view and user's command, (2) assess the human's attention level and attract their attention, (3) disambiguate references to objects using human feedback and (4) learn novel objects after clarification from the user. System evaluation results are presented. The framework is modular and can be applied to different robotic platforms.**

I. INTRODUCTION

Social referencing is a powerful mechanism that helps humans adaptively change their behavior based on the feedback of their social interaction partners. Particularly in human infants, social referencing is important to aid infants in resolving ambiguity and learning their surroundings [1]. Social referencing has two necessary components: 1- information-seeking, where one makes requests for clarification directed at the social partner, and 2- behavior-regulatory, where one can alter one's behaviors based on the feedback received from others [2]. The successful modeling of social referencing could aid domestic service robots in functioning more effectively and naturally within the human social environment.

Despite the fact that social referencing is often employed as a means of disambiguation and discovery in human infants, current computational models of social referencing in robotics often emphasize the emotional modeling aspect [3], [4], attempting to make the robot associate simulated emotion to objects based on human feedback. Complementary to these approaches, in this paper, we propose a social referencing framework that focuses on the utilitarian purpose of disambiguation in the context of object selection, as the proper manipulation of common objects is critical to the practical application of domestic service robots.

Let us consider the scenario depicted in Figure 1 where the human user instructs the service robot to "pass the salt". The robot then looks for objects on the table and finds multiple containers. However, without a specifically trained object detection model, it cannot understand which one is the salt container. Additionally, the robot is encountering ambiguity due to detecting multiple potential targets. After determining the presence and the degree of ambiguity, the robot needs to seek clarification from the human partner to disambiguate. The robot can then proceed to execute the task, adapting its behaviors based on the human's feedback. Importantly, if the ambiguity stems from the novelty of an object, the robot needs to learn that object's characteristics to avoid ambiguity in the future.

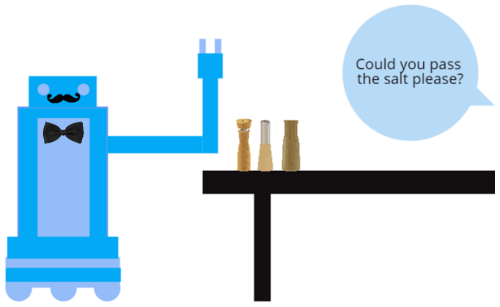

Fig. 1. A domestic service robot encounters ambiguity in an object selection task. The robot can only recognize generic containers, and it does not know which one is the "salt"; disambiguation and learning are required.

The contribution of this work can therefore be summarized as follows: (1) We propose a novel social referencing disambiguation framework that enables a robot to determine and analyze task ambiguity, evaluate human attention, and disambiguate objects through human feedback. Additionally, the framework allows the service robot to learn novel objects in real-time by remembering the novel example. (2) We implemented the proposed framework on a physical robot for system validation.

II. RELATED WORK

Object selection is one of the most common and fundamental tasks that a service robot needs to achieve. The resolution of ambiguity involved in object selection has, therefore, been extensively explored, by exploiting various human signals, such as gestures or speech.

Gesture is one of the most commonly used methods for disambiguation [5], [6], [7]. Early work, such as [5] utilized the visual appearance of objects (color and shape) for discrimination, and the system relied heavily on deictic (pointing) gestures from the human demonstrator to disambiguate similar objects.

*This work was supported, in part, thanks to funding from Canada 150 Research Chair Program.

[1] The authors are with the Dept. of Electrical & Computer Engineering, University of Waterloo, N2L 3G1, ON, Canada [2]the Dept. of Systems Design Engineering, University of Waterloo, N2L 3G1, ON, Canada
Contact: k36fan@uwaterloo.ca

Although human gestures can be interpreted relatively easily and have been proven to be successful [5], [6], more recent systems rely more on natural language, due to the rising popularity and capability of deep learning. The system proposed in [7] formulated the object selection task as a Partially Observable Markov Decision Process (POMDP) that accepted human deictic gestures and speech to determine the object desired by the user. The system can ask for verbal clarification when the object cannot be obtained from the initial human desire estimation. The system in [8] combined speech understanding of user instruction and visual display of the potential targets to allow the user to provide additional information.

