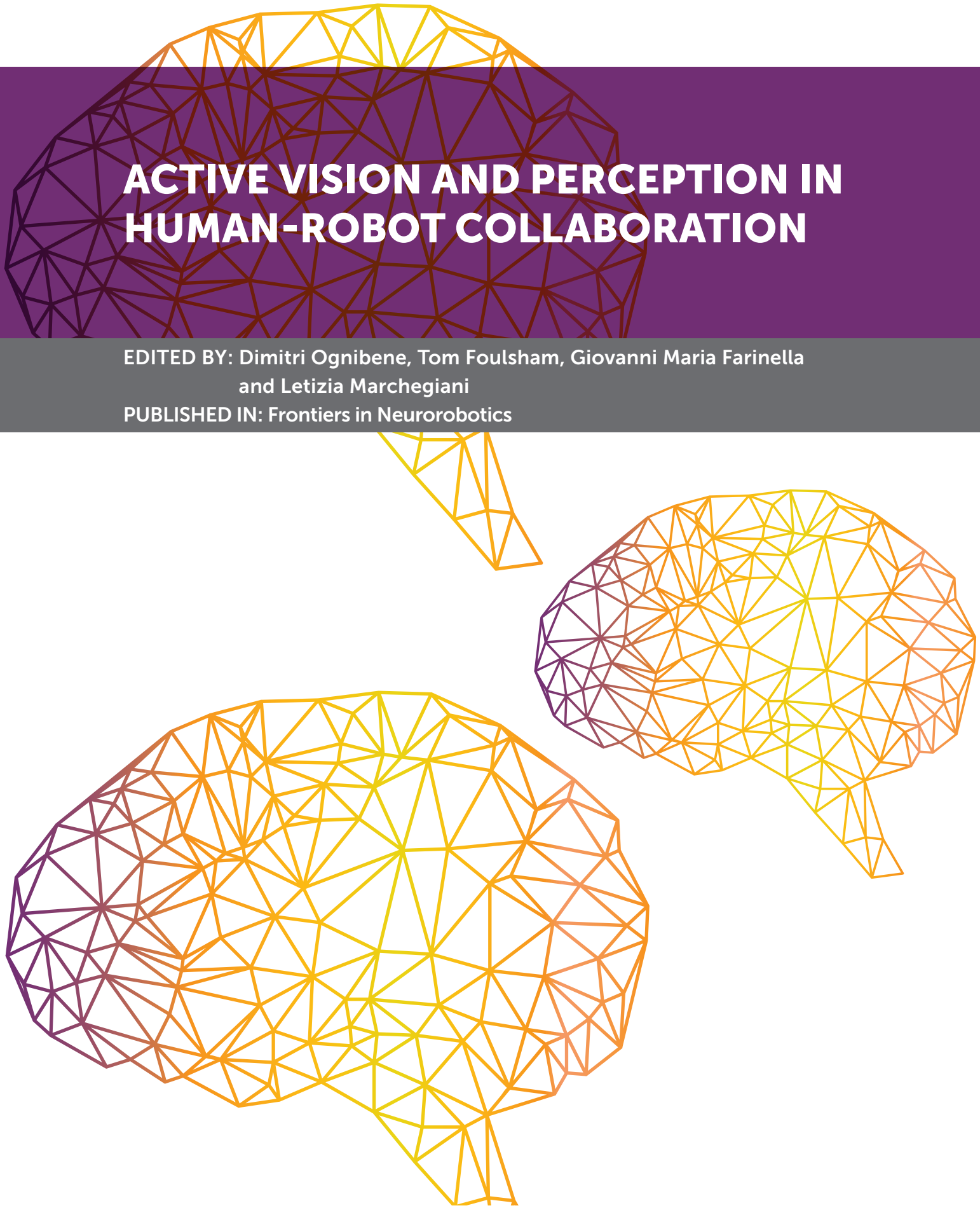
ACTIVE VISION AND PERCEPTION IN HUMAN-ROBOT COLLABORATION

EDITED BY: Dimitri Ognibene, Tom Foulsham, Giovanni Maria Farinella and Letizia Marchegiani

PUBLISHED IN: Frontiers in Neurorobotics



frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-599-9
DOI 10.3389/978-2-88974-599-9

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

ACTIVE VISION AND PERCEPTION IN HUMAN-ROBOT COLLABORATION

Topic Editors:
Dimitri Ognibene, University of Milano-Bicocca, Italy
Tom Foulsham, University of Essex, United Kingdom
Giovanni Maria Farinella, University of Catania, Italy
Letizia Marchegiani, Aalborg University, Denmark

Citation: Ognibene, D., Foulsham, T., Farinella, G. M., Marchegiani, L., eds. (2022). Active Vision and Perception in Human-Robot Collaboration. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-599-9

Table of Contents

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 15 October 2020
doi: 10.3389/fnbot.2020.567571

Toward Shared Autonomy Control Schemes for Human-Robot Systems: Action Primitive Recognition Using Eye Gaze Features

*Xiaoyu Wang, Alireza Haji Fathaliyan and Veronica J. Santos**

Biomechatronics Laboratory, Mechanical and Aerospace Engineering, University of California, Los Angeles, Los Angeles, CA, United States

The functional independence of individuals with upper limb impairment could be enhanced by teleoperated robots that can assist with activities of daily living. However, robot control is not always intuitive for the operator. In this work, eye gaze was leveraged as a natural way to infer human intent and advance action recognition for shared autonomy control schemes. We introduced a classifier structure for recognizing low-level action primitives that incorporates novel three-dimensional gaze-related features. We defined an action primitive as a triplet comprised of a verb, target object, and hand object. A recurrent neural network was trained to recognize a verb and target object, and was tested on three different activities. For a representative activity (making a powdered drink), the average recognition accuracy was 77% for the verb and 83% for the target object. Using a non-specific approach to classifying and indexing objects in the workspace, we observed a modest level of generalizability of the action primitive classifier across activities, including those for which the classifier was not trained. The novel input features of gaze object angle and its rate of change were especially useful for accurately recognizing action primitives and reducing the observational latency of the classifier.

Keywords: action primitive recognition, activities of daily living, eye gaze, gaze-object angle, human-robot systems, recurrent neural network, shared autonomy

OPEN ACCESS

Edited by:
Dimitri Ognibene,
University of Essex, United Kingdom

Reviewed by:
Hong Zeng,
Southeast University, China
Giacinto Barresi,
Italian Institute of Technology (IIT), Italy

***Correspondence:**
Veronica J. Santos
vjsantos@ucla.edu

Received: *30 May 2020*
Accepted: *13 August 2020*
Published: *15 October 2020*

Citation:
Wang X, Haji Fathaliyan A and Santos VJ (2020) Toward Shared Autonomy Control Schemes for Human-Robot Systems: Action Primitive Recognition Using Eye Gaze Features. Front. Neurorobot. 14:567571. doi: 10.3389/fnbot.2020.567571

INTRODUCTION

Activities of daily living (ADLs) can be challenging for individuals with upper limb impairment. The use of assistive robotic arms is an active area of research, with the aim of increasing an individual's functional independence (Groothuis et al., 2013). However, current assistive robotic arms, such as the Kinova arm and Manus arm, are controlled by joysticks that require operators to frequently switch between several modes for the gripper, including a position mode, an orientation mode, and an open/close mode (Driessen et al., 2001; Maheu et al., 2011). Users need to operate the arm from the gripper's perspective, in an unintuitive Cartesian coordinate space. Operators would greatly benefit from a control interface with a lower cognitive burden that can accurately and robustly inference human intent.

The long-term objective of this work is to advance shared autonomy control schemes so that individuals with upper limb impairment can more naturally control robots that assist with activities of daily living. Toward this end, the short-term goal of this study is to advance the use of eye gaze for action recognition. Our approach is to develop a neural-network based algorithm that exploits eye gaze-based information to recognize action primitives that could be used as modular, generalizable

building blocks for more complex behaviors. We define new gaze-based features and show that they increase recognition accuracy and decrease the observational latency (Ellis et al., 2013) of the classifier.

This article is organized as follows. Section Related Work outlines related work with respect to user interfaces for assistive robot arms and action recognition methods. Section Materials and Methods introduces the experimental protocol and proposed structure of an action primitive recognition model, whose performance is detailed in section Results. Section Discussion addresses the effects of input features on classifier performance and considerations for future real-time implementation. Contributions are summarized in section Conclusion.

RELATED WORK

User Interfaces for Assistive Robot Arms

Many types of non-verbal user interfaces have been developed for controlling assistive robot arms that rely on a variety of input signals, such as electrocorticographic (ECoG) (Hochberg et al., 2012), gestures (Rogalla et al., 2002), electromyography (EMG) (Bi et al., 2019), and electroencephalography (EEG) (Bi et al., 2013; Salazar-Gomez et al., 2017). Although ECoG has been mapped to continuous, high-DOF hand and arm motion (Chao et al., 2010; Wang et al., 2013), a disadvantage is that an invasive surgical procedure is required. Gesture-based interfaces often require that operators memorize mappings from specific hand postures to robot behaviors (Rogalla et al., 2002; Ghobadi et al., 2008; Raheja et al., 2010), which is not natural. EMG and EEG-based interfaces, although non-invasive and intuitive, require users to don and doff EMG electrodes or an EEG cap, which may be inconvenient and require a daily recalibration.

In this work, we consider eye gaze-based interfaces, which offer a number of advantages. Eye gaze is relatively easy to measure and can be incorporated into a user interface that is non-verbal, non-invasive, and intuitive. In addition, with this type of interface, it may be possible to recognize an operator's intent in advance, as gaze typically precedes hand motions (Hayhoe et al., 2003).

Numerous studies have reported on the use of eye gaze for robot control. In the early 2000's, the eyetracker was used as a *direct substitute* for a handheld mouse such that the gaze point on a computer display designates the cursor's position, and blinks function as button clicks (Lin et al., 2006; Gajwani and Chhabria, 2010). Since 2015, eye gaze has been used to communicate a 3D *target position* (Li et al., 2015a, 2017; Dziemian et al., 2016; Li and Zhang, 2017; Wang et al., 2018; Zeng et al., 2020) for directing the movement of the robotic end effector. No action recognition was required, as these methods assumed specific actions in advance, such as reach and grasp (Li et al., 2017), write and draw (Dziemian et al., 2016), and pick and place (Wang et al., 2018). Recently, eye gaze has been used to *recognize an action* from an a priori list. For instance, Shafti et al. developed an assistive robotic system that recognized subjects' intended actions (including reach to grasp, reach to drop, and reach to pour) using a finite state machine (Shafti et al., 2019).

In this work, we advance the use of eye gaze for action recognition. We believe that eye gaze control of robots is promising due to the non-verbal nature of the interface, the rich information that can be extracted from eye gaze, and the low cognitive burden on the operator during tracking of natural eye movements.

Action Representation and Recognition

Moeslund et al. described human behaviors as a composition of three hierarchical levels: (i) activities, (ii) actions, and (iii) action primitives (Moeslund et al., 2006). At the highest level, activities involve a number of actions and interactions with objects. In turn, each action is comprised of a set of action primitives. For example, the activity "making a cup of tea" is comprised of a series of actions, such as "move the kettle to the stove." This specific action can be further divided into three action primitives: "dominant hand reaches for the kettle," "dominant hand moves the kettle to the stove," and "dominant hand sets down the kettle onto the stove."

A great body of computer vision-based studies has already contributed to the recognition of activities of daily living such as walk, run, wave, eat, and drink (Lv and Nevatia, 2006; Wang et al., 2012; Vemulapalli et al., 2014; Du et al., 2015). These studies detected joint locations and joint angles as input features from external RGB-D cameras and classified ADLs using algorithms such as hidden Markov models (HMMs) and recurrent neural networks (RNNs).

Other studies leveraged egocentric videos taken by head-mounted cameras or eyetrackers (Yu and Ballard, 2002; Yi and Ballard, 2009; Fathi et al., 2011, 2012; Behera et al., 2012; Fathi and Rehg, 2013; Matsuo et al., 2014; Li et al., 2015b; Ma et al., 2016). Video preprocessing methods necessitated first subtracting the foreground and then detecting human hands and activity-relevant objects. Multiple features related to hands, objects, and gaze were then used as inputs for the action recognition using approaches such as HMMs, neural networks, and support vector machines (SVMs). Hand-related features included hand pose, hand location, relationship between left and right hand, and the optical flow field associated with the hand (Fathi et al., 2011; Ma et al., 2016). Object-related features included pairwise spatial relationships between objects (Behera et al., 2012), state changes of an object (open vs. closed) (Fathi and Rehg, 2013), and the optical flow field associated with objects (Fathi et al., 2011). The "visually regarded object," defined by Yi and Ballard (2009) as the object being fixated by the eyes, was widely used as the gaze-related feature (Yu and Ballard, 2002; Yi and Ballard, 2009; Matsuo et al., 2014). Some studies additionally extracted features such as color and texture near the visually regarded object (Fathi et al., 2012; Li et al., 2015b).

Due to several limitations, state-of-the-art action recognition methods cannot be directly applied to the intuitive control of an assistive robot via eye gaze. First, computer vision-based approaches to the automated recognition of ADLs have focused on the activity and action levels according to Moeslund's description of action hierarchy (Moeslund et al., 2006). Yet, state-of-the-art robots are not sophisticated enough to autonomously plan and perform these high-level behaviors. Second, eye

movements are traditionally used to estimate gaze point or gaze object alone (Yu and Ballard, 2002; Yi and Ballard, 2009; Matsuo et al., 2014). More work could be done to extract other useful features from spatiotemporal eye gaze data, such as time histories of gaze object angle and gaze object angular speed, which are further described in section Gaze-Related Quantities.

MATERIALS AND METHODS

Experimental Set-Up

This study was approved by the UCLA Institutional Review Board. The experimental setup and protocol were previously reported in our prior paper (Haji Fathaliyan et al., 2018). Data from 10 subjects are reported [nine males, one female; aged 18–28 years; two pure right-handers, six mixed right-handers, two neutral, per a handedness assessment (Zhang, 2012) based on the Edinburgh Handedness Inventory (Oldfield, 1971)]. Subjects were instructed to perform three bimanual activities involving everyday objects and actions: make instant coffee, make a powdered drink, and prepare a cleaning sponge (**Figure 1**). The objects involved in these three activities were selected from the benchmark Yale-CMU-Berkeley (YCB) Object Set (Calli et al., 2015). We refer to these objects as *activity-relevant objects* since they would be grasped and manipulated as subjects performed specific activities.

For Activity 1, subjects removed a pitcher lid, stirred the water in the pitcher, and transferred the water to a mug using two different methods (scooping with a spoon and pouring). For Activity 2, subjects were instructed to remove a coffee can lid, scoop instant coffee mix into a mug, and pour water from a pitcher into the mug. For Activity 3, subjects unscrewed a spray bottle cap, poured water from the bottle into a mug, sprayed the water onto a sponge, and screwed the cap back onto the bottle. In order to standardize the instructions provided to subjects, the experimental procedures were demonstrated via a prerecorded video. Each activity was repeated by the subject four times; the experimental setup was reset prior to each new trial.

A head-mounted eyetracker (ETL-500, ISCAN, Inc., Woburn, MA, USA) was used to track the subject's gaze point at 60 Hz with respect to a built-in egocentric scene camera. Per calibration data, the accuracy and precision of the eyetracker were ∼1.4 deg and 0.1 deg, respectively. The motion of the YCB objects, eyetracker, and each subject's upper limb were tracked at 100 Hz by six motion capture cameras (T-Series, Vicon, Culver City, CA, USA). A blackout curtain surrounded the subject's field of view in order to minimize visual distractions. A representative experimental trial is shown in **Supplementary Video 1**.

Gaze-Related Quantities

We extract four types of gaze-related quantities from natural eye movements as subjects performed Activities 1–3. The quantities include the *gaze object* (GO) (Yu and Ballard, 2002; Yi and Ballard, 2009; Matsuo et al., 2014) and *gaze object sequence* (GOS) (Haji Fathaliyan et al., 2018). This section describes how these quantities are defined and constructed. As described in section Input Features for the Action Primitive Recognition Model, these gaze-related quantities are used as inputs to a long-short term memory (LSTM) recurrent neural network in order to recognize action primitives.

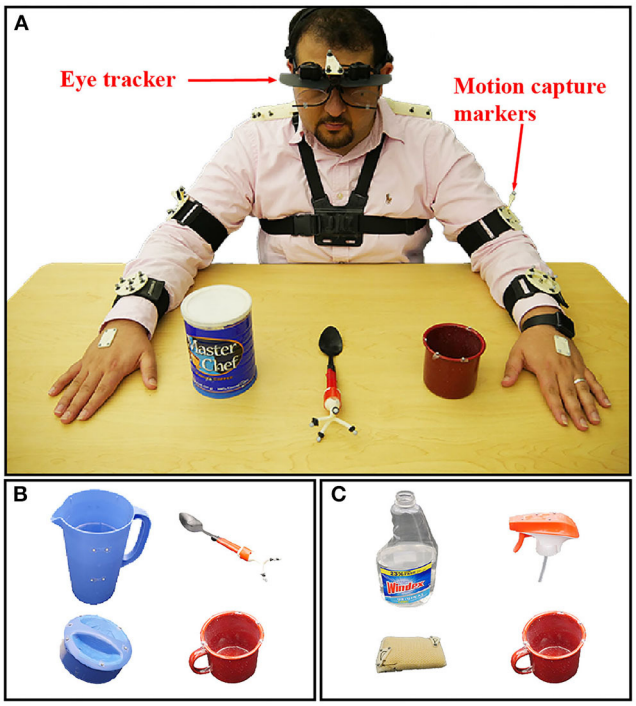


FIGURE 1 | (A) A subject prepares to perform Activity 2 (make instant coffee) while eye gaze and kinematics are tracked with a head-mounted eyetracker and motion capture system (not shown). Activity 2 involves a coffee can, spoon and mug. **(B)** Activity 1 (make a powdered drink) involves a coffee can, spoon and mug. **(C)** Activity 3 (prepare a cleaning sponge) involves a spray bottle and cap, sponge, and mug. The subject shown in panel **(A)** has approved of the publication of this image.

The raw data we obtain from the eyetracker is a set of 2D pixel coordinates. The coordinates represent the perspective projection of a subject's gaze point onto the image plane of the eyetracker's egocentric scene camera. In order to convert the 2D pixel coordinate into a 3D gaze vector, we use camera calibration parameters determined using a traditional chessboard calibration procedure (Heikkila and Silven, 1997) and the MATLAB Camera Calibration Toolbox (Bouguet, 2015). The 3D gaze vector is constructed by connecting the origin of the egocentric camera frame with the gaze point location in the 2D image plane that is now expressed in the 3D global reference frame.

The *gaze object (GO)* is defined as the first object to be intersected by the 3D gaze vector, as the gaze vector emanates from the subject. Thus, if the gaze vector pierces numerous objects, then the object that is closest to the origin of the 3D gaze vector (within the head-mounted eyetracker) is labeled as the gaze object.

As defined in our prior paper, the *gaze object sequence (GOS)* refers to the identity of the gaze objects in concert with the sequence in which the gaze objects are visually regarded (Haji Fathaliyan et al., 2018). Specifically, the gaze object sequence time history $GOS(t_i)$ is comprised of a sequence of gaze objects

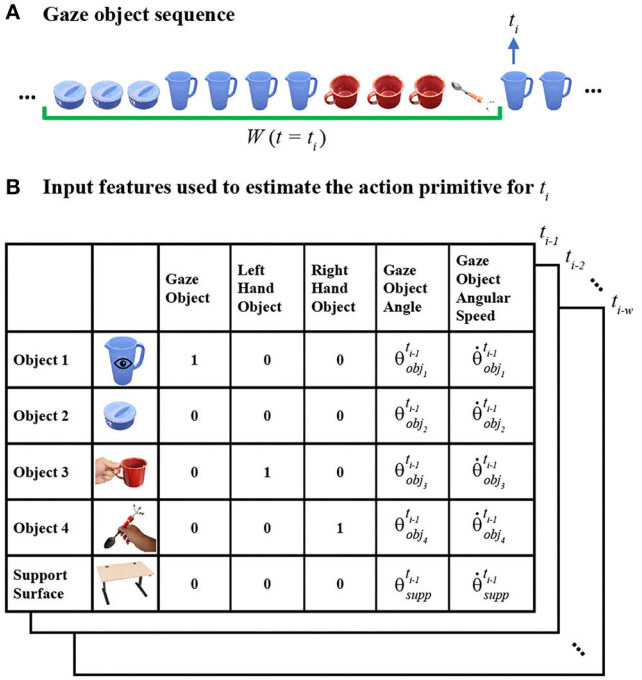


FIGURE 2 | (A) The gaze object sequence time history GOS(t_i) within a window of time W(t_i) (green bracket) is shown for Activity 1 (make a powdered drink). **(B)** To predict the action primitive at time step t_i, input feature vectors (shown as 5 × 5 matrices for clarity) are created for each of the times from t_{i-w} to t_{i-1}. Activity-relevant objects are sorted according to their frequency of occurrence in the GOS(t_i).

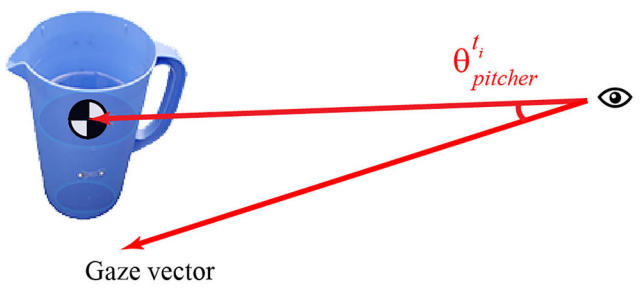


FIGURE 3 | Gaze object angle is defined as the angle between the gaze vector and the eye-object vector (ending at the object's center of mass).

sampled at 60 Hz within a given window of time $W(t_i)$ (**Figure 2**). The time window $W(t_i)$ contains w time steps from t_{i-w} to t_{i-1}.

In this work, we use a value of $w = 75$ time steps, equivalent to 1.25 s. This time window size was determined from a pilot study whose results are presented in section Effect of Time Window Size on Recognition Accuracy. The pilot study was motivated by the work of Haseeb et al. in which the accuracy of an LSTM RNN was affected by time window size (Haseeb and Parasuraman, 2017).

The *gaze object angle* (GOA) describes the spatial relationship between the gaze vector and each gaze object. The GOA is defined as the angle between the gaze vector and the eye-object vector (**Figure 3**). The eye-object vector shares the same origin as the gaze vector but ends at an object's center of mass. Each object's center of mass was estimated by averaging the 3D coordinates of the points in the object's point cloud. Each object's point cloud was scanned with a structured-light 3D scanner (Structure Sensor, Occipital, Inc., CA, USA) and custom turntable apparatus. Containers, such as the pitcher and mug, are assumed to be empty for center of mass estimation.

The *gaze object angular speed* (GOAS) is calculated by taking the time derivative of the GOA. We use the GOAS to measure how the gaze vector moves with respect to other activity-relevant objects. Previously, the gaze object and gaze object sequence have been used to recognize actions (Yi and Ballard, 2009; Matsuo et al., 2014). To our knowledge, this is the first work to leverage the gaze object angle and gaze object angular speed for action primitive recognition.

Action Primitive Recognition Model

Action Primitive Representation

We represent each action primitive as a triplet comprised of a *verb*, *target object (TO)*, and *hand object (HO)*. Each action primitive can be performed by either the dominant hand or non-dominant hand. When both hands are active at the same time, hand-specific action primitives can occur concurrently.

The verb can be one of four classes: *Reach, Move, Set down, or Manipulate*. The classes Reach, Move, and Set down describe hand movements toward an object or support surface, with or without an object in the hand. Notably, these verbs are not related to or dependent upon object identity. In contrast, the class Manipulate includes a list of verbs that are highly related to object-specific affordances (Gibson, 1977). For instance, in Activity 1, the verb "scoop" and "stir" are closely associated with the object "spoon" (**Table 1**). We refer to these verbs as *manipulate-type verbs*.

In addition to a verb, the action primitive triplet includes the identity of two objects. The *target object* TO refers to the object that will be directly affected by verbs such as Reach, Move, Set down, and Manipulate. The *hand object* HO refers to the object that is currently grasped. For instance, when the dominant hand grasps a spoon and stirs inside a mug, the triplet of the action primitive for the dominant hand is: manipulate (verb), mug (TO), and spoon (HO). A hierarchical description of activities, actions, and action primitives for Activities 1–3 are presented in **Table 1**.

In order to develop a supervised machine learning model for action primitive recognition, we manually label each time step with the action primitive triplet for either the dominant or non-dominant hand. The label is annotated using video recorded by an egocentric scene camera mounted on the head-worn eyetracker. We annotate each time step with the triplet of a subject's dominant hand as it is more likely the target of the subject's attention. For instance, when the dominant hand (holding a spoon) and the non-dominant hand (holding a mug) move toward each other simultaneously, we label the action

TABLE 1 | Each of three activities is divided into actions that are further decomposed into action primitives. Each action primitive is defined as a triplet comprised of a verb, target object (TO), and hand object (HO).

Activities		Activity 1: make a powdered drink	Activity 2: make instant coffee	Activity 3: prepare a cleaning sponge
Actions		Remove pitcher lid Stir liquid inside pitcher Scoop liquid into mug Close pitcher lid Pour liquid into mug	Remove coffee can lid Scoop coffee insider can Transfer coffee into mug Stir liquid inside mug Close coffee can lid	Remove spray bottle cap Transfer cleanser into mug Close spray bottle cap Spray cleanser onto sponge
Action primitives	Verb	Reach, Move, Set down, Manipulate (open, close, stir, scoop, drop, pour)	Reach, Move, Set down, Manipulate (open, close, stir, scoop, drop, pour)	Reach, Move, Set down, Manipulate (screw, unscrew, lift, pour, insert, spray)
	TO	Pitcher, pitcher lid, mug, spoon, table	Coffee can, coffee lid, mug, spoon, table	Spray bottle, spray cap, mug, sponge, table
	HO	Pitcher, pitcher lid, mug, spoon	Coffee can, coffee lid, mug, spoon	Spray bottle, spray cap, mug, sponge

primitive as "move the spoon to the mug," where the verb is "move" and the target object is "mug." However, when the dominant hand is not performing any action primitive, we refer to the non-dominant hand instead. If neither hand is moving or manipulating an object, we exclude that time step from the RNN training process.

Input Features for the Action Primitive Recognition Model

Given that the identity of gaze objects will vary across activities, we substitute the specific identities of gaze objects with numerical indices. This is intended to improve the generalizability of our action primitive recognition algorithm across different activities. For each time step t_i, the n activity-relevant objects are sorted in descending order according to their frequency of occurrence in $GOS(t_i)$. Once sorted, the objects are indexed as Object 1 to Object n, such that Object 1 is the object that most frequently appears in the gaze object sequence at t_i. If two or more objects appear in the gaze object sequence with the same frequency, the object with the smaller gaze object angle is assigned the smaller numerical index, as it is aligned most closely to the gaze vector and will be treated preferentially.

Figure 2 exemplifies how activity-relevant objects in a gaze object sequence would be assigned indices at a specific time step t_i. The activity-relevant objects ($n = 4$) in Activity 1 were sorted according to their frequency of occurrence in $GOS(t_i)$, which is underlined by a green bracket in **Figure 2A**. Based on frequency of occurrence, the activity-relevant objects were indexed as follows: pitcher (Object 1), pitcher lid (Object 2), mug (Object 3), and spoon (Object 4).

We introduce here the idea of a "support surface," which could be a table, cupboard shelf, etc. In this work, we do not consider the support surface (experiment table) as an activity-relevant object, as it cannot be moved or manipulated and does not directly affect the performance of the activity. Nonetheless, the support surface still plays a key role in the action primitive recognition algorithm due to the strong connection with the verb Set down. In addition, the support surface frequently appears in the *GOS*.

To predict the action primitive at time step t_i, input feature vectors are created for each of the time steps from time t_{i-w} to t_{i-1}, as shown in **Figure 2B**. For Activity 1, each input feature vector consists of five features for each of four activity-relevant objects and a support surface. For clarity, each resulting 25×1 feature vector is shown as a five-by-five matrix in **Figure 2B**. Gaze object, left-hand object, and right-hand object are encoded in the form of one-hot vectors while gaze object angle and angular speed are scalar values.

Gaze object identity was included as an input feature because it supported action recognition in prior studies (Yu and Ballard, 2002; Yi and Ballard, 2009; Matsuo et al., 2014). We included the hand object as an input feature although it is a component of the action primitive triplet that we seek to recognize. Considering the application of controlling a robotic arm through eye gaze, we expect the robotic system to determine an object's identity before it plans any movements with respect to the object. As a result, we assume that the hand object's identity is always accessible to the classification algorithm. We included the GOA and GOAS as input features because we hypothesized that spatiotemporal relationships between eye gaze and objects would be useful for action primitive recognition. The preprocessing pipeline for the input features is shown in **Supplementary Video 1**.

