frontiers in Neurorobotics

EDITORIAL
published: 08 February 2022
doi: 10.3389/fnbot.2022.848065

Editorial: Active Vision and Perception in Human-Robot Collaboration

Dimitri Ognibene[1,2]*, *Tom Foulsham*[3]*, *Letizia Marchegiani*[4] *and Giovanni Maria Farinella*[5]

[1] *Department of Psychology, Università degli Studi di Milano-Bicocca, Milan, Italy,* [2] *School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom,* [3] *Department of Psychology, University of Essex, Colchester, United Kingdom,* [4] *Department of Electronic Systems, Aalborg University, Aalborg, Denmark,* [5] *Department of Mathematics and Computer Science, University of Catania, Catania, Italy*

Keywords: active vision, social perception, intention prediction, egocentric vision, natural human-robot interaction, human-robot collaboration

Editorial on the Research Topic

Active Vision and Perception in Human-Robot Collaboration

1. APPLYING PRINCIPLES OF ACTIVE VISION AND PERCEPTION TO ROBOTICS

Finding the underlying design principles which allow humans to adaptively find and select relevant information (Tistarelli and Sandini, 1993; Findlay and Gilchrist, 2003; Krause and Guestrin, 2007; Friston et al., 2015; Ognibene and Baldassare, 2015; Bajcsy et al., 2017; Jayaraman and Grauman, 2018; Ballard and Zhang, 2021) is important for Robotics and related fields (Shimoda et al., 2021; Straub and Rothkopf, 2021). Active inference, which has recently become influential in computational neuroscience, is a normative framework proposing one such principle: action, perception, and learning are the result of minimization of variational free energy, a form of prediction error. Active vision and visual attention must involve balancing long and short-term predictability and have been the focus of several previous modeling efforts (Friston et al., 2012, 2015; Mirza et al., 2016). Parr et al. review several probabilistic models which are needed for different aspects of biological active vision. They propose a mapping between the involved operations and particular brain structures.

Van de Maele et al. use deep neural networks to implement an active inference model of active perception, working in a rendered 3D environment similar to a robotics setting. Their network learns the necessary generative model of visual data and when tested shows interesting exploratory behavior. However, they also highlight the many computational challenges that must be solved before such a system can be tested on real robots with tasks to perform and humans to interact with.

Due to this high computational complexity, in practice, robotics scenarios often substitute optimal active perception strategies with flexible architectures that allow the development of behaviors for different tasks. Martin et al. introduce a scalable framework for service robots that efficiently encodes precompiled perceptual needs in a distributed knowledge graph.

OPEN ACCESS

Edited and reviewed by:
Florian Röhrbein,
Technische Universität Chemnitz, Germany

***Correspondence:**
Dimitri Ognibene
dimitri.ognibene@unimib.it
Tom Foulsham
foulsham@essex.ac.uk

Received: *03 January 2022*
Accepted: *12 January 2022*
Published: *08 February 2022*

Citation:
Ognibene D, Foulsham T, Marchegiani L and Farinella GM (2022) Editorial: Active Vision and Perception in Human-Robot Collaboration. Front. Neurorobot. 16:848065. doi: 10.3389/fnbot.2022.848065

2. THE CHALLENGE OF SOCIAL INTERACTIONS

Social interactions involve non trivial tasks, such as intention prediction (Sebanz and Knoblich, 2009; Ognibene and Demiris, 2013; Donnarumma et al., 2017a), activity recognition (Ansuini et al., 2015; Lee et al., 2015; Sanzari et al., 2019) or even simple gesture recognition (e.g., pointing at a target), which may require perceptual policies that are difficult to precompile. This is because they are contingent on previous observations, hierarchically organized (Proietti et al., 2021), and must extend over time, space and scene elements which may not be always visible (Ognibene et al., 2013). While some active recognition systems and normative models for action and social interactions have already been proposed (Ognibene and Demiris, 2013; Lee et al., 2015; Donnarumma et al., 2017a; Ognibene et al., 2019b), it is not completely clear what strategy humans adopt in such tasks, not least because of the heterogeneity of the stimuli. Salatiello et al. introduce a validated generative model of social interactions that can generate highly-controlled stimuli useful for conducting behavioral and neuroimaging studies, but also for the development and validation of computational models.

An alternative approach is to simplify the challenges posed by social interactions by adopting a strict signaling and interaction protocol. Papanagiotou et al. investigate a collaborative human-robot industrial assembly task powered by an egocentric perspective (where the camera shares the user's viewpoint) and where the system must recognize gestures.

