JA-EGO: The Egocentric Dimension of Joint Attention in HRI

Hendry F. Chame[1] and Rachid Alami[2]

***Abstract*— Despite important progress in recent years, there is a long way to go for including social robots capable of adaptation, interaction and communication in our environments. Our research is concerned with the study of social cognition in HRI, in particular with communication skills relying on joint attention (JA) and knowledge sharing. Since JA involves low-level cognitive skills, we take into account the implications addressed by Moravec's Paradox and focus on the aspect of knowledge representation. By embracing embodiment and 4E cognition principles, we investigate the notion of egocentric localization and propose a neural network representation suited for joint attention research named JA-EGO. Inspired by *dynamic fields theory*, the model consists in a dynamical system representation which fuses information from immediate sensation and provides the means for attention selection and working memory under the influence of top-down and bottom-up modulation processes of attention in HRI. We firstly studied the selection model in simulations and analyzed some application scenarios. We then conducted a real experiment with the robot Pepper considering propioception, vision and basic proxemics. Results showed that JA-EGO is a convenient representation for HRI situations allowing the human and the robot to share attention and knowledge about objects in the environment.**

I. INTRODUCTION

According to Moravec's paradox, although machines can perform tasks at adults' level of intelligence such as inductive and deductive reasoning, they have tremendous difficulty with sensory-motor or social skills, as demonstrated by a one-year-old child. Behind this paradox remains the question in artificial intelligence research of what sort of knowledge representation would be suitable for allowing a machine to accomplish cognitive tasks, which has important philosophical implications. Thus, recent studies have contrasted the Cartesian (traditional) view of social cognition as a process confined to the brain, to the notion of an *embodied*, *embedded*, *enacted* and *extended* process, unfolding between the brain, the body and environment in interaction; a perspective known as *4E cognition* [8].

In agreement with the perspective of 4E cognition, we believe that for social robots to leave the lab and adapt to human environments, it is crucial to provide them with forms of behavior regulation which take into account the dynamics of human low-level social skill processes such as the capacity of engaging in *joint attention* (JA), and the possibility of such processes be modulated in direct interaction. Moreover, we study JA as a multi-dimensional construct involving cognitive skills so constituting forms of social attention at distinct levels of interaction and knowledge sharing [14].

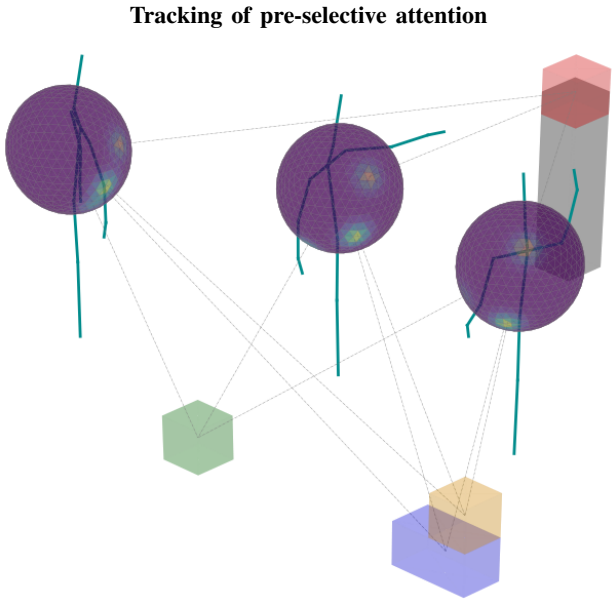

Fig. 1. The scene shows a situation where two agents are interacting about the orange object, whereas another agent is leaving the scene. Three other objects are present the effect of each object agents ego-spherical localization at a pre-selective stage is shown.

As a continuation of previous research in which we proposed a model for tracking JA in HRI within a topology-based representation constituting a *scale of jointness* [4], in this work we investigate the more fundamental aspect of attention selection and how such mechanism could allow the emergence of JA in human-robot interaction from top-down and bottom-up modulation processes. For this, as shown in Fig. 1, we suppose unconstrained situations where agents can become interested in objects on the environment and eventually share attention and knowledge about it (e.g. situations like asking someone for direction or commenting about salient stimulus like a noise or an object).

Tacking into account the considerations above, we propose the model named JA-EGO for tracking the attention focus of agents as represented from egocentric perspective, and resulting from acquisitions of robot's on-board sensory. For this, we model the evolution of attention selection as a dynamical system process represented by a recurrent neural network inspired on *dynamic neural fields* (DNF) theory [1]. Thus, JA-EGO can be used for track egocentric focus of attention from spherical localization references.