Action Primitive Recognition Model Architecture

We train a long short-term memory (LSTM) recurrent neural network to recognize the verb and the target object *TO* for each time step t_i. With this supervised learning method, we take as inputs the feature vectors described in section Input Features for the Action Primitive Recognition Model. For the RNN output, we label each time step t_i with a pair of elements from a discrete set of verbs and generic, indexed target objects:

$$\text{Verb}\,(t_i) \in \mathcal{V} = \{Reach,\ Move, Set\ down,\ Manipulate\} \quad (1)$$

$$\text{TO}\,(t_i) \in \mathcal{O} = \{Object_1,\ Object_2, Object_3, Support\ surface\} \quad (2)$$

The target object class Object 4 was excluded from the model output since its usage accounted for <1% of the entire dataset. The four verb labels and four TO labels are combined as 16 distinct verb-TO pairs, which are then taken as output classes when we train the RNN.

$$\begin{aligned}&\left(Verb\,(t_i)\,, TO\,(t_i)\right) \in \mathcal{O}\times\mathcal{V}\\ &= \left\{\ \left(Reach, Object_1\right), \ldots, \left(Manipulate, Support\ surface\right)\ \right\}\end{aligned} \quad (3)$$

As a result, verb-TO pairs that never occur during the training process, such as (Manipulate, Support surface), can be easily eliminated.

In order to evaluate the RNN's performance on the verb and target object individually, we split the verb-TO pairs after

recognition. A softmax layer was used as the final layer of the RNN.

$$Verb\,(t_i) = argmax_{v\in\mathcal{V}}\left(\sum_{o\in\mathcal{O}} softmax\left(\left(Verb\,(t_i) = v, TO\,(t_i) = o\right)\right)\right) \tag{4}$$

$$TO\,(t_i) = argmax_{o\in\mathcal{O}}\left(\sum_{v\in\mathcal{V}} softmax\left(\left(Verb\,(t_i) = v, TO\,(t_i) = o\right)\right)\right) \tag{5}$$

The RNN was comprised of one LSTM layer, three dense layers, and one softmax layer. The LSTM contained 64 neurons and each of the three dense layers contained 30 neurons. The RNN was trained with an Adaptive Momentum Estimation Optimization (Adam), which was used to adapt the parameter learning rate (Kingma and Ba, 2015). A dropout rate of 0.3 was applied in order to reduce overfitting and improve model performance. The batch size and epoch number were set as 128 and 20, respectively. The RNN was built using the Keras API in Python with a TensorFlow (version 1.14) backend, and in the development environment of Jupyter Notebook.

Class imbalance is a well-known problem that can result in a classification bias toward the majority class (Japkowicz, 2000). Since our dataset was drawn from participants naturally performing activities, the training set of samples was not balanced among various verb and TO classes (see sample sizes in **Figure 5**). An imbalance in TO classes might also result from sorting and indexing the objects as described in section Input Features for the Action Primitive Recognition Model. For instance, Object 1 occurs most frequently in the GOS by definition. Thus, Object 1 is more likely to be the target object than Objects 2 or 3. In order to compensate for the class imbalance, each class' contribution in the cross-entropy loss function was weighted by its corresponding number of samples (Aurelio et al., 2019).

The temporal sequence of the target object and verb recognized by the RNN can contain abrupt changes, as shown in the top rows of **Figures 5A,B**. These abrupt changes occur for limited time instances and make the continuous model prediction unsmooth. Such unstable classifier results might cause an assistive robot to respond unexpectedly. Thus, we implemented a one-dimensional mode filter with an order of m (in our work, $m = 12$ time steps, equivalent to 0.2 s) to smooth out these sequences (Wells, 1979):

$$\text{verb}\,(t_i) = \text{mode}\left(\left\{verb\,(t_{i-m}), verb\,(t_{i-m+1}), \ldots, verb\,(t_{i-1})\right\}\right) \tag{6}$$

$$\text{TO}\,(t_i) = \text{mode}\left(\left\{TO\,(t_{i-m}), TO\,(t_{i-m+1}), \ldots, TO\,(t_{i-1})\right\}\right) \tag{7}$$

The sequences after filtering are shown in the middle rows of **Figures 5A,B**.

Considering that 10 subjects participated in our study, we adopted a leave-one-out cross-validation method. That is, when one subject's data were reserved for testing, the other nine subjects' data were used for training.

Performance Metrics for Action Recognition

In order to evaluate the performance of the action primitive classification, we assessed overall accuracy, precision, recall, and the F1-score. Overall accuracy is the number of correctly classified samples divided by the total size of the dataset. For each class of verb or target object, precision represents the fraction of correctly recognized time steps that actually belong to the given class, and recall represents the fraction of the class that are successfully recognized. We use TP, TN, and FP to represent the number of true positives, true negatives, and false positives when classifying a verb or target object class.

$$\text{overall accuracy} = \frac{\sum TP}{total\ size\ of\ dataset} \tag{8}$$

$$\text{precision} = \frac{TP}{TP + FP} \tag{9}$$

$$\text{recall} = \frac{TP}{TP + TN} \tag{10}$$

The F1-score is the harmonic mean of precision and recall.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{11}$$

We also used performance metrics that were related to the temporal nature of the data. In order to evaluate how early an action primitive was successfully recognized, we adopted the terminology "observational latency," as defined in Ellis et al. (2013). The term was defined as "the difference between the time a subject begins the action and the time the classifier classifies the action," which translates to the amount of time that a correct prediction lags behind the start of an action primitive. It should be noted that the observational latency does not include the computation time that the recognition algorithm requires to preprocess the input data and recognize the actions by the model.

We conservatively judged the success of an action primitive's classification by checking whether more than 75% of its time period was predicted correctly. Summary statistics for observational latency are reported for action primitives that were deemed correct according to this 75% threshold. Observational latency is negative if the action primitive is predicted before it actually begins.

RESULTS

Recall our aim of specifying the three components of the action primitive triplet: verb, target object, and hand object. Given that the hand object is already known, as described in section Input Features for the Action Primitive Recognition Model, we report on the ability of the RNN to recognize the verb and target object. A demonstration of the trained RNN is included in **Supplementary Video 1**.

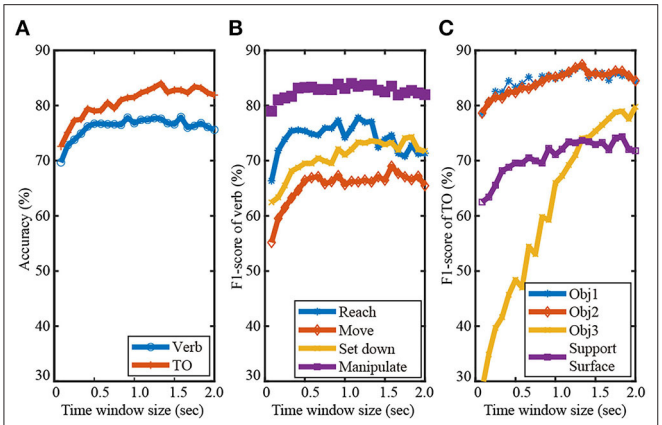


FIGURE 4 | The effect of time window size (ranging from 83 ms to 2 s) on recognition performance is shown for Activity 1. The overall recognition accuracy for verb and target object are shown in **(A)**. F1-scores for the verb and target object classes are shown in **(B,C)**, respectively.

Effect of Time Window Size on Recognition Accuracy

In order to set the time window size, we conducted a pilot study inspired by Haseeb and Parasuraman (2017). We tested how the F1-scores of the verb and TO classes varied as the time window size was increased from five time steps (equivalent to 83 ms) to 2 s in increments of five time steps (**Figure 4**). Considering the average duration of an action primitive was only 1.2 s, we did not consider time window sizes beyond 2 s.

As seen in **Figure 4A**, time window size had a more substantial effect on the recognition of TO than that of verb. This is due to the fact that time window size can greatly affect the data sample distributions among target object classes as a result of sorting and indexing the activity-relevant objects. **Figure 4C** shows that the TO class Object 3 was especially sensitive to the window size. The corresponding F1-score continuously increased from ∼30% to 80% until the window size reached 1.8 s. Recognition performance of the other three TO classes Object 1, Object 2, and Support surface were also improved as the time-window size was increased from 80 ms to 1.25 s. The increased F1-scores of the TO classes can be partly attributed to alleviated class imbalance problem as the time window was lengthened, especially for the class Object 3. The number of data samples of Object 3 greatly increased due to the nature of sorting and indexing objects according to their frequency of occurrence in gaze object sequence.

As seen in **Figure 4B**, the F1-scores of the verb classes Reach, Move, and Manipulate increased as the time-window size increased from 80 ms to 0.5 s. Little improvement in the F1-scores was observed for time window sizes > 0.5 s, except for Set down. This suggested that a memory buffer of 0.5 s might be sufficient for predicting the verb class based on eye gaze. Gaze-related information collected long before the start of an action primitive was very likely to be irrelevant to the verb.

Considering the effect of the time window size on the classification accuracy of both the verb and target object (**Figure 4**), we decided to use a time window size of 1.25 s. A time window longer than 1.25 s might slightly improve recognition performance, but with additional computational cost.

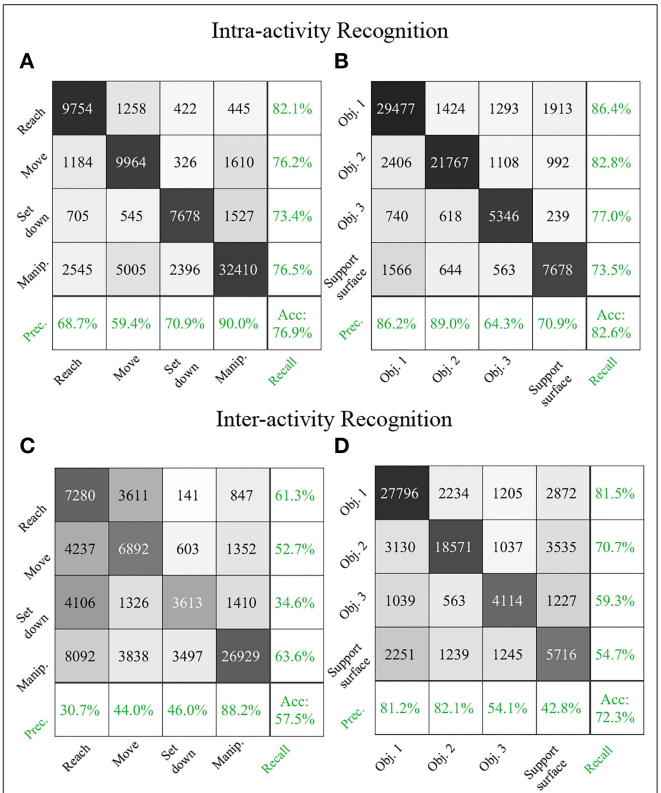


FIGURE 5 | Intra-activity recognition results for Activity 1 are shown in confusion matrix form for **(A)** verb and **(B)** target object. Inter-activity recognition results for an RNN trained on Activity 2 and tested on Activity 1 are shown for **(C)** verb and **(D)** target object. Integers in the confusion matrices represent numbers of samples. The confusion matrices are augmented with precision, recall, and accuracy results (green).

Intra-Activity Recognition

We report results for intra-activity recognition, in which we trained and tested the recurrent neural network on the same activity. These results describe how well the RNN recognized novel instances of each activity despite variability inherent to activity repetition. Intra-activity recognition results for Activity 1 are shown in **Figure 5** in the traditional form of confusion matrices. The rows correspond to the true class and the columns correspond to the predicted class. For brevity, intra-activity recognition results for Activities 1 and 2 are also shown in **Table 2** in the form of F1-scores. The weighted averages of F1-scores for verb and target object were each calculated by taking into account the number of data samples for each class. The RNN was not trained on Activity 3 due to its smaller dataset as compared to Activities 1 and 2. Thus, no intra-activity recognition results were reported for Activity 3.

We augmented the traditional confusion matrix used to report results according to true and predicted classes with additional

TABLE 2 | The RNN performance for intra- and inter-activity recognition is reported via F1-scores (%). Weighted averages of F1-scores that account for the number of data samples in each class are reported for both verb and target object (TO).

Intra- or Inter-activity recognition	Intra	Inter	Inter	Intra	Inter	Inter
Activity # (training)	1	1	1	2	2	2
Activity # (testing)	1	2	3	2	1	3
F1-scores for verb recognition (%)						
Reach	74.8	52.9	54.8	56.5	40.9	55.6
Move	66.8	36.6	61.1	59.5	48.0	60.5
Set down	72.1	49.3	45.3	59.6	39.5	44.4
Manipulate	82.7	73.7	72.7	81.4	73.9	71.8
Verb Average	77.4	60.3	63.6	68.6	59.9	63.1
F1-scores for target object recognition (%)						
Object 1	86.3	72.1	78.0	80.2	81.3	77.4
Object 2	85.8	80.7	83.6	87.2	76.0	80.8
Object 3	70.1	41.7	52.5	55.2	56.6	56.8
Support surface	72.2	56.9	49.8	69.3	48.0	46.6
TO Average	82.8	73.0	74.9	81.1	72.8	73.4

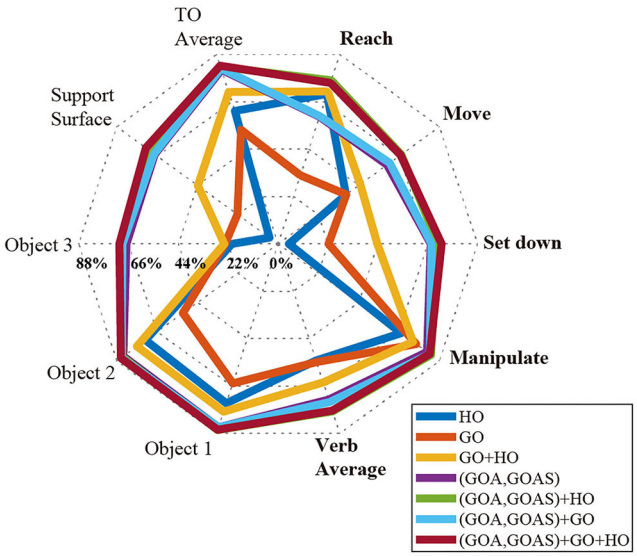


FIGURE 6 | For Activity 1, RNN performance is reported by F1-scores for different combinations of input features (HO, GO, GOA, GOAS) using a radar chart. Axes represent the verb (bold) and target object classes. F1-score gridlines are offset by 22%. Each of the polygons corresponds to one combination of input features. The combined use of HO, GO, GOA, and GOAS features resulted in the best performance; HO alone performed the worst.

metrics of precision and recall (**Figure 5**). Precision and recall were reported as percentages (in green) in the far right column and bottom-most row, respectively. The cell in the lower-right corner represented the overall recognition accuracy.

The data samples were not balanced among various verb and TO classes since our dataset was drawn from participants naturally performing activities. The proportion of each verb and TO class in Activity 1 was the sum of the corresponding row in **Figures 5A,B** divided by the total size of the dataset (77,774 time step samples). The proportions for the verb classes were 15% for Reach, 17% for Move, 13% for Set down, and 55% for Manipulate. The proportions for the target object classes were 44% for Object 1, 34% for Object 2, 9% for Object 3, and 13% for Support surface.

The RNN achieved a good performance in recognizing the majority verb class Manipulate (precision: 90%, recall: 77%) and the TO class Object 1 (precision: 86%, recall: 86%), which laid a solid foundation for its overall accuracy (verb: 77%, TO: 83%).

Inter-activity Recognition

We report results for inter-activity recognition, in which we trained and tested the recurrent neural network on different activities. These results describe how well the RNN can recognize verbs and target objects despite variability across different activities. To evaluate the algorithm's cross-activity generalizability, an RNN trained on Activity 2 (make instant coffee) was tested on Activity 1 (make a powdered drink), and vice versa. RNNs trained on Activity 1 and Activity 2 were additionally tested on Activity 3 (prepare a cleaning sponge). The confusion matrices of an RNN trained on Activity 2 and tested on Activity 1 are shown in **Figures 5C,D** for verb and target object estimation, respectively. For brevity, additional inter-activity recognition results are presented in **Table 2** in the form of F1 scores.

We also compared intra-activity and inter-activity performance of RNN models tested on the same activity. For this, we subtracted the average F1-scores for inter-activity recognition from those of the appropriate intra-activity recognition for RNNs tested on Activity 1 and Activity 2. As expected, when testing with an activity that differed from the activity on which the RNN was trained, the classification performance decreased. The average F1-scores of verb and target object each dropped by 8% when the RNN was trained on Activity 1 and tested on Activity 2. The average F1-scores of verb and target object dropped by 18 and 10%, respectively, when the RNN was trained on Activity 2 and tested on Activity 1. The average F1-score decreases were no larger than 20%, which suggested that the classification algorithm was able to generalize across activities to some degree. In addition, despite the fact that Activity 3 shared only one common activity-relevant object (mug) with the other two activities, the average F1-scores of verb and TO achieved for Activity 3 were slightly higher than those of the other inter-activity recognition tests (**Table 2**).

Effect of Input Features on Recognition Accuracy

In order to evaluate feature importance, we compared the classification performance achieved in Activity 1 with various combinations of input features using a radar chart (**Figure 6**). Axes represented the verb and target object classes. Gridlines marked F1-scores in increments of 22%. Classification using HO alone was poor, with F1-scores for "Set down" and "Object 3" being $< 10\%$. Only slightly better, classification using GO alone

was still not effective, with F1-scores of the "Set down," "Object 3," and "Support surface" only reaching values near 22%. In contrast, GOA-based features (GOA, GOAS) alone outperformed both HO and GO on their own in every verb and target object class. With the exception of "Reach," GOA-based features alone also outperformed the use of HO and GO together.

Although the feature HO alone did not provide good recognition result, it could substantially improve the classification performance when used in concert with GOA-based features. For every class, the F1-scores achieved with the combination of GOA-based feature and HO were equal to or higher than with the GOA-based feature alone.

Effect of Input Features on Observational Latency

The time histories of the verb and target object recognition for a representative Activity 1 trial are shown in **Figures 7A,B**. In each of **Figures 7A,B**, the top colorbar represents a time history of raw prediction results. The middle colorbar shows the output of the mode filter that smooths the raw prediction results. The bottom colorbar represents the ground truth. White gaps in the ground truth correspond to instances when neither hand was moving or manipulating an object. The observational latency is obtained by comparing the middle and bottom colorbars.

While **Figure 7** shows the observational latency for a single representative trial, the observational latencies for all trials and participants are presented in **Figure 8**. Specifically, **Figures 8A,B**, summarize results for the recognition of verb and target object, respectively, for an RNN trained and tested on Activity 1. **Figure 8** illustrates the effect of input features on observational latency by comparing the results of an RNN that only used GO and HO as input features to those of an RNN that additionally used GOA, and GOAS as input features.

We hypothesized that the incorporation of GOA-based input features could significantly decrease observational latency. To test this, we conducted a Wilcoxon signed-rank test (following a Lilliefors test for normality) with a total of 714 action primitives. The one-tailed p-values for the verbs and target objects were all less than the α level of 0.05 except for the target object of pitcher lid. Thus, we concluded that the use of GOA and GOAS as input features in addition to GO and HO resulted in a reduction in observational latency (**Figure 8**).

DISCUSSION

Features Based on Gaze Object Angle Improve Action Primitive Recognition Accuracy

The long-term objective of this work is to advance shared autonomy control schemes so that individuals with upper limb impairment can more naturally control robots that assist with activities of daily living. One embodiment of such a teleoperated system could include both a joystick and eyetracker as user input devices. The short-term goal of this study was to improve action primitive recognition accuracy and observational latency. We pursued this goal by (i) focusing on the recognition of low-level action primitives, and (ii) defining eye gaze-based input features that improve action primitive recognition.

Previous studies leveraged egocentric videos to recognize actions when a subject was naturally performing ADLs. The features reported in these studies can be divided into three categories: features based on human hands, objects, or human gaze. Examples of hand-based features include hand location,

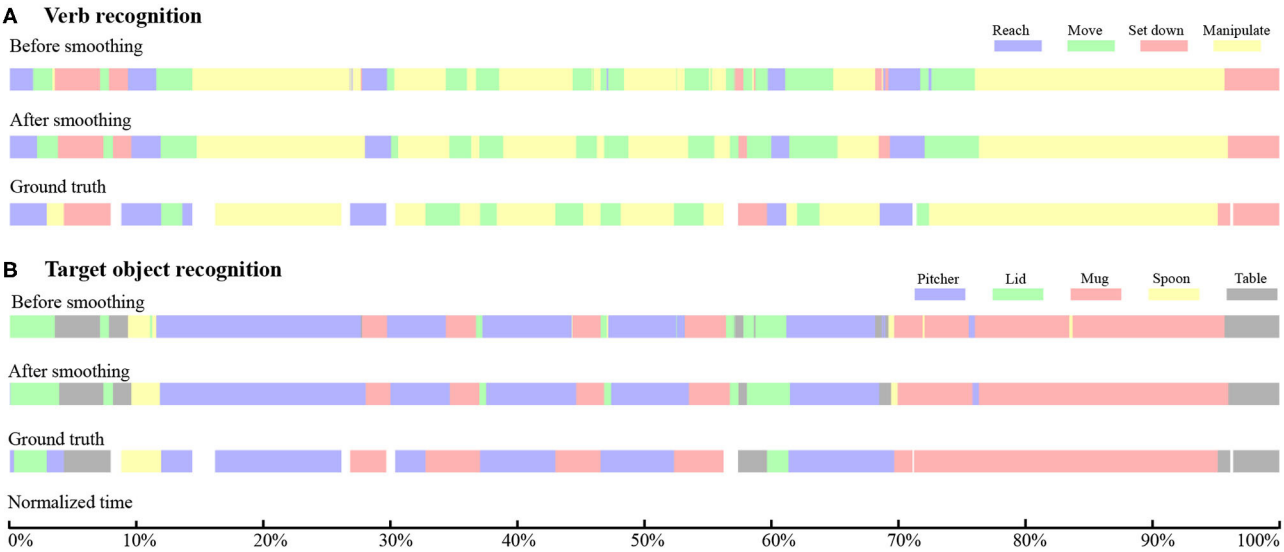


FIGURE 7 | For a representative trial of Activity 1, temporal sequences of recognition results and ground truth are presented for **(A)** verb and **(B)** target object. In both **(A,B)**, the top, middle, and bottom color bars represent the raw RNN output, RNN output smoothed by a mode filter, and hand-labeled ground truth, respectively. The total duration of this trial is 36 s.

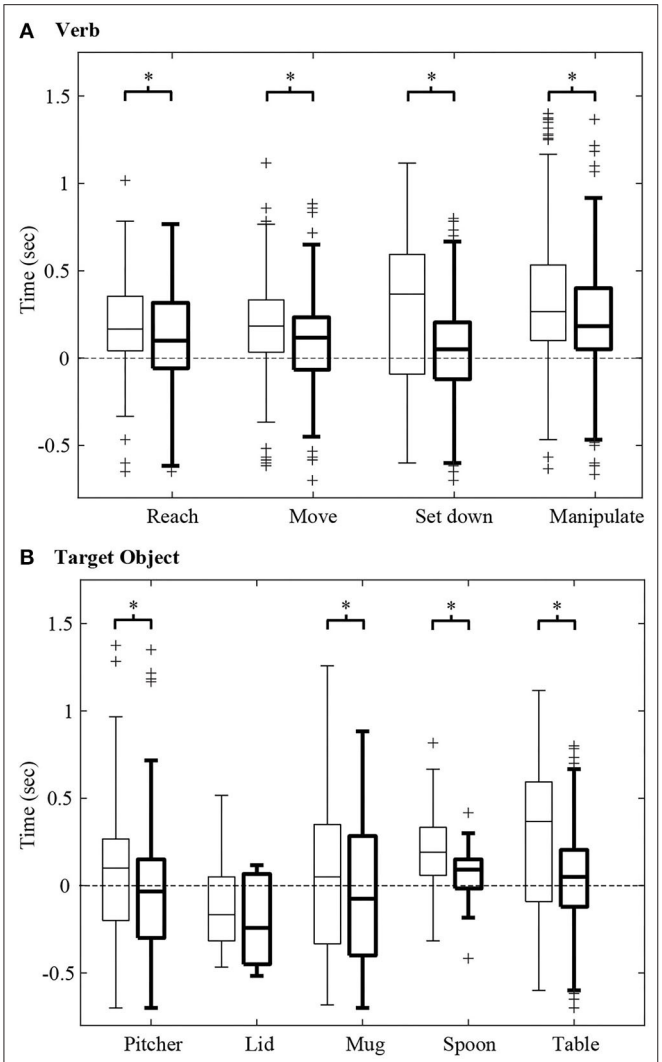


FIGURE 8 | For Activity 1, the observational latency for recognition of **(A)** verb and **(B)** target object are shown using box and whisker plots. A negative latency value indicates that a verb or target object is identified before the start of the action primitive. For each boxplot pair, the observational latency without using GOA and GOAS (thin lines) is compared with that using GOA and GOAS (thick lines). Each boxplot indicates the 25, 50, and 75th percentiles. The whiskers extend to the most extreme data points that are not considered outliers ("+") having values of more than 1.5 times the interquartile range from the top or bottom of the box. Asterisks indicate $p < \alpha = 0.05$.

hand pose, and relative location between left and right hands (Fathi et al., 2011; Ma et al., 2016). Fathi et al. relied on changes in the state of objects, such as the state of the "coffee jar" (open vs. closed) (Fathi and Rehg, 2013), to recognize actions. Behera et al. used spatiotemporal relationships between objects as classifier inputs (Behera et al., 2012). Features related to human gaze included the gaze-object, which was widely used to classify actions (Yi and Ballard, 2009; Matsuo et al., 2014). The use of object appearance (histogram of color and texture) in the neighborhood of the gaze point was also effective in improving recognition accuracy (Fathi et al., 2012; Li et al., 2015b).

Considering the long-term objective of this work, we elected not to rely solely on features based on human hands or objects for action primitive recognition. Features based on human hands are only available when subjects use their own hands to directly grasp and manipulate objects. For the assistive robot application we envision, features of human hands such as hand location, hand pose, and relative location between left and right hands (Fathi et al., 2011; Ma et al., 2016) will not be available. Features based on objects are consequence of hand motions, such as changes in the states of objects or spatiotemporal relationships between objects. Such object-based features would only be available in hindsight and cannot be collected early enough to be useful for the proposed assistive robot application.

We aim to exploit observations that gaze behavior is a critical component of sighted grasp and manipulation activities, and that eye movements precede hand movements (Johansson et al., 2001; Land, 2006). In particular, it has been reported that eye gaze often shifts to a target object before any hand movement is observed (Land and Hayhoe, 2001). As such, we adopted the gaze-based feature GO from the literature (e.g., Yi and Ballard, 2009) and supplemented it with two new features that we defined: GOA and GOAS.

As reported in section Effect of Input Features on Recognition Accuracy, models that included GOA and GOAS as input features outperformed models that relied primarily on GO or HO for every verb and target object class. The addition of GOA and GOAS substantially improved the average F1-score from 64% to 77% for verb and from 71 to 83% for target object (**Figure 6**).

The advantages of using features based on gaze object angle for action primitive recognition are 2-fold. First, the gaze object angle quantifies the spatiotemporal relationship between the gaze vector and every object in the workspace, including objects upon which the subject is not currently gazing. In contrast, the gaze object only captures the identity of the object upon which the subject is gazing at that particular instant. Considering that daily activities generally involve a variety of objects, it is vital for the classifier to collect sufficient information related to gaze-object interactions. The feature GOA could indirectly provide information similar to that of GO. For example, a GOA value that is close to zero would result if the gaze vector is essentially pointing at the gaze object. When GOA, GOAS, and HO have already been included as input features, the addition of GO as an input feature has little to no impact on classification accuracy (**Figure 6**). Also, classifier performance improves when using GOA and GOAS as input features as compared to using GO, HO, or their combination (**Figure 6**).

Second, the input feature GOAS contains GOA rate information. To some extent, GOAS also captures directional information, as positive and negative GOAS values reflect whether the gaze vector is approaching or departing from each object in the workspace, respectively. We believe that approach/departure information can be leveraged to predict the target object for a given action primitive because gaze is used to gather visual information for planning before and during manual activities (Land, 2006). An object being approached by the gaze vector is not necessarily the target object, as the object could simply be in the path of the gaze vector during its movement.

However, objects are less likely to be labeled as the "target object" when the gaze vector moves away from them.

Features Based on Gaze Object Angle Improve Observational Latency

While recognition accuracy is important, human-robot systems also require low observational latency (Ellis et al., 2013). Even an action primitive that is correctly recognized 100% of the time will cease to be useful if the delay in recognition prohibits an effective response or adds to the cognitive burden of the operator. The earlier that a robotic system can infer the intent of the human operator or collaborator, the more time will be available for computation and the planning of appropriate robot movements.

Previous studies have focused on classifying actions in videos that have already been segmented in time (e.g., Fathi et al., 2012). However, these methods that were designed to recognize actions in hindsight would be less effective for real-time use. We desire the intended action primitive to be predicted in advance of robot movement and with as low an observational latency as possible.

Hoffman proposed several metrics to evaluate fluency in human-robot collaborative tasks. For instance, the robot's functional delay was defined as the amount of time that the human spent waiting for the robot (Hoffman, 2019). This concept of fluency reflects how promptly a robot can respond correctly to an operator's commands. A high observational latency will degrade the fluency of a human-robot system and increase the operator's cognitive burden, effort, and frustration levels. A user interface that requires operators to intentionally gaze at specific objects or regions for a fixed period of time may be less natural and have lower fluency than a user interface that leverages natural eye gaze behaviors (Li et al., 2017; Wang et al., 2018).