3. TRANSPOSING ACTIVE PERCEPTION STRATEGIES FROM ECOLOGICAL INTERACTIONS TO HUMAN ROBOT COLLABORATION

However, a better understanding of active vision and eye movements during social interaction may lead to more natural interfaces. Of course one of the most important ways in which humans interact is through speech. While there is a long tradition of studying the relationship between speech and gaze for behavior analysis, there is much less investigation with modern computational tools. Aydin et al. take a step in this direction by providing a multimodal analysis and predictors of eye contact data. This analysis reveals patterns in real conversation - such as the tendency for speakers to look away from their partner (Ho et al., 2015). In a similar context, D'Amelio and Boccignone introduce a novel computational model replicating visual attention behaviors while observing groups speaking on video. The model is based on a foraging framework where individuals must seek out socially relevant information. Testing these models with social robots would enable principled and natural conversational interaction but also determine if humans would find it effective (Palinko et al., 2016).

In ecological conditions where participants act in the world, gaze dynamics can also be highly informative about intentions (Land, 2006; Tatler et al., 2011; Borji and Itti, 2014; Ballard and Zhang, 2021). Wang et al. verify this hypothesis in a manipulation and assembly task to create a gaze-based intentions predictor covering multiple levels of the action hierarchy (action primitives, actions, activities) and study the factors that affect response time and generalization over different layouts.

4. SPECIFICITY OF GAZE BEHAVIORS DURING HUMAN ROBOT INTERACTION

When Fuchs and Belardinelli studied the impact of a similar ecological approach to perform an actual teleoperation task, they found that gaze dynamics are still informative and usable. Interestingly, the patterns observed might partially differ from those in natural eye-hand coordination, probably due to limited confidence in robot behavior. While they expect that users would eventually learn an effective strategy, they suggest that more adaptive and personalized models of the effect of robot behavior on user gaze would further improve the interaction.

Eldardeer et al. developed a biologically inspired multimodal framework for emergent synchronization and joint attention in human-humanoid-robot interaction. The resulting interaction was robust and close to natural, but the robot showed slower audio localization due to ambient noise. While specific audio processing methods (Marchegiani and Newman, 2018; Tse et al., 2019) may ameliorate this issue, it highlights the importance of a detailed understanding of the temporal aspects of active perception and attention resulting from the interplay between exploration and communication demands in the human robot collaboration context (Donnarumma et al., 2017b; Ognibene et al., 2019a).

As these works show, human attentional and active perception strategies while interacting with a robot are interesting in their own right (Rich et al., 2010; Moon et al., 2014; Admoni and Scassellati, 2017). In ecological conditions, behavior with a robot will be different from performing the task alone (free manipulation), using a tool and even from collaborating with a human partner. At the same time, aspects of each situation will be reproduced, since robots can be perceived as body extensions, tools or companions. Following Fuchs and Belardinelli, we should expect the balance between these factors to shift after experience with a particular design of robot (Sailer et al., 2005).

To understand how humans and robots interact (and how they can interact better), a sensible place to start is by comparing this to how humans interact with each other. Czeszumski et al. report differences in the way that participants respond to errors in a collaborative task, depending on whether they are interacting with a robot or another person. Moreover, there were differences in neural activity in the two situations. This is an example of how researchers can begin to understand communication between humans and robots, while also highlighting potential brain based interfaces which could improve this communication.

5. CONCLUSIONS

Ultimately this collection of articles highlights the potential benefits of deepening our understanding of active perception

and the resulting egocentric behavior in the context of human robot collaboration. Some of the challenges for future research are to:

1. Scale normative frameworks to deal with realistic tasks and environments (see Van de Maele et al. and Ognibene and Demiris, 2013; Lee et al., 2015; Donnarumma et al., 2017a; Ognibene et al., 2019b).
2. Enable scalable frameworks to deal with the uncertain, multimodal, distributed, and dynamic nature of social interactions (see Eldardeer et al., Martin et al., and Ognibene et al., 2013; Schillaci et al., 2013).
3. Deepen the integration of user state, e.g., beliefs (Bianco and Ognibene, 2019; Perez-Osorio et al., 2021), inference, into predictive models.
4. Improve egocentric perception (Grauman et al., 2021) and interfaces (see Papanagiotou et al.) to build advanced wearable assistant and to balance usability and robustness.
5. Understand and exploit the peculiarities of Human AI interactions (see Fuchs and Belardinelli, Czeszumski et al., and Paletta et al., 2019).
6. Provide new benchmarks and datasets (see Salatiello et al. and Ammirato et al., 2017; Damen et al., 2018; Calafiore et al., 2021).