The work introduced in [9] analyzed detailed natural language instructions with a multimodal target-source classifier model with attention branches. The model took in visual information, spatial locations, and linguistic instruction. The likely target-source pairs are then generated, using an attention map and the user could select the most desirable pair for the robot to pick up without ambiguity.

Furthermore, the interactive dialogue system in [10] analyzed unstructured text or speech instructions with Long Short-Term Memory (LSTM) recurrent network and visual data with Single Shot MultiboxDetector and a Convolutional Neural Network (CNN). The system disambiguated references by providing speech queries to the user. Similarly, in [11] the system could analyze unstructured speech instructions and disambiguate by providing more informative queries to the user. In [11], the system performed a relevancy clustering operation to extract the target object after the LSTM and CNN analysis. Zhang et al. suggested an improved interactive dialogue system by using a POMDP model to track the history of observations and ask appropriate informative clarification questions to disambiguate [12]. More similar language-based interactive dialogue systems are discussed in [13], [14] with improved language understanding and more detailed disambiguation queries.

While natural language-based approaches are visibly gaining popularity over time and offering gradually more sophisticated speech understanding, heavily relying on speech analysis for object selection clarification can be unreliable, due to, e.g., the unconstrained size of human language vocabulary, the frequent usage of synonyms (different words with the same meaning), and homonyms (words that have the same pronunciation but completely different meaning) in natural expressions, as well as the possible use of irony or sarcasm [15]. Moreover, many current systems lack the ability to learn novel objects after disambiguation, aside from [5]. However, [5] assumes objects have a monotone color and does not explain the retrieval of unfamiliar objects.

To address those issues, our system combines both human speech and gesture feedback and utilizes simple robot speech queries and robot gestures to disambiguate the correct reference. In addition, our system also incorporates a short-term/long-term memory learning scheme to learn novel objects after disambiguation, so that the robot can execute the task without ambiguity, if the same instruction is given again. One example is described in the setting shown in Figure 1, where the robot can locate the salt container by adapting to human speech or gesture feedback, so that when the user asks for salt again, the robot can pass over the correct container with no need to repeat disambiguation, thanks to its long-term memory.

III. PROPOSED METHOD

A. System Overview

The proposed social referencing disambiguation framework performs and analyzes both speech and gesture signals to disambiguate references. The framework is highly modular and can be broken down into the following functional components:

- User command interface;
- Fuzzy task ambiguity level analysis;
- Fuzzy human attention level assessment;
- Social referencing disambiguation;
- Short-term long-term memory object learning.

The framework[1] takes in natural language commands or text-based input for task instruction. It extracts the object label plus any descriptive terms of the object, e.g., "*salt* container" and "*blue* cup". The framework then analyzes the visual scene of the environment, determines if the object is known and computes the ambiguity level of the task via fuzzy logic [16]. If the ambiguity level is high (i.e., there are multiple target candidates or the requested object is unknown), the framework notifies the user about the ambiguity via speech and assesses human attention level, based on visual data, and it tries to attract human attention via speech or robot gesture (i.e., waving) before the disambiguation process. The framework allows the robot to perform deictic gestures, gaze communication, and simple verbal queries to inquire about the potential target candidate. The human's speech and head gestures are analyzed to provide feedback and to enable the framework to obtain the correct reference. If the primary ambiguity source originates from the novelty of the object label, the framework will store the clarified item's visual example into the long-term memory for future retrieval. Lastly, the robot fetches the determined target item.

The social referencing disambiguation framework integrates multiple data-driven, deep-learning solutions, fuzzy logic inference and expert systems to perform task, speech, and gesture analysis. The high-level flow diagram of the entire framework is shown in Figure 2.

B. User Instruction Analysis

The user can either type in the name of the object with a keyboard or verbally instruct the robot to pick up an item. However, analyzing verbal instructions is challenging, as users cannot be overly restricted in terms of the way they issue their instructions.