[1]LORIA-CNRS (NeuroRhythms team). Address: Campus Scientifique, 615 Rue du Jardin-Botanique, 54506 Vandœuvre-lès-Nancy, France. E-mail: hendry.ferreira-chame@loria.fr.

[2]LAAS-CNRS (Robotics and InteractionS team). Address: 7 Av. du Colonel Roche, 31400 Toulouse, France. E-mail: rachid.alami@laas.fr.

This document is organized as follows: Section II discusses previous works and argues how our contribution would help to advance the state of the art in the field. Section III presents the mathematical definition of the model and discusses theoretical assumptions behind it. Section IV presents the methodology of the work, consisted in: a) studying in simulation the attention tracking mechanism proposed, which relies on the selection process, and showing potential applications, and b) conducting an experiment with the robot Pepper including propioception, vision and basic proxemics about objects in the environment. Section V reports the study's results, and Section VI presents conclusions and future perspectives.

II. PREVIOUS WORK

Contributions in the field have achieved impressive results by exploring attention saliency in sensory egocentric representations under the effect of modulation processes (e.g. bottom-up [10] and top-down [2]). Although most research have considered environment exploration tasks, so the robot can focus on learning new things based on novelty, some have studied social cognition or joint attention tasks inspired on psychological theories of attention (e.g. [10]).

Overall, the representation proposed for attention modeling has neglected bio-inspiration on neural systems, consisting in storage arrays for data indexed by spherical tessellation mapping (see [9]). Consequently, the dynamics of attention has tended to be modeled as a process governed by knowledge represented in the form of production rules, where the possibility of compositionality from low-level sensory to higher-level decision space has been of less importance.

Another limitation of previous research is considering the robot as the only one in interaction given with embodied ego-sensory mapping representations, so data coming from human agents is mostly represented in the robot ego-sensory sphere. In our opinion, this would be a too much egocentric view of joint attention in HRI. We believe that when keeping track of embodied relations between agents and objects in the environment as a distributed dynamical system, the robot would be able to take more informed decisions for operating in such environment. Hence, we propose that attention selection should be tracked in egocentric representations simultaneously for participants in interaction.

Our previous research has also constituted relevant steps in the direction of developing the current study which is worth mentioning. In [3] an embodied egocentric cylindrical localization mechanism for attention selection was proposed for autonomous positioning with respect to object in the environment. In [4] the model TOP-JAM was proposed as a means for JA tracking in HRI from allocentric references. In [5] joint attention in HRI is studied for a providing guidance task in a shopping mall. [Many papers from the RIS team could be cited: human-aware planning, joint action, etc. To be confirmed with Rachid].

To summarize ...

III. THE MATHEMATICAL MODEL

Feature integration theory (FIT) [15] of visual attention studies the role attention plays in selecting and fusing complex information. According to FIT, at a pre-attention level the perceptual system is constituted by separate maps, each encoding feature salience (e.g., color, edges, shapes) which are lately combined at a attention stage. A biologically plausible architecture has been proposed from FIT in [7], an the implementation by [6] has been employed for visually-guided autonomous navigation (e.g. [13]).

Let the agent's peripersonal space be represented by a localization topology defined by a vertex set, resulting from the tessellation of an icosahedron polyhedron, which approximates a spherical region around the agent. From above theoretical considerations, we propose a model of attention selection inspired by FIT and DNF theory [1]. For this we consider a pre-attention phase where proprioceptive and exteroceptive stimuli excite the ego-space topology, encoded by a recurrent neural network modeling a dynamical system, under the influence of top-down and bottom-up modulation processes. In a posterior stage, attention results from a competition process between received stimulation.

A. Pre-attention phase

Let the activation of the i^{th} neuron encode the dynamics of stimulation affecting a location $\mathbf{x}_i$ in 3D Cartesian coordinates at a polyhedron surface representing the agent's ego-space, such that

$$\tau_{\text{u}}\dot{\mathbf{u}}_{i(t)} = -\mathbf{u}_{i(t-1)} + h_{\text{u}} + \sum_j (\mathbf{U}_{ij} + \epsilon)\mathbf{u}_{i(t-1)} + \mathbf{s}_{i(t,\Xi)} \quad (1)$$

According to the principle of local interconnections [11], the interaction strength $\mathbf{U}_{ij}$ between neurons i and j is selected so proximal locations have stronger interaction. Hence, multivariate Gaussian weights are selected, such that

$$\mathbf{U}_{ij}(|\mathbf{x}_i - \mathbf{x}_j|) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\text{t}}\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)\right)}{\sqrt{2\pi^3|\mathbf{\Sigma}|}} \quad (2)$$

The term $\mathbf{s}_{i(t,\Xi)}$ represents the stimulation received at time instant t affecting the ego-space locations Ξ. As it is going to be detained in Section IV, this term included stimulation from both top-down and bottom-up processes. Finally, in (1) h_{u} corresponds to the activation resting state, τ_{u} is the a time constant, and ϵ is a global inhibition factor affecting lateral interactions between neurons.

