2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Bio-Inspired Object Reference Recognition in Human-Robot Interaction under Ambiguous Non-Verbal Cues

Daiju Kanaoka[1,2], Hakaru Tamukoh[1,3] and Hendry F. Chame[4]

Abstract— The object reference recognition task in human-robot interaction (HRI) consists of identifying the object to which a human is referring, based on communicative cues, including gaze and pointing, which is particularly challenging under ambiguous non-verbal behavior. This paper proposes a bio-inspired multimodal fusion algorithm to enable robots to recognize object references based on human gaze and pointing gestures. The proposed method integrates and encodes sensory inputs into a dynamic neural field, allowing the robot to adaptively resolve ambiguities in object referencing. The model was evaluated in an experimental setting where the participants interacted with a Furhat robot. The results showed that the system identified referenced objects with higher accuracy when both gaze and pointing cues were combined. Additionally, subjective evaluations using the Godspeed questionnaire indicated that participants perceived the robot more favorably when it engaged in joint attention behaviors. These results highlight the potential of dynamic neural models in improving intuitive and seamless HRI by addressing non-verbal ambiguity in shared workspaces. Future work will explore improved gaze-tracking techniques and closed-loop interaction models to enhance system robustness and adaptability.

I. INTRODUCTION

The development of robots that can operate in shared spaces and interact with humans should necessarily address the problem of communicating and expressing intention toward objects. Moreover, for such skills to be effective, human social dynamics should be considered by the robot. In this sense, according to research on developmental and social cognition, most communication and intention grasping in object-mediated interactions emerge early in childhood and are based on elementary sensory-motor skills (e.g., gazing, pointing, automatic attunement, mimicry) [1]. Such skills are believed to persist throughout life, allowing individuals to cope with most ordinary situations [2]. Thus, social robots should be able to handle non-verbal behavior and correctly recognize and convey intention toward objects by sharing attention to them with humans.

[1]Daiju Kanaoka and Hakaru Tamukoh are with Graduate School of Life Science and Systems Engineering, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0196, Japan. {kanaoka.daiju604@mail.kyutech.jp, tamukoh@brain.kyutech.jp}
[2]Daiju Kanaoka is with Guardian Robot Project, RIKEN, 2-2-2 Hikaridai, Seika-cho, Sorakugun, Kyoto 619-0288, Japan.
[3]Hakaru Tamukoh is with Research Center for Neuromorphic AI Hardware, Kyushu Institute of Technology, 2-4 Hibikino, Wakamatsu, Kitakyushu, Fukuoka 808-0196, Japan
[4]Hendry F. Chame is with Université de Lorraine, CNRS, LORIA, F-54000 Nancy, France. hendry.ferreira-chame@loria.fr

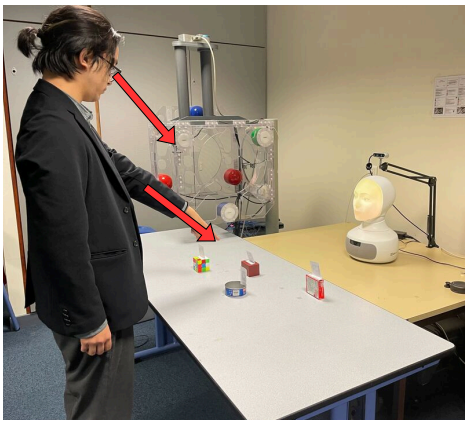

Fig. 1. The interaction scene. The robot Furhat welcomes visitors and explains the utility of objects as the person shows interest to them.

In the human-robot interaction (HRI) literature, the object reference recognition problem includes the study of perception, decision-making, and behavior schemes to provide the robot with the capacity to recognize and respond appropriately to human communication related to objects. The goal is to identify the object to which a person is referring based on communication cues such as gazing and pointing. Several studies have highlighted the difficulty of reliably performing this task with verbal and non-verbal communication, emphasizing the importance of being able to count on flexible and adaptive means of interpreting human gestures in real-life situations.

In less structured interactions, variability of personal style makes it difficult for a robot to rely on a single behavioral modality to correctly recognize the designated object. Spontaneity of nonverbal behavior impacts the robot's sensory systems as a dynamic process, resulting in multimodal ambiguous salience. Thus, the robot must be able to fuse contextual information in order to act congruently with the reference intention of the person by conveying attention to that particular object.

Dynamic Field Theory (DFT) originates from the hypothesis that interacting patterns in cortical neural excitation play an essential role in information processing. Such patterns can be described mathematically as a dynamic neural system in which units in the field represent the activity of a population of related neurons at a macro-level of abstraction. The theory explains how

978-1-6654-5784-2/26/$31.00 ©2026 IEEE

stable states, known as attractors, emerge in cognitive processes, which could account for the formation of memories, spatial representations, and motor behaviors.