In this work, the use of gaze-related features enabled the recognition of action primitives at an early stage. The average observational latency for verb recognition was 120 ms, ∼10% of the average duration of an action primitive (1.2 s). The average observational latency for target object was −50 ms; the negative latency value indicates that the target object was sometimes identified before the start of the action primitive. Unfortunately, pooled across all classes, the observational latency for the target object was not statistically significantly less than zero ($p = 0.075$; $\alpha = 0.05$). Nonetheless, the fact that some of the trials resulted in negative observational latency values was surprising and encouraging.

Among gaze-related input features, the use of GOA and GOAS decreased the observational latency as compared with using GO alone (**Figure 8**). Per a Wilcoxon signed rank test, observational latency was statistically significantly smaller when GOA and GOAS were used as input features than when they were excluded ($p < \alpha = 0.05$). This was true for all verb classes and all target object classes, with the exception of lid. For the verb and target object, the observational latency dropped by an average of 108 and 112 ms, respectively. One reason for this could be that GOA-based features may encode the tendency of the gaze vector to approach an object once the eyes start to move. In contrast, the GO feature does not capture the identity of any object until the gaze vector reaches the object.

The sub-second observational latency values that we report likely resulted from the fact that eye movement generally precedes hand movement for manual activities (Johansson et al., 2001; Land, 2006). Land et al. reported that the gaze vector typically reached the next target object before any visible signs of hand movement during the activity of making tea (Land and Hayhoe, 2001). The small observational latency values may also result from the fact that our classifier was designed to recognize action primitives, which are much simpler than actions or activities (Moeslund et al., 2006). Action primitives often involve a single object, a single hand, and occur over a shorter period of time than actions and activities. The recognition of actions and activities for ADLs would require observations over a longer period of time and would necessarily involve more complex eye behaviors, more complex body movements, and gaze interactions with multiple objects.

Ryoo predicted activities of daily living and defined the "observation ratio" as the ratio between the observational latency and the activity duration (Ryoo, 2011). Ryoo reported that a minimum observation ratio of ∼45% was needed to classify activities with at least 60% accuracy. In this work, we found that minimum observation ratios of 18 and 5% were needed to achieve an accuracy of 60% for each the verb and the target object, respectively. This suggests that recognition of low-level action primitives can be achieved at lower observation ratios and within shorter time periods than high-level activities, which require the passage of more time and collection of more information for similar levels of accuracy.

One limitation of this work is that the action primitive recognition algorithm has not yet been tested in real-time. This is an area of future work and considerations for real-time implementation are discussed in section Comparisons to State-of-the-Art Recognition Algorithms. Based on our experience, we expect that the overall latency will be dominated by observational latency and less affected by computational latency. This is due to the relatively simple structure of the proposed RNN architecture and the fact that the RNN model would be trained offline a priori.

Segmenting Objects Into Regions According to Affordance Could Improve Recognition Performance

The distribution of gaze fixations can be concentrated on certain regions of an object, such as those associated with "object affordances." An object affordance describes actions that could be performed on an object (Gibson, 1977). For example, Belardinelli et al. showed human subjects a 2D image of a teapot and instructed them to consider lifting, opening, or classifying the teapot as an object that could or could not hold fluid (Belardinelli et al., 2015). It was observed that subjects' gaze fixations were focused on the teapot handle, lid, and spout for lifting, opening, and classifying, respectively. In addition, in a prior study, we reported 3D gaze heat maps for the activity "make a powdered drink" (Haji Fathaliyan et al., 2018). We observed that gaze fixations were focused on the top and bottom of pitcher during the action unit "reach for pitcher" and "set down pitcher."

Inspired by these findings, we hypothesized that information about the action primitive can, in theory, be encoded by gaze behavior with respect to specific regions of objects. This would provide a classification algorithm with information at a finer spatial resolution than when considering each object as a whole. In a *post hoc* study, we segmented the point clouds of each of the four activity-relevant objects in Activity 1 (make a powdered drink) into several regions according to object affordances (**Figure 9**). For instance, the spoon was segmented into the upper and bottom faces for the bowl, the handle, and the tip of the handle. Notably, the inner and outer wall of containers (pitcher and mug) were treated as different regions since the inner and outer walls were often fixated upon differently depending on the action primitive.

After the segmentation, we augmented the gaze-related features (GO, GOA, GOAS) by treating each region as an independent object while keeping the features left-hand object and right-hand object unchanged. We then retrained the RNN with the new augmented features. The recognition accuracy for verb increased slightly from 77 to 79% and accuracy for the target object increased from 83 to 86%. By increasing the total number of object regions from 4 to 20, the time taken for the trained RNN to produce one classifier output increased by 26%. Depending on the consequences of an incorrect classification and the minimum acceptable accuracy level, one could decide which objects to segment and how finely the objects should be segmented. For instance, one may still be able to improve recognition performance if the mug were segmented into inner wall, outer wall, and handle, as opposed to the five segments that we tested.

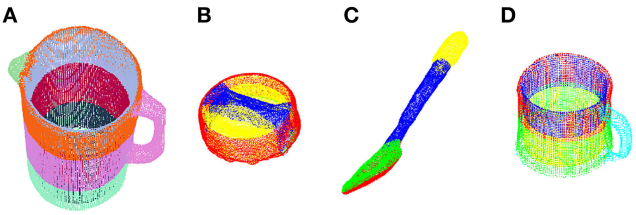


FIGURE 9 | Point clouds of the four activity-relevant objects involved in Activity 1 were segmented into multiple regions for finer spatial resolution: **(A)** pitcher, **(B)** pitcher lid, **(C)** spoon, and **(D)** mug.

Comparisons to State-of-the-Art Recognition Algorithms

In the evaluation of our proposed gaze-based action primitive recognition method, we were unable to identify suitable benchmarks for a direct quantitative comparison. First, our approach is designed to recognize low-level action primitives that could be used as modular, generalizable building blocks for more complex levels of the action hierarchy (Moeslund et al., 2006). The literature on action recognition provides methods for recognition at the level of actions and activities, but not at the level of action primitives that are investigated in our work. For instance, the public dataset "GTEA+" and "EGTEA Gaze+" provided by Fathi et al. (2012) Li et al. (2018) involve actions such as "take bread." This action would need to be split into two separate action primitives: "reach bread," and "set down bread onto table." Likewise, the public dataset "CMU-MMAC" provided by De la Torre et al. (2009) involves actions such as "stir egg." This action would need to be split into three action primitives: "reach fork," "move fork into bowl," and "stir egg in the bowl using fork." Many state-of-the-art recognition methods for ADLs (whether leveraging gaze behavior or not) are based on these publicly available datasets at the action level.

Second, action recognition models in the literature rely on computer-vision based approaches to analyze 2D videos recorded by an egocentric camera, e.g., (Fathi et al., 2011, 2012; Fathi and Rehg, 2013; Matsuo et al., 2014; Soran et al., 2015; Ma et al., 2016; Li et al., 2018; Furnari and Farinella, 2019; Sudhakaran et al., 2019; Liu et al., 2020). Whether using hand-crafted features (Fathi et al., 2011, 2012; Fathi and Rehg, 2013; Matsuo et al., 2014; Soran et al., 2015; Ma et al., 2016; Furnari and Farinella, 2019) or learning end-to-end models (Li et al., 2018; Sudhakaran et al., 2019; Liu et al., 2020), the computer vision-based approaches to action recognition must also address the challenges of identifying and tracking activity-relevant objects. In contrast, we bypassed the challenges inherent in 2D image analysis by combining an eyetracker with a marker-based motion capture system. This experimental set-up enabled the direct collection of 3D gaze-based features and object identity and pose information so that we could focus on the utility of 3D gaze features, which are unattainable from 2D camera images. Our method could be introduced into non-lab environments by combining an eyetracker with 2D cameras and ArUco markers, for example, in place of a marker-based motion capture system.

Considerations for Real-Time Implementation of an Action Primitive Recognition Algorithm in Human-Robot Systems

As an example of how our action primitive recognition model could be applied in a human-robot shared autonomy scenario, consider the action "stir contents inside a mug." First, as a subject's eye gaze vector moves toward the spoon, the probability of the potential action primitive "reach spoon" increases until it exceeds a custom threshold. The crossing of the threshold triggers the robotic end effector to move autonomously toward the spoon handle in order to grasp the spoon. The robot would use its real-time 3D model of the scene to plan its low-level movements in order to reduce the cognitive burden on the human operator. Second, as the subject's eye gaze switches to the mug after a successful grasp of the spoon, the model would recognize the highest probability action primitive as "move spoon to mug." Again the crossing of a probability threshold, or confidence level, would trigger the autonomous placement of the grasped spoon within the mug for a subsequent, allowable manipulate-type action primitive, which would be limited to a set of allowable manipulate-type action primitives based on the gaze object and hand object. Third, as the subject fixates their gaze on the

mug, the model would recognize the highest probability action primitive as "stir inside mug" and autonomous stirring would begin. The stirring trajectory could be generated using parametric dynamic motion primitives (Schaal, 2006), for example. Lastly, as the subject's gaze saccades to a support surface and the action primitive is recognized as "set down spoon," the system would proceed to determine a location on the table at which to place the spoon. This exact location could be extracted from filtered eye gaze signals as introduced in Li et al. (2015a).

As described in the above example, we envision that our model could be used to recognize subjects' intended action primitives through their natural eye gaze movements while the robot handles the planning and control details necessary for implementation. In contrast to some state-of-the-art approaches to commanding robot movements (Li and Zhang, 2017; Wang et al., 2018; Shafti et al., 2019; Zeng et al., 2020), subjects would not be forced to unnaturally, intentionally fixate their gaze at target objects in order to trigger pre-programmed actions. Of course, much work is necessary to implement the proposed shared autonomy control scheme and this is the subject of future work.

Concerning the practical implementation of the proposed action primitive recognition method, several limitations must be addressed.

Specificity of the Action Primitive

The proposed recognition method is intended to assign generalized labels to each time step as one of the four verb classes (reach, move, set down, and manipulate). The current method does not distinguish between subclasses of manipulate-type verbs, such as "pour" and "stir." Recognition of subclasses of a verb could enable assistive robots to provide even more specific assistance than that demonstrated in this work.

Recognition specificity could be advanced by incorporating additional steps. One idea is to create a lookup table based on the affordances of the objects involved in the activities. For example, the action primitive triplet of (verb = manipulate, TO = mug, HO = pitcher) is associated with the verb subclass "pour." However, the triplet (verb = manipulate, TO = pitcher, HO = spoon) is associated with both verb subclasses "stir" and "scoop." As an alternative, we suggest the use of gaze heat maps to facilitate the classification of verb subclasses since action primitives are activity-driven and the distribution of gaze fixations can be considerably affected by object affordance (Belardinelli et al., 2015; Haji Fathaliyan et al., 2018).

Distracted or Idle Eye Gaze States

The proposed recognition method does not recognize human subjects' distracted or idle states. For example, a subject's visual attention can be distracted by environmental stimuli. In this study, we minimized visual distractions through the use of black curtains and by limiting the objects in the workspace to those required for the instructed activity. The incorporation of distractions (audio, visual, cognitive, etc.) is beyond the scope of this work, but would need to be addressed before transitioning the proposed recognition method to natural, unstructured environments.

Idle states are not currently addressed in this work. Hands are not used for every activity and subjects may also wish to rest. If the gaze vector of a daydreaming or resting subject happens to intersect with an activity-relevant object, an assistive robot may incorrectly recognize an unintended action primitive and perform unintended movements. This is similar to the "Midas touch" problem in the field of human-computer interaction, which faces a similar challenge of "how to differentiate 'attentive' saccades with intended goal of communication from the lower level eye movements that are just random" (Velichkovsky et al., 1997). This problem can be addressed by incorporating additional human input mechanisms, such as a joystick, which can be programmed to reflect the operator's agreement or disagreement with the robot's movements. The inclusion of "distracted" and "idle" verb classes would be an interesting area for future advancement.

Integration With Active Perception Approaches

The proposed recognition method could be combined with active perception approaches that could benefit a closed-loop human-robot system that leverages the active gaze of both humans and robots. In this work, the 3rd person cameras comprising the motion capture system passively observed the scene. However, by leveraging the concept of "joint attention" (Huang and Thomaz, 2010), one could use an external and/or robot-mounted camera set-up to actively explore a scene and track objects of interest, which could be used to improve the control of a robot in a human-robot system.

As discussed in section Comparisons to State-of-the-Art Recognition Algorithms, for the purposes of this work, we bypassed the process of identifying and locating activity-relevant objects by implementing a marker-based motion capture system in our experiment. Nonetheless, the perception of activity-relevant objects in non-laboratory environments remains a challenge due to object occlusions and limited field of view. Active perception-based approaches could be leveraged in such situations. In multi-object settings, such as a kitchen table cluttered with numerous objects, physical camera configurations could be actively controlled to change 3rd person perspectives and more accurately identify objects and estimate their poses (Eidenberger and Scharinger, 2010). Once multiple objects' poses are determined, a camera's viewpoint could then be guided by a human subject's gaze vector to reflect the subject's localized visual attention. Since humans tend to align visual targets with the centers of their visual fields (Kim et al., 2004), one could use natural human gaze behaviors to control camera perspectives (external or robot-mounted) in order to keep a target object, such as one recognized by our proposed recognition method, in the center of the image plane for more stable computer vision-based analysis and robotic intervention (Li et al., 2015a). When realized by a visible robot-mounted camera, the resulting bio-inspired centering of a target object may also serve as an implicit communication channel that provides feedback to a human collaborator. Going further, the camera's perspective could be controlled actively and autonomously to focus on the affordances of a target object after a verb-TO pair is identified using our proposed recognition method. Rather than changing

the physical configuration of a camera to center an affordance in the image plane, one could instead focus a robot's attention on an affordance at the image processing stage (Ognibene and Baldassare, 2015). For instance, the camera's foveal vision could be moved to a pitcher's handle in order to guide a robot's reach-to-grasp movement. Such focused robot attention, whether via physical changes in camera configuration or via digital image processing methods, could be an effective way to maximize limited computational resources. The resulting enhanced autonomy of the robot could help to reduce the cognitive burden on the human in a shared autonomy system.

Considering the goal of our work to infer human intent and advance action recognition for shared autonomy control schemes, one could also integrate our proposed methods with the concept of "active event recognition," which uses active camera configurations to simultaneously explore a scene and infer human intent (Ognibene and Demiris, 2013). Ognibene and Demiris developed a simulated humanoid robot that actively controlled its gaze to identify human intent while observing a human executing a goal-oriented reaching action. Using an optimization-based camera control policy, the robot adjusted its gaze in order to minimize the expected uncertainty over numerous prospective target objects. It was observed that the resulting robot gaze gradually transitioned from the human subject's hand to the true target object before the subject's hand reached the object. As future work, it would be interesting to investigate whether and how the integration of 1st person human gaze information, such as that collected from an ego-centric camera, could enhance the control of robot gaze for action recognition. For instance, the outputs of our proposed action primitive recognition method (verb-TO pairs) could be used as additional inputs to an active event recognition scheme in order to improve recognition accuracy and reduce observational latency.

Effects of the Actor on Eye Gaze Behavior

The proposed recognition model was trained using data in which non-disabled subjects were performing activities with their own hands instead of subjects with upper-limb impairment who were observing a robot that was performing activities. In our envisioned human-robot system, we seek to identify operator intent via their natural gaze behaviors before any robotic movements occur. It is known that gaze behaviors precede and guide hand motions during natural hand-eye coordination (Hayhoe et al., 2003). In contrast, we hypothesize that the eye gaze behaviors of subjects observing robots may be reactive in nature. Aronsen et al. have shown that subjects' gaze behaviors are different in human-only manipulation tasks and human-robot shared manipulation tasks (Aronson et al., 2018). The further investigation of the effect of a robot on human eye gaze is warranted, but is beyond the scope of this work. We propose that the eye gaze behaviors reported in this work could be used as a benchmark for future studies of human-robot systems that seek to recreate the seamlessness of human behaviors.

The direct translation of the model to a human-robot system may not be possible. For one, the robot itself would need to be considered as an object in the shared workspace, as it is likely to receive some of the operator's visual attention. Fortunately, as suggested by Dragan and Srinivasa in Dragan and Srinivasa (2013), the action primitive prediction does not need to be perfect since the recognition model can be implemented with a human in the loop. The robotic system could be designed to wait until a specific confidence level for its prediction of human intent has been achieved before moving.

Another important consideration is that the recognition of action primitives via human eye gaze will necessarily be affected by how the robot is programmed to perform activities. For example, eye gaze behaviors will depend on experimental variables such as manual teleoperation vs. preprogrammed movements, lag in the robot control system and processing for semi-autonomous behaviors (e.g., object recognition), etc. Recognizing that there are innumerable ways in which shared autonomy could be implemented in a human-robot system, we purposely elected to eliminate the confounding factor of robot control from this foundational work on human eye-hand coordination.

Integration of Low-Level Action Primitive Recognition Models With Higher Level Recognition Models

This work focused on the recognition of low-level action primitives. However, the envisioned application to assistive robots in a shared autonomy schema would require recognition at all three hierarchical levels of human behavior (action primitives, actions, activities) (Moeslund et al., 2006) in order to customize the degree of autonomy to the operator (Kim et al., 2012; Gopinath et al., 2017). For instance, the outputs of the low-level action primitive recognition models (such as in this work) could be used as input features for the mid-level action recognition models (e.g., Haji Fathaliyan et al., 2018), that would then feed into the high-level activity recognition models (Yi and Ballard, 2009). Simultaneously, knowledge of the activity or action can be leveraged to predict lower level actions or action primitives, respectively.

CONCLUSION

The long-term objective of this work is to advance shared autonomy by developing a user-interface that can recognize operator intent during activities of daily living via natural eye movements. To this end, we introduced a classifier structure for recognizing low-level action primitives that incorporates novel gaze-related features. We defined an action primitive as a triplet comprised of a verb, target object, and hand object. Using a non-specific approach to classifying and indexing objects, we observed a modest level of generalizability of the action primitive classifier across activities, including those for which the classifier was not trained. We found that the gaze object angle and its rate of change were especially useful for accurately recognizing action primitives and reducing the observational latency of the classifier. In summary, we provide a gaze-based approach for recognizing action primitives that can be used to infer the intent of a human operator for intuitive control of a robotic system. The method can be further advanced by combining classifiers across multiple levels of the action hierarchy (action primitives, actions,

activities) (Moeslund et al., 2006) and finessing the approach for real-time use. We highlighted the application of assistive robots to motivate and design this study. However, our methods could be applied to other human-robot applications, such as collaborative manufacturing.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because they were not intended for public dissemination as raw data. Requests to access the datasets should be directed to Veronica J. Santos, vjsantos@ucla.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of California, Los Angeles Institutional Review Board. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

XW and AH supervised data collection. XW performed the data analysis, interpretation, and assisted by AH. XW created the first draft of the manuscript, which was further edited by VS and AH. All authors have read and approved the submitted manuscript and contributed to the conception and design of the study.

FUNDING

This work was supported in part by Office of Naval Research Award #N00014-16-1-2468. Any opinions, findings, conclusions, or recommendations are those of the authors and do not necessarily represent the official views, opinions, or policies of the funding agencies.

ACKNOWLEDGMENTS

The authors thank Daniela Zokaeim, Aarranon Bharathan, Kevin Hsu, and Emma Suh for assistance with data analysis. The authors thank Eunsuk Chong, Yi Zheng, and Eric Peltola for discussions on early drafts of this manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2020.567571/full#supplementary-material

Supplementary Video 1 | The Supplemental Video shows (i) a description of the experimental set-up and gaze vector reconstruction, (ii) a description of the preparation of input features for the recurrent neural network (RNN), and (iii) a demonstration of the RNN recognition of the verb and target object.

REFERENCES

Aronson, R. M., Santini, T., Kübler, T. C., Kasneci, E., Srinivasa, S., and Admoni, H. (2018). "Eye-hand behavior in human-robot shared manipulation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL: ACM), 4–13. doi: 10.1145/3171221.3171287

Aurelio, Y. S., de Almeida, G. M., de Castro, C. L., and Braga, A. P. (2019). Learning from imbalanced data sets with weighted cross-entropy function. *Neural Process. Lett.* 18, 1–13. doi: 10.1007/s11063-018-09977-1

Behera, A., Hogg, D. C., and Cohn, A. G. (2012). "Egocentric activity monitoring and recovery," in *Asian Conference on Computer Vision*, eds K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu (Berlin; Heidelberg: Springer), 519–532. doi: 10.1007/978-3-642-37431-9_40

Belardinelli, A., Herbort, O., and Butz, M. V. (2015). Goal-oriented gaze strategies afforded by object interaction. *Vis. Res.* 106, 47–57. doi: 10.1016/j.visres.2014.11.003

Bi, L., Fan, X., and Liu, Y. (2013). EEG-based brain-controlled mobile robots: a survey. *IEEE Trans. Hum. Machine Syst.* 43, 161–176. doi: 10.1109/TSMCC.2012.2219046

Bi, L., Feleke, A. G., and Guan, C. (2019). A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration. *Biomed. Signal. Process. Control* 51, 113–127. doi: 10.1016/j.bspc.2019.02.011

Bouguet, J.-Y. (2015). *Camera Calibration Toolbox for MATLAB*. Available online at: http://www.vision.caltech.edu/bouguetj/calib_doc/ (accessed September 2, 2020).

Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., and Dollar, A. M. (2015). Benchmarking in manipulation research: the ycb object and model set and benchmarking protocols. *arXiv preprint arXiv*. Available online at: http://arxiv.org/abs/1502.03143 (accessed September 2, 2015).

Chao, Z. C., Nagasaka, Y., and Fujii, N. (2010). Long-term asynchronous decoding of arm motion using electrocorticographic signals in monkey. *Front. Neuroeng.* 3:3. doi: 10.3389/fneng.2010.00003

De la Torre, F., Hodgins, J., Bargteil, A., Martin, X., Macey, J., Collado, A., et al. (2009). *Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) Database*. Technical Report. CMU-RI-TR-08-22.

Dragan, A. D., and Srinivasa, S. S. (2013). A policy-blending formalism for shared control. *Int. J. Robot. Res.* 32, 790–805. doi: 10.1177/0278364913490324

Driessen, B., Evers, H., and Woerden, J. (2001). MANUS—a wheelchair-mounted rehabilitation robot in proceedings of the institution of mechanical engineers, part H. *J. Eng. Med.* 215, 285–290. doi: 10.1243/0954411011535876

Du, Y., Wang, W., and Wang, L. (2015). "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Boston, MA: IEEE), 1110–1118.

Dziemian, S., Abbott, W. W., and Faisal, A. A. (2016). "Gaze-based teleprosthetic enables intuitive continuous control of complex robot arm use: Writing & drawing," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)* (IEEE), 1277–1282. doi: 10.1109/BIOROB.2016.7523807

Eidenberger, R., and Scharinger, J. (2010). "Active perception and scene modeling by planning with probabilistic 6D object poses," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 1036–1043. doi: 10.1109/IROS.2010.5651927

Ellis, C., Masood, S. Z., Tappen, M. F., LaViola, J. J., and Sukthankar, R. (2013). Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vis.* 101, 420–436. doi: 10.1007/s11263-012-0550-7

Fathi, A., Farhadi, A., and Rehg, J. M. (2011). "Understanding egocentric activities," in *Proceedings of the IEEE International Conference on Computer Vision* (Barcelona: IEEE), 407–414. doi: 10.1109/ICCV.2011.6126269

Fathi, A., Li, Y., and Rehg, J. M. (2012). "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*, eds A. Fitzgibbon,

S. Lazebnik, P. Perona, Y. Sato, and C. Schmid (Berli; Heidelberg: Springer), 314–327. doi: 10.1007/978-3-642-33718-5_23

Fathi, A., and Rehg, J. M. (2013). "Modeling actions through state changes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Portland, OR: IEEE), 2579–2586. doi: 10.1109/CVPR.2013.333

Furnari, A., and Farinella, G. (2019). "What would you expect? anticipating egocentric actions with rolling-unrolling LSTMs and modality attention," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul: IEEE), 6251–6260. doi: 10.1109/ICCV.2019.00635

Gajwani, P. S., and Chhabria, S. A. (2010). Eye motion tracking for wheelchair control. *Int. J. Inf. Technol.* 2, 185–187. Available online at: http://csjournals.com/IJITKM/PDF%203-1/2.pdf

Ghobadi, S. E., Loepprich, O. E., Ahmadov, F., Hartmann, K., Loffeld, O., and Bernshausen, J. (2008). Real time hand based robot control using multimodal images. *IAENG Int. J. Comput. Sci.* 35, 110–121. Available online at:http://www.iaeng.org/IJCS/issues_v35/issue_4/IJCS_35_4_08.pd

Gibson, J. J. (1977). "The theory of affordances," in *Perceiving, Acting, and Knowing: Towards an Ecological Psychology*, eds R. Shaw and J. Bransford (Hoboken, NJ: John Wiley & Sons Inc.), 127–143.

Gopinath, D., Jain, S., and Argall, B. D. (2017). Human-in-the-Loop Optimization of Shared Autonomy in Assistive Robotics. *IEEE Robot. Autom. Lett.* 2, 247–254. doi: 10.1109/LRA.2016.2593928

Groothuis, S. S., Stramigioli, S., and Carloni, R. (2013). Lending a helping hand: toward novel assistive robotic arms. *IEEE Robot. Autom. Magaz.* 20, 20–29. doi: 10.1109/MRA.2012.2225473

Haji Fathaliyan, A., Wang, X., and Santos, V. J. (2018). Exploiting three-dimensional gaze tracking for action recognition during bimanual manipulation to enhance human–robot collaboration. *Front. Robot. AI* 5:25. doi: 10.3389/frobt.2018.00025

Haseeb, M. A. A., and Parasuraman, R. (2017). Wisture: RNN-based Learning of Wireless signals for gesture recognition in unmodified smartphones. *arXiv:1707.08569*. Available online at: http://arxiv.org/abs/1707.08569 (accessed September 5, 2019).

Hayhoe, M. M., Shrivastava, A., Mruczek, R., and Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *J. Vis.* 3:6. doi: 10.1167/3.1.6

Heikkila, J., and Silven, O. (1997). "A four-step camera calibration procedure with implicit image correction," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (San Juan, PR: IEEE), 1106–1112. doi: 10.1109/CVPR.1997.609468

Hochberg, L. R., Bacher, D., Jarosiewicz, B., Masse, N. Y., Simeral, J. D., Vogel, J., et al. (2012). Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* 485, 372–375. doi: 10.1038/nature11076

Hoffman, G. (2019). Evaluating Fluency in Human–Robot Collaboration. *IEEE Trans. Hum. Machine Syst.* 49, 209–218. doi: 10.1109/THMS.2019.2904558

Huang, C.-M., and Thomaz, A. L. (2010). *Joint Attention in Human-Robot Interaction. in 2010 AAAI Fall Symposium Series*. Available online at: https://www.aaai.org/ocs/index.php/FSS/FSS10/paper/view/2173 (accessed August 5, 2020).

Japkowicz, N. (2000). "The class imbalance problem: significance and strategies," in *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)* (Las Vegas, NV), 111–117.

Johansson, R. S., Westling, G., Bäckström, A., and Flanagan, J. R. (2001). Eye–hand coordination in object manipulation. *J. Neurosci.* 21, 6917–6932. doi: 10.1523/JNEUROSCI.21-17-06917.2001

Kim, D.-J., Hazlett-Knudsen, R., Culver-Godfrey, H., Rucks, G., Cunningham, T., Portee, D., et al. (2012). How autonomy impacts performance and satisfaction: results from a study with spinal cord injured subjects using an assistive robot. *IEEE Trans. Syst. Man Cybern. Part A* 42, 2–14. doi: 10.1109/TSMCA.2011.2159589

Kim, J. H., Abdel-Malek, K., Mi, Z., and Nebel, K. (2004). *Layout Design using an Optimization-Based Human Energy Consumption Formulation*. Warrendale, PA: SAE International. doi: 10.4271/2004-01-2175

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *International Conference for Learning Representations* (San Diego, CA), 1–13. Available online at: https://dblp.org/db/conf/iclr/iclr2015.html; https://arxiv.org/abs/1412.6980

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

Land, M. F., and Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vis. Res.* 41, 3559–3565. doi: 10.1016/S0042-6989(01)00102-X

Li, S., and Zhang, X. (2017). Implicit intention communication in human–robot interaction through visual behavior studies. *IEEE Trans. Human Machine Syst.* 47, 437–448. doi: 10.1109/THMS.2017.2647882

Li, S., Zhang, X., Kim, F. J., Donalisio da Silva, R., Gustafson, D., and Molina, W. R. (2015a). Attention-aware robotic laparoscope based on fuzzy interpretation of eye-gaze patterns. *J. Med. Dev.* 9:041007. doi: 10.1115/1.4030608

Li, S., Zhang, X., and Webb, J. D. (2017). 3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments. *IEEE Trans. Biomed. Eng.* 64, 2824–2835. doi: 10.1109/TBME.2017.2677902

Li, Y., Liu, M., and Rehg, J. M. (2018). "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 619–635. doi: 10.1007/978-3-030-01228-1_38

Li, Y., Ye, Z., and Rehg, J. M. (2015b). "Delving into egocentric actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Boston, MA), 287–295. doi: 10.1109/CVPR.2015.7298625

Lin, C.-S., Ho, C.-W., Chen, W.-C., Chiu, C.-C., and Yeh, M.-S. (2006). Powered wheelchair controlled by eye-tracking system. *Opt. Appl.* 36, 401–412. Available online at: http://opticaapplicata.pwr.edu.pl/article.php?id=2006230401

Liu, M., Tang, S., Li, Y., and Rehg, J. (2020). Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. *arXiv:1911.10967 [cs]*. Available at: http://arxiv.org/abs/1911.10967 (accessed July 21, 2020).