Human language is productive and creative. There are different ways to express the same idea with different vocabulary and sentence structures. For example, here are some

[1]Source code available:
https://github.com/KevinFan9729/socialReferencingDisambiguation

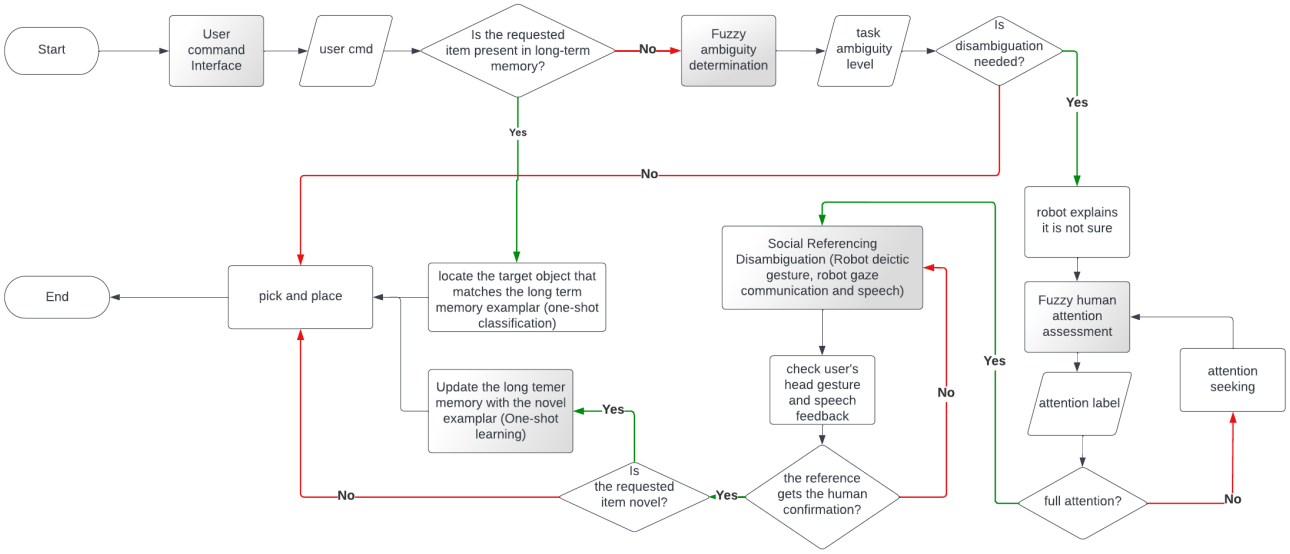

Fig. 2. The high-level flow diagram of the social referencing disambiguation framework.

different expressions for acquiring a cup: “Pass me the cup”, “Hand me over the cup”, “The cup, please”, etc. We utilize Google web API to transcribe speech to text. Then, the obtained text is tokenized into individual words, and Hunpos tagger [17] classifies tokens into word classes. We extract adjectives and nouns and reconstruct the object label from the tokens.

For example, if the task instruction is “Could you please pass me the blue cup?”, the extracted object label is “blue cup”. In this study, we only focus on object fetching tasks, but the proposed approach can be extended to other tasks.

C. Visual Scene Ambiguity Analysis

After obtaining the object label, the robot needs to analyze its visual scene and determine if the current task is ambiguous. We utilize the you-only-look-once (YOLO) real-time object detection algorithm [18] trained with Microsoft COCO dataset [19] to detect objects and create bounding boxes. Based on the visual data, we implemented a fuzzy inference system that analyzes the number of candidate objects, detection confidence score and object label novelty to compute the ambiguity level of the situation. Since we deal with dynamic environments, language-based instructions, and real-time object detection, fuzzy logic [20], [21] is applied here to mimic the human decision process to handle the imprecision of real-world data. There are different kinds of ambiguity involved in the robot object selection tasks:

- Ambiguity due to multiple target candidates.
- Ambiguity due to misclassification and low confidence in the object detection algorithm.
- Ambiguity due to object novelty.