Inspired by DFT, in this study, we propose a novel object reference recognition approach for HRI. We consider a scenario where a robot welcomes visitors and explains the utility of objects as the person shows interest to them by pointing or gazing (seen Fig. 1). To do this, the locations that people point to and look at are provided as salience maps to a dynamic neural field encoding the shared interaction space, where behavioral and object-related information are fused. By carefully considering compositionality in our architecture, we show through an experiment how the robot is able to identify the object being referenced in most cases. Moreover, according to the subjects' responses to a measure on their perception of the robot, the results suggested that participants rated the robot more positively when it conveyed attention by gazing at the object being referenced, as compared to only addressing the object verbally.

In summary, the main contributions of this work are:

(i) The proposal of a bio-inspired multimodal fusion algorithm to provide the robot with attention selection for object reference recognition under non-verbal ambiguity and fluctuations in sensory salience.

(ii) Evaluating the model with a sample of participants, where the robot Furhat provided information about object addressed by subjects through pointing and gazing.

This manuscript is organized as follows. Section II presents theoretical principles and research related to our study. In Section III we describe our architecture that allows combining information about objects and subject's behavior from a lightweight two-camera sensors arrangement. Section IV presents the details of the experiment design, including the acquisition of self-reported and behavioral data. The results of the study are discussed in Section V, while conclusions and perspectives are given in Section VI.

II. PREVIOUS WORK

A. Object reference recognition

Object reference recognition (ORR) is a fundamental research problem in the field of human-robot interaction. ORR enables robots to acknowledge and respond appropriately to human instructions regarding objects in shared environments. Here, we review in a non-exhaustive manner previous research that has addressed the challenges of ambiguity and alignment in ORR in online HRI.

The study of [3] proposed to combine verbal and non-verbal gestures to model communicative behavior around object references. This method integrates bottom-up saliency and top-down context cues to predict and guide user's attention toward a referenced object and maximize the likelihood of recognizing the correct object reference. Another study focused on how human-robot alignment (emphasizing or using the same gesture and words) can enhance shared understanding in object reference recognition [4].

A significant challenge in ORR is ambiguity resolution. The work of [5] introduced a fuzzy logic-based system that determines ambiguity in robot object selection tasks and assesses human attention before engaging in disambiguation. This approach ensures that the robot only seeks clarification when necessary, reducing redundant interactions. Another study leveraged augmented reality-powered techniques to resolve ambiguous object references by employing graph-matching methods to fuse visual and spatial information in order to achieve ORR in shared workspaces [6].

Human gaze and attention estimation have also been studied as key factors in ORR. A pipeline is proposed in [7] to estimate human attention toward objects with on-board cameras on the iCub humanoid robot. This system integrates face detection, gaze tracking, and object detection to predict which object a human is focusing on, thereby enhancing robot responsiveness in collaborative tasks.

Beyond gaze estimation, lexical entrainment plays a role in reference resolution. The work of [8] investigated whether humans adapt their vocabulary to robots during interactions, which has implications for how robots should process and generate object references. In addition, referent identification strategies have incorporated ontological reasoning and perspective-taking mechanisms to enhance the robustness of reference resolution [9].

In general, previous work highlights the importance of a multimodal approach that integrates visual perception, linguistic alignment, ambiguity management, and attention estimation; to obtain effective recognition of object references. Thus, most of the research has focused on improving the fluidity and naturalness of human-robot communication in shared environments. However, the research on how to handle information ambiguity of non-verbal modalities in ORR has gained, in our opinion, less attention in the HRI literature, an issue on which we focus in this work.

B. Dynamic field theory

DFT is a framework that allows the study and model of cognitive processes through the temporal evolution of neural population activity in continuous feature spaces. It describes how neural activation patterns emerge and stabilize as peaks in dynamic fields, representing cognitive states such as perception, working memory, and decision-making.