Lv, F., and Nevatia, R. (2006). "Recognition and segmentation of 3-D human action using HMM and multi-class AdaBoost," in *Computer Vision – ECCV Lecture Notes in Computer Science*, eds A. Leonardis, H. Bischof, and A. Pinz (Berlin; Heidelberg: Springer), 359–372. doi: 10.1007/11744085_28

Ma, M., Fan, H., and Kitani, K. M. (2016). "Going deeper into first-person activity recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV: IEEE), 1894–1903. doi: 10.1109/CVPR.2016.209

Maheu, V., Frappier, J., Archambault, P. S., and Routhier, F. (2011). "Evaluation of the JACO robotic arm: clinico-economic study for powered wheelchair users with upper-extremity disabilities," in *2011 IEEE International Conference on Rehabilitation Robotics* (Zurich: IEEE), 1–5. doi: 10.1109/ICORR.2011.5975397

Matsuo, K., Yamada, K., Ueno, S., and Naito, S. (2014). "An attention-based activity recognition for egocentric video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. (Columbus), 551–556. doi: 10.1109/CVPRW.2014.87

Moeslund, T. B., Hilton, A., and Krüger, V. (2006). A survey of advances in vision-based human motion capture and analysis. *Comp. Vision Image Understand.* 104, 90–126. doi: 10.1016/j.cviu.2006.08.002

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. Mental Devel.* 7, 3–25. doi: 10.1109/TAMD.2014.2341351

Ognibene, D., and Demiris, Y. (2013). "Towards active event recognition". in *Twenty-Third International Joint Conference on Artificial Intelligence*. Available online at: https://www.aaai.org/ocs/index.php/IJCAI/IJCAI13/paper/view/6705 (accessed August 5, 2020).

Oldfield, R. C. (1971). The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia* 9, 97–113. doi: 10.1016/0028-3932(71)90067-4

Raheja, J. L., Shyam, R., Kumar, U., and Prasad, P. B. (2010). "Real-time robotic hand control using hand gestures," in *2010 Second International Conference on Machine Learning and Computing* (Bangalore), 12–16. doi: 10.1109/ICMLC.2010.12

Rogalla, O., Ehrenmann, M., Zollner, R., Becher, R., and Dillmann, R. (2002). "Using gesture and speech control for commanding a robot assistant," in *Proceedings of IEEE International Workshop on Robot and Human Interactive Communication* (Berlin), 454–459. doi: 10.1109/ROMAN.2002.1045664

Ryoo, M. S. (2011). "Human activity prediction: early recognition of ongoing activities from streaming videos," in *2011 International Conference on Computer Vision* (Barcelona), 1036–1043. doi: 10.1109/ICCV.2011.6126349

Salazar-Gomez, A. F., DelPreto, J., Gil, S., Guenther, F. H., and Rus, D. (2017). "Correcting robot mistakes in real time using EEG signals," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Singapore), 6570–6577. doi: 10.1109/ICRA.2017.7989777

Schaal, S. (2006) "Dynamic movement primitives - a framework for motor control in humans and humanoid robotics," in *International Symposium on Adaptive Motion of Animals and Machines*, eds H. Kimura, K. Tsuchiya, A. Ishiguro, and H. Witte (Tokyo: Springer), 261–280. doi: 10.1007/4-431-31381-8_23

Shafti, A., Orlov, P., and Faisal, A. A. (2019). "Gaze-based, context-aware robotic system for assisted reaching and grasping," in *2019 International Conference on Robotics and Automation (ICRA)* (Montreal, QC), 863–869. doi: 10.1109/ICRA.2019.8793804

Soran, B., Farhadi, A., and Shapiro, L. (2015). "Action Recognition in the Presence of One Egocentric and Multiple Static Cameras," in *Computer Vision – ACCV 2014 Lecture Notes in Computer Science*, eds D. Cremers, I. Reid, H. Saito, and M.-H. Yang (Cham: Springer International Publishing), 178–193. doi: 10.1007/978-3-319-16814-2_12

Sudhakaran, S., Escalera, S., and Lanz, O. (2019). "LSTA: Long Short-Term Attention for Egocentric Action Recognition," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Long Beach), 9946–9955. doi: 10.1109/CVPR.2019.01019

Velichkovsky, B., Sprenger, A., and Unema, P. (1997). "Towards gaze-mediated interaction: Collecting solutions of the Midas touch problem," in *Human-Computer Interaction INTERACT '97* (Boston, MA: Springer), 509–516. doi: 10.1007/978-0-387-35175-9_77

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," in *2014 IEEE Conference on Computer Vision and Pattern Recognition* (Columbus, OH: IEEE), 588–595. doi: 10.1109/CVPR.2014.82

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). "Mining actionlet ensemble for action recognition with depth cameras," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (Providence, RI), 1290–1297. doi: 10.1109/CVPR.2012.6247813

Wang, M.-Y., Kogkas, A. A., Darzi, A., and Mylonas, G. P. (2018). "Free-view, 3D gaze-guided, assistive robotic system for activities of daily living," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 2355–2361. doi: 10.1109/IROS.2018.8594045

Wang, W., Collinger, J. L., Degenhart, A. D., Tyler-Kabara, E. C., Schwartz, A. B., Moran, D. W., et al. (2013). An electrocorticographic brain interface in an individual with tetraplegia. *PLoS ONE* 8:e55344. doi: 10.1371/journal.pone.0055344

Wells, D. C. (1979). "The mode filter: a nonlinear image processing operator," in *Instrumentation in Astronomy III* (Tucson, AZ: International Society for Optics and Photonics), 418–421. doi: 10.1117/12.957111

Yi, W., and Ballard, D. (2009). Recognizing behavior in hand-eye coordination patterns. *Int. J. Hum. Robot.* 06, 337–359. doi: 10.1142/S0219843609001863

Yu, C., and Ballard, D. H. (2002). "Understanding human behaviors based on eye-head-hand coordination," in *International Workshop on Biologically Motivated Computer Vision*, eds H. H. Bülthoff, C. Wallraven, S. W. Lee, and T. A. Poggio (Berlin: Springer), 611–619. doi: 10.1007/3-540-36181-2_61

Zeng, H., Shen, Y., Hu, X., Song, A., Xu, B., Li, H., et al. (2020). Semi-autonomous robotic arm reaching with hybrid gaze–brain machine interface. *Front. Neurorobot.* 13:111. doi: 10.3389/fnbot.2019.00111

Zhang, Y. (2012). *Edinburgh Handedness Inventory*. Available online at: http://zhanglab.wikidot.com/handedness (accessed October 1, 2017).

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 04 March 2021
doi: 10.3389/fnbot.2021.598895

Speech Driven Gaze in a Face-to-Face Interaction

Ülkü Arslan Aydin[1], *Sinan Kalkan*[2] *and Cengiz Acartürk*[1,3]*

[1] *Cognitive Science Department, Middle East Technical University, Ankara, Turkey,* [2] *Computer Engineering Department, Middle East Technical University, Ankara, Turkey,* [3] *Cyber Security Department, Middle East Technical University, Ankara, Turkey*

Gaze and language are major pillars in multimodal communication. Gaze is a non-verbal mechanism that conveys crucial social signals in face-to-face conversation. However, compared to language, gaze has been less studied as a communication modality. The purpose of the present study is 2-fold: (i) to investigate gaze direction (i.e., aversion and face gaze) and its relation to speech in a face-to-face interaction; and (ii) to propose a computational model for multimodal communication, which predicts gaze direction using high-level speech features. Twenty-eight pairs of participants participated in data collection. The experimental setting was a mock job interview. The eye movements were recorded for both participants. The speech data were annotated by ISO 24617-2 Standard for Dialogue Act Annotation, as well as manual tags based on previous social gaze studies. A comparative analysis was conducted by Convolutional Neural Network (CNN) models that employed specific architectures, namely, VGGNet and ResNet. The results showed that the frequency and the duration of gaze differ significantly depending on the role of participant. Moreover, the ResNet models achieve higher than 70% accuracy in predicting gaze direction.

Keywords: face-to-face interaction, gaze analysis, deep learning, speech annotation, multimodal communication

Edited by:
Tom Foulsham,
University of Essex, United Kingdom

Reviewed by:
Erwei Yin,
Tianjin Artificial Intelligence Innovation Center (TAIIC), China
Dimitri Ognibene,
University of Milano-Bicocca, Italy

***Correspondence:**
Cengiz Acartürk
acarturk@metu.edu.tr

Received: *25 August 2020*
Accepted: *25 January 2021*
Published: *04 March 2021*

Citation:
Arslan Aydin Ü, Kalkan S and Acartürk C (2021) Speech Driven Gaze in a Face-to-Face Interaction. Front. Neurorobot. 15:598895. doi: 10.3389/fnbot.2021.598895

INTRODUCTION

Our skills of conversation by means of language, along with the accompanying non-verbal signals, set us apart from other species. Hence, conversation is considered to be one of the important indicators of humanness and human interaction. Recently, Embodied Conversational Agents (ECAs) that allow face-to-face communication are becoming more common. Face-to-face communication implies that interaction should be characterized as an inherently multimodal phenomenon, instead of speech in isolation (e.g., Levinson and Holler, 2014; Kendon, 2015; Mondada, 2016). This is because humans have an ability to send and receive information by means of non-verbal cues such as facial expressions, gestures, gaze, and posture, during a face-to-face conversation. In particular domains, they even correspond to 50–70% of the entire messages that the speaker conveyed (Holler and Beattie, 2003; Gerwing and Allison, 2009).

Gaze is an important non-verbal cue that conveys crucial social signals in face-to-face communication. Although its characteristics depend on individuals and cultural backgrounds, we usually make eye contact with the interlocutor, which, for instance, facilitates joint and shared attention. Even though we have such a tendency, face-to-face conversation is not just an interactive communication where partners constantly sustain eye contact; instead, it involves a sort of transition between gazing toward and away from the communication partner(s). Compared

to non-human primates, the specialized morphology of the human eyes, which have a sharp contrast between the white sclera and darker pupil, indicates the special role of revealing gaze direction by the sender and, thus, enables those around the sender to acknowledge about the direction of his gaze. These findings have been well-recognized since the past several decades (e.g., Kobayashi and Kohshima, 1997). We have the ability to make a distinction between directed and averted gaze from a very young age. Even an infant can make such a distinction in the first days of his life (Farroni et al., 2002). The present study focuses on gaze within language context, thus proposing a multimodal approach to computational analysis of face-to-face conversation. In the following section, we present the related work and technical background for the rest of the paper.

Related Work

Gaze in Social Interaction Settings

There exist various functions that the gaze fulfills in social interaction. Expressing emotions is one of the well-known function of gaze (Izard, 1991). An individual should perform eye movements in an appropriate way for the aim of conveying emotional states to an addressee successfully (Fukayama et al., 2002). In addition, gaze takes part in regulation of conversation, transmitting the intention, coordination of turn taking, asserting uncertainty or dissatisfaction, regulation of intimacy, and signaling the dominance and conversational roles (Kendon, 1967; Duncan, 1972; Argyle et al., 1974; Ho et al., 2015).

In recent decades, the development of eye-tracking technologies has enabled robust measurements and novel experimental designs in this field (Gredebäck et al., 2010). However, most of the studies have been performed in laboratory settings by adopting static eye-tracking methods (Pfeiffer et al., 2013), in which participants often monitor the stimulus presented on a computer screen. Although such experimental designs are advantageous in allowing one to provide a controlled procedure, the findings lack generalizability. Eye movements in the field might be different from those in studies conducted with static stimuli in a highly controlled laboratory environment (Risko et al., 2016). This difference can be explained by the two-way function of gaze in social communication. While gaze sends messages about, for instance, floor management or the desire to work together, we also gather information on emotions, intention, or attentional states of others by gazing on them.

Advances in mobile eye-tracking technology have opened the door to researchers who study social interaction in daily-life settings. Broz et al. (2012) studied mutual gaze in a face-to-face conversation with participants wearing mobile eye-tracking devices. They observed a mutual face gaze occurring for about 46% of a conversation. Rogers et al. (2018) also conducted a dual eye-tracking study and reported that the mutual face gaze comprised 60% of the conversation with 2.2 s duration on average.

An important characteristic of gaze in communication is that it is closely connected to speech acts. Accordingly, an analysis of communication in daily settings has to address speech in relation to gaze. In the following section, we introduce systematic approaches to study speech in communication.

Speech Annotation

The studies of Natural Language Processing (NLP) involving text mining, automated question answering, and machine translation have gained momentum as a reflection of the developments in Machine Learning (ML) technology (Meyer and Popescu-Belis, 2012; Sharp et al., 2015; Popescu-Belis, 2016). Hence, researchers' attention to discourse analysis has increased in parallel. In the last few decades, a variety of discourse annotation schemas were proposed involving RST (Rhetorical Structure Theory), RST Treebank (Carlson et al., 2001), SDRT, ANNODIS, and PDTB (Penn Discourse Treebank) (Prasad et al., 2008). Although there were some common communicative functions in those schemes, there were also inconsistencies between. In order to overcome mapping difficulty between proposed schemes, in the late 1990s, a domain-independent and multi-layered scheme, DAMSL[1] (Dialogue Act Markup using Several Layers) was proposed. Subsequently, many studies were carried out until the establishment of ISO standard for dialogue act annotation. Eventually, ISO standard 24617-2 "Semantic annotation framework (SemAF)—Part 2: Dialogue acts" was developed (ISO 24617-2, 2012).

The dialogue act is the act that the speaker is performing during a dialogue. In a simplified sense, it is a speech act used in a conversation. A dialogue act has a particular semantic content that specifies the objects, events, and their relations. Furthermore, it maintains a communicative function intended to change the state of mind of an addressee by means of its semantic content. In practice, dialogue act annotation generally depends on the communicative function. A turn represents the duration that the speaker is talking, and it is an important organizational tool in spoken discourse. Turns can be rather long and complex; in this case, they cannot be taken as units to determine communicative functions. They need to be cut into smaller parts called functional segments. Functional segments supply information to determine both the semantic content, namely, "dimensions" (see **Table 1**), and communicative functions of a dialogue act; for detailed information, see ISO 24617-2 (2012) and Bunt et al. (2017a), and for sample annotations, see DialogBank[2] (Bunt et al., 2019).

Dialogue act annotation can be achieved in three main steps: (i) the dialogue is divided into two or more functional segments, (ii) every single functional segment is associated with one or more dialogue acts, and lastly (iii) annotation components are assigned to dialogue acts (see **Table 2** for the related components). Although ISO 24617-2 does not provide any specific set for Rhetorical Relations (RRs), for this purpose, it suggests a specific standard, namely, Semantic Relations in discourse, core annotation schema (DR-Core) (ISO 24617-8, 2016).

A multimodal analysis of gaze and speech allows an intuitive understanding of their accompanying role in face-to-face conversation. However, a systematic analysis requires the specification of the relationship between gaze and speech in

[1]For Draft of DAMSL: Dialog Act Markup in Several Layers, see https://www.cs.rochester.edu/research/speech/damsl/RevisedManual/.

[2]You can find a collection of dialogues annotated according to international standard ISO 24617-2 under https://dialogbank.uvt.nl/.

TABLE 1 | Dimensions and communicative functions defined in ISO 24617-2.

Dimension		Communicative functions
Task	Category of dialogue acts that helps to carry out the tasks or activities that inspire the dialogue	General Purpose Functions (GPFs)
Auto-feedback	Category of dialogue acts that take place, in which the sender addresses his processing of past dialogue	AutoPositive, AutoNegative, GPFs
Allo-feedback	Category of dialogue acts that take place, in which the sender argues about the addressee's processing of past dialogue	AlloPositive, AlloNegative, FeedbackElicitation, GPFs
Turn management	Category of dialogue acts that are intended to coordinate the role of the speaker	TurnAccept, TurnAssign, TurnGrab, TurnKeep, TurnRelease, TurnTake, GPFs
Time management	Category of dialogue acts that deal with the allocation of time during the speech	Stalling, Pausing, GPFs
Own communication management	Category of dialogue acts where in the ongoing turn the speaker alters his own speech	SelfCorrection, SelfError, Retraction, GPFs
Partner communication management	Category of dialogue acts where in the ongoing turn the speaker alters the speech of the previous speaker	Completion, CorrectMisspeaking, GPFs
Discourse structuring	Category of dialogue acts that organize the dialogue directly	InteractionStructuring, Opening, GPFs
Social obligations management	Category of dialogue acts carried out to meet social responsibilities such as welcoming, thanking, and apologizing	InitialGreeting, ReturnGreeting, InitialSelfIntroduction, ReturnSelfIntroduction, Apology, AcceptApology, Thanking, AcceptThanking, InitialGoodbye, ReturnGoodbye, GPFs

TABLE 2 | Annotation components.

Component	Number
Dimension	1..1
Communicative function	1..1
Qualifier	0..N
Rhetorical relation*	0..N
Participant	
Sender	1..1
Addressee	1..1
Other	0..N
Dependence relation	
Feedback**	0..N
Functional*	0..N

*One and only one-dimension, communicative function, sender, and addressee should be attached to a single dialogue act. On the other hand, there might be zero, one, or more qualifiers, rhetorical relations, dependence relations, and participants other than sender and addressee. *Relation is between dialogue acts. **Relation is between either dialogue acts or a dialogue act and a functional segment.*

terms of the identification of specific patterns, which would allow making certain predictions about the interplay of gaze and speech in dialogue. This requires the development of computational models that characterize gaze-speech patterns that emerge during the course of communication. In the following section, we introduce the concept of computational modeling that we employed in the present study.

Computational Model

The deep learning approach has greatly improved many artificial intelligence tasks including machine translation, object detection, and speech recognition. In addition to classical AI tasks, researchers have adapted deep learning to various areas. Wang et al. (2017) performed sentiment analysis with data from multiple modalities; Gatys et al. (2016) utilized neural models to produce images in different styles; and Osako et al. (2015) eliminated noise from speech signals.

Convolution Neural Networks (CNNs) are localized versions of fully connected networks (LeCun et al., 1998; Goodfellow et al., 2016). It is based on an important operation, namely, convolution, which integrates the product of two functions. Convolution is useful for calculating change in signals, finding patterns, detecting edges, applying blurs, etc. CNN models that essentially learn the right convolution operations for the task at hand can produce high-accuracy results, especially in the areas of image classification and recognition. A basic CNN architecture

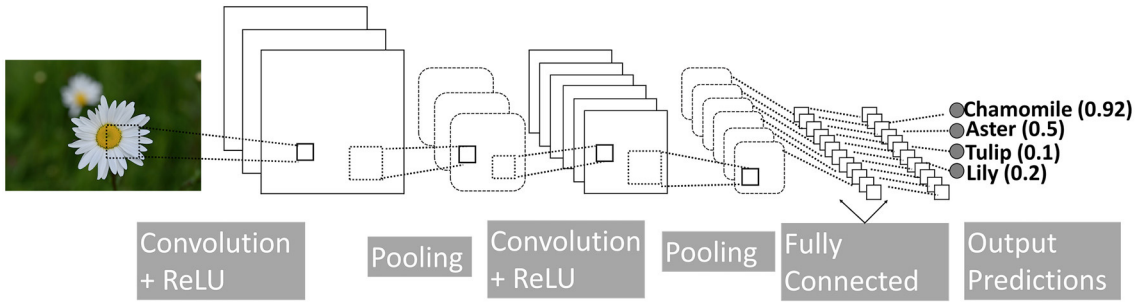


FIGURE 1 | A simple CNN model, composed of convolution, non-linearity (ReLU), pooling, and fully connected layers. The final output is one of the four flowers. Drawing adapted from LeCun et al. (1998).

includes four fundamental operations: (i) convolution, (ii) non-linearity (e.g., ReLU), (iii) pooling or subsampling, and (iv) classification (Fully Connected), see **Figure 1**.

Although CNN models are mostly used for image processing, they can be used in the same manner for time series (Fawaz et al., 2019). In this study, we collected the gaze data in the form of a time series and trained 1D CNN networks.

The Present Study

As reviewed in other articles (e.g., Admoni and Scassellati, 2017; Stefanov et al., 2019), research on the relationship between gaze and speech revealed their close coupling in communication settings (Prasov and Chai, 2008; Qu and Chai, 2009; Andrist et al., 2014). In the present study, we investigated the relation between speech (particularly high-level features of it) and gaze direction (i.e., face gaze or aversion) in a dyadic conversation.

The research into how speech and eye gaze are linked lead to a better understanding of the underlying cognitive mechanisms, but also this relation has been studied for practical applications in Educational Science (e.g., Jarodzka et al., 2017), human robot interaction (e.g., Chidambaram et al., 2012; Ham et al., 2015), web-based conferencing (e.g., Ward et al., 2016), and virtual reality (VR) systems (e.g., Garau et al., 2003; Batrinca et al., 2013). Some of those studies hold under operational assumptions such as simulating gaze aversion through head movements alone, conducting research under highly controlled conditions, which does not reflect real-life settings, or encoding just the presence of human speech rather than exhaustive speech analysis.

The main motivation of the present study is to explore eye gaze and speech relation in a more nuanced and comprehensive manner through employing state-of-the-art technologies and by taking into account the limitations of the previous studies in the field. Moreover, by using the data gathered experimentally, we trained the simplified versions of two deep networks, the ResNets (He et al., 2016) and VGGNet (Simonyan and Zisserman, 2015) that predict gaze direction based on high-level speech features.

Stefanov et al. (2019) showed that listener's gaze direction could be modeled from low-level speech features without considering semantic information, and they concluded that different methods are required for modeling speaker's gaze direction. In successful communication, the listener understands what the speaker says the way the speaker desires. In doing so, the listener takes into account the basic characteristics of the speaker's utterances, as well as the motivation behind the initiation and the history of the dialogue, and even his/her assumptions about the opinions and goals of the interlocutor. We cannot derive the communicative function of a dialogue by considering only the surface form of utterances since the same utterance forms can have different meanings in different conversations between different people. In the present study, to model states of both listening and speaking, we used high-level speech features.

It has been reported that (e.g., Dbabis et al., 2015; Bunt et al., 2017b) as high-level speech features, the dimensions and dialogue acts of ISO 24617-2 standard could be automatically recognized with fairly high accuracy. Therefore, even in case of a fully automated analysis, which can be conducted as a further study, ISO 24617-2 standard is a good candidate for extracting high-level speech features. The analysis of gaze and its ties to co-occurring speech is not a new topic of inquiry (e.g., Ekman, 1979; Zoric et al., 2011; Ho et al., 2015); however, as mentioned above, speech analysis was performed based on syntactic features or just for specific communicative function(s) such as turn taking, instead of adopting comprehensive semantic annotation frameworks. To the best of our knowledge, ISO 24617-2 standard has not been adopted in predicting gaze direction, so far.

In the present study, the speech annotation was handled in two ways: (i) ISO 24617-2 and ISO 24617-8 for annotating discourse and rhetorical relations, respectively, and (ii) an alternative set of speech tags that we proposed based on the roles attributed specifically to the gaze in social communication. The reason of annotating speech with two different methods is to investigate which characteristics of speech will produce better performance in modeling social gaze. In the following section, we present experimental investigation with analysis results.

EXPERIMENTAL INVESTIGATION

Materials and Design

Participants

Twenty-eight pairs involved seven professional interviewers, 4 females (mean age = 33.8, SD = 4.72) and 3 males (mean age

= 35.7, *SD* = 0.58), with the mean age of 34.6 (*SD* = 3.51); and 28 interviewees, 14 females (mean age = 25.1, *SD* = 2.57), and 14 males (mean age = 25.4, *SD* = 2.68), with the mean age of 25.3 (*SD* = 2.58) took part in the study. Interviewers took part in multiple interviews (*M* = 4, *SD* = 0.93). Participants in each pair did not know each other beforehand. All the participants were native speakers and had a normal or corrected-to-normal vision.

Apparatus

Both participants in a pair wore monocular Tobii eye-tracking glasses, which had a sampling rate of 30 Hz with a 56° × 40° recording visual angle capacity for the visual scene. The glasses recorded the video of the scene camera and the sound, in addition to gaze data. Interviewers read the questions and evaluated the interviewee's response on a Wacom PL-1600 15.6 Inch Tablet, which enables users to interact with the screen by using a digital pen.

Procedures

The task was a mock job interview. It is adopted from the previous studies (i.e., Andrist et al., 2013, 2014). Eight common job interview questions, adopted from Villani et al. (2012), were translated into Turkish and presented to interviewers beforehand. The interviewer was instructed to ask given questions and also to evaluate the interviewee for each question right after the response. A beeping sound was generated to indicate the beginning of a session. The participants stayed alone in the room throughout the sessions.

Data and Analysis

Data analysis consists of three main steps. In the first one, we extracted gaze directions of each participant. As the next step, we analyzed audio data for extracting high-level speech features. In the final step, we synchronized gaze direction data with speech annotations.

We have developed an open source application that provides an environment for researchers working in the field without requiring a technical background (Arslan Aydin et al., 2018). It is capable of detecting and tracking conversation partner's face automatically, overlaying gaze data on top of the face video, and incorporating speech through speech tag annotation. It automatically detects whether the extracted raw gaze data is face gaze of an interlocutor or an aversion. In addition, it provides interfaces for speech analysis involving segmentation, synchronization of pair recordings, and annotation of segments. It significantly reduces the time and effort required for manual annotation of eye and audio recording data. Manual annotation is vulnerable to human-related errors, and in addition, automatic annotation with the state-of-the-art methods provide further information that may not be extracted manually such as detecting the coordinates of facial landmarks, taking into account the error margins while annotating the gaze direction or segmentation of the speech at milliseconds precision. The application employs OpenFace (Baltrusaitis et al., 2016) for gaze direction analysis, CMUSphinx[3] for audio recording analysis, and dlib[4] for training custom face detector. We generally used interfaces of the developed application in the gaze and the speech tag set analysis.

Speech Analysis

Audio stream from each participant's recordings was extracted before performing the speech analysis. The mean duration of the recordings was 09:41.543 (*SD* = 04:05.418) (in mm:ss.ms format). We performed speech analysis with two methods both including segmentation and annotation sub-steps. As the first step of speech tag set analysis, the audio files of sessions were segmented into smaller chunks including sub-words and pauses. The number of segments (*M* = 737.4, *SD* = 414.1) varied depending on the length and the content of the audio. Since the developed application called Sphinx4 libraries for the segmentation of audio files, each segment had a maximum temporal resolution of 10 ms.

Then, in order to determine session intervals and provide synchronization between the pair recordings, we listened to audio segments and identified the ones containing beeping sound. The time offset between the pair's recordings was calculated by using the application interface. Lastly, for improving segmentation quality, the synchronized pair recordings were re-segmented via merging the time interval information of both participants' segments (see resources[5] for an example and usage of developed application).

At the annotation stage of speech tag set analysis, segments were annotated with the predefined speech labels that we decided to use by benefiting from the founding of previous social gaze studies (e.g., Kendon, 1967; Emery, 2000; Rogers et al., 2018) and also by examining the data we have collected. We considered the following factors while creating the tag set including 14 labels:

- Separate labels were identified for *Speech, Asking a Question*, and *Confirmation*.
- We classified pauses by their duration as proposed by Heldner and Edlund (2010) (*Pre-Speech, Speech Pause, Micro Pause*).
- In parallel with the turn management role of speech, we defined separate label for *Signaling End of Speech*.
- We named the conversation segment as *Thinking* when it included filler sounds, such as uh, er, um, eee, and drawls.
- As the interviewer reads the questions from the screen, the interviewer's gaze would evidently be directed toward the screen, so we tagged this case separately (*Read Question*).
- A separate label for repeating the question was identified (*Repetition of the Question*).
- We assumed that gaze direction would be affected by laughter (*Laugh, Speech While Laughing*).
- We handled *Greeting* apart from *Speech*, because we assumed that the sender would aim to signal intimacy while greeting and this might have an effect on gaze direction.

[3]The Sphinx4 is a speech recognition system jointly designed by Carnegie Mellon University, Sun Microsystems Laboratories, Mitsubishi Electric Research Labs, and Hewlett-Packard's Cambridge Research Lab. The Official website is: http://cmusphinx.sourceforge.net/.

[4]It is a C++ Library, http://dlib.net/ (accessed on April 15, 2017).

[5]See the MAGiC App Channel under YouTube, https://www.youtube.com/channel/UC2gvq0OluwpdjVKGSGg-vaQ, and MAGiC App Wiki Page under Github, https://github.com/ulkursln/MAGiC/wiki.

- The interviewers evaluated the interviewee's answer before proceeding to the next question. This evaluation process was performed by looking at the screen. If it did not meet one of the above conditions, the interval from the end of the interviewee's answer to the beginning of the new question was labeled as *Questionnaire Filling*.

The second method is dialogue act analysis using the ISO 24617-2 standard. The closer the microphone was to the participant, the cleaner and the better the gathered audio recording was. Therefore, in order to not miss any data, we transcribed the conversations by listening to the audio streams of both the interviewer and the interviewee in a pair separately. We first opened a Google Document and enabled speech to text feature, then started to articulate audio while listening to the interviewee's audio stream. After that, we listened to the same recording once more to add non-verbal vocalizations to the transcribed texts, such as Unfinished Word, Filler Sound, Laugh, Drawl, Warm-up, and so on. Adding non-verbal vocalizations is recommended by the ISO 24617-2 standard depending on their effect on the choice of communicative function, or qualifiers (ISO 24617-2, 2012; Bunt et al., 2017a). Then, while we were listening to the interviewer's audio stream for the same pair, we completed missing words in the transcription text file of a session. Thus, we reviewed the transcription of a session twice in this phase. Lastly, we divided the transcription text file into two separate files based on the source. As a result, at the end of the Transcription phase, two files per session were created in total, one for the interviewer's transcription and other for the interviewee's.