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

DO and TF were supported by the European Union's Horizon 2020 research and innovation programme under grant agreement (No. 824153 POTION).

REFERENCES

Admoni, H., and Scassellati, B. (2017). Social eye gaze in human-robot interaction: a review. *J. Human Robot Interact*. 6, 25–63. doi: 10.5898/JHRI.6.1.Admoni

Ammirato, P., Poirson, P., Park, E., Košecká, J., and Berg, A. C. (2017). "A dataset for developing and benchmarking active vision," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE), Singapore. 1378–1385.

Ansuini, C., Cavallo, A., Bertone, C., and Becchio, C. (2015). Intentions in the brain: the unveiling of mister hyde. *Neuroscientist* 21, 126–135. doi: 10.1177/1073858414533827

Bajcsy, R., Aloimonos, Y., and Tsotsos, J. K. (2017). Revisiting active perception. *Auton. Robots*. 42, 177–196. doi: 10.1007/s10514-017-9615-3

Ballard, D. H., and Zhang, R. (2021). The hierarchical evolution in human vision modeling. *Top. Cogn. Sci*. 13, 309–328. doi: 10.1111/tops.12527

Bianco, F., and Ognibene, D. (2019). "Functional advantages of an adaptive theory of mind for robotics: a review of current architectures," in *2019 11th Computer Science and Electronic Engineering (CEEC)* (Colchester: IEEE), 139–143.

Borji, A., and Itti, L. (2014). Defending yarbus: eye movements reveal observers' task. *J. Vis*. 14, 29–29. doi: 10.1167/14.3.29

Calafiore, C., Foulsham, T., and Ognibene, D. (2021). Humans select informative views efficiently to recognise actions. *Cogn. Proc*. 22, 48–48. doi: 10.1007/s10339-021-01058-x

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., et al. (2018). "Scaling egocentric vision: the epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Cham: Springer), 720–736

Donnarumma, F., Costantini, M., Ambrosini, E., Friston, K., and Pezzulo, G. (2017a). Action perception as hypothesis testing. *Cortex* 89:45–60. doi: 10.1016/j.cortex.2017.01.016

Donnarumma, F., Dindo, H., and Pezzulo, G. (2017b). Sensorimotor communication for humans and robots: improving interactive skills by sending coordination signals. *IEEE Trans. Cogn. Dev. Syst*. 10, 903–917. doi: 10.1109/TCDS.2017.2756107

Findlay, J. M., and Gilchrist, I. D. (2003). *Active Vision: The Psychology of Looking and Seeing*. Oxford: Oxford University Press.

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., and Pezzulo, G. (2015). Active inference and epistemic value. *Cogn. Neurosci*. 6, 187–214. doi: 10.1080/17588928.2015.1020053

Friston, K., Thornton, C., and Clark, A. (2012). Free-energy minimization and the dark-room problem. *Front. Psychol*. 3:130. doi: 10.3389/fpsyg.2012.00130

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., et al. (2021). Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint* arXiv:2110.07058.

Ho, S., Foulsham, T., and Kingstone, A. (2015). Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PLoS ONE* 10:e0136905. doi: 10.1371/journal.pone.0136905

Jayaraman, D., and Grauman, K. (2018). "Learning to look around: intelligently exploring unseen environments for unknown tasks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE), 1238–1247. doi: 10.1109/CVPR.2018.00135

Krause, A., and Guestrin, C. (2007). "Near-optimal observation selection using submodular functions," in *AAAI'07: Proceedings of the 22nd National Conference on Artificial Intelligence, Vol. 7*, Vancouver. 1650–1654.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Prog. Retin Eye Res*. 25, 296–324. doi: 10.1016/j.preteyeres.2006.01.002

Lee, K., Ognibene, D., Chang, H. J., Kim, T.-K., and Demiris, Y. (2015). Stare: spatio-temporal attention relocation for multiple structured activities detection. *IEEE Trans. Image Process*. 24, 5916–5927. doi: 10.1109/TIP.2015.2487837

Marchegiani, L., and Newman, P. (2018). Listening for sirens: locating and classifying acoustic alarms in city scenes. *arXiv preprint* arXiv:1810.04989.