In DFT, a typical neuron model encoding spatial-temporal relations in a 2D topological space is derived from the formalism proposed by [10], such that:

$$\tau \dot{u}(x,y,t) = -u(x,y,t) + h + S(x,y,t) + \int w(x-x', y-y') f(u(x',y',t)) dx'dy' + C_{mem} u_{mem}(x,y,t). \quad (1)$$

Here, $u(x,y,t)$ represents the neural activation level at position (x,y) and time t, while τ is the time constant governing the speed of neural dynamics, h is the resting level, at which the activation level settles in the absence of external stimulation. The term $S(x,y,t)$ represents external stimulus input, encoding sensory or motor information. The lateral interaction kernel $w(x-x', y-y')$ defines local excitation and global inhibition, influencing the stability of activation peaks. Ricker wavelet, as shown in Eq. 2, is one of the commonly used interaction kernels. This interaction kernel induces strong excitation at the location where the stimulation is applied, strong inhibition in the vicinity of the stimulation, and weak inhibition at larger distances.

$$w(x-x', y-y') = \frac{2}{\sqrt{3\sigma}\pi^{1/4}} \left(1 - \left(\frac{d}{\sigma}\right)^2\right) e^{-\frac{d}{\sigma}}, \quad \text{where } d = \sqrt{(x-x')^2 + (y-y')^2}. \quad (2)$$

The nonlinear activation function $f(u)$ is often selected as the sigmoid function as shown in Eq. 3, ensuring threshold-dependent responses.

$$f(u) = \frac{1}{1 + e^{-\beta(u-\theta)}} \quad (3)$$

In DFT, it is possible to define hierarchical models by layering fields in order to account for distinct spatial-temporal dynamics. Hence, a memory field $u_{mem}(x,y,t)$ (see Eq. 4 [11]) can be defined to persist neuron activity as a short-time memory. This allows to maintain the field activation for a while after the external stimulation has ceased.

$$\tau_{\text{mem}} \dot{u}_{\text{mem}}(x,y,t) = -u_{\text{mem}}(x,y,t) + f(u(x,y,t)) \quad (4)$$

DFT has been applied in robotics, particularly in perceptual decision-making, action selection, and sensorimotor control. In human-robot interaction, it enables the integration of sensory cues for adaptive responses [12]. In object grasping and motion planning, DFT allows for the dynamic representation of targets, facilitating real-time grasp adaptation [13]. In autonomous navigation, decision fields regulate movement selections based on environmental sensory inputs [14]. Due to its continuous and adaptive nature, DFT is well-suited for real-time robotic applications that require flexible decision-making under uncertainty.

C. Joint attention and DFT

Joint attention (JA) is a fundamental social skill that allows partners to share and establish common ground about objects mediating interaction. Although some theoretical perspectives have circumscribed coordination in JA to the alignment of inferred mental or psychological states (e.g. [15]), we explore an embodied cognition perspective for JA, according to which coordination in attention sharing is grounded in sensory-motor reciprocal movements, where perceptual processes are not considered to be inferential, but emerge online in direct interaction [16].

Previous works have employed DFT to study JA skills in HRI. In [17], a model was designed to track JA typologies online, organized in a continuum from lack of jointness (or individual attention) to a highest degree of knowledge and attention sharing. In [18], a neural network architecture was proposed to track individual attention selection in sensory ego–spheres, which allows adaptation and perspective taking in object referencing by pointing gestures. These previous researches did not explore the problem of object referencing under ambiguity, which is our main focus in this work.

III. PROPOSED METHOD

We propose a model for object reference recognition that can process ambiguous non-verbal information. In this section, the components of the ORR model are detailed, as well as the robot behavior model for the interaction task.

A. Proposed model

Figure 2 shows the architectural view of our ORR model that included the five information processing components detailed next.

1) Gesture stimulation: This component processes the non-verbal information a person conveys and computes attention saliency to locations in the field. First, from sensory acquisitions on body posture, the robot estimates a projection in the field. This estimation is subject to stochasticity, which is assumed to follow a Gaussian distribution with distinct variance per gesture modality. In the example shown in Fig. 2, the variance in gaze location estimation is expected to be higher than that of pointing location estimation. As a result, the stimulation map for the gaze location is applied more broadly across the field than the stimulation map for the pointing location. Finally, a general formulation for computing the stimulation map for m behavior modalities ($m \in \{\text{gaze}, \text{pointing}\}$ in this study), with the normalization function $f.$ is obtained such that

$$S_{\text{gesture}} = f\left(\sum_m S_m\right) \quad (5)$$

This fusion component enables the flexible inclusion of additional non-verbal modalities into the gesture stimulation saliency map.