Secondly, by using the Praat[6] program, three students marked the time interval of a total of 16,716 words in 15 out of 25 sessions. When selecting these 15 sessions, we have given priority to long sessions in which dialogue act and RR tagging might be more frequent. Praat is a free application for speech analysis in phonetics. We employed only the "Transcribing speech with Praat function." As we have already transcribed audio stream, the word or non-word vocalization was copied from the transcription file and pasted into the related area in an interface. Then, the time interval of a word was specified by marking the beginning and the end. Even though we reviewed the transcript text twice in the previous phase, there would still be some missed words or non-word vocalizations. In such cases, the transcription file was updated with the missing word and/or non-word vocalization. In addition to that, after each word was processed, a controller checked if it was necessary to update the time intervals of words and transcribed texts. Thus, the transcribed text file was reviewed four times in total since its creation and word intervals were checked twice. As a result, at the end of this phase, we are left with a single transcription file and two files storing time intervals of words, one for the interviewer's transcription and the other for the interviewee's.

We segmented speech utterances into dialogue act units. As proposed by Prasad and Bunt (2015), dialogue act units were determined based on the meaning rather than the syntactic features. Dialogue act represents the communicative function that serves in a dialogue to change the state of mind of an addressee by means of its semantic content.

Since we were investigating the relation between dialogue act units and gaze direction, which was able to change quite fast, we specified dialogue act units in smaller intervals that differed from the previous and the subsequent dialogue act units in terms of communicative function, qualifiers, and RRs. Even though ISO 24617-2 supports RR annotation, it does not specify any particular set for RR. Thus, we employed another standard recommended by ISO 24617-2 for the annotation of discourse relation. ISO 24617-8, also known as ISO DR-Core, was proposed as an international standard for the annotation of discourse relations (Prasad and Bunt, 2015; Bunt and Prasad, 2016; ISO 24617-8, 2016). To understand the discourse as a whole, the relation between the sentences or clauses in the discourse (i.e., Rhetorical Relations) should be considered.

Lastly, dialogue act units were annotated on the human-friendly excel file in DiAML-MultiTab format; the workflow is presented in **Figure 2**. According to DiAML-Multitab representation, an annotator has to assign the unique ID to each dialogue act. Moreover, if there is a functional or feedback dependence between two dialogue acts, intending to represent this relation, the ID of the preceding dialogue act should be referenced by the succeeding one. We developed an excel macro[7] to automatize the process of assigning unique ID's and updating references. As suggested in the annotation guideline, whatever the way the speaker expressed himself, the following questions were considered during annotation: (i) why the speaker said it, (ii) what the purpose of the speaker in using this utterance is, and (iii) what the speaker's assumptions about the person he was addressing are. ISO 24617-2 indicates that labeling should be based on the speaker's intention, instead of what he or she says literally. Therefore, this standard proposes to think functionally rather than relying on linguistic cues, which are useful but focusing only on them could make us miss what the speaker really wants to say and that would cause false labeling[8].

ISO 24617-2 proposed nine dimensions based on the type of semantic content: Task, Turn Management, Time Management, Auto Feedback, Own Communication Management, Discourse Structuring, Social Obligation Management, Allo Feedback, Partner Communication Management, and 56 communicative functions. In the present study, we encountered 43 out of 56 communicative functions, except the following ones: Correction, Accept Offer, Decline Offer, Decline Request, Decline Suggestion, Auto Negative, Allo Negative, Feedback Elicitation, Return Self Introduction, Question, Address Offer, Address Request, and Address Suggest. Moreover, ISO DR-Core recommends 18 labels for RR annotation. In the present study, all 18 labels were included.

We calculated the intra-annotator agreement via Cohen's Kappa score to measure annotation (or annotator) reliability. More than 6 months after the first annotation, the same

[6]For detailed information, see the website: http://www.fon.hum.uva.nl/praat/.

[7]It is available under https://gist.github.com/ulkursln.

[8]A binary decision tree that can be used while determining the communicative functions and the dimensions is available for annotation of Turkish dialogues, under https://github.com/ulkursln/Dialogue-Act-Annotation.

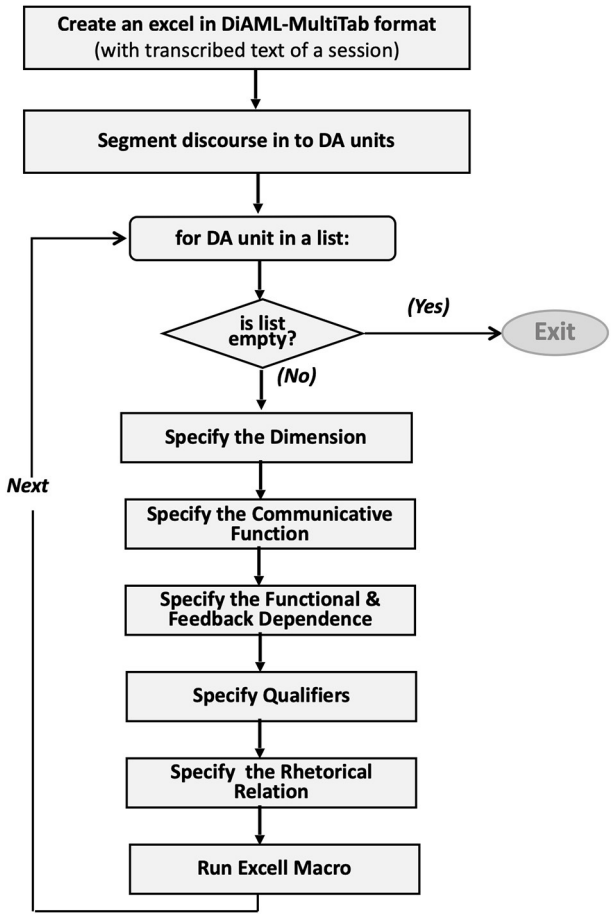


FIGURE 2 | The workflow of segmentation and annotation.

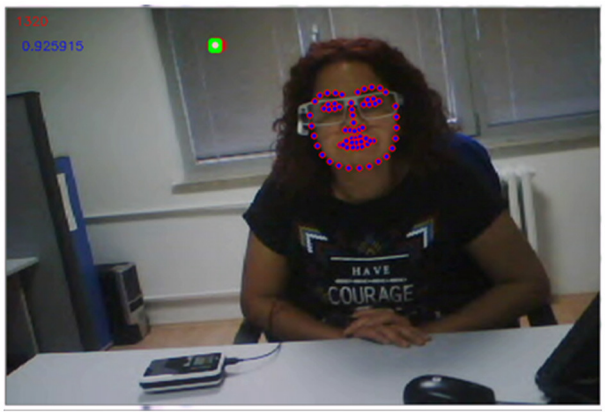


FIGURE 3 | The set of facial landmarks are presented around the face of the interviewer with pink circles. The green dot represents the gaze location of the conversation partner. Color should be used in print.

annotator annotated ∼25% of the data (corresponding to six sessions out of 25 sessions for annotations with speech tag set and four sessions out of 15 sessions for annotations with ISO 24617-2 standard). The Cohen's Kappa scores were observed to be equal to 0.85, 0.80, and 0.89 for dimensions of ISO 24617-2, communicative functions of ISO 24617-2 and speech tags, respectively ($p < 0.0001$).

Gaze Analysis

We performed gaze analysis by using the related interfaces of developed application (Arslan Aydin et al., 2018). Firstly, we exported raw data of eye movements as an output file storing x and y positions of the right eye at a resolution of 33.3 ms.

Then, in order to interpolate missing gaze data, first the scaling factor was calculated via Equation 1 (where t represents timestamp), and then the location of the first sample after gap was multiplied by the scaling factor, and lastly the result was added to the location of the last sample before the gap. The max gap length that would be filled with interpolation was chosen to be shorter than a normal blink, which was 75 ms as proposed by previous studies (e.g., Ingre et al., 2006; Komogortsev et al., 2010; Benedetto et al., 2011).

$$s_{scaling\ factor} = \frac{t_{sample\ to\ be\ replaced} - t_{first\ sample\ after\ gap}}{t_{last\ sample\ before\ gap} - t_{first\ sample\ after\ gap}}, \quad (1)$$

[taken from Olsen, 2012]

Secondly, we extracted face boundaries with the default detector proposed by the developed application. Video recordings of 28 pairs consisting of a total of 828,618 frame images were processed for gaze analysis. The face boundaries over 68 2D facial landmarks were automatically detected and stored under text files as an outcome of face-tracking process. Thirdly, we extracted Area of Interest (AOI) labels corresponding to the frame image, along with the input parameters: (i) 2D landmarks of faces; and (ii) linearly interpolated raw gaze data. AOIs provided information of whether, at a particular time, a participant was looking at the interlocutor's face, i.e., face gaze, or looking away from it, i.e., aversion. Also, the relative positions of gaze data with respect to the face on each particular frame image were stored. If the gaze position was outside the face boundary, one of eight character values, a, b, c, d, f, g, h, and i, was assigned in order to denote gaze aversion; otherwise, an e character was assigned as an AOI label to denote face gaze (see **Figure 3**).

Fourthly, we monitored the efficiency of face detection by looking at the number and percentage of extracted AOI labels in frame images. The detection of AOI labels failed due to undetected faces and/or the missing gaze data. Fifthly, we trained a custom face detector via training interface of the developed application for the video streams in case more than 30% of frame images could not be assigned to an AOI label. Then, we extracted face boundaries with the custom detector and, after that, monitored the performance. The detection percentage of AOIs that were extracted by employing either default or custom face detector were compared, and we continued the analysis with the AOIs that got the higher detection ratio. We carried on analysis for 11 records of interviewees and a single record of interviewers with AOI labels extracted by

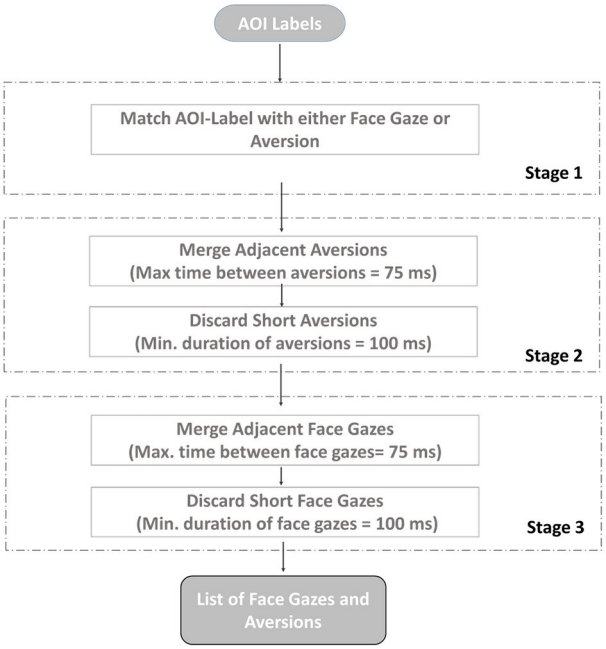


FIGURE 4 | Process flow for fixation detection of gaze directions.

employing trained detectors. For all remaining recordings, the ones extracted by employing default detectors were adopted. Sixthly, we assigned AOI labels to the frame images manually for the following cases:

- The face of the interlocutor was on frame image, yet it could not be detected automatically.
- The face of the interlocutor was on frame image, but it was not detected correctly.
- The face of the interlocutor did not exist for that particular frame image. This happens especially when an interviewer was looking at the monitor while evaluating the response or reading the question. In such cases with respect to the location of monitor, we easily inferred AOI label.

After reviewing and updating the extracted AOI labels manually, we re-monitored the performance and eliminated three pairs in which the amount of assigned AOI labels correspond to <70% of interviewers' and/or interviewees' recordings in a pair. Hence, we continued analysis with the remaining 25 pairs.

Lastly, in order to get rid of noise, saccadic movements, or blinks in the data, fixations were extracted in order to group the raw gaze data. In line with the literature (e.g., Manor and Gordon, 2003; Camilli et al., 2008; Komogortsev et al., 2010; Benedetto et al., 2011), we followed the consequent steps, as illustrated in **Figure 4**.

Multimodal Data

For each speech annotation method, the data obtained in speech and gaze analyses were merged into a single summary file. As a result, we obtained a series of gaze direction and related features taken at successive intervals of 33.3 ms.

Gaze and Speech Tag Set

The columns of the summary file were the speech tag, sender, gaze direction of sender, and of an interlocutor on the particular frame image.

Gaze and Dialogue Act

We first found the time interval of a particular dialogue unit by concatenating the time intervals of each word that produced a dialogue unit together. In the summary file, each line represented the gaze direction of a sender and of an interlocutor on the particular frame with the corresponding communicative function(s), dimension(s), sender information, and, if exist, RR(s), functional dependence(s), feedback dependence(s), certainty, and sentiment qualifier.

ANALYSIS RESULTS

All statistical analyses were carried out in R programming language (R Core Team, 2016) and publicly available[9]. We first screened data and removed outliers. After that, we checked the assumptions of analysis and consequently decided whether we should transform data and run the parametric test or the non-parametric one. We handled individual differences by employing mixed models.

Frequency

We calculated the normalized frequency by dividing the count of extracted AOIs of a particular session by the duration of that session. The paired sample *t*-test was performed to compare the frequencies of face gaze and aversion per role. The analysis revealed that there was no significant difference between the frequencies of gaze aversion ($M = 20.8$, $SE = 2.62$) and face contact ($M = 23.2$, $SE = 1.86$) for interviewers, $t_{(22)} = -1.82$, $p = 0.08$. On the other hand, interviewees' gaze aversion frequency ($M = 44.7$, $SE = 3.6$) was significantly higher than their face contact frequency ($M = 35$, $SE = 3.13$), $t_{(24)} = 2.49$, $p = 0.02$. Moreover, interviewees performed aversion ($M = 44.7$, $SE = 3.60$) and face gaze ($M = 35$, $SE = 3.13$) more frequently compared to the interviewers (aversion: $M = 20.8$, $SE = 2.62$; face gaze: $M = 23.2$, $SE = 1.86$) and the differences were significant for both aversion, $t_{(23)} = -5.03$, $p < 0.000$, and face gaze, $t_{(22)} = -3.28$, $p = 0.003$ (see **Figure 5**). It is possible for an interviewer to perform higher frequency in both gaze directions. Because there was also significant difference in the duration of gaze directions between roles, see section Duration.

We conducted analysis with the fixations instead of raw gaze data. Raw gaze data include noise and saccadic movements, which are rapid and designed to direct the fovea to the vision of interest. Saccadic behavior might be important for particular research questions like searching for visual targets, but in the present study, since we focused on maintaining gaze on the

[9]Please see https://github.com/ulkursln/R-scripts for R scripts.

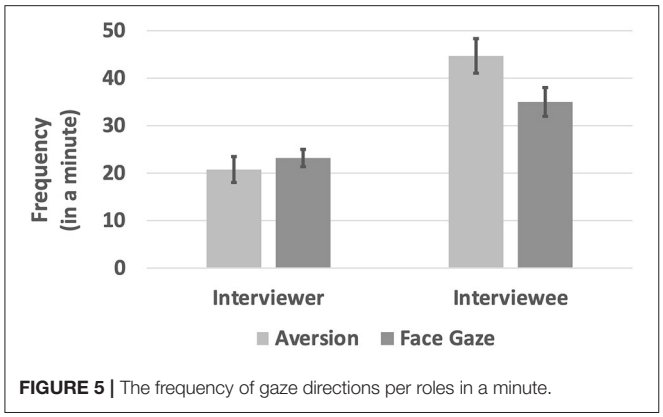


FIGURE 5 | The frequency of gaze directions per roles in a minute.

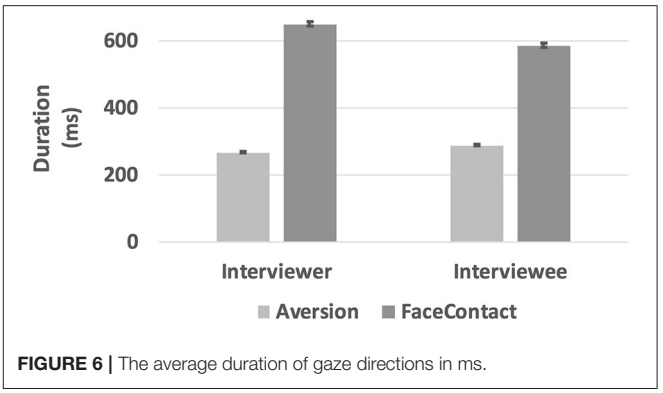


FIGURE 6 | The average duration of gaze directions in ms.

interlocutor's face or out of the face, we should eliminate jumping behaviors as well as noise from the data.

In this study, if we had worked with raw gaze data instead of fixations, we could not observe a significant effect of the role on the frequency of gaze directions. The average frequency of face gaze comprised 53% of the sessions for interviewers, whereas it was 58% for interviewees. We also examined the frequency and duration of when two participants look at each other's face at the same time, i.e., mutual face gaze. The mutual face gaze averagely comprised 27.7% ($SE = 4.51$) of the entire session, and its average duration was 517.7 ms ($SE = 0.23$).

Duration

We first screened data and removed outliers, and then tested the assumptions of the linear mixed model. Since the data were non-normal and violated the homogeneity assumption, we performed penalized quasi-likelihood (PQL) instead of linearity test. PQL is a flexible model that can deal with unbalanced design, non-linear data, and random effects.

We compared the potential models by ANOVA test to find out which one fits best. The statistical model for the duration of gaze aversion is given in Equation 2 below. Fixed effects were *Gender*, *Partner Gender*, *Role*, and their two-way and three-way interactions. In addition to that, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups. Lastly, we considered varying the slope of the interaction between *Gender* and *Partner Gender* differing across interviewers' groups.

$$\begin{aligned} \text{Fixedeffects} &= \text{Role} \times \text{Gender} \times \text{PartnerGender}, \\ \text{Randomeffects} &= 1 + \text{Gender} \times \text{PartnerGender} \\ &\quad |\text{InterviewerID/IntervieweeID}. \end{aligned} \tag{2}$$

The statistical model for the duration of face gaze is given in Equation 3. We compared the potential models by ANOVA test to find out which one fits best. Fixed effects were *Gender*, *Partner Gender*, *Role*, and their interactions. In addition to that, the mixed effect term was added for varying intercepts by interviewers, and by interviewees that are nested within interviewers' groups.

$$\begin{aligned} \text{Fixed effects} &= \text{Role} \times \text{Gender} \times \text{PartnerGender}, \\ \text{Random effects} &= 1| \text{InterviewerID} /\text{IntervieweeID}. \end{aligned} \tag{3}$$

The interviewer's face gaze duration ($M = 648.9$ ms, $SE = 7.06$) was significantly higher than the interviewee's face gaze duration ($M = 585.8$ ms, $SE = 6.06$), $t_{(10,434)} = -1.977$, $p = 0.048$. There was a significant effect of the role, i.e., being an interviewer or an interviewee, on the duration of gaze aversion (see **Figure 6**). The *post-hoc* tests revealed that a significant difference between the aversion durations of interviewers ($M = 258.2$ ms, $SE = 5.25$) and interviewees ($M = 313.2$ ms, $SE = 3.43$) was observed when the partner gender was female, $t_{(9,760)} = 5.75$, $p < 0.0001$.

Multimodal Analysis of Gaze and Speech

In multimodal analysis, we examined the relation of gaze direction with either speech tags or communicative functions. The statistical analyses were conducted on the top five labels for both annotation schemes. In this section, we will describe the analysis steps via speech tag set. Similar calculations were also performed for dialogue act analysis.

Primarily, we extracted the ratio of gaze behavior observed during an instance of speech tag set. Each instance of speech tag set might be assigned several times during a session. In Equation 4, let B be a set including percentages of aversion (A) and face gaze (FG) during occurrences of speech tags, for session x and participant p, where i is the element of F, which is a set of frame IDs labeled with particular speech tag. D function gets the frame IDs and type of gaze direction, namely, A or FG, as input parameters and returns the durations of that specified gaze direction among those frames.

$$B_{x,p}(S, A) = \left\{ i \in F_s : \frac{D(i, A)}{D(i, A) + D(i, FG)} \right\}, \tag{4}$$

The process details are given in **Table 3**. A sample implementation of Equation 4 for **Table 3** would be as follows:

Frame Set:

$F_{s1} = \{[1–9], [46–95]\}$

Gaze Directions:

$D([1-9], A) = \{6\}$; $D([46-95], A) = \{25, 14\}$

TABLE 3 | Illustration of calculating the ratio of gaze direction (GD) to the particular speech-tags, $S_{1,\#index}$ and $S_{2,\#index}$.

Frame no.	Speech-Tag(S) ($_{id,\#index}$)	GD	Ratio of GD duration
1	$S_{1,1}$	A	$\|A\| / \|S_{1,1}\| = 6/9$
2		A	
3		A	
4		A	
5		A	
6		A	
7		FG	$\|FG\| / \|S_{1,1}\| = 3/9$
8		FG	
9		FG	
26	$S_{2,1}$	FG	$\|FG\| / \|S_{2,1}\| = 10/20$
27		FG	
28–35		FG	
36–44		A	$\|A\| / \|S_{2,1}\| = 10/20$
45		A	
46	$S_{1,2}$	A	$\|A\| / \|S_{1,2}\| = 25/50$
47		A	
48–70		A	
71		FG	$\|FG\| / \|S_{1,2}\| = 11/50$
72–81		FG	
82		A	$\|A\| / \|S_{1,2}\| = 14/50$
83–95		A	

Only the interviewer's gaze behavior is considered. A similar calculation is also performed for interviewees. We intentionally skipped the frames between 10 and 25 to simulate realistic data. During the analysis, we excluded the frames in which there was no extracted gaze direction for the interviewer or interviewee.

$$D([1-9], FG) = \{3\}; D([46-95], FG) = \{11\}$$
Set of Aversion Percentages, during S_1:
$$B_{1,interviewer}(S_1, A) = \{i \in \{[1\text{–}9], [46\text{–}95]\}: \frac{D(i,A)}{(D(i,A)+D(i,FG))}\}$$
$$B_{1,interviewer}(S_1, A) = \{6/9, 25/50, 14/50\}$$
Set of Face Gaze Percentages, during S_1:
$$B_{1,interviewer}(S_1, FG) = \{i \in \{[1\text{–}9], [46\text{–}95]\}: \frac{D(i,FG)}{(D(i,A)+D(i,FG))}\}$$
$$B_{1,interviewer}(S_1, FG) = \{3/9, 11/50\}$$

As well as the duration, we also calculated the frequency of gaze directions during a particular speech tag. This time, we just consider the fixation counts of related gaze direction. For instance, in **Table 3**, the frequency of face gaze was one for $S_{1,2}$, whereas the frequency of aversion was two. Thus, the percentages were 1/3 and 2/3, respectively.

Speech Tag Set Annotation

The data were non-normal and violated the homogeneity assumption; thus, we performed PQL. The statistical model is described by Equation 5. Fixed effects were *Role*, *Speech tag*, their mutual interaction, *Interviewer's Gender*, *Interviewee's Gender*, and their mutual interaction. Besides, the mixed effect term was added for varying intercepts by interviewers and by interviewees that are nested within interviewers' groups. Lastly, we added the *Speech tag ID*, which was a unique identifier for each occurrence of speech tag, as a mixed effect term.

$$\begin{aligned}\text{Fixed effects} &= \text{Role} \times \text{SpeechTag} + \text{Interviewer's Gender} \\ &\quad \times \text{Interviewee's Gender}, \\ \text{Random effects} &= 1 \mid \text{InterviewerID/IntervieweeID} + 1 \mid \\ &\quad \text{Speech tag ID}. \end{aligned} \tag{5}$$

There was a significant difference in the frequency of gaze direction ratios between the interviewers and interviewees when the speech tag was *Thinking* [$t_{(6,840)} = 13$, $p < 0.0001$], *Speech* [$t_{(6,840)} = 12.9$, $p < 0.0001$], *Speech Pause* [$t_{(6,840)} = 10.8$, $p < 0.0001$], or *Micro Pause* [$t_{(6,840)} = 7.23$, $p < 0.0001$] (see **Figure 7**).

We also examined the difference in duration of gaze direction between the interviewers and interviewees. Similarly, results revealed that when the speech tag was *Thinking* [$t_{(6,840)} = 13.3$, $p < 0.0001$], *Speech* [$t_{(6,840)} = 12.9$, $p < 0.0001$], *Speech Pause* [$t_{(6,840)} = 10.7$, $p < 0.0001$], or *Micro Pause* [$t_{(6,840)} = 7.8$, $p < 0.0001$], interviewee's gaze aversion duration was significantly longer than the interviewer's.

Dialogue Act Annotation

The data were non-normal and violated the homogeneity assumption; thus, we performed PQL. The statistical model is described by Equation 6. Fixed effects were *Role*, *Communicative Function*, their mutual interaction, *Interviewer's Gender*, *Interviewee's Gender*, and their mutual interaction. In addition, the mixed effect term was added for varying intercepts by interviewers and by interviewees that are nested within interviewers' groups. Lastly, we also added the *Communicative Function ID*, which was a unique identifier for each occurrence of communicative functions, as a mixed effect term.

$$\begin{aligned}\text{Fixed effects} &= \text{Role} \times \text{Communicative Function} \\ &+ \text{Interviewer's Gender} \times \text{Interviewee's Gender} \\ \text{Random effects} &= 1 \mid \text{InterviewerID/IntervieweeID} + 1 \mid \\ &\text{Communicative Function ID}. \end{aligned} \tag{6}$$

There was a significant difference in the frequency of gaze direction ratios between the interviewers and interviewees when the communicative function was *Answer* [$t_{(5,334)} = 13.1$, $p < 0.0001$], *Stalling* [$t_{(5,334)} = 19.9, p < 0.0001$], or *Turn Take* [$t_{(5,334)} = 5.69, p < 0.0001$] (see **Figure 8**).

We also examined the difference in the duration of gaze direction between the interviewers and interviewees. Similarly, results revealed that when the communicative function was *Answer* [$t_{(5,334)} = 14.2$, $p < 0.0001$], *Stalling* [$t_{(5,334)} = 19.8$, $p < 0.0001$], or *Turn Take* [$t_{(5,334)} = 5.58$, $p < 0.0001$], interviewee's gaze aversion duration was significantly longer than the interviewer's.

A DEEP COMPUTATIONAL MODEL

For computational modeling, we use CNNs. CNNs are specialized versions of fully connected networks with localized receptive fields. In the present study, we adapted simplified

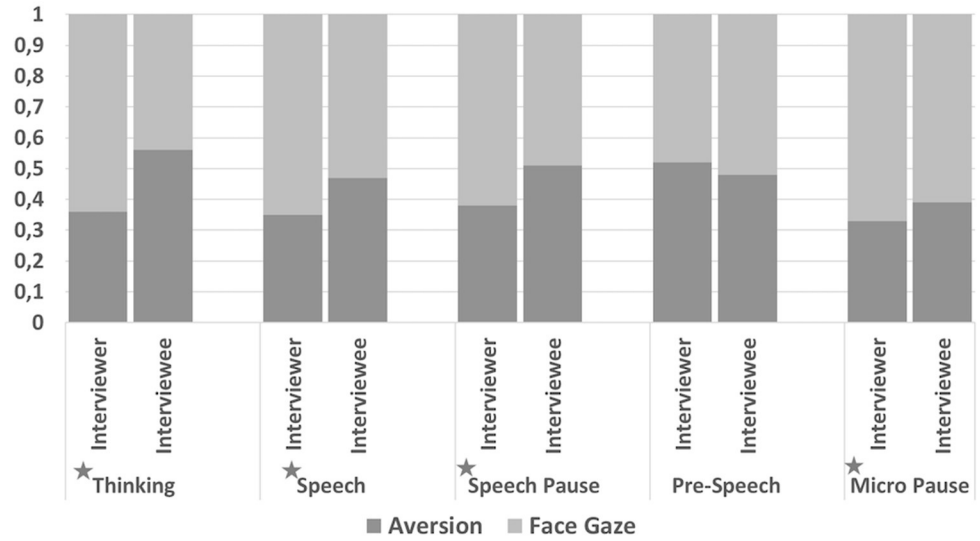


FIGURE 7 | Frequency of gaze direction ratios for the top five speech tags observed in the collected data. Since the gaze direction can be either face gaze or aversion, a total ratio for all bars are 1. Significant differences are presented with * character.