Mirza, M. B., Adams, R. A., Mathys, C. D., and Friston, K. J. (2016). Scene construction, visual foraging, and active inference. *Front. Comput. Neurosci*. 10:56. doi: 10.3389/fncom.2016.00056

Moon, A., Troniak, D. M., Gleeson, B., Pan, M. K., Zheng, M., Blumer, B. A., et al. (2014). "Meet me where i'm gazing: how shared attention gaze affects human-robot handover timing," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*, Bielefeld. 334–341.

Ognibene, D., and Baldassare, G. (2015). Ecological active vision: four bioinspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Trans. Auton. Ment. Dev*. 7, Beijing, 3–25. doi: 10.1109/TAMD.2014.2341351

Ognibene, D., Chinellato, E., Sarabia, M., and Demiris, Y. (2013). Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspirat. Biomimet*. 8, 035002. doi: 10.1088/1748-3182/8/3/035002

Ognibene, D., and Demiris, Y. (2013). "Towards active event recognition," in *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2495–2501.

Ognibene, D., Giglia, G., Marchegiani, L., and Rudrauf, D. (2019a). Implicit perception simplicity and explicit perception complexity in sensorimotor comunication. *Phys. Life Rev*. 28, 36–38. doi: 10.1016/j.plrev.2019.01.017

Ognibene, D., Mirante, L., and Marchegiani, L. (2019b). "Proactive intention recognition for joint human-robot search and rescue missions through monte-carlo planning in pomdp environments," in *International Conference on Social Robotics* (Berlin; Heidelberg: Springer), 332–343.

Paletta, L., Pszeida, M., Ganster, H., Fuhrmann, F., Weiss, W., Ladstätter, S., et al. (2019). "Gaze-based human factors measurements for the evaluation of intuitive human-robot collaboration in real-time," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)* (Zaragoza: IEEE), 1528–1531.

Palinko, O., Rea, F., Sandini, G., and Sciutti, A. (2016). "Robot reading human gaze: why eye tracking is better than head tracking for human-robot collaboration," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Daejeon: IEEE), 5048–5054.

Perez-Osorio, J., Wiese, E., and Wykowska, A. (2021). "Theory of mind and joint attention," in *The Handbook on Socially Interactive Agents: 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 1: Methods, Behavior, Cognition, 1st Edn* (New York, NY: Association for Computing Machinery), 311–348.

Proietti, R., Pezzulo, G., and Tessari, A. (2021). An active inference model of hierarchical action understanding, learning and imitation. *PsyArXiv*. doi: 10.31234/osf.io/ms95f

Rich, C., Ponsler, B., Holroyd, A., and Sidner, C. L. (2010). "Recognizing engagement in human-robot interaction," in *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Osaka: IEEE), 375–382.

Sailer, U., Flanagan, J. R., and Johansson, R. S. (2005). Eye-hand coordination during learning of a novel visuomotor task. *J. Neurosci.* 25, 8833–8842. doi: 10.1523/JNEUROSCI.2658-05.2005

Sanzari, M., Ntouskos, V., and Pirri, F. (2019). Discovery and recognition of motion primitives in human activities. *PLoS ONE* 14:e0214499. doi: 10.1371/journal.pone.0214499

Schillaci, G., Bodiroža, S., and Hafner, V. V. (2013). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *Int. J. Soc. Rob.* 5, 139–152. doi: 10.1007/s12369-012-0174-7

Sebanz, N., and Knoblich, G. (2009). Prediction in joint action: what, when, and where. *Top. Cogn. Sci.* 1, 353–367. doi: 10.1111/j.1756-8765.2009.01024.x

Shimoda, S., Jamone, L., Ognibene, D., Nagai, T., Sciutti, A., Costa-Garcia, A., et al. (2021). What is the role of the next generation of cognitive robotics? *Adv. Rob.* 1–14. doi: 10.1080/01691864.2021.2011780

Straub, D., and Rothkopf, C. A. (2021). Looking for image statistics: Active vision with avatars in a naturalistic virtual environment. *Front. Psychol.* 12:431. doi: 10.3389/fpsyg.2021.641471

Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: reinterpreting salience. *J. Vis.* 11, 5–5. doi: 10.1167/11.5.5

Tistarelli, M., and Sandini, G. (1993). On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 401–410. doi: 10.1109/34.206959

Tse, T. H. E., De Martini, D., and Marchegiani, L. (2019). "No need to scream: robust sound-based speaker localisation in challenging scenarios," in *International Conference on Social Robotics* (Berlin; Heidelberg: Springer), 176–185.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.