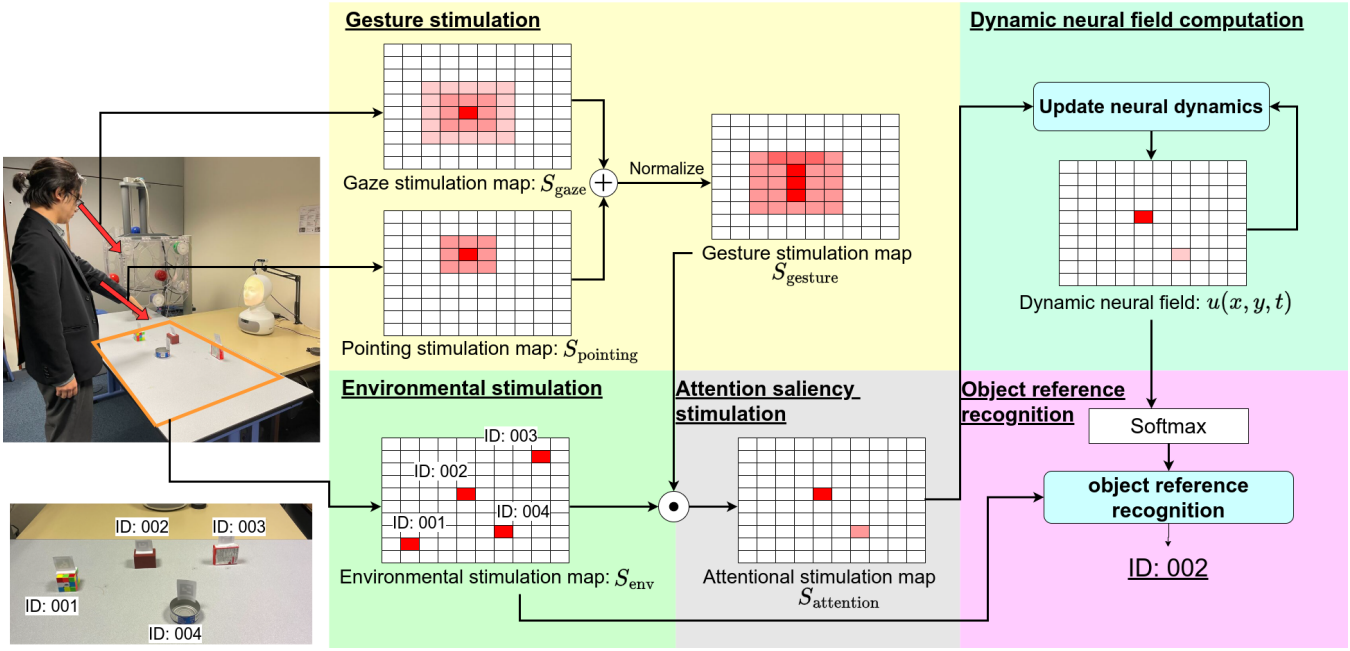

Fig. 2. Architecture of the proposed method and interaction scenario. DFT is used for multimodal saliency information fusion and attention selection.

2) Environmental stimulation: In this component, the robot observes the surrounding environment and computes the S_{env} saliency map from the estimation of the localization of stimuli in the field. In this study, this component is used to track object spatial references. This calculation produces a map showing which objects in the environment attract the most attention. For more implementation detail, please refer to Sec . IV-B.

3) Attention saliency stimulation: This component takes the Hadamard product of the calculated gesture stimulation map S_{gesture} and the environmental stimulation map S_{env} as shown in Eq. 6. The calculated map provides information on attention behavior correlates in relation to object locations.

$$S_{\text{attention}} = S_{\text{gesture}} \odot S_{\text{env}} \tag{6}$$

4) Dynamic neural field computation: The neuron model in Eq. 1 is computed by receiving as input the attention saliency stimulation map $S_{\text{attention}}$. As a dynamical system, the neural field is a causal system that depends on the current and past stimulation received. Due to this property, it allows tracking the current attention under fluctuations in saliency maps computations. Also, by considering the memory component in Eq. 4, the dynamic neural field activation is preserved for a while and slowly fades away. This helps to improve consistency in ORR.

5) Object reference recognition: We carefully considered the aspect of compositionality from sub-symbolic representations to a process of emergent recognition of the referenced object. Thus, ORR is computed by applying the softmax function to the dynamic neural field and obtaining the location $\hat{u}_{x,y}$ of the most activated unit encoding a region in the interaction space, such that

$$\hat{u}_{x,y} = \arg\max_{(x,y)} \big(\text{softmax}(u(x, y, t))\big). \tag{7}$$

In case there is an object related to the unit's location, it will be selected as a candidate for the actual focus of attention. For practical purposes, ORR is considered to happen after n consecutive selections.

B. The robot behavior model

The robot was programmed to perform four behaviors.

1) Greeting the human: The robot turns to face the human and greets them. It then asks the human if there are any objects on the table that interest they.

2) Tracking the human: While gazing at the human, the robot tracks human gestures from images obtained from the RealSense D435 sensor (see Fig. 4). Then, it computes saliency stimulation maps and updates the neural field state for recognizing attention to objects on the table.