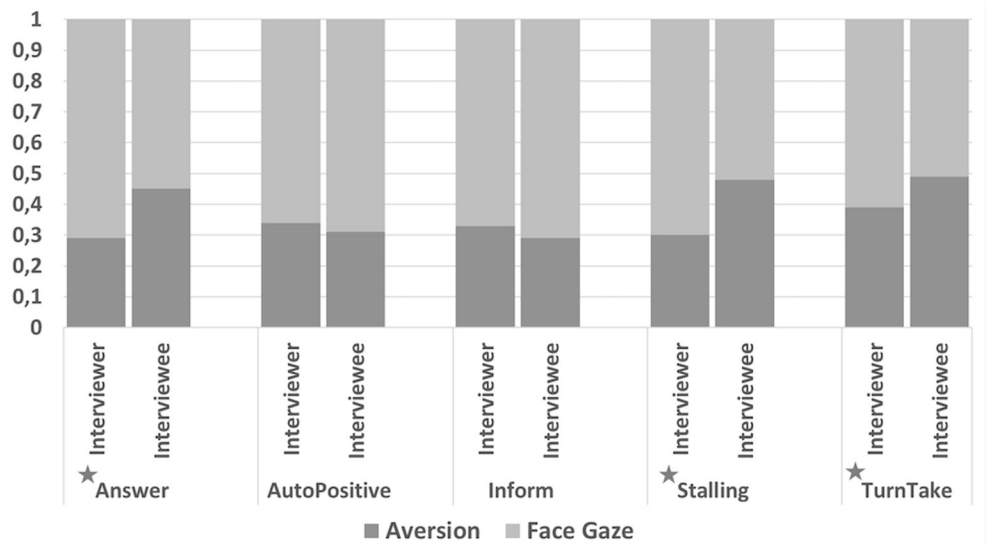


FIGURE 8 | Frequency of gaze direction ratios for the top five communicative functions observed in the collected data. Since the gaze direction can be either face gaze or aversion, a total ratio for all bars are 1. Significant differences are presented with * character.

versions of two state-of-the-art CNN architectures, namely, ResNet (He et al., 2016) and VGGNet (Simonyan and Zisserman, 2015).

We collected gaze data in the form of a time series and trained two 1D CNN networks. In 1D CNNs, data points in time series are generally introduced to the network as a window of instances. The window is slid in time by a number of time steps, which is called stride. For instance, for a two-channel signal consisting of eight time steps, a window size of four and stride of two would yield three input samples with a size of 4×2 (see **Figure 9**).

We adapt two CNN architectures (VGGNet and ResNet) and called them gazeVGG and gazeResNet (see **Figure 10**). Batch normalization, pooling, weight regularization, and dropout were applied to both networks for handling overfitting.

Data Presentation Details

In the present study, we obtained a series of gaze direction and related features at successive intervals of 33.3 ms. According to the data obtained from the human–human experiment (see section Experimental Investigation), the average gaze aversion

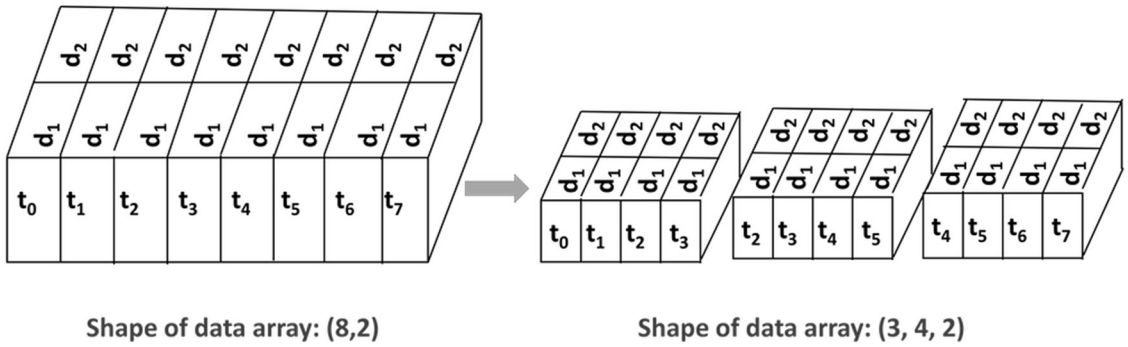


FIGURE 9 | The time-series data and how they are prepared before processing with the deep networks. On the left, the input has a size of 8 × 2 where the number of time steps is eight and the number of channels is two. On the right-hand side, the shape has changed to 3 × 4 × 2 where the window size is four and the stride is two.

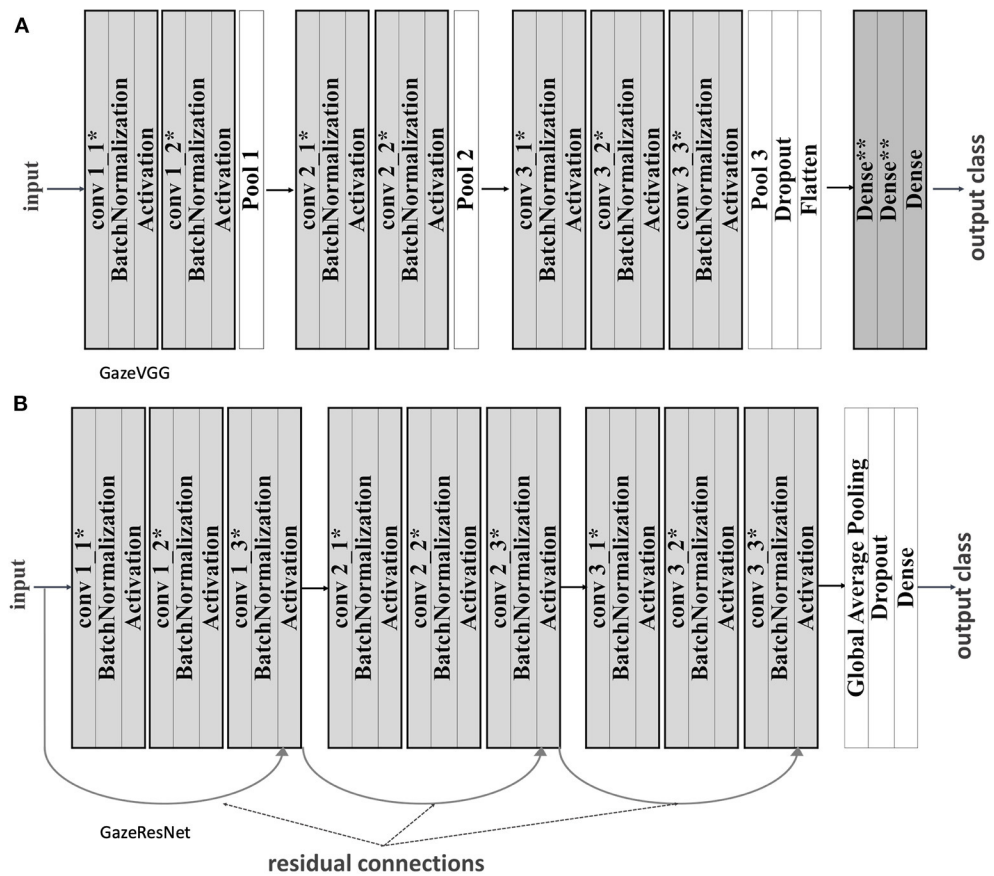


FIGURE 10 | (A) GazeVGG architecture with batch normalization, regularization, and pooling. **(B)** GazeResNet architecture with batch normalization, regularization, and pooling. There is convolution between blocks and a residual connection between the last item of the previous block and the current one. *L2 weight and bias regularizers were applied. **L2 weight regularizer was applied.

duration was ∼300 ms. Therefore, we used nine as the window size as single frame took 33.3 ms, and since the minimum fixation duration was 100 ms, we set stride to three.

In our experimental design, while the interviewees participated in a single session, interviewers took part in multiple interviews. We designed our computational models for predicting gaze direction of interviewers. At first, we applied One-Hot-Encoding to convert categorical data into numbers. For the input data including speech annotation with the speech tag set, we used a total of 20 features including

Sender, *Speech Instance*, *Gender*, *Is the Same Person*, and *Interviewee's Gaze direction*. On the other hand, a total of 137 channels involving *Sender*, *Gender*, *Is the Same Person*, *Interviewee's Gaze direction*, *Communicative Function*, *Dimension*, *Certainty*, *Sentiment*, *Functional Dependence*, *Feedback Dependence*, *Rhetorical Relation*, and *Argument Number of Rhetorical Relations* were utilized for the dialogue act models. Therefore, for a window of size nine, a single input to a CNN had 180 dimensions for data annotated with the speech tag set and 1,233 dimensions for data annotated with ISO 24617-2 standard.

Training and Implementation Details

For both CNNs, binary-cross entropy was used as the objective function, which was minimized using Adam optimizer. Moreover, we used dropout with a value of 0.2 and L2 regularization with a value of 0.001. We trained gazeResNet models for 100 epochs with a batch size of 64. Similarly, we trained gazeVGG models for 100 epochs with a batch size of 64 and pool sizes of 2. We have empirically changed and evaluated the different settings for L1, L2, epoch count, window size, stride, etc., and we have provided the best settings. In the hyper-parameter tuning phase, we used backtesting that is specific to the time series as a cross-validation method. We trained gazeVGG and gazeResNet models with 16 or 32 filters in the first block and, taking input data, annotated either with an ISO 24617-2 standard (i.e., dialogue acts) or speech tag set. For the data annotated with dialogue acts with 32 filters in ResNet and 16 filters in VGG, and for the data annotated with a speech tag set, 32 filters in both VGG and ResNet achieved better accuracies.

In the *n*-fold back-testing, the ratio of data provided for the training and validation is different at each split. It is five for the fifth split and one for the first split. When the training data are not big enough, the network might not quite learn about the underlying trend of the data. For instance, in the present study, the second and the third interviewer had a greater tendency to aversion whereas the sixth one had a tendency in the opposite direction. Hence, especially for the second and the fourth splits, the distribution of data for training and testing was different, which resulted in validation fluctuations. Particular orders of interviewers in the input data result in specific orders of interviewers in splits used for training and validation. This might cause testing the network with a different distribution than the one used in training. The classical cross-validation method enables one to handle such distribution issues by randomly dividing the set of input data into training and test sets. However, time-series data have temporal relations that prevent randomized division. In order to overcome this issue, we trained and evaluated the models by building 5-fold cross-validation with data sets created by shuffling the orders of interviewers in the input data while preserving temporal order within each session.

Training was performed on Google Colab, which is a free Jupyter notebook environment provided by Google. Colab offers Tesla K80 GPU. The training codes were implemented in Python 3.0 by Keras libraries with Tensorflow backend.

Results

We analyzed the gaze prediction performances of the two CNN architectures. **Table 4** lists the performances with both dialogue act and speech tag inputs. We see that models running on the data annotated with speech tags generally perform better than the ones running on the data annotated with dialogue acts.

In order to examine the quantitative differences between classification accuracy of the models, we also analyzed confusion matrices in **Table 5**, which contain the percentages of false and correct estimations. We notice that models with both speech tag and dialogue act could predict the direction of face gaze with similar and relatively high accuracies (i.e., speech tag set model achieved 85.1% accuracy and dialogue act model achieved 94.8% accuracy), whereas there was a difference in the prediction accuracies of aversion between the models. Speech tag set model could predict aversions better than Dialogue act model.

The performances of GazeResNet models were also assessed via calculating the recall, precision, and *F* scores. In predicting aversions, a precision of 0.69, a recall of 0.63, and an *F* score of 0.65 were obtained for the data annotated with speech tag scheme, while dialogue act scheme yielded a precision value of 0.65, a recall of 0.22, and an *F* score of 0.33.

DISCUSSION

Face-to-face communication is inherently multimodal. Gaze provides an effective way to receive and send information in a face-to-face interaction as a non-verbal communication channel accompanying speech. When studying gaze and speech, it is necessary to decide from which level both models will be addressed. Low-level eye movements, anatomic features of the eye, and kinematics of eye movements have been extensively studied by physiologists. However, although there exist studies in the related fields, eye movements have some other high-level characteristics that are still waiting to be resolved, like when they occur, how long they last, and what their roles are in communication (Ruhland et al., 2015). As in the gaze studies, researchers have dealt with the speech at different levels for modeling non-verbal communication components driven by speech (Cassell et al., 1999; Zoric et al., 2011; Marsella et al., 2013).

Experimental Analysis

In the present study, we investigated the roles of the high-level characteristic of eye movements driven by high-level features of speech in a face-to-face interaction. The two major research questions of the study were: "What are the underlying features of gaze direction among humans" and "What is the relation between gaze and speech to achieve conversational goals in a specified face-to-face interaction?" To examine these questions, we conducted a mock job interview task. Twenty-eight pairs consisted of seven professional interviewers and 28 interviewees took part in the study. They wore Tobii glasses throughout the study.

We automated the analysis mostly by utilizing the state of the art methods. That way, we aimed to overcome some methodological problems and reduce the amount of human-related errors and the time necessary for annotation. We

TABLE 4 | Performances of computational models with 5-fold cross-validation.

Tagging scheme	CNN architecture	Avg. training accuracy (%)	Test accuracy of folds (%)	Avg. test accuracy (%)
Dialogue act	*VGG*	83.2 (*SD*: 1.20)	89.5, 76.7, 70.6, 60.3, 57.1	69.6 (*SD*: 11.3)
	ResNet	83.1 (*SD*: 0.88)	87.7, 77.1, 70.8, 59.8, 58	70.7 (*SD*: 12.3)
Speech tag set	*VGG*	81.1 (*SD*: 0.18)	83.2, 68.6, 81.5, 76.9, 74.6	76.9 (*SD*: 5.82)
	ResNet	81.1 (*SD*: 0.14)	**84.6, 69.6, 81.4, 82, 76.2**	**78.8** (*SD*: 5.94)

The highest test accuracy was obtained with the GazeResNet model when applied on data annotated with the speech tags. Those accuracy values are presented in bold.

TABLE 5 | Confusion matrix of the GazeResNet models with the highest performances for each tagging scheme.

		Predicted class	
		Speech tag set/Dialogue act (%)	
		Face gaze	Aversion
Actual class	Face gaze	85.1/94.8	14.9/5.2
	Aversion	23.7/46.0	76.3/54.0

It represents the percentages of true and false predictions made on actual classes, i.e., aversion and face gaze. The percentage of true aversion predictions is 76.3% for the Speech tag set model, while it is 54% for the dialogue act model.

used an open source project (Arslan Aydin et al., 2018) that provided interfaces for the analysis of gaze involving face detection and identification of gaze direction. Moreover, it enabled speech analysis including segmentation, annotation, and synchronization of pair's recordings.

Gaze direction was identified as either face gaze or gaze aversion based on the decision whether the participant was looking at the other person's face or not. The gaze analysis was carried out in three steps: (i) determining the boundaries of the face, i.e., face detection; (ii) deciding whether the partner's gaze was within those boundaries, i.e., identification of gaze direction; and (iii) fixation detection.

We monitored the ratio of unidentified gaze direction on frame images of recordings. We observed that the AOI identification rate on the frame images of the 11 interviewees' recordings and two interviewers' recordings was <70%. By visualizing the recordings frame by frame, we realized that even there exist gaze raw data of interviewees, the interlocutors' (i.e., interviewers') face might not be detected while they were reading a question or evaluating the responses of an interviewee by turning their head and accordingly face to the screen. For such cases, we trained a custom face detector instead of using Haar-Cascade classifiers, which were provided by the OpenFace software, as the default detector. Moreover, in order to minimize data loss, we manually determined the gaze direction on frame images if they could not be detected automatically, but it was possible to identify their AOI labels, like in the cases when the face of the interlocutor was on frame image but could not be tracked automatically.

We observed that interviewees performed face gaze and aversion significantly more frequently when compared to interviewers. Moreover, the gaze aversion durations of interviewers were significantly longer than those of interviewees. On the other hand, face gaze durations of interviewees were significantly longer than that of interviewers. When we examined gaze direction per role, we found that there was no difference between the frequencies of gaze aversion and face gaze for interviewers, while a significant difference was observed for interviewees. Interviewees avert their gaze more frequently compared to performing face gaze. These findings are in line with the conclusions summarized by Kendon (1967) in his detailed study investigating the function of gaze in a face-to-face conversation. Kendon (1967) stated that individuals tend to look at others more frequently when listening compared to speaking and the glances of speakers would be shorter than the listeners. He had grouped the roles in the conversation as speakers and listeners. In the present study, due to the role of interviewees, they spoke more frequently than the interviewers. Comparing interviewers and interviewees, the gaze direction of the latter was more similar to that of the speakers mentioned in Kendon (1967).

Broz et al. (2012) studied mutual gaze in a face-to-face conversation with participants wearing eye-tracking devices. They observed a mutual face gaze occurring for about 46% of a conversation. Rogers et al. (2018) also conducted a dual eye-tracking study and reported that the mutual face gaze comprised 60% of the conversation with 2.2 s duration on average. On the other hand, when cumulative data of all sessions are taken into account, we found a lower ratio in the present study, which was 27.7% ($SE = 4.51$), and the average duration was 517.7 ms ($SE = 0.23$), possibly due to differences in data collection settings and analysis methods as reviewed below.

There are two crucial steps in determining mutual face gaze: (i) deciding whether the gaze of an individual was inside the face boundaries of an interlocutor, and (ii) synchronization of recordings exported from eye-trackers. Broz et al. (2012) and Rogers et al. (2018) manually annotated gaze direction

in each frame. However, in the present study, interlocutor's face boundaries were detected based on 68 facial landmark points and gaze direction was generally decided automatically. Manual coding of gaze direction might be open to human-related errors. Compared to the previous studies, we employed state-of-the-art technologies for face boundary detection. Moreover, because of the hardware or operational constraints, eye-tracking devices might estimate gaze positions with deviations. Eye tracker manufacturers provide the estimated error that is specific to device in degrees for the visual angle. In the present study, we utilized the developed application (Arslan Aydin et al., 2018), which automatically considers such error margins to estimate gaze direction, to visualize gaze and face boundaries overlaid on a frame image. It is not possible to take exact error margin into account just by visualizing data without benefiting from proper scripts. For instance, Rogers et al. (2018) used 15 pixels for the size of the circle that represents the gaze position. They decided on a size of 15 pixels to achieve a balance between comfort in the manual coding process while providing distinguishable regions. In addition, using fixations instead of raw gaze data and the methods adapted for fixation extraction and synchronization of pair recordings might also affect the findings. Also, differences in eye-tracking equipment, cultures, spoken language, and experimental procedures might have an impact on the variety of the reported ratio of mutual face gaze and its duration. For instance, we performed a mock job interview task; on the other hand, the ratio of gaze directions of participants might be different in conversations without a predetermined topic.

We handled speech analysis by employing two annotation methods. In the first one, discourse and rhetorical relations were annotated with standards of ISO 24617-2 and ISO 24617-8, respectively. As a second method, we used an alternative set of speech tags that we produce based on studies in the role of eye movements in social communication and also based on our observations on the data that we collected. Our aim of annotating speech with the produced speech tag set is not to propose an alternative scheme for speech annotation but instead to investigate the characteristics of speech that produce better performance in modeling social gaze. Then we conducted analysis, to see the relation between gaze and speech. There was a significant difference in the frequency of gaze directions between the interviewers and interviewees when the speech tag was *Thinking*, *Speech*, *Speech Pause*, and *Micro Pause*. Interviewees' gaze aversion frequency was higher for all those cases. We performed similar analysis for dialogue acts. This time, we found that, there was a significant difference between the interviewer and interviewee when the communicative functions were *Answer*, *Stalling*, and *Turn Take*. Similarly, for all these three communicative functions, gaze aversion frequency was higher for interviewees compared to interviewers.

Computational Models

The present study investigated further research questions to improve the methodology of multimodal analysis of communication, as follows: "How can we computationally model gaze direction with the high-level features of speech" and "How appropriate is employing discourse analysis scheme, namely, ISO 24617-2 standard, in a computational model of gaze direction?" To this aim, we trained two common Convolutional Neural Network (CNN) architectures, namely, VGGNet and ResNet. According to the experimental design, each interviewee took part in a single session whereas an interviewer attended more than one session. Therefore, we collected more data for each individual interviewer compared to an interviewee. We trained computational models to predict the gaze direction of interviewers.

We trained GazeVGG and GazeResNet models with 16 or 32 filters in the first block and, taking input data, annotated with either ISO 24617-2 standard or speech tag set. We observed that GazeResNet models achieved better accuracies for both annotation methods due to VGG bottleneck, which causes loss of generalization capability after some depth whereas ResNet handles this vanishing gradient problem by using residual connections. Moreover, we found that the speech tag set gave rise to better performances compared to dialogue act annotations. Although both GazeResNet models predicted face gaze with higher accuracies, ISO 24617-2 standard was not good at predicting aversions. Compared to data annotated with dialogue acts, Speech tags are more constant over time. Therefore, attributing the difference in the accuracy of models to that would not be a correct interpretation. The probable reasons might be the differences in the number of features and the number of input data. In addition, speech tag set involves *Pre-Speech*, *Speech Pause*, and *Micro Pause* for annotation of pauses whereas ISO 24617-2 standard does not handle pauses.

We obtained a series of gaze direction and related features at successive intervals of 33.3 ms in the present study. According to the human–human experiment data (section Experimental Investigation) the average gaze aversion duration was ∼300 ms. Therefore, we used nine as the window size since a single frame took 33.3 ms. However, different values of window-size and stride may lead to differences in the success ratio of the models. Moreover, we just used the previous features in the training. For instance, to predict the gaze direction at t_i, the features between t_{i-8} and t_i were presented to the network. However, we could get information from the subsequent frames since we conducted an offline analysis. For instance, it might be necessary to evaluate the entire speech up to t_{i+10} to decide whether the speech label at t_i was a *Question*. This constraint should be addressed in an online system. We think that one way to address this concern is as follows: Based on available data at the time of a prediction, confidence values might be assigned to all potential labels.

As presented in **Table 4**, even though we applied pooling, weight, and dropout regularizations, there was still a difference of around 10% between training and test accuracy performances of the models that receive the data annotated by ISO 24617-2 standard. To get a more robust estimation about how accurately models make predictions on unseen data, we then performed 10-fold cross-validation on those data by splitting the last 10% of data for testing in each iteration. We obtained accuracy performances similar to the 5-fold validation. Early stopping

and increasing the size of input data might improve the model's generalization capability.

CONCLUSION

We investigated gaze accompanying speech in a face-to-face interaction. Firstly, we studied the characteristics of gaze and its relations with speech with an experimental research conducted via mobile eye-tracking devices. The results indicate that the frequency and duration of gaze differ significantly depending on the role. We showed that these differences could not be observed in the analysis performed with raw gaze data instead of detected fixations. As in some of the previous studies, performing gaze analysis with raw gaze data or with detected fixations by using black box solutions is inadequate to obtain comparable results. Moreover, in multimodal analysis, it is important to automate annotations with the state-of-the-art methods. Manual annotation is vulnerable to human-related errors, and in addition, automatic annotation with the state-of-the-art methods provide further information that may not be extracted manually, such as detecting the coordinates of facial landmarks, taking into account the error margins while annotating the gaze direction or segmentation of the speech at milliseconds precision. In the multimodal analysis, we find the significant effect of speech tag set instances and communicative functions, those related with time and turn management, in the gaze directions.

Secondly, we developed CNN models of gaze direction in a face-to-face interaction. At the computational model of gaze, we observed that annotation with a simple tag set leads to a better performance despite the higher effort spent for making the dialogue act annotation on the same data. It might be due to the differences in the number of features and input data, but also a specific difference between the two annotation methods is whether *Pauses* are addressed. The speech tag set involves *Pre-Speech* (i.e., warming up the voice), *Micro Pause* (i.e., gaps up to 200 ms, as proposed by Heldner and Edlund, 2010), and *Speech Pause* (i.e., pauses that are not included in the other two categories) for annotation of pauses. However, the dialogue act annotation does not handle pauses. This suggests that multimodality should be taken into account when proposing automatic speech annotation schemes. Even though there was no verbal communication, *Pauses* during a conversation had an impact on non-verbal signals and, thus, on the interaction. This finding may be justified by the fact that in natural settings, listeners comprehend the speakers' messages by integrating both non-verbal and verbal channels in multiple channels (Kelly et al., 2015). In addition, results showed that CNN, especially ResNet models, allows us to predict high-level features of eye movement with high-level features of speech.

As future work, other non-verbal cues accompanying speech might be experimentally investigated to examine their characteristics, roles, and relations in social communication. In addition, the effect of language, culture, and personal differences might be investigated to assess the generalizability of the result. Moreover, neural network models mimic humanly cognitive faculty at the behavioral level. Thus, such models do not represent the process that take place in the brain. There exist articles discussing the capabilities of DNNs (e.g., Cichy and Kaiser, 2019). Despite the advances and rapid adaptation of deep neural networks in various fields, their lack of interpretability remains a problem. In particular, the visualization of 1D-CNN models that take the input data as 1D vector is relatively new; however, considering its explanatory power, future studies can be done to explore the effect of input features.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found at: http://dx.doi.org/10.17632/7v728yymmk.2.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Applied Ethics Research Center, METU. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

AUTHOR CONTRIBUTIONS

ÜA: conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, and visualization. SK: conceptualization, methodology, writing—review and editing, and supervision. CA: conceptualization, methodology, writing—review and editing, supervision, project administration, and funding acquisition. All authors contributed to the article and approved the submitted version.

FUNDING

This project has been supported by TÜBITAK 117E021 a gaze-mediated framework for multimodal Human Robot Interaction.

REFERENCES

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

Andrist, S., Mutlu, B., and Gleicher, M. (2013). "Conversational gaze aversion for virtual agents," in *Intelligent Virtual Agents. IVA 2013. Lecture Notes in Computer Science, Vol. 8108*, eds R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira (Berlin; Heidelberg: Springer), 249–262. doi: 10.1007/978-3-642-40415-3_22

Andrist, S., Tan, X. Z., Gleicher, M., and Mutlu, B. (2014). "Conversational gaze aversion for humanlike robots," in *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction* (Bielefeld), 25–32. doi: 10.1145/2559636.2559666

Argyle, M., Lefebvre, L., and Cook, M. (1974). The meaning of five patterns of gaze. *Eur. J. Soc. Psychol.* 4, 125–136. doi: 10.1002/ejsp.2420040202

Arslan Aydin, Ü., Kalkan, S., and Acarturk, C. (2018). MAGiC: a multimodal framework for analysing gaze in dyadic communication. *J. Eye Mov. Res.* 11. doi: 10.16910/jemr.11.6.2

Baltrusaitis, T., Robinson, P., and Morency, L. P. (2016). "OpenFace: an open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016* (Lake Placid, NY: Institute of Electrical and Electronics Engineers Inc.). doi: 10.1109/WACV.2016.7477553

Batrinca, L., Stratou, G., Shapiro, A., Morency, L. P., and Scherer, S. (2013). "Cicero - Towards a multimodal virtual audience platform for public speaking training," in *Intelligent Virtual Agents. IVA 2013. Lecture Notes in Computer Science, Vol. 8108*, eds R. Aylett, B. Krenn, C. Pelachaud, and H. Shimodaira (Berlin; Heidelberg: Springer), 116–128. doi: 10.1007/978-3-642-40415-3_10

Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., and Montanari, R. (2011). Driver workload and eye blink duration. *Transp. Res. Part F Traffic Psychol. Behav.* 14, 199–208. doi: 10.1016/j.trf.2010.12.001

Broz, F., Lehmann, H., Nehaniv, C. L., and Dautenhahn, K. (2012). "Mutual gaze, personality, and familiarity: dual eye-tracking during conversation," in *2012 IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication* (Paris), 858–864. doi: 10.1109/ROMAN.2012.6343859

Bunt, H., Petukhova, V., and Fang, A. C. (2017a). Revisiting the ISO standard for dialogue act annotation. *Jt. ISO-ACL Work. Interoper. Semant. Annot.* Available online at: https://www.iso.org/standard/76443.html (accessed August 25, 2020).

Bunt, H., Petukhova, V., Malchanau, A., Fang, A., and Wijnhoven, K. (2019). The DialogBank: dialogues with interoperable annotations. *Lang. Resour. Eval.* 53, 213–249. doi: 10.1007/s10579-018-9436-9

Bunt, H., Petukhova, V., Traum, D., and Alexandersson, J. (2017b). "Dialogue act annotation with the ISO 24617-2 standard," in *Multimodal Interaction with W3C Standards: Toward Natural User Interfaces to Everything* (Cham: Springer International Publishing), 109–135. doi: 10.1007/978-3-319-42816-1_6

Bunt, H., and Prasad, R. (2016). "ISO DR-Core (ISO 24617-8): core concepts for the annotation of discourse relations," in *Proceedings 12th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-12)* (Portoroz), 45–54.