B. Attention selection phase

Let the activation of the i^{th} neuron represent the dynamics of an attention selection process at a particular location in the ego-space, such that

$$\tau_{\text{a}}\dot{\mathbf{a}}_{i(t)} = -\mathbf{a}_{i(t-1)} + h_{\text{a}} + \sum_j \mathbf{A}_{ij} f\left(\mathbf{a}_{(t-1)}, \mathbf{u}_{(t)}\right) \quad (3)$$

Since inhibitory neural interaction has been associated with selection mechanisms [12], lateral interaction $\mathbf{A}_{ij}$ between neuron i and j is modeled so

$$\mathbf{A}_{ij}(|\mathbf{x}_i - \mathbf{x}_j|) = 1 - \varphi \mathbf{U}_{ij} \tag{4}$$

with $\varphi = \max(\mathbf{U}_{i:})^{-1}$ a scaling factor. The activation function f in (3) is defined such that

$$f\left(\mathbf{a}_{(t-1)}, \mathbf{u}_{(t)}\right) = \text{sigmoid}\left(\alpha\left(\mathbf{a}_{i(t-1)} + \varsigma \mathbf{u}_{i(t)}\right)\right) \tag{5}$$

IV. METHODOLOGY

A. Simulations

B. Experiment

V. RESULTS

VI. CONCLUSIONS

ACKNOWLEDGMENT

This research was only possible with the collaboration of colleagues from the robotics teams of both LAAS-CNRS (project ANITI) and LORIA-CNRS (project Creativ'Lab).

REFERENCES

[1] AMARI, S.-I. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological cybernetics 27*, 2 (1977), 77–87.

[2] BODIROZA, S., SCHILLACI, G., AND HAFNER, V. V. Robot ego-sphere: An approach for saliency detection and attention manipulation in humanoid robots for intuitive interaction. In *2011 11th IEEE-RAS International Conference on Humanoid Robots* (2011), IEEE, pp. 689–694.

[3] CHAME, H. F., AND CHEVALLEREAU, C. Grounding humanoid visually guided walking: From action-independent to action-oriented knowledge. *Information Sciences 352* (2016), 79–97.

[4] CHAME, H. F., CLODIC, A., AND ALAMI, R. 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)* (2023).

[5] HEIKKILÄ, P., LAMMI, H., NIEMELÄ, M., BELHASSEIN, K., SARTHOU, G., TAMMELA, A., CLODIC, A., AND ALAMI, R. Should a robot guide like a human? a qualitative four-phase study of a shopping mall robot. In *Social Robotics: 11th International Conference, ICSR 2019, Madrid, Spain, November 26–29, 2019, Proceedings 11* (2019), Springer, pp. 548–557.

[6] ITTI, L., KOCH, C., AND NIEBUR, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence 20*, 11 (1998), 1254–1259.

[7] KOCH, C., AND ULLMAN, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology 4*, 4 (1985), 219–227.

[8] NEWEN, A., GALLAGHER, S., AND DE BRUIN, L. 34E Cognition: Historical Roots, Key Concepts, and Central Issues. In *The Oxford Handbook of 4E Cognition*. Oxford University Press, 09 2018.

[9] PETERS, R. A., HAMBUCHEN, K. A., AND BODENHEIMER, R. E. The sensory ego-sphere: a mediating interface between sensors and cognition. *Autonomous Robots 26*, 1 (2009), 1–19.

[10] RUESCH, J., LOPES, M., BERNARDINO, A., HORNSTEIN, J., SANTOS-VICTOR, J., AND PFEIFER, R. Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub. In *2008 IEEE International Conference on Robotics and Automation* (2008), IEEE, pp. 962–967.

[11] SAMSONOVICH, A., AND MCNAUGHTON, B. L. Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience 17*, 15 (1997), 5900–5920.

[12] SCHÖNER, G., AND SPENCER, J. P. *Dynamic thinking: A primer on dynamic field theory*. Oxford University Press, 2016.

[13] SIAGIAN, C., CHANG, C. K., AND ITTI, L. Autonomous mobile robot localization and navigation using a hierarchical map representation primarily guided by vision. *Journal of Field Robotics 31*, 3 (2014), 408–440.

[14] SIPOSOVA, B., AND CARPENTER, M. A new look at joint attention and common knowledge. *Cognition 189* (2019), 260–274.

[15] TREISMAN, A. M., AND GELADE, G. A feature-integration theory of attention. *Cognitive psychology 12*, 1 (1980), 97–136.