3) Cuing attention: Once object reference recognition is achieved, the robot turns its head towards the object to signal its state of attention to the human.

4) Describing the object: The robot turns to face the human and provides a description of the recognized object.

IV. EXPERIMENT

We conducted an experiment in the LORIA lab with 9 participants, all students aged between 21-28 years old and right handed, who voluntarily participated in the study, no remuneration was provided. The experimental

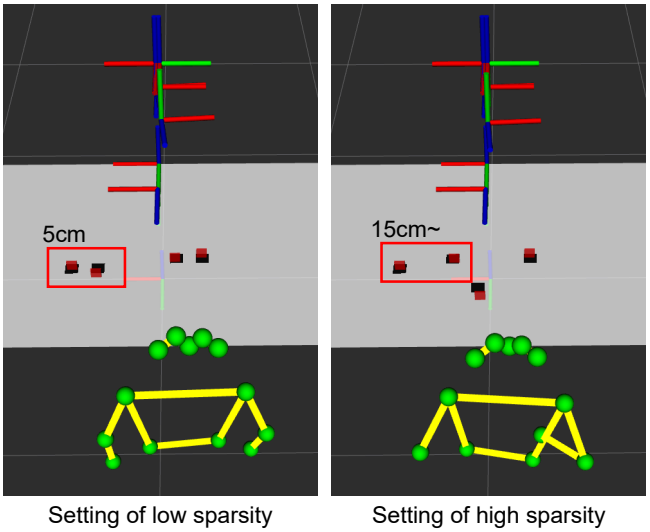

Fig. 3. Settings for the object sparsity experimental variable.

protocol strictly followed the lab's code of ethics and complied with the European General Data Protection Regulation (GDPR).

A. Experimental design

The participants were provided with a short story about a fictional robotic character named Junior (the robot Furhat [19] was considered in the study) who is learning to interact with humans for the first time and is willing to show its knowledge on some objects on the table (see Fig. 5). The participants were instructed to stand up in front of the table to interact with the robot (see Fig. 2 left for a third-person perspective of the scene, and Fig. 4 for a first-person view). The duration of the experiments was about 30 minutes, followed by a 10-minute session, where subjects responded to a Godspeed [20] measure on how they perceived the robot.

Three experimental variables were studied.

1) Object sparsity: In the condition of low-sparsity, the Rubik's Cube and Foam Brick objects where placed side-by-side at 5 cm of each other. The same proximity relation was adopted for Gelatin Box and Tuna Fish Can (see Fig. 3). The condition of high-sparsity is shown in Fig 4, in which there is a minimum distance of 15 cm between objects.

2) Gesture modality: In the condition pointing-and-gazing, subjects were asked to freely reference an object. In the condition gazing, subjects were required to only direct the head towards an object of interest. A pointing-only condition was deliberately excluded. This decision is based on the premise that when humans intentionally point to an object, their gaze naturally follows, making a pointing-only gesture a not realistic and less ecologically valid task for participants. Our goal was to evaluate the system in the context of more natural human communicative behaviors.

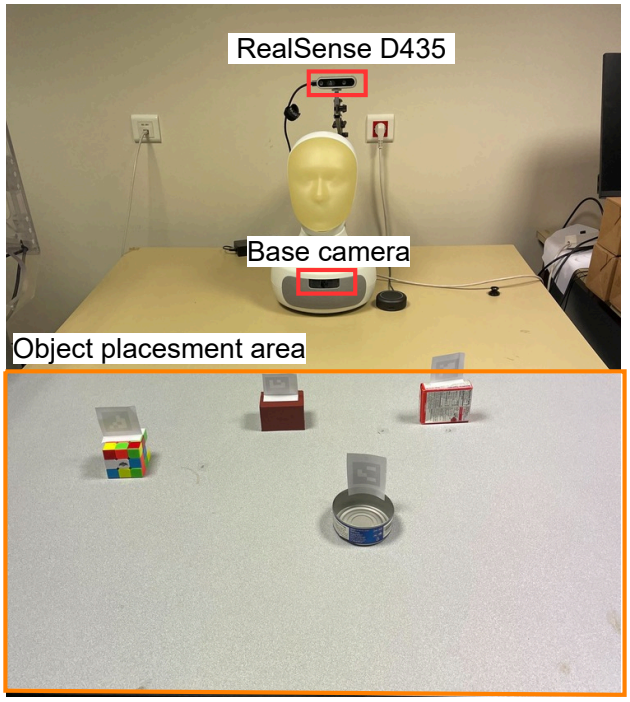

Fig. 4. First-person perspective view of the experimental scene.