Instead of just giving a binary ambiguity analysis (ambiguous vs. non-ambiguous), our fuzzy logic inference system can compute a continuous ambiguity level to enable the robot to adapt to those levels and address different kinds of ambiguity.

The fuzzy system membership functions are constructed with smooth Gaussian functions in the form of:

$$f(x) = \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{1}$$

the triangular function in the general form of:

$$f(x) = \begin{cases} \frac{(x-a)}{(b-a)}, & a < x < b \\ \frac{(c-x)}{(c-b)}, & b < x < c \\ 0, & x \leq a \text{ or } x \geq c, \end{cases} \tag{2}$$

the translatable Kronecker delta function:

$$\delta(x-T) = \begin{cases} 1, & x = T \\ 0, & otherwise, \end{cases} \tag{3}$$

and the half trapezoidal function which can be specified:

$$f(x) = \begin{cases} \frac{(x-a)}{(b-a)}, & a < x < b \\ 1, & b < x. \end{cases} \tag{4}$$

Because the fuzzy membership function describes the degree of truth that an element belongs to a subset (partial truthiness), the maximum value of the membership function is 1. The antecedent membership functions describe the visual scene, and the consequent membership function describes the overall ambiguity level. A network of if-else statements constructed the knowledge base of the fuzzy ambiguity unit to infer decisions. Some examples of the fuzzy rules are as follows:

- **IF** (*object confidence* is very low **OR** low) **AND** (*object count* has no object **OR** multiple potential targets) **THEN** *ambiguity level* is very high
- **IF** (*object confidence* is high **OR** very high)**AND**(*object count* is one) **THEN** *ambiguity level* is very low

As shown, our fuzzy rules are intuitive in describing events (a highly explainable system). Furthermore, rules

are also flexible for modifications and additions when new domain knowledge is obtained. The ambiguity level is a crisp real number that is defuzzified using the centroid method described in [22], where the center of mass of the activated membership function is calculated as the output value. More detailed descriptions of the fuzzy inference system can be found in our previous work [16].

D. Human Attention Assessment

The system cannot disambiguate effectively if the human user is not paying adequate attention to the robot. Therefore, the framework needs to assess the human's attention level before engaging in any disambiguation process. The system utilizes both head orientation and gaze direction to evaluate the attention level of the user. The framework applies the method proposed in [23] to extract user facial key-points, iris coordinates and performs Perspective-n-Point (PnP) solving with Levenberg-Marquardt optimization [24]. The head pose can then be determined by the thresholding logic of the head rotational values in the x and y dimensions. To estimate eye gaze direction, we apply another fuzzy inference system due to the fact that eyes are small visual features that can be noisy in terms of detection. In addition, eye movements are also considerably more subtle compared to head rotation, and the fuzzy inference system can aid the estimation to be more robust. The framework employs Gaussian membership functions in the form of equation (1) to describe the gaze direction in each eye and the activated memberships in each eye are then projected to the consequent Gaussian membership function for overall gaze direction estimation. The membership functions in each eye are also continuously calibrated with the positions of the respective eye's inner and outer corners to accommodate different eye shapes and sizes that can vary due to head movements. The rules of the if-else network for inference remain highly intuitive; one of the examples is as follows:

- **IF** (*left eye position level* is medium) **OR** (*right eye position level* is medium) **THEN** *overall gaze direction* is center

Subsequently, the attention level can be assessed with simple logic rules that incorporate the head rotation and eye gaze label, some examples of these are as follows:

- **IF** (*head orientation* is center) **AND** (*gaze focus* is eye contact) **THEN** *attention* is full attention
- **IF** (*head orientation* is center) **AND** (*gaze focus* is eye away) **THEN** *attention* is semi-attention

The defuzzification method is also the centroid method mentioned above [22], and further explanation of this functional component is elaborated in depth in our previous work [16].