Camilli, M., Nacchia, R., Terenzi, M., and Di Nocera, F. (2008). ASTEF: a simple tool for examining fixations. *Behav. Res. Methods* 40, 373–382. doi: 10.3758/BRM.40.2.373

Carlson, L., Marcu, D., and Okurowski, M. E. (2001). "Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue* (Aalborg: Association for Computational Linguistics (ACL)), 1–10. doi: 10.3115/1118078.1118083

Cassell, J., Torres, O. E., and Prevost, S. (1999). "Turn taking vs. discourse structure," in *Machine Conversations. The Springer International Series in Engineering and Computer Science, Vol. 511*, ed Y. Wilks (Boston, MA: Springer). doi: 10.1007/978-1-4757-5687-6_12

Chidambaram, V., Chiang, Y. H., and Mutlu, B. (2012). "Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues," *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction.* (Boston, MA). doi: 10.1145/2157689.2157798

Cichy, R. M., and Kaiser, D. (2019). Deep neural networks as scientific models. *Trends Cogn. Sci.* 23, 305–317. doi: 10.1016/j.tics.2019.01.009

Dbabis, S. B., Ghorbel, H., Belguith, L. H., and Kallel, M. (2015). "Automatic dialogue act annotation within Arabic debates," in *Computational Linguistics and Intelligent Text Processing. CICLing 2015. Lecture Notes in Computer Science, Vol. 9041*, ed A. Gelbukh (Cham: Springer), 467–478. doi: 10.1007/978-3-319-18111-0_35

Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *J. Pers. Soc. Psychol.* 23, 283–292. doi: 10.1037/h0033031

Ekman, P. (1979). "About brows: emotional and conversational signals," in *Human Ethology*, eds D. In von Cranach, M. Foppa, K. Lepenies, and W. Ploog (Cambridge: Cambridge University Press), 169–249.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.* 24, 581–604. doi: 10.1016/S0149-7634(00)00025-7

Farroni, T., Csibra, G., Simion, F., and Johnson, M. H. (2002). Eye contact detection in humans from birth. *Proc. Natl. Acad. Sci. U.S.A.* 99, 9602–9605. doi: 10.1073/pnas.152159999

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., and Muller, P. A. (2019). Deep learning for time series classification: a review. *Data Min. Knowl. Discov.* 33, 917–963. doi: 10.1007/s10618-019-00619-1

Fukayama, A., Ohno, T., Mukawa, N., Sawaki, M., and Hagita, N. (2002). "Messages embedded in gaze of interface agents - impression management with agent's gaze," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* (Minneapolis, MN). doi: 10.1145/503384.503385

Garau, M., Slater, M., Vinayagamoorthy, V., Brogni, A., Steed, A., and Sasse, M. A. (2003). "The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* (Ft. Lauderdale, FL). doi: 10.1145/642700.642703

Gatys, L., Ecker, A., and Bethge, M. (2016). A neural algorithm of artistic style. *J. Vis.* 16:326. doi: 10.1167/16.12.326

Gerwing, J., and Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: a comparison of three methods. *Gesture* 9, 312–336. doi: 10.1075/gest.9.3.03ger

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. Available online at: http://www.deeplearningbook.org (accessed August 25, 2020).

Gredebäck, G., Johnson, S., and Von Hofsten, C. (2010). Eye tracking in infancy research. *Dev. Neuropsychol.* 35, 1–19. doi: 10.1080/87565640903325758

Ham, J., Cuijpers, R. H., and Cabibihan, J.-J. (2015). Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. *Int. J. Soc. Robot.* 7, 479–487. doi: 10.1007/s12369-015-0280-4

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.* (Las Vegas, NV). doi: 10.1109/CVPR.2016.90

Heldner, M., and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *J. Phon.* 38, 555–568. doi: 10.1016/j.wocn.2010.08.002

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

Holler, J., and Beattie, G. (2003). How iconic gestures and speech interact in the representation of meaning: are both aspects really integral to the process? *Semiotica* 146, 81–116. doi: 10.1515/semi.2003.083

Ingre, M., Åkerstedt, T., Peters, B., Anund, A., i, and Kecklund, G. (2006). Subjective sleepiness, simulated driving performance and blink duration: examining individual differences. *J. Sleep Res.* 15, 47–53. doi: 10.1111/j.1365-2869.2006.00504.x

ISO 24617-2 (2012). *Language Resource Management - Semantic Annotation Framework (SemAF) - Part 2: Dialogue Acts.*

ISO 24617-8 (2016). *Language resource management – Semantic annotation framework (SemAF), Part 8: Semantic Relations in discourse, core annotation schema (DR-Core).* Available online at: https://www.iso.org/standard/60780.html (accessed August 25, 2020).

Izard, C. E. (1991). *The Psychology of Emotions*, *1. Edn*. New York, NY: Plenum Press New York.

Jarodzka, H., Holmqvist, K., and Gruber, H. (2017). Eye tracking in educational science: theoretical frameworks and research agendas. *J. Eye Mov. Res.* 10. doi: 10.16910/jemr.10.1.3

Kelly, S., Healey, M., Özyürek, A., and Holler, J. (2015). The processing of speech, gesture, and action during language comprehension. *Psychon. Bull. Rev.* 22, 517–523. doi: 10.3758/s13423-014-0681-7

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol.* 26, 22–63. doi: 10.1016/0001-6918(67)90005-4

Kendon, A. (2015). *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press. doi: 10.5860/choice.42-5687

Kobayashi, H., and Kohshima, S. (1997). Unique morphology of the human eye. *Nature* 387, 767–768. doi: 10.1038/42842

Komogortsev, O. V., Gobert, D. V., Jayarathna, S., Koh, D. H., and Gowda, S. M. (2010). Standardization of automated analyses of oculomotor fixation and saccadic behaviors. *IEEE Trans. Biomed. Eng.* 57, 2635–2645. doi: 10.1109/TBME.2010.2057429

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Levinson, S. C., and Holler, J. (2014). The origin of human multi-modal communication. *Philos. Trans. R. Soc. B Biol. Sci.* 369:20130302. doi: 10.1098/rstb.2013.0302

Manor, B. R., and Gordon, E. (2003). Defining the temporal threshold for ocular fixation in free-viewing visuocognitive tasks. *J. Neurosci. Methods* 128, 85–93. doi: 10.1016/S0165-0270(03)00151-1

Marsella, S., Xu, Y., Lhommet, M., Feng, A., Scherer, S., and Shapiro, A. (2013). "Virtual character performance from speech," in *Proceedings - SCA 2013: 12th ACM SIGGRAPH / Eurographics Symposium on Computer Animation*. (Anaheim, CA). doi: 10.1145/2485895.2485900

Meyer, T., and Popescu-Belis, A. (2012). "Using sense-labeled discourse connectives for statistical machine translation," in *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)* (Avignon: Association for Computational Linguistics), 129–138.

Mondada, L. (2016). Challenges of multimodality: language and the body in social interaction. *J. Socioling.* 20, 336–366. doi: 10.1111/josl.1_12177

Olsen, A. (2012). *The Tobii I-VT Fixation Filter*. Copyright © Tobii Technology AB. Available online at: https://www.tobiipro.com/learn-and-support/learn/steps-in-an-eye-tracking-study/data/how-are-fixations-defined-when-analyzing-eye-tracking-data/

Osako, K., Singh, R., and Raj, B. (2015). "Complex recurrent neural networks for denoising speech signals," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015*. (New Paltz, NY). doi: 10.1109/WASPAA.2015.7336896

Pfeiffer, U. J., Vogeley, K., and Schilbach, L. (2013). From gaze cueing to dual eye-tracking: novel approaches to investigate the neural correlates of gaze in social interaction. *Neurosci. Biobehav. Rev.* 37, 2516–2528. doi: 10.1016/j.neubiorev.2013.07.017

Popescu-Belis, A. (2016). *Manual and Automatic Labeling of Discourse Connectives for Machine Translation (Keynote Paper)*. Available online at: http://infoscience.epfl.ch/record/223763 (accessed August 25, 2020).

Prasad, R., and Bunt, H. (2015). "Semantic relations in discourse: the current state of ISO 24617-8," in *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)* (London), 80–92.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., and Joshi, A. (2008). "The penn discourse treebank 2.0," in *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008* (Marrakech).

Prasov, Z., and Chai, J. Y. (2008). "What's in a Gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces," in *Proceedings of the 13th International Conference on Intelligent User Interfaces*. Gran Canaria. doi: 10.1145/1378773.1378777

Qu, S., and Chai, J. Y. (2009). "The role of interactivity in human-machine conversation for automatic word acquisition," in *Proceedings of the SIGDIAL 2009 Conference: 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (London, UK). doi: 10.3115/1708376.1708404

R Core Team (2016). *R: A Language And Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Availavle online at: http://www.R-project.org/ (accessed August 25, 2020).

Risko, E. F., Richardson, D. C., and Kingstone, A. (2016). Breaking the fourth wall of cognitive science: real-world social attention and the dual function of gaze. *Curr. Dir. Psychol. Sci.* 25, 70–74. doi: 10.1177/0963721415617806

Rogers, S. L., Speelman, C. P., Guidetti, O., and Longmuir, M. (2018). Using dual eye tracking to uncover personal gaze patterns during social interaction. *Sci. Rep.* 8:4271. doi: 10.1038/s41598-018-22726-7

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., and Gleicher, M. (2015). A review of eye gaze in virtual agents, social robotics and HCI: behaviour generation, user interaction and perception. *Comput. Graph. Forum.* 34, 299–326. doi: 10.1111/cgf.12603

Sharp, R., Jansen, P., Surdeanu, M., and Clark, P. (2015). "Spinning straw into gold: Using free text to train monolingual alignment models for non-factoid question answering," in *Proceedings of the Conference on NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Denver, CO). doi: 10.3115/v1/n15-1025

Simonyan, K., and Zisserman, A. (2015). "Very deep convolutional networks for large-scale image recognition," in *Proceedings of 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track* (San Diego, CA).

Stefanov, K., Salvi, G., Kontogiorgos, D., Kjellström, H., and Beskow, J. (2019). Modeling of human visual attention in multiparty open-world dialogues. *ACM Trans. Hum. Robot Interact.* 8, 1–22. doi: 10.1145/3323231

Villani, D., Repetto, C., Cipresso, P., and Riva, G. (2012). May I experience more presence in doing the same thing in virtual reality than in reality? An answer from a simulated job interview. *Interact. Comput.* 24, 265–272. doi: 10.1016/j.intcom.2012.04.008

Wang, H., Meghawat, A., Morency, L. P., and Xing, E. P. (2017). "Select-additive learning: Improving generalization in multimodal sentiment analysis," in *Proceedings - IEEE International Conference on Multimedia and Expo* (Hong Kong). doi: 10.1109/ICME.2017.8019301

Ward, N. G., Jurado, C. N., Garcia, R. A., and Ramos, F. A. (2016). "On the possibility of predicting gaze aversion to improve video-chat efficiency," in *Eye Tracking Research and Applications Symposium (ETRA)* (Charleston, SC). doi: 10.1145/2857491.2857497

Zoric, G., Forchheimer, R., and Pandzic, I. S. (2011). On creating multimodal virtual humans-real time speech driven facial gesturing. *Multimed. Tools Appl.* 54, 165–179. doi: 10.1007/s11042-010-0526-y

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 05 March 2021
doi: 10.3389/fnbot.2021.642780

Active Vision for Robot Manipulators Using the Free Energy Principle

Toon Van de Maele, Tim Verbelen, Ozan Çatal, Cedric De Boom and Bart Dhoedt*

IDLab, Department of Information Technology, Ghent University—imec, Ghent, Belgium

Occlusions, restricted field of view and limited resolution all constrain a robot's ability to sense its environment from a single observation. In these cases, the robot first needs to actively query multiple observations and accumulate information before it can complete a task. In this paper, we cast this problem of active vision as active inference, which states that an intelligent agent maintains a generative model of its environment and acts in order to minimize its surprise, or expected free energy according to this model. We apply this to an object-reaching task for a 7-DOF robotic manipulator with an in-hand camera to scan the workspace. A novel generative model using deep neural networks is proposed that is able to fuse multiple views into an abstract representation and is trained from data by minimizing variational free energy. We validate our approach experimentally for a reaching task in simulation in which a robotic agent starts without any knowledge about its workspace. Each step, the next view pose is chosen by evaluating the expected free energy. We find that by minimizing the expected free energy, exploratory behavior emerges when the target object to reach is not in view, and the end effector is moved to the correct reach position once the target is located. Similar to an owl scavenging for prey, the robot naturally prefers higher ground for exploring, approaching its target once located.

Keywords: active vision, active inference, deep learning, generative modeling, robotics

OPEN ACCESS

Edited by:
Dimitri Ognibene,
University of Milano-Bicocca, Italy

Reviewed by:
Thomas Parr,
University College London,
United Kingdom
Yinyan Zhang,
Jinan University, China

****Correspondence:***
Toon Van de Maele
toon.vandemaele@ugent.be

Received: *16 December 2020*
Accepted: *03 February 2021*
Published: *05 March 2021*

Citation:
Van de Maele T, Verbelen T, Çatal O, De Boom C and Dhoedt B (2021) Active Vision for Robot Manipulators Using the Free Energy Principle. Front. Neurorobot. 15:642780. doi: 10.3389/fnbot.2021.642780

1. INTRODUCTION

Despite recent advances in machine learning and robotics, robot manipulation is still an open problem, especially when working with or around people, in dynamic or cluttered environments (Billard and Kragic, 2019). One important challenge for the robot is building a good representation of the workspace it operates in. In many cases, a single sensory observation is not sufficient to capture the whole workspace, due to restricted field of view, limited sensor resolution or occlusions caused by clutter, human co-workers, or other objects. Humans on the other hand tackle this issue by actively sampling the world and integrating this information through saccadic eye movements (Srihasam and Bullock, 2008). Moreover, they learn a repertoire of prior knowledge of typical shapes and objects, allowing them to imagine "what something would look like" from a different point of view. For example, when seeing a coffee mug, we immediately reach for the handle, even though the handle might not be directly in sight. Recent work suggests that active vision and scene construction in which an agent uses its prior knowledge about the scene and the world can be cast as a form of active inference (Mirza et al., 2016; Conor et al., 2020), i.e., that actions are selected that minimize surprise.

Active inference is a corollary of the free energy principle, which casts action selection as a minimization problem of expected free energy or surprise (Friston et al., 2016). The paradigm states that intelligent agents entail a generative model of the world they operate in (Friston, 2013). The expected free energy naturally unpacks as the sum of an information-seeking (epistemic) and an utility-driven (instrumental) term, which matches human behavior of visual search and "epistemic foraging" (Mirza et al., 2018). Furthermore it is also hypothesized that active inference might underpin the neurobiology of the visual perception system in the human brain (Parr and Friston, 2017).

Recent work has illustrated how active vision emerges from active inference in a number of simulations (Mirza et al., 2016; Daucé, 2018; Conor et al., 2020). However, these approaches typically define the agent's generative model upfront, in terms of small, often discrete state and observation spaces. Most similar is the work by Matsumoto and Tani (2020), which also considers a robot manipulator that must grasp and move an object by minimizing its free energy. Their approach differs from ours in the sense that they use an explicitly defined state space, containing both the robot state and the object locations. In order to be applicable for real-world robot manipulation, the generative model should work with realistic sensory observations such as camera inputs. Therefore, in this paper, we explore the use of deep neural networks to learn expressive generative models, and evaluate to what extent these can drive active vision using the principles from active inference. We consider the active vision problem of finding and reaching a certain object in a robotic workspace.

While a lot of research on learning generative models of the environment has been performed, most of them only consider individual objects (Sitzmann et al., 2019b; Häni et al., 2020), consider scenes with a fixed camera viewpoint (Kosiorek et al., 2018; Kulkarni et al., 2019; Lin et al., 2020) or train a separate neural network for each novel scene (Mildenhall et al., 2020; Sitzmann et al., 2020). We tackle the problem of an active agent that can control the extrinsic parameters of an RGB camera as an active vision system. Both camera viewpoint and its RGB observation are therefore available for our approach. To leverage the available information, our learned generative model is based on the Generative Query Network (GQN) (Eslami et al., 2018). This is a variational auto-encoder that learns a latent space distribution to encode the appearance of the environment through multiple observations from various viewpoints. Whereas, Eslami et al. (2018) integrates information of these different viewpoints by simply adding feature vectors, we show that this does not scale well for many observations, and propose a novel Bayesian aggregation scheme. The approximate posterior is computed through Gaussian multiplication and results in a variance that properly encodes uncertainty.

We evaluate our approach on three specific scenarios. First, we validate our generative model and Bayesian latent aggregation strategy on plane models of the ShapeNet v2 dataset (Chang et al., 2015). In addition, we provide an ablation study on the different aspects of our model architecture and compare different aggregation methods. Second, we evaluate action selection through active inference on observations of 3D coffee cups with and without handles. We evaluate the interpretation of the uncertainty about the cup from the variance of the latent distributions. Finally, we consider a robotic manipulator in a simulated workspace. The robot can observe its workspace by an RGB camera that is mounted to its gripper and is tasked to find and reach an object in the workspace. In order to solve the reach task, the robot must first locate the object and then move toward it. We show that exploratory behavior emerges naturally when the robot is equipped with our generative model and its actions are driven through active inference.

To summarize, the contributions of this paper are three-fold:

- We develop a deep neural network architecture and training method to learn a generative model from pixel data consistent with the free energy principle, based on Generative Query Networks (GQN).
- We propose a novel Bayesian aggregation strategy for GQN-based generative models which leverages the probabilistic nature of the latent distribution.
- We show that we can use a learned generative model to partake in active inference and that natural behavior emerges, first searching before attempting to reach it.

This paper is structured as follows: the proposed method is explained in section 2, where the generative model (section 2.1) and the active inference framework (section 2.2) are introduced first. Section 2.3.1 then explains how the approximation of the expected free energy can be achieved using the learned distributions. Section 2.3.2 finally elaborates on how these distributions are learned using deep neural networks through pixel-based data. Section 3 considers the results from applying the proposed method on numerous scenes of increasing complexity. First, the proposed model architecture is evaluated on a subset of the ShapeNet dataset (section 3.1). Next, the learned distributions are evaluated on whether they can be used within the active inference framework on the use case three dimensional cup (section 3.2). Finally the robot manipulator in simulation is used for the reaching problem (section 3.3). A discussion on the results, related work and possible future prospects is provided in section 4. A conclusion is provided in section 5.

2. METHOD

In this section we first discuss how the artificial agent interacts with the world through a Markov blanket, and that its internal generative model can be described by a Bayesian network. Next, we further unpack the generative model and describe how the internal belief over the state is updated. In the second section the theoretical framework of active vision and how this relates to an agent choosing its actions is elaborated on. Finally, we show how a learned generative model can be used to compute the expected free energy to drive the action-perception system known as active inference. We also go into the details of the neural network architecture and how it is learned exclusively

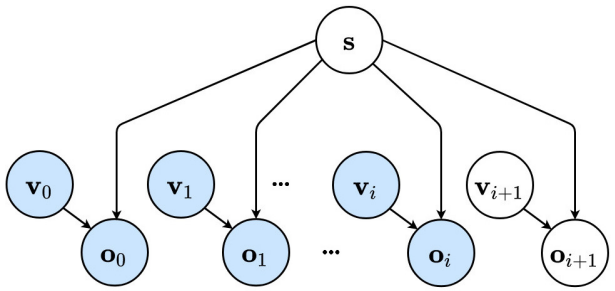


FIGURE 1 | The internal generative model of the artificial agent is represented as a Bayesian network. The environment is considered unchanging and is described by a latent variable $\mathbf{s}$. The observations $\mathbf{o}_k$ depend on both the environment described by $\mathbf{s}$ and the agent's viewpoint denoted by $\mathbf{v_k}$. The first i viewpoints have been visited and are used to infer a belief over the joint distribution. Future viewpoint $\mathbf{v}_{i+1}$ has not been visited or observed yet. Observed variables are shown in blue, while unobserved variables are shown in white.

from pixel-based observations by minimizing the variational free energy.

2.1. The Generative Model

We model the agent as separated from the true world state through a Markov blanket, which means that the agent can only update its internal belief about the world by interacting with the world through its chosen actions and its observed sensory information (Friston et al., 2016). In the case of active vision, the actions the agent can perform consist of moving toward a new viewpoint to observe its environment. We thus define the action space as the set of potential viewpoints the agent can move to. The sensory inputs of the agents in this paper are a simple RGB camera and the observations are therefore pixel-based. In this paper, we limit ourselves to an agent observing and reaching toward objects in the environment, but not interacting with them. Hence, we assume the environment is static and its dynamics should not be modeled in our generative model as we do not expect an object on the table to suddenly change color, shape, or move around without external interaction. However, one might extend the generative model depicted here to also include dynamics, similar to Çatal et al. (2020).

More formally, we consider the generative model to take the shape of a Bayesian network (**Figure 1**) in which the agent can not observe the world state directly, but has to infer an internal belief through sensory observations $\mathbf{o}_k$ and chosen viewpoints $\mathbf{v}_k$. The environment or world which can be observed from different viewpoints is described by the latent factor $\mathbf{s}$. When a viewpoint $\mathbf{v}_k$ is visited, an observation $\mathbf{o}_k$ is acquired which depends on the chosen viewpoint and environment state $\mathbf{s}$. The agent uses the sequence of observations to infer a belief about the world through the latent distribution $\mathbf{s}$.

The generative model describes a factorization of the joint probability $P(\mathbf{o}_{0:i}, \mathbf{s}, \mathbf{v}_{0:i})$ over a sequence of observations $\mathbf{o}_{0:i}$, states $\mathbf{s}$ and viewpoints $\mathbf{v}_{0:i}$. In the remainder of this paper, the colon notation $0:i$ is used to represent a sequence going from element 0 until the ith element. The generative model is factorized as:

$$P(\mathbf{o}_{0:i}, \mathbf{s}, \mathbf{v}_{0:i}) = P(\mathbf{s}) \prod_{k=1}^{i} P(\mathbf{o}_k|\mathbf{v}_k, \mathbf{s}) P(\mathbf{v}_k) \tag{1}$$

As the artificial agent can only interact with the world through its Markov blanket, the agent has to infer the posterior belief $P(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$. For high dimensional state spaces, computing this probability becomes intractable and approximate inference methods are used (Beal, 2003). The approximate posterior Q is introduced, which is to be optimized to approximate the true posterior distribution. The approximate posterior is factorized as:

$$Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i}) = \prod_{k=0}^{i} Q(\mathbf{s}|\mathbf{o}_k, \mathbf{v}_k), \tag{2}$$

This approximate posterior corresponds to the internal model that the agent uses to reason about the world. In the next section, we will discuss how variational methods can be used to optimize the approximate posterior by minimizing the variational free energy.

2.2. The Free Energy Principle

According to the free energy principle, agents minimize their variational free energy (Friston, 2010). This quantity describes the difference between the approximate posterior and the true distribution or equivalently, the surprise. The free energy F for the graphical model described in **Figure 1** can be formalized as:

$$\begin{aligned}
F &= \mathbb{E}_{Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})}[\log Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i}) - \log P(\mathbf{o}_{0:i}, \mathbf{s}, \mathbf{v}_{0:i})] \\
&= \underbrace{-\log P(\mathbf{o}_{0:i}, \mathbf{v}_{0:i})}_{\text{Evidence}} + \underbrace{\mathrm{D}_{\mathrm{KL}}[Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})||P(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})]}_{\text{Approximate vs true posterior}} \\
&= \underbrace{\mathbb{E}_{Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})}[-\log P(\mathbf{o}_{0:i}|\mathbf{v}_{0:i}, \mathbf{s})]}_{\text{Accuracy}} \\
&\quad + \underbrace{\mathrm{D}_{\mathrm{KL}}[Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})||P(\mathbf{s})]}_{\text{Complexity}} \\
&= \mathbb{E}_{Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})}\left[-\sum_{k=0}^{i} \log P(\mathbf{o}_k|\mathbf{v}_k, \mathbf{s})\right] \\
&\quad + \mathrm{D}_{\mathrm{KL}}[Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})||P(\mathbf{s})]
\end{aligned} \tag{3}$$

This formalization can be unpacked as the sum of the Kullback-Leibler divergence between the approximate posterior and the true belief over $\mathbf{s}$, and the expected negative log likelihood over the observed views $\mathbf{o}_{0:i}$ given their viewpoints $\mathbf{v}_{0:i}$. It is clear that if both distributions are equal, the KL-term will evaluate to zero and the variational free energy F equals the log likelihood. Minimizing the free energy therefore maximizes the evidence.

We can further interpret Equation (3) as an accuracy term, encouraging better predictions for an observation $\mathbf{o}_k$ given a viewpoint $\mathbf{v}_k$ and the state $\mathbf{s}$, and a complexity term promoting

"simpler" explanations, i.e., closer to the prior belief over $\mathbf{s}$. The approximate posterior can then be acquired by:

$$Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i}) = \underset{Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})}{\operatorname{argmin}} \; F \approx P(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i}), \tag{4}$$

However, the agent does not only want to minimize its surprise for past observations, but also for the future. Minimizing the free energy with respect to the future viewpoints will drive the agent to observe the scene in order to further maximize its evidence, and can therefore be used as a natural approach to exploration. The next viewpoints to visit can hence be selected by evaluating their free energy. However, it is impossible to compute this free energy, as observations from the future are not yet available. Instead, similar to Conor et al. (2020), the *expected* free energy G can be computed for the next viewpoint $\mathbf{v}_{i+1}$. This quantity is defined as the free energy expected to encounter in the future when moving to a potential viewpoint $\mathbf{v}_{i+1}$. The probability distribution over the considered future viewpoints can be computed with respect to G as:

$$P(\mathbf{v}_{i+1}) = \sigma(-G(\mathbf{v}_{i+1})), \tag{5}$$

Where $G(\mathbf{v}_{i+1})$ is the expected free energy for the future visited viewpoint, σ is the softmax operation which transforms the expected free energy G for every considered viewpoint $\mathbf{v}_{i+1}$ into a categorical distribution over these viewpoints. The expected free energy is then obtained by computing the free energy for future viewpoint $\mathbf{v}_{i+1}$:

$$\begin{aligned}
&G(\mathbf{v}_{i+1}) \\
&= \mathbb{E}_{Q(\mathbf{s},\mathbf{o}_{i+1}|\mathbf{o}_{0:i},\mathbf{v}_{0:i+1})}[\log Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i+1}) - \log P(\mathbf{o}_{0:i+1}, \mathbf{s}|\mathbf{v}_{0:i+1})] \\
&= \mathbb{E}_{Q(\mathbf{s},\mathbf{o}_{i+1}|\mathbf{o}_{0:i},\mathbf{v}_{0:i+1})}[\log Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i+1}) - \log P(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1}) \\
&\quad - \log P(\mathbf{o}_{0:i+1}|\mathbf{v}_{0:i+1})] \\
&\approx \underbrace{-\mathbb{E}_{Q(\mathbf{o}_{i+1}|\mathbf{o}_{0:i},\mathbf{v}_{0:i+1})}\big[\mathrm{D}_{\mathrm{KL}}[Q(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1})||Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})]\big]}_{\text{Epistemic value}} \\
&\quad \underbrace{-\mathbb{E}_{Q(\mathbf{o}_{i+1}|\mathbf{o}_{0:i},\mathbf{v}_{0:i+1})}[\log P(\mathbf{o}_{0:i+1})]}_{\text{Instrumental value}}
\end{aligned} \tag{6}$$

This expected free energy can be reformulated as the sum of an instrumental and an epistemic term. The epistemic value is the KL-divergence between the posterior belief over $\mathbf{s}$ after observing the future viewpoint, and before visiting this viewpoint. As the true posterior is not available, we approximate $P(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1})$ using the approximate posterior distribution $Q(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1})$. Please note that in the final step, the condition on the viewpoints in the instrumental value can be omitted. Which can be interpreted as an intelligent agent creating a preferred prior in advance that is not dependent on the corresponding viewpoints. Intuitively, this KL-term represents how much the posterior belief over $\mathbf{s}$ will change given the new observation. An agent that minimizes free energy will thus prefer viewpoints that change the belief over $\mathbf{s}$, or equivalently, to learn more about its environment. The instrumental value represents the prior likelihood of the future observation. This can be interpreted as a goal that the agent wants to achieve. For example in a reaching task, the agent expects to see the target object in its observation.

2.3. Active Vision and Deep Neural Networks

To apply active inference in practice, a generative model that describes the relation between different variables in the environment, i.e., actions, observations, and the global state, is required. When using this paradigm for complex tasks, such as reaching an object with a robot manipulator, it is often difficult to define the distributions over these variables upfront. In this paper, we learn the mapping of observations and viewpoints to a posterior belief directly from data using deep neural networks. We model the approximate posterior $Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$ and likelihood $P(\mathbf{o}_k|\mathbf{s}, \mathbf{v}_k)$ as separate neural networks that are optimized simultaneously, similar to the variational auto-encoder approach (Kingma and Welling, 2014; Rezende et al., 2014).

The approximate posterior $Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$ is modeled through a factorization of the posteriors after each observation. The belief over $\mathbf{s}$ can then be acquired by multiplying the posterior beliefs over $\mathbf{s}$ for every observation. We learn an encoder neural network with parameters ϕ to learn the posterior $q_\phi(\mathbf{s}|\mathbf{o}_k, \mathbf{v}_k)$ over $\mathbf{s}$ given a single observation and viewpoint pair $(\mathbf{o}_k, \mathbf{s}_k)$. The posterior distributions over $\mathbf{s}$ given each observation and viewpoint pair are combined through a Gaussian multiplication. We acquire the posterior distribution as a Normal distribution proportional to the product of the posteriors:

$$Q_\phi(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i}) \propto \prod_{k=0}^{i} q_\phi(\mathbf{s}|\mathbf{o}_k, \mathbf{v}_k). \tag{7}$$

Secondly, we create a neural network with parameters ψ that estimates the pixel values of an observation $\hat{\mathbf{o}}_k$, given the selected viewpoint $\mathbf{v}_k$ and a state vector $\mathbf{s}$. The likelihood over the observation $p_\psi(\hat{\mathbf{o}}_k|\mathbf{v}_k, \mathbf{s})$ is modeled as an image where every pixel is an independent Gaussian distribution with the pixel value being the mean and a fixed variance.