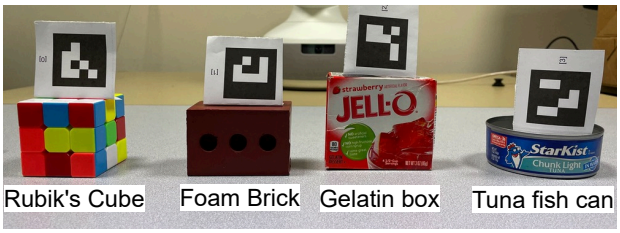

Fig. 5. Objects used in the experiment.

3) Attention behavior: In the condition true, the full robot behavior model was active, whereas in the condition false the behavior cuing attention was deactivated (see Section III-B).

Two experimental groups were formed to evaluate the previous set of variables. Subjects were given, according to their order of arrival, an integer index starting from 1, so odd numbers were assigned to Group A and even numbers to Group B. The full conditions of the variables object sparsity and gesture modality were presented to both groups, whereas attention behavior=true was presented to Group A (which had 5 participants) and attention behavior=false was presented to Group B (which had 4 participants).

B. Estimation of object location

The robot uses its base camera (see Fig. 4) to recognize AR markers attached to objects (see Fig. 5) and to estimate the location of objects [21] within the object placement area.

TABLE I

Parameters used in the experiment

Parameter	value
h	-0.05
τ	0.2
C_{mem}	0.2
τ_{mem}	2.0
dt	0.1
σ	0.7
β	50.0
θ	0.5
Number of neurons	800
$\sigma^2_{\text{pointing}}$	0.06
σ^2_{gaze}	0.03

C. Settings for the robot behavior model

Here, we provide more details on the implementation of the behaviors described in Sec III-B. For the accomplishment of the behavior tracking the human. the robot has to estimate the pointing and gaze direction, which is detailed below.

1) Pointing estimation: We estimated the pointing location based on the 3D human skeleton segmentation. From images acquired from the RealSense D435 sensor (see Fig. 4), first, we used OpenPifPaf [22] to perform 2D pose estimation. The obtained posture was then extended to a 3D pose using depth information. Finally, the pointing position was estimated by calculating an intersection of the extending line from the elbow to the wrist and the field.

2) Gaze estimation: In this research, we estimated the gaze position assuming that the gaze direction aligns with the head direction. We employed HopeNet [23] as a head direction estimation. The estimated face direction is then used to calculate the intersection point with the field, which is considered the mean gaze position in a Gaussian distribution.

3) Parameters of dynamic neural field: Table I shows the parameters of the proposed method used in the experiments. The dynamic neural field encodes the range of the object placement area. As shown in Fig. 4, the objects were placed in a 2.0 m width $\times$ 1.0 m height surface. Thus, each neuron encodes a region of a 5 cm $\times$ 5 cm in a rectangular grid, resulting in a total of 800 neurons (see Fig. 6, the resolution has been down-sampled in Fig. 2 for illustration purposes). In Table I, σ_{pointing} and σ_{gaze} denote the variance of the Gaussian distribution for the pointing and gaze stimulation maps, respectively. All parameters were manually selected for this study.

4) Resources: The experiment was conducted on a system with dual Intel Xeon Gold 5218R CPUs, 128 GB DDR4 RAM (3200 MHz), and NVIDIA RTX A2000 (12 GB VRAM). The software environment was Ubuntu 20.04-based virtualized ROS Noetic environment created using Apptainer, and Python 3.8 was used for implementation.

TABLE II

Trials Success rate (SR) for object sparsity (OS), gesture modality (GM) and attention behavior(AB)

Group	Trial	OS	GM	AB	SR
A	1	high-sparsity	pointing/gazing	True	93.33%
A	2	high-sparsity	gazing	True	20.00%
A	3	low-sparsity	pointing/gazing	True	86.67%
A	4	low-sparsity	gazing	True	66.67%
B	1	high-sparsity	pointing/gazing	False	100.00%
B	2	high-sparsity	gazing	False	50.00%
B	3	low-sparsity	pointing/gazing	False	100.00%
B	4	low-sparsity	gazing	False	41.67%

TABLE III

Average Godspeed scores per group

Category	Group A	Group B
Anthropomorphism	2.60 ± 0.346	2.15 ± 0.300
Animacy	3.10 ± 0.190	2.37 ± 0.343
Likeability	4.24 ± 0.589	3.40 ± 0.673
Perceived Intelligence	3.48 ± 0.540	3.00 ± 0.673
Perceived Safety	3.66 ± 0.781	4.00 ± 0.272

D. Results

Table II presents the success rate calculated between the object reference recognized by the robot and the self-reported reference collected after each trial of the experiment. Table III shows the average and standard deviation of subjects' evaluation per category defined in the Godspeed measure.