E. Disambiguation

When the robot detects that there is ambiguity in the task, and after ensuring that it has adequate human attention, the robot verbally informs the user about the ambiguity. The robot uses a deictic gesture (the robot closes its gripper to form a pointing finger) to point at a candidate object and

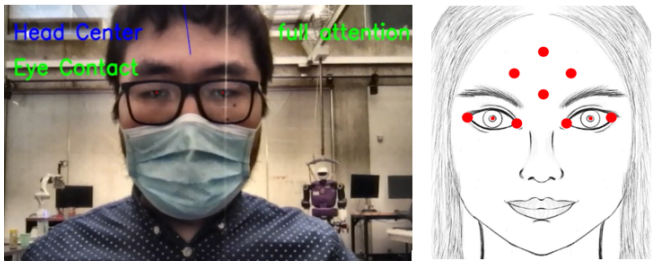

Fig. 3. Attention assessment. Left shows attention assessment of a user with heavy facial obstruction, right shows facial key-points utilized in the system.

verbally inquires for feedback. At the same time, the robot alternates its gaze between the object in question and the user to establish gaze communication. The framework estimates human head gestures by employing the Caffe model [25] to detect and locate the face, and it exploits the Lucas-Kanade optical flow method to track the motion of the head [26]. Furthermore, users can provide verbal feedback to the robot. We employ a multi-layer, bi-directional LSTM recurrent neural network trained with the Twitter US Airline Sentiment dataset [27] to classify speech feedback. The feedback network is illustrated as part of Figure 4. If the robot receives negative feedback on the selected item, the robot points to the next most confident potential target candidate (confidence scores by YOLO) to continue the disambiguation process. After either the user gesture or the speech feedback is analyzed and the correct reference item is clarified, the ambiguity is eliminated and the robot proceeds to execute the task and engage in learning, if necessary.

F. Learning and Object Selection

When the framework recognizes that the label of the object is novel, it is important to learn the object, so that when the same object is requested again, the robot can proceed without ambiguity. The YOLO network trained with COCO is satisfactory at detecting common objects. However, the network's knowledge of the object is often not granular enough. Though the YOLO model can detect different shapes, textures, and colors, for generalization purposes and because of the COCO dataset's labels, many visually distinct objects are classified into the same class. If the user asks for a "blue cup", the robot will run into task-halting confusion due to a lack of knowledge, as it recognizes all cups with the generic "cups" label. It is, therefore, important to have the means to update the robot's knowledge base through interactions.

The framework employs a short-term/long-term memory scheme to facilitate learning. When an unfamiliar object label is supplied and the robot has no prior experience with the object label, the robot looks at all available pickable objects. The framework stores all images of pickable objects in the short-term memory along with other information (names of objects, base-frame/map-frame coordinates of objects). The short-term memory is volatile, so the memory elements will be destroyed after usage. After all target candidates are stored, the robot disambiguates through social referencing as described in subsection III-E. Once the correct reference

is clarified, the referenced object's short-term memory item is transferred to the persistent long-term memory for later recall. The robot approaches the target, turns toward the target object and picks up the appropriate object. Note that the learning process described only occurs when the framework determines that the object label is unfamiliar. It would be inefficient to engage in learning if the base object detection model already knows the object.

When the robot is asked for an item that has been learned previously, the framework extracts all pickable items seen by the robot's camera and transfers them into the short-term memory, and it invokes the robot's relevant long-term memory element. Short-term memory images are compared against the long-term memory exemplar with a custom siamese network for one-shot learning [28]. The correct item in the short-term memory should have the highest similarity score with the exemplar in the long-term memory and can be retrieved easily. The siamese network we implemented employs CNN twin networks in the well-tested VGG16 architecture[29] and we trained our network with the empirically chosen (among mean squared error, cross-entropy loss and contrastive loss) contrastive loss described below:

$$L = mean((1-Y)*(Y_{pred})^2 + Y*(max(M-Y_{pred},0)^2)), \quad (5)$$

where M is the margin, empirically set to 1, and Y_{pred} is based on the L1 distance. The siamese network works in parallel with the YOLO model as an auxiliary; thus, the original objects from YOLO are not modified, and major issues like catastrophic forgetting in online learning are successfully prevented. The learning network is represented as part of Figure 4.