We jointly train these models using a dataset of tuples $\{(\mathbf{o}_k, \mathbf{v}_k)\}_{k=0}^{k=i}$ for a number of environments by minimizing the free energy loss function:

$$\mathcal{L} = \sum_{k=0}^{i} ||\hat{\mathbf{o}}_k - \mathbf{o}_k||_2 + \mathrm{D}_{\mathrm{KL}}[Q_\phi(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})||\mathcal{N}(\mathbf{0}, \mathbf{I})] \tag{8}$$

This loss function is reformulated as a trade-off between a reconstruction term and a regularization term. The reconstruction term computes the summed mean squared error between the reconstructed observations $\hat{\mathbf{o}}_{0:i}$ and ground-truth observations $\mathbf{o}_{0:i}$. This term corresponds with the accuracy term of Equation (3), as minimization of the mean squared error is equivalent to minimizing log likelihood when the likelihood is a Gaussian distribution with a fixed variance. The regularization

term is identical to the complexity term of Equation (3) and computes the KL-divergence between the belief over the state $\mathbf{s}$ and a chosen prior, which we choose to be an isotropic Gaussian with unit variance.

2.3.1. Approximating the Expected Free Energy for Active Vision

Under active inference, the agent chooses the next viewpoint to visit in order to minimize its expected free energy as described in section 2.2. The agent selects the viewpoint by sampling from the categorical distribution $P(\mathbf{v}_{i+1})$. As described by Equation (5), this categorical distribution is acquired by computing the expected free energy G for every potential viewpoint $\mathbf{v}_{i+1}$, and applying the softmax function on the output. The expected free energy is computed by separately evaluating the epistemic and instrumental term from Equation 6. Calculating these expectations for every possible viewpoint is intractable, hence we resort to Monte Carlo methods to approximate the expected free energy through sampling.

A schematic overview of our method is shown in **Figure 2**. For a target viewpoint $\mathbf{v}_{i+1}$, the epistemic term is the expected value of the KL divergence between the belief over state $\mathbf{s}$ after observing $\mathbf{o}_{i+1}$ (i.e., $Q(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1})$) and prior to observing $\mathbf{o}_{i+1}$ (i.e., $Q(\mathbf{s}|\mathbf{v}_{0:i}, \mathbf{o}_{0:i})$). The latter distribution is the output after feeding all previous observations $\mathbf{o}_{0:i}$ and their corresponding viewpoints $\mathbf{v}_{0:i}$ through the neural network $q_\phi(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$. This is shown on the left of **Figure 2** and provides the agent with the current belief over $\mathbf{s}$. To estimate the posterior distribution $Q(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1})$, an imagined observation $\hat{\mathbf{o}}_{i+1}$ must be sampled. The likelihood model is used to do this, conditioned on the potential viewpoint $\mathbf{v}_{i+1}$ and a sampled state vector from $Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$, an estimate of the observed view $\hat{\mathbf{o}}$ is made. Together with the initial observations $\mathbf{o}_{0:i}$ and viewpoints $\mathbf{v}_{0:i}$, the imagined view is encoded through the posterior model to approximate $Q(\mathbf{s}|\mathbf{o}_{0:i+1}, \mathbf{v}_{0:i+1})$ as shown on the right of **Figure 2**. As both prior and posterior distributions are approximated by a Multivariate Gaussian with a diagonal covariance matrix, the KL divergence can be computed analytically. To approximate the expected value over $Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$, we repeat this process for multiple state samples and average the obtained values.

The instrumental term, as described in Equation 6, is the expected negative log likelihood of the observed view $\mathbf{o}_{i+1}$ for the future viewpoint $\mathbf{v}_{i+1}$. Again, we approximate this value by sampling from the state distribution, and forwarding this through the likelihood model. We calculate the negative log likelihood of each imagined observation $\hat{\mathbf{o}}_{i+1}$ according to a prior distribution over this observation. This process is repeated for numerous samples from $Q(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$, and the computed log likelihood is averaged to calculate the instrumental term. In the case of a robotic reaching task, this prior distribution takes the form of a desired goal observation, and computing log likelihood reduces to computing the mean squared error between an imagined observation $\hat{\mathbf{o}}_{i+1}$ and a reference goal observation.

2.3.2. Model and Training Details

Both neural networks are directly optimized end-to-end through pixel data, using a dataset consisting of different scenes. We define a scene as a static environment or object in or around which the agent's camera can move to different viewpoints. The agent has observed the set of i observations and viewpoints from a scene $\mathcal{S} = \{(\mathbf{o}_k, \mathbf{v}_k)\}_{k=0}^{k=i}$. The view $\mathbf{o}_k$ is an RGB image scaled down to a resolution of 64×64 pixels and the viewpoint $\mathbf{v}_k$ is represented by a seven dimensional vector that consists of both the position coordinates and the orientation quaternion.

The generative model we consider belongs to the family of variational auto-encoders (Kingma and Welling, 2014; Rezende et al., 2014). It most resembles the Generative Query Network (GQN) (Eslami et al., 2018). This variational auto-encoder variant encodes information for each observation separately and aggregates the acquired latent codes. Similarly to the GQN, our encoder generates a latent distribution for each observation separately and combines them to form the current scene representation. From this scene representation, the decoder has to render the expected observations given a target viewpoint.

We deviate from the GQN presented by Eslami et al. (2018) in two ways. First, whereas GQNs concatenate the viewpoint parameters somewhere in the encoder and use an auto-regressive decoder architecture, we use convolutional neural networks for both encoding and decoding, and use FiLM layers (Perez et al., 2018) for conditioning. The encoder is conditioned on the viewpoint parameters and the decoder is conditioned on both the query viewpoint $\mathbf{v}_{i+1}$ and the scene representation vector. Secondly, whereas GQNs aggregate the extracted representations from the encoder by mere addition, we use a Bayesian inspired aggregation scheme. We consider the distributions from the model described in section 2.1. Instead of the addition used in the GQN, we use a factorization of the posterior $Q(\mathbf{s}|\mathbf{v}_{0:i}, \mathbf{o}_{0:i})$ to aggregate the acquired representations through Gaussian multiplication. When a new observation $\mathbf{o}_i$ is available, the current belief distribution $\mathcal{N}(\boldsymbol{\mu}_{cur}, \boldsymbol{\sigma}^2_{cur}\mathbf{I})$ is updated with the output of the encoder network $q_\phi(\mathbf{o}_i|\mathbf{v}_i)$, a Normal distribution $\mathcal{N}(\boldsymbol{\mu}_{obs}, \boldsymbol{\sigma}^2_{obs}\mathbf{I})$, using Gaussian multiplication:

$$\boldsymbol{\mu} = \frac{\boldsymbol{\sigma}^2_{cur} \cdot \boldsymbol{\mu}_{obs} + \boldsymbol{\sigma}^2_{obs} \cdot \boldsymbol{\mu}_{cur}}{\boldsymbol{\sigma}^2_{cur} + \boldsymbol{\sigma}^2_{obs}}, \tag{9}$$

$$\frac{1}{\boldsymbol{\sigma}^2} = \frac{1}{\boldsymbol{\sigma}^2_{cur}} + \frac{1}{\boldsymbol{\sigma}^2_{obs}} \tag{10}$$

This way of refining belief of the acquired representations is equivalent to the update step found in Bayesian filtering systems such as the Kalman filter (Kalman, 1960). As the variance in each dimension reflects the spread over that state vector, it can be interpreted as the confidence of the model. The belief over the state is therefore updated based on their uncertainty in each dimension. Additionally, using this type of aggregation has the benefit that the operation is magnitude-preserving. This results in a robust system that is invariant to the amount of received observations, unlike an addition-based system. For

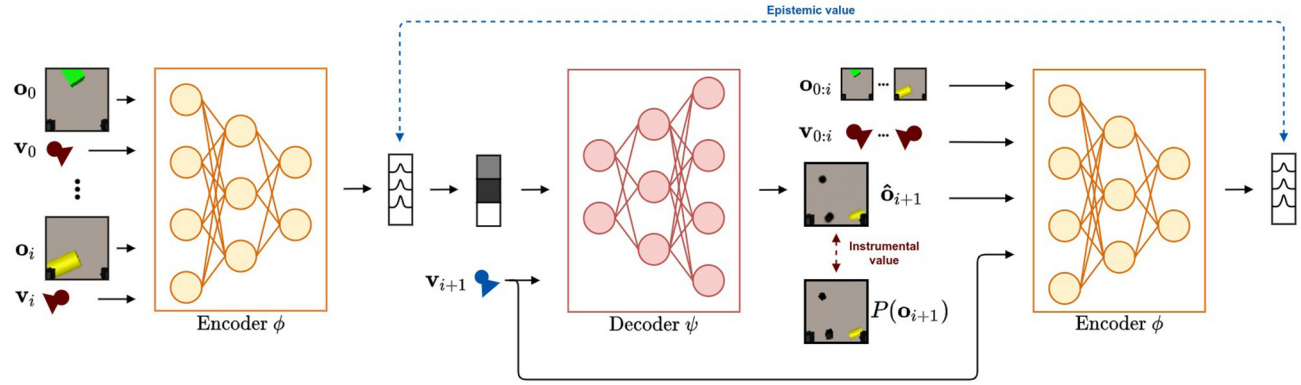


FIGURE 2 | The flow followed when evaluating the expected free energy using deep neural networks for a given potential new viewpoint $\mathbf{v}_{i+1}$. Starting on the left of the figure, the encoder neural network that approximates the posterior $q_\phi(\mathbf{s}|\mathbf{o}_{0:i}, \mathbf{v}_{0:i})$ encodes the observations $\mathbf{o}_{0:i}$ and corresponding viewpoints $\mathbf{v}_{0:i}$ until now into a belief over the state $\mathbf{s}$. From this belief, a state vector is sampled and is used together with viewpoint $\mathbf{v}_{i+1}$ to predict the imagined view from this viewpoint. The instrumental value is computed as the log likelihood that the target image is generated from the distribution over the predicted image. This is marked by the red arrow. This imagined view $\hat{\mathbf{o}}_{i+1}$ is passed through the approximate posterior model to acquire the expected belief over $\mathbf{s}$ after selecting viewpoint $\mathbf{v}_{i+1}$. The epistemic value is computed as the KL divergence between the approximate posterior before observing the imagined view $\hat{\mathbf{o}}_{i+1}$ and after. This is marked by the blue arrow. Finally, the expected free energy is approximated by averaging over a number of samples.

stability reasons, we clip the variance of the resulting distribution to a value of 0.25.

We parameterize our model as follows. The inputs are first expanded by using a 1×1 convolution that maps the RGB channels to a higher dimensional space of 64 channels. The encoder consists of four convolutional layers with a stride of 2, a kernel size of 3×3 and feature maps that increase with a factor 2 every layer (16, 32, 64, 128). They are interleaved with FiLM layers (Perez et al., 2018) that learn a transform for the extracted features based on the viewpoint pose. The extracted feature representation is then transformed in two feature vectors that represent the mean and variance of the latent state $\mathbf{s}$. In each considered experiment this latent size is different. The decoder mirrors this architecture with four convolution blocks, each convolution block first applies a convolution that halves the amount of feature maps, after which a convolution is applied which preserves the amount of feature channels (128, 128, 64, 64, 32, 32). Here, the FiLM layers are conditioned on the concatenated latent code and query pose. Between every convolution block in the decoder, the image is linearly upsampled. LeakyReLU activations are used after every convolutional layer. The output of the decoder is finally processed using a 1x1 convolution that maps the 64 channels back to RGB channels. For the specifics of the neural network, the reader is referred to **Supplementary Material**.

This model is optimized end-to-end by minimizing the free energy loss with respect to the model parameters, as described in Equation (8) using Adam (Kingma and Ba, 2015), a gradient-based optimizer. Additionally, we use the constraint-based GECO algorithm (Rezende and Viola, 2018) that balances the reconstruction and regularization term by optimizing Lagrangian multipliers using a min-max scheme.

3. RESULTS

Three experiments were designed to evaluate both our model and the proposed active vision system. In a first experiment, we consider a subset of the ShapeNet dataset (Chang et al., 2015) to evaluate model performance. We conduct an ablation study on different aggregation methods for the state encodings produced by the generative model. We show that our model exhibits performance similar to other aggregation strategies, while being more resistant to the number of observations and better leveraging the Bayesian character of the extracted distributions. In a second experiment, we consider scenes consisting of a 3D coffee cup that potentially has a handle. We investigate the learned approximate posterior distribution and its behavior when observing different views. We analyze the behavior that emerges in our artificial agent when driving viewpoints selection using the epistemic term. In the final experiment, we consider a realistic robotic workspace in CoppeliaSim (Rohmer et al., 2013). Scenes are created with an arbitrary amount of random toy objects with random colors. A task is designed in which the robot manipulator must find and reach a target object. First, we investigate the exploratory behavior when no preferred state is provided and see that the agent explores the workspace. We then provide the agent with a goal by specifying a preferred observation and computing the full value of G. We observe that the agent explores the workspace until it has found and reached its target.

3.1. ShapeNet

In the first experiment we want to evaluate the proposed neural network architecture on a subset of the ShapeNet dataset (Chang et al., 2015). We focus on whether the neural architecture is capable of learning to implicitly encode the three dimensional structure of a scene from purely pixel-based observations by minimization of the free energy loss function. Additionally,

we want to validate our novel aggregation strategy which uses a factorization of the approximate posterior to combine the extracted representations for all observations. The novel aggregation method ensures that the resulting distribution will always be in the same order of magnitude, independently of the number of observations, in contrast to the addition method from the original work by Eslami et al. (2018). We expect to see that our approach outperforms the GQN baseline when provided with a large amount of observations.

To separate the influence of the overall network architecture from the used aggregation method to combine extracted latent distributions from all separate observations into a belief over the state **s**, we perform an ablation study. Besides the proposed approach, we also introduce three variants to combine latent distributions, while using the same encoder-decoder architecture with a latent size of 64 dimensions. We compare our approach to the addition method from the original GQN paper (Eslami et al., 2018), a mean operation (Garnelo et al., 2018), or a max-pooling (Su et al., 2015) operation. As these ablations do not propose a method to integrate the variance of the individual reconstructions, the variance of the new observation is set to a fixed value of 1 for every dimension. We also compare the results with the original GQN architecture.

All models in this experiment are trained on the same data using the free energy loss function from Equation (8). The observations are RGB images with a resolution of 64×64. The viewpoints are a 7-dimensional vector, that correspond to the position in Euclidean coordinates and the orientation in quaternion representation. The model is optimized end-to-end as described in section 2.3.2. A batch size of 100 sequences per mini-batch is used. Similar to the approach used by the GQN, between 3 and 10 observations are randomly provided during training to enforce independence on the amount of observed data. These models are then trained until convergence. The GQN baseline is optimized using the traditional ELBO loss as described in the original paper by Eslami et al. (2018).

Table 1 shows the average mean squared error (MSE) of novel views generated for all objects in the test set for a varying number of observations. We observe that our model outperforms the others for 30 and 60 observations, whereas GQN has similar performance on 10 observations. Also note that our model has an order of magnitude fewer parameters than the GQN model. From the ablation study, we can indeed note that the GQN suffers from the addition aggregation method. Max-pooling seems to perform better with more than ten observations, but has an overall higher MSE compared to our approach. The same is true for the mean-pool ablation, which improves as more observations are added. This improvement can be attributed to the reduction of noise on the representation vector by having more observations.

Examples of the reconstructions generated from the aggregated latent space are shown in **Figure 3**. Clearly the GQN achieves the best performance when operating in the trained range, but when more observations are added the quality of the decoded image decays rapidly and the object is no longer recognizable. The same behavior can be noticed for the addition ablation. Our model yields comparable reconstructions as the GQN for 10 observations, but achieves to uphold this quality level as well after 60 observations, and is even able to improve its reconstruction. Both the max-pool and the mean-pool ablation are less affected after 60 observations, but the overall reconstructions are less detailed.

TABLE 1 | Average MSE over all objects in the selected test set of ShapeNet planes data.

Model	# param	MSE (10 obs)	MSE (30 obs)	MSE (60 obs)
GQN	49.5M	**0.0143 ± 0.0110**	0.0354 ± 0.0228	0.0438 ± 0.0275
Ours	3.6M	0.0151 ± 0.0138	**0.0148 ± 0.0133**	**0.0147 ± 0.0133**
Addition ablation	3.6M	0.0169 ± 0.0122	0.1222 ± 0.1102	0.2409 ± 0.1599
Max-pool ablation	3.6M	0.0175 ± 0.0112	0.0170 ± 0.0110	0.0176 ± 0.0101
Mean-pool ablation	3.6M	0.0182 ± 0.0110	0.0175 ± 0.0103	0.0175 ± 0.0094

The bold value indicates the lowest MSE for every column.

3.2. The Cup

In active inference, viewpoints are selected by minimizing the agent's expected free energy. It is essential that the predicted distributions through our learned generative model are well-behaved and thus are able to properly represent ambiguity when it has no, or incomplete, information about the scene. In this experiment, we evaluate the distributions produced by the learned generative model and analyze whether they are able to capture the ambiguity provided by the scene. We expect to see dubiety in both the reconstructed imagined views of the cup, as well as in the variance of the produced distributions. We also investigate the behavior that emerges when viewpoints are selected by minimizing the epistemic term of the expected free energy and expect exploratory behavior to surface.

We consider simple scenes that consist of a 3D model of a coffee cup that can vary in size and orientation. It can potentially be equipped with a handle. For each created scene, 50 views of 64×64 pixels are randomly sampled from viewpoints around the object. A dataset of 2,000 different scenes containing a cup were created in Blender (Blender Online Community, 2018), of which half are equipped with a handle. One thousand eight hundred of these scenes were used to train the generative model. The parameters of the neural network are optimized in advance using this prerecorded dataset by minimizing the free energy over the acquired observations as explained in section 2.3. For each scene, between 3 and 5 images are provided to the model during training. The model for this experiment is the same as described in section 2.3, but with a latent dimension size of 9. The following experiments were conducted on scenes of cups in the validation set that were not seen during training.

To evaluate whether the generative model is able to capture the ambiguity of a cup when not all information is gathered through observations yet, we consider two nearly identical cups, both positioned in the same orientation and scaled with the same factor. The only difference between these cups is that one has a handle, while the other one does not. We provide our learned model with a single observation that does not resolve the

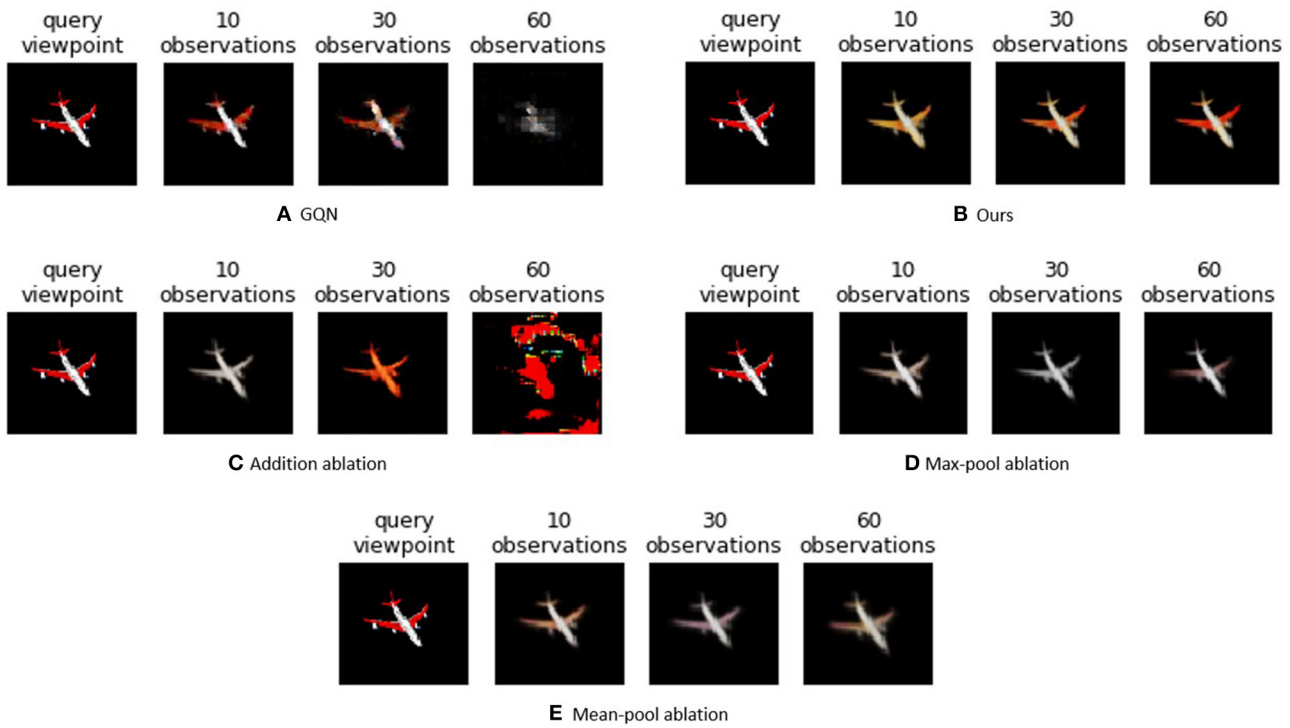


FIGURE 3 | Evolution of imagined observations for different models on an unseen ShapeNet example when 10, 30, or 60 observations are provided. Our Observer model maintains a good reconstruction quality even if more observations are considered than during training.

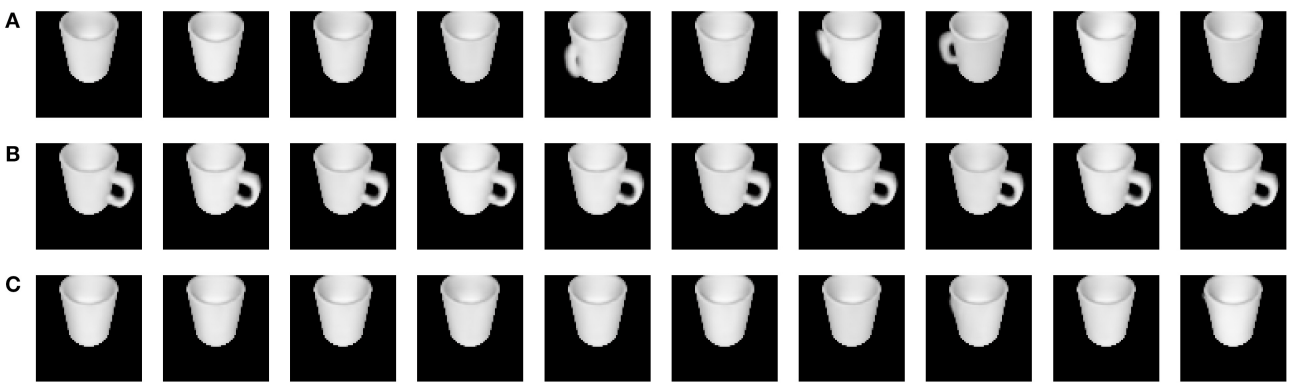


FIGURE 4 | Generated observations are shown in this figure. **(A)** An ambiguous observation is provided to the generative model, and this is reflected in ambiguous reconstructions after observing the cup from the other side. **(B)** The model is provided with an additional unambiguous observation of a handle. **(C)** The model is provided with an additional observation of the cup from the other side which does not contain a handle.

ambiguity about the location and does not reveal the presence of a handle. We now use the likelihood model over the observation $\mathbf{o}_{k+1}$ to generate the expected observation, which is shown in **Figure 4A**. When looking at these generated cups, it shows both cups with and without handle, with the handle at a random position. This can be attributed to the fact that the orientation of the cup is not known, and the model therefore does not know at what position to draw a handle. This ambiguity is also reflected in the high variance shown in the extracted latent distribution (**Figure 5A**).

When a new observation from a different viewpoint around the cup is added to the model, the ambiguity can be noticed to clearly drop. **Figure 4B** shows the reconstructed cups in case the handle is observed. These reconstructions are sharp and draw the handle consistently at the same position. This consistency is also reflected by the lower variance of its latent distribution shown in **Figure 5B**. The same observation without a handle was provided as a second observation for the cup without a handle. The generated cups of this scene are shown in **Figure 4C**. In **Figure 5C**, a lower variance compared to the

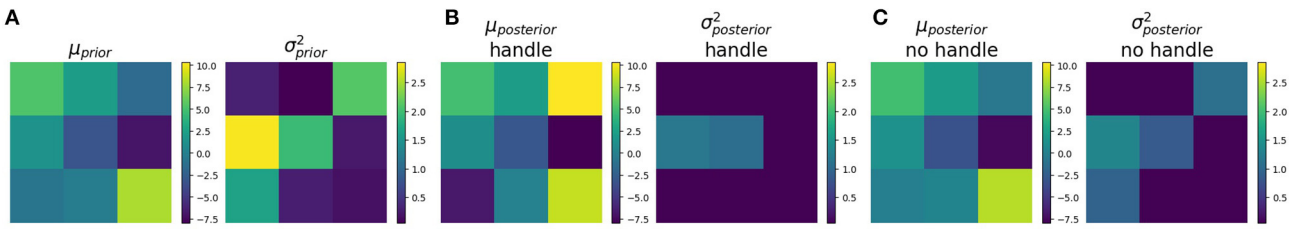


FIGURE 5 | This figure is a representation of the distributions over the approximate posterior with a latent state of nine dimensions. As the approximate posterior is represented by a multivariate Gaussian distribution with a diagonal covariance matrix, each dimension can be considered independently. **(A)** The distribution after an ambiguous observation is represented. **(B)** The approximate distribution for an unambiguous observations of a handle is represented. The variance is on average 9.08 times lower in each dimension than prior to observation in **(A)**. **(C)** The same unambiguous observations without a handle is provided. In this case, the variance is on average 2.83 times smaller in each dimension than in **(A)**.

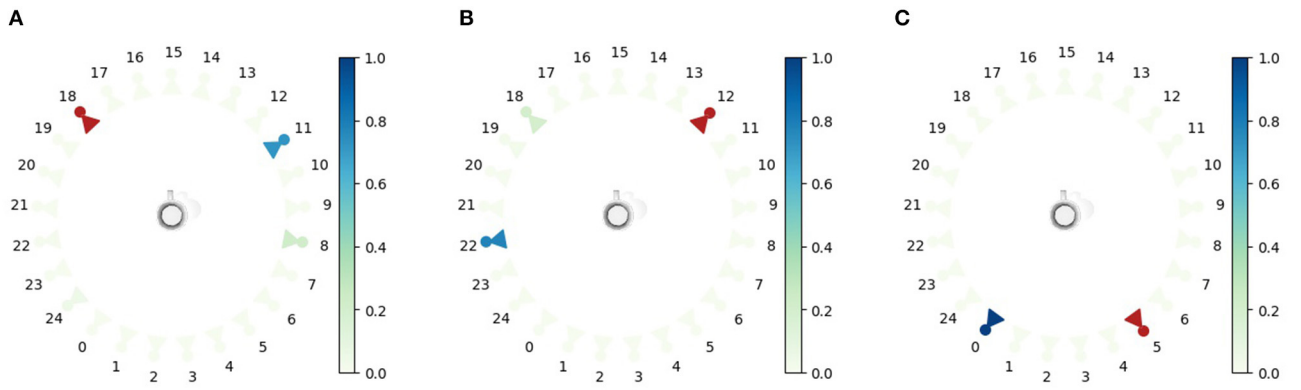


FIGURE 6 | The epistemic values are shown for all potential viewpoints. This value is computed as the KL divergence between the belief over latent state **s** with the observations marked by a red observer and the expected posterior belief when choosing the next viewpoint $\mathbf{v}_{i+1}$. The epistemic values are normalized by the softmax operation with a temperature of 1, as described in section 2. The red color marks the observation viewpoint that has been observed. The color of the camera represents the likelihood that this viewpoint is selected next. **(A–C)** shown for three scenarios with a different initial viewpoint in **(A)** through **(C)**.

one shown in **Figure 5A** can again be noticed. We thus conclude that optimizing the generative model through a minimization of expected free energy results in well-behaved latent distributions.

Additionally, we want to evaluate whether using the expected free energy as a viewpoint selection policy is a valid approach for active vision. We hypothesize that if the agent observes the cup from one viewpoint, it will prefer policies that move the agent to observe the cup from the other side, to gain as much information as possible in the least amount of observations. The potential viewpoints are uniformly spaced in a circle around the cup at a fixed height, and with an orientation toward the cup. **Figure 6** shows the probability distribution over the potential viewpoints $P(\mathbf{v}_{i+1})$ for three different initial observations. It is clear that in general, the agent will choose a viewpoint far away from the current observation to maximize the information gain with respect to the cup.

3.3. Robot Manipulator

In the final experiment, a robotic environment in CoppeliaSim (Rohmer et al., 2013) is considered. The workspace is equipped with a robot manipulator on a fixed table, which has an RGB camera mounted to its gripper. Some toy objects are placed on the table within reach of the manipulator. These objects are randomly chosen and can take the shape of a cube, a sphere, a cylinder or a bar that could either be standing up or laying down. These objects have a Lambertian surface with a uniform color. An example of such a scene is shown in **Figure 7**. The agent is able to manipulate the extrinsic camera parameters through robotic actuation of the gripper. It can then observe different areas of the workspace. Similar to the previous experiment, we first learn the neural network parameters from a prerecorded dataset, which is then used in the proposed active vision scheme for viewpoint selection. The model architecture is identical to the one in the previous experiments, but with 256 latent state space dimensions.

In order to learn the model parameters, a prerecorded dataset was created using the same environment in CoppeliaSim. Up to five randomly selected toy objects are spawned in the workspace. The orientation and position of the objects within the workspace are determined randomly by sampling from a uniform distribution. A dataset of 8,000 such scenes is created, in which the robot end-effector is moved along a trajectory that covers the entire workspace at different heights. We constrain the end-effector to look in a downwards orientation. This facilitates