V. DISCUSSION

The system proposed was able to achieve a considerably high success rate in the free referencing behavior for both low- and high-sparsity conditions. Interestingly, as shown in Fig. 6, the model is able to handle ambiguous saliency maps from gaze and pointing tracking, while providing stable recognition of the object referencing behavior in the relatively difficult condition of low object sparsity.

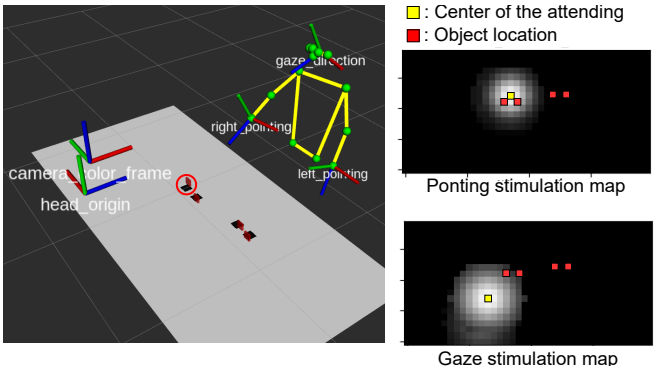

Fig. 6. RViz visualization of an experimental trial. On the left, the human's frames show the pointing gesture toward the red-circled object with the right arm, at the opposite side of the table it is shown the robot's head and real-sense reference frames. On the right, it is shown how the model is able to handle ambiguity between pointing and gazing object referencing behavior under a low-sparsity scenario.

Overall, the success rate obtained by pointing-and-gazing gesture modality is considerably higher than gazing alone. Since gaze tracking could only be implemented through head direction alignment, we hypothesize that this resulted in a more difficult gesture to perform by the human.

When analyzing the attention behavior variable, it is observed that pointing-and-gazing gesture modality performed slightly better when the cuing attention behavior was off. A possible explanation for this result is the fact that, although the robot performed mostly well in trials, since behavior control was only possible in open-loop, due to restrictions on the platform; the participants may have attempted to compensate for perceived misalignment when engaging in JA with the robot. Adaptation can also be observed in group A's higher performance on trial 4 over trial 2. Furthermore, since this effect is not observed in group B, a plausible explanation would be that subjects felt more motivated to adapt to the robot in condition A.

Finally, when comparing the perception of the robot, the participant's rates in Godspeed's measured categories are consistently higher for group A compared to group B, with the exception of perceived safety. We believe that the latter can be explained by the fact that in group B the robot moved less, to only gesticulate speech and gaze the human, so the higher motion on group A could have impacted the perception of safety in participants. Overall, these results suggest that engaging in JA exerts a positive influence on participant's perception of the robot.

We believe these result can be attributed to the inherent characteristics of the DNF. Human non-verbal cues, such as gaze and pointing, are not static but fluctuate over time. The DNF can naturally handle these temporal dynamics by integrating current stimuli while retaining past input information, thereby forming stable activity peaks that are robust against transient noise and fluctuations. Furthermore, when presented with multiple modalities of varying precision and characteristics, as in this study (gaze and pointing), the DNF can represent them on a common saliency map and fuse them non-linearly. This allows the system to resolve ambiguity and make robust decisions by integrating all available information, even when one modality is ambiguous. Therefore, we conclude that the DNF is a highly suitable framework for object reference recognition based on ambiguous non-verbal cues, which is the focus of this research.

VI. CONCLUSION

In this study, we proposed a bio-inspired multimodal fusion algorithm for object reference recognition in HRI, designed to handle ambiguous non-verbal cues such as pointing and gazing. Inspired by DFT, our model integrates contextual information and adapts to fluctuations in sensory salience, while allowing spontaneous and effective object reference recognition in human-robot interactions.

Experimental results shown that our approach achieves high accuracy in object reference recognition, particularly when both pointing and gazing cues are utilized. The system performed robustly under varying object sparsity conditions, effectively managing ambiguity in non-verbal referencing behaviors. Additionally, user perception of the robot, assessed through the Godspeed questionnaire, indicated that participants responded more positively when the robot actively engaged in joint attention behaviors.

Despite these promising results, limitations remain. The accuracy of gaze tracking was constrained by the assumption that head direction aligns with gaze, which may not always be the case. Furthermore, the system's open-loop behavior control may have influenced participants' adaptation strategies. Another significant limitation is the small sample size of our experiment, which precludes inferential statistical analysis. Consequently, our evaluation is confined to the scope of descriptive statistics. Future work will explore improved gaze estimation techniques and real-time closed-loop interaction models to enhance adaptability and responsiveness. Future studies should also involve a larger participant pool and more realistic experimental scenarios to allow for more comprehensive analysis.