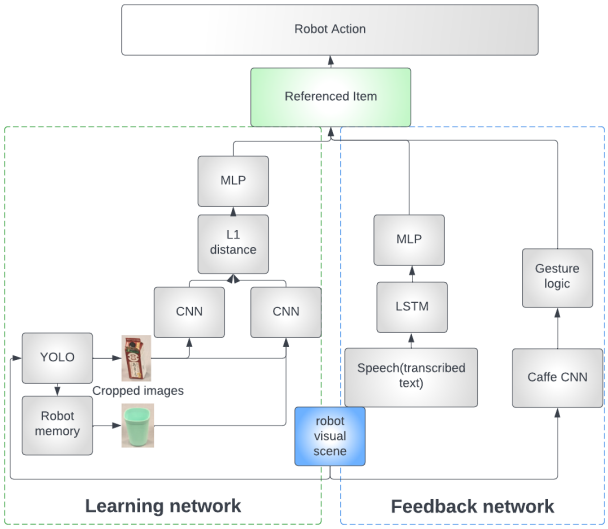

Fig. 4. Network overview of the social referencing disambiguation framework.

IV. Experiment

A. Experiment setup

Our experiment includes three tests to validate our proposed framework.

- **Test 1**: A set of objects is placed on the table and the target item is unique among all other items. The name of the target item is *known* to the base object detection model. Test 1 is designed to verify if the framework can successfully recognize unambiguous situations and proceed with task execution without the disambiguation process.
- **Test 2**: There are multiple candidates for the potential target, the name of the target item is *known* to the base object detection model. Test 2 is designed to address the ambiguity involved with multiple potential targets and low confidence/misclassification in the base object detection model.
- **Test 3**: There are multiple candidates for the potential target, the name of the target item is *unknown* to the base object detection model. Test 3 examines the learning ability of the framework. Can the framework learn novel object labels after the interaction? Moreover, can the framework successfully retrieve the correct item based on its memory?

The object positions are periodically randomized between trials, and objects are also added/removed between trials for better generalization.

A trial is considered successful if the correct target item is selected by the robot (the success of object pick up is out of the scope of this paper). All tests are repeated 20 times to compute the success rate of the framework. In addition, test 3 consists of pairs of sub-trials for each run. In the first sub-trial, the user needs to supply a novel descriptive object label to the robot, and the robot will disambiguate to locate the target item and learn the item. In the second sub-trial, the user will provide the same object label as in the previous sub-trial, and the robot should remember the target item and proceed to pick up the item without ambiguity. Therefore, in test 3, success is considered if and only if both sub-trials of the pair are successful. The general setup of the experiment can be seen in Figure 5.

Fig. 5. Experiment setup and the Fetch Robot.

B. Robotic System

We use the Fetch mobile manipulator robot, as shown in Figure 5. An additional Google Pixel 4a camera is mounted on the head of the robot to provide a higher resolution/frame rate video for the human attention assessment functional block III-D. The framework is integrated onto the robot with

Robot Operating System (ROS) Melodic. All deep learning networks, data processing pipelines, robot motion planning and navigation modules are handled by a PC equipped with an NVIDIA RTX 2060 GPU and an i7 Intel Core CPU (11th Gen Intel i7-11700 @ 2.5GHz x 16) running Ubuntu 18.04.

V. Results

We performed all tests described on the physical robot. The success rate of test 1 is 95%, 90% for test 2 and 80% for test 3. The framework is effective based on our experimental results. However, it is worth mentioning the major causes of failure: errors in robot navigation, faulty/low confidence object detection, unstable bounding boxes, and variable lighting conditions.