Overall, our findings highlight the importance of multimodal fusion and bio-inspired cognitive modeling in improving HRI. By refining these mechanisms, we can further bridge the gap between human and robotic communication, fostering more intuitive and seamless interactions in shared environments.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Numbers 23H03468 and 23K18495, and JST JPMJAP2305. We also thank Slim Ouni, responsible of the MULTISPEECH team at LORIA for providing us with the Furhat robot for the experiments.

References

[1] C. Trevarthen and P. Hubley, "Secondary intersubjectivity," Action, gesture and symbol: The emergence of language, pp. 183–229, 1978.

[2] S. Gallagher, "Understanding others: embodied social cognition," in Handbook of cognitive science. Elsevier, 2008, pp. 437–452.

[3] H. Admoni, T. Weng, and B. Scassellati, "Modeling communicative behaviors for object references in human-robot interaction," in 2016 IEEE International Conference on Robotics and Automation (ICRA). IEEE, May 2016, pp. 3352–3359.

[4] M. Kimoto, T. Iio, M. Shiomi, I. Tanev, K. Shimohara, and N. Hagita, "Improvement of object reference recognition through human robot alignment," in 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, Aug. 2015.

[5] 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, Nov. 2022, pp. 140–145.

[6] P. Gao, B. Reily, S. Paul, and H. Zhang, "Visual reference of ambiguous objects for augmented reality-powered human-robot communication in a shared workspace," in Virtual, Augmented and Mixed Reality. Design and Interaction, ser. Lecture notes in computer science. Cham: Springer International Publishing, 2020, pp. 550–561.
[7] S. Hanifi, E. Maiettini, M. Lombardi, and L. Natale, "A pipeline for estimating human attention toward objects with on-board cameras on the iCub humanoid robot," Front. Robot. AI, vol. 11, p. 1346714, Oct. 2024.
[8] T. Iio, M. Shiomi, K. Shinozawa, K. Shimohara, M. Miki, and N. Hagita, "Lexical entrainment in human robot interaction: Do humans use their vocabulary to robots?" Int. J. Soc. Robot., vol. 7, no. 2, pp. 253–263, Apr. 2015.
[9] R. Ros, S. Lemaignan, E. A. Sisbot, R. Alami, J. Steinwender, K. Hamann, and F. Warneken, "Which one? grounding the referent based on efficient human-robot interaction," in 19th International Symposium in Robot and Human Interactive Communication. IEEE, Sept. 2010, pp. 570–575.
[10] S.-i. Amari, "Dynamics of pattern formation in lateral-inhibition type neural fields," Biological cybernetics, vol. 27, no. 2, pp. 77–87, 1977.
[11] G. Schöner and J. P. Spencer, Dynamic thinking: A primer on dynamic field theory. Oxford University Press, 2016.
[12] O. Lomp, M. Richter, S. K. U. Zibner, and G. Schöner, "Developing dynamic field theory architectures for embodied cognitive systems with cedar," Front. Neurorobot., vol. 10, p. 14, Nov. 2016.
[13] G. Knips, S. K. U. Zibner, H. Reimann, I. Popova, and G. Schoner, "A neural dynamics architecture for grasping that integrates perception and movement generation and enables on-line updating," in 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, Sept. 2014, pp. 646–653.
[14] M. Richter, Y. Sandamirskaya, and G. Schöner, "A robotic architecture for action selection and behavioral organization inspired by human cognition," in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2012, pp. 2457–2464.
[15] M. Tomasello, The cultural origins of human cognition. Harvard university press, 2009.
[16] S. Gallagher, "Joint attention, joint action, and participatory sense making," Alter. Revue de phénoménologie, no. 18, pp. 111–123, 2010.
[17] H. F. Chame, A. Clodic, and R. Alami, "Top-jam: A bio-inspired topology-based model of joint attention for human-robot interaction," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 7621–7627.
[18] H. F. Chame and R. Alami, "Aego: Modeling attention for hri in ego-sphere neural networks," in 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 2549–2555.
[19] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in Cognitive Behavioural Systems, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 114–130.
[20] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," International journal of social robotics, vol. 1, pp. 71–81, 2009.
[21] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," Pattern Recognition, vol. 47, no. 6, pp. 2280–2292, 2014.
[22] S. Kreiss, L. Bertoni, and A. Alahi, "OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association," IEEE Trans. Intell. Transp. Syst., vol. 23, no. 8, pp. 13 498–13 511, Aug. 2022.
[23] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2018.