We now provide a detailed step-by-step description of the results generated by the framework for the most complex setting: test 3 (novel object label). First, the user asked, "Can you please pass me the green cup?" The robot extracted "green cup" as the target label from the instruction and proceeded to analyze the ambiguity of the task. The robot recognized that the object label was novel; therefore, it could not utilize YOLO to locate the "green cup" among all objects. The robot saved all pickable objects in view in the short-term memory for future disambiguation. As shown in Figure 6, the resultant ambiguity is classified as "very high", and the final ambiguity level is as high as 0.918.

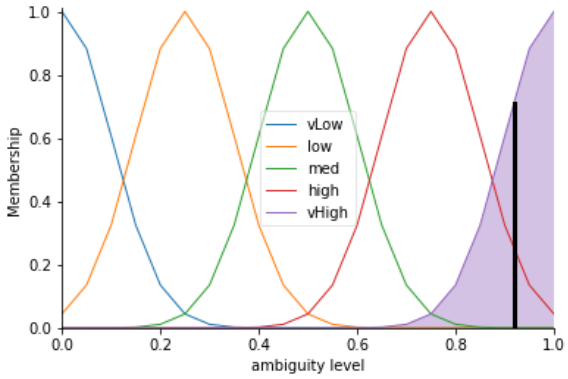

Fig. 6. Ambiguity level result, ambiguity level can be classified as "vHigh" which means very high, "high" which is high, "med" as medium, "low" as low and "vLow" as very low.

The robot then proceeded to assess the user's attention before disambiguation. The robot determined the user was semi-attentive because the user's head was facing the robot, but the user had no eye contact with the robot; thus, the robot verbally requested more attention and the user shifted his gaze toward the robot to provide full attention for the interaction 7.

The robot disambiguated through its pointing gesture, gaze communication, and verbal queries. The robot recognized head gestures and it listened to verbal feedback from the user. The user sometimes would say phrases that were irrelevant to the interaction, such as "Nice weather we have today." The language model classified those phrases as irrelevant and ignored them. After clarifying the target item, the robot saved

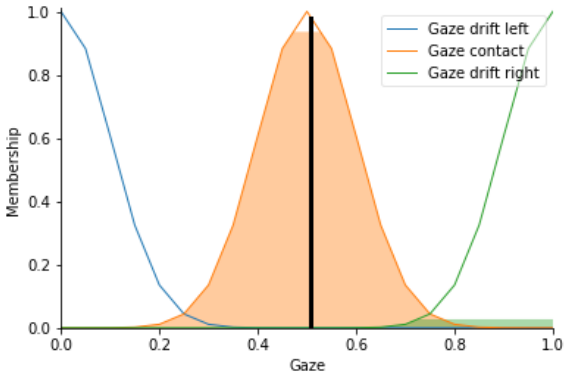

Fig. 7. Gaze level after attention attraction.

the target item in the long-term memory to associate the new label with the object for future retrieval. When the learned object was requested again, the robot retrieved the relevant memory in the long-term memory and utilized the siamese network to distinguish the target item from all objects in view. Example comparisons can be seen in Figure 8.

Fig. 8. The siamese similarity comparison; the left vertical pair has a similarity score of 0.72, and the right vertical pair has a similarity score of 0.95. The top query images are in the robot's view, and the bottom anchor images are in the robot's long-term memory.

Finally, the robot successfully located and fetched the target object based on the highest similarity score with the memorized sample.

VI. Conclusions

The proposed social referencing disambiguation framework explored the practical utilitarian aspects of social referencing. It exploited modern deep learning methods to resolve various ambiguities in robot object selection. In the future, we plan to conduct a full human user study to further validate our framework. Additionally, we are interested in incorporating more sophisticated language understanding models to enrich the framework's utility; we also want to expand the learning scheme so it can cluster and associate similar examples in long-term memory. Furthermore, we wish to add the emotional aspect of social referencing to the framework, so that a complete computational framework of social referencing can be established.

REFERENCES

[1] A. Bandura, "Social cognitive theory of social referencing," in *Social Referencing and the Social Construction of Reality in Infancy*. Springer, 1992, pp. 175–208.

[2] J. J. Campos, S. Thein, and D. Owen, "A Darwinian legacy to understanding human infancy: Emotional expressions as behavior regulators," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 110–134, 2003.

[3] A. L. Thomaz, M. Berlin, and C. Breazeal, "An embodied computational model of social referencing," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.* IEEE, 2005, pp. 591–598.

[4] S. Boucenna, P. Gaussier, and L. Hafemeister, "Development of first social referencing skills: Emotional interaction as a way to regulate robot behavior," *IEEE Transactions on Autonomous Mental Development*, vol. 6, no. 1, pp. 42–55, 2013.

[5] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer, "A multi-modal object attention system for a mobile robot," in *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2005, pp. 2712–2717.

[6] A. Utsumi, N. Tetsutani, and S. Igi, "View-based detection of 3-d interaction between hands and real objects," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 4. IEEE, 2004, pp. 961–964.

[7] D. Whitney, E. Rosen, J. MacGlashan, L. L. Wong, and S. Tellex, "Reducing errors in object-fetching interactions through social feedback," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1006–1013.

[8] E. Sibirtseva, D. Kontogiorgos, O. Nykvist, H. Karaoguz, I. Leite, J. Gustafson, and D. Kragic, "A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction," in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2018, pp. 43–50.

[9] A. Magassouba, K. Sugiura, and H. Kawai, "A multimodal target-source classifier with attention branches to understand ambiguous instructions for fetching daily objects," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 532–539, 2020.

[10] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken language instructions," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3774–3781.

[11] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," *arXiv preprint arXiv:1806.03831*, 2018.

[12] H. Zhang, Y. Lu, C. Yu, D. Hsu, X. La, and N. Zheng, "Invigorate: Interactive visual grounding and grasping in clutter," *arXiv preprint arXiv:2108.11092*, 2021.

[13] P. Pramanick, C. Sarkar, S. Paul, R. dev Roychoudhury, and B. Bhowmick, "Doro: Disambiguation of referred object for embodied agents," *IEEE Robotics and Automation Letters*, 2022.

[14] P. Pramanick, C. Sarkar, S. Banerjee, and B. Bhowmick, "Talk-to-resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot," *Robotics and Autonomous Systems*, vol. 155, p. 104183, 2022.

[15] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multimedia Tools and Applications*, pp. 1–32, 2022.

[16] K. Fan, M. Jouaiti, K. Dautenhahn, and C. L. Nehaniv, "Fuzzy object ambiguity determination and human attention assessment for domestic service robots," in *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2022 in press.

[17] P. Halácsy, A. Kornai, and C. Oravecz, "HunPos-an open source trigram tagger," *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pp. 209–212, 07 2007.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 779–788.

[19] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[20] L. Zadeh, "Fuzzy logic," *Computer*, vol. 21, no. 4, pp. 83–93, Apr. 1988.

[21] L. A. Zadeh, "Fuzzy logic = computing with words," *IEEE Transactions on Fuzzy Systems*, vol. 4, no. 2, pp. 103–111, May 1996.

[22] "JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2," Nov. 2019. [Online]. Available: https://zenodo.org/record/3541386

[23] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann, "MediaPipe: A Framework for Building Perception Pipelines," *arXiv:1906.08172 [cs]*, Jun. 2019, arXiv: 1906.08172.

[24] E. Eade, "Gauss-Newton / Levenberg-Marquardt optimization," Mar. 2013. [Online]. Available: https://www.ethaneade.com/optimization.pdf

[25] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM International Conference on Multimedia*, 2014, pp. 675–678.

[26] B. D. Lucas, T. Kanade *et al.*, *An iterative image registration technique with an application to stereo vision*. Vancouver, 1981, vol. 81.

[27] F. Eight, "Twitter us airline sentiment," Oct 2019. [Online]. Available: https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment

[28] G. Koch, R. Zemel, R. Salakhutdinov *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, vol. 2, no. 1. Lille, 2015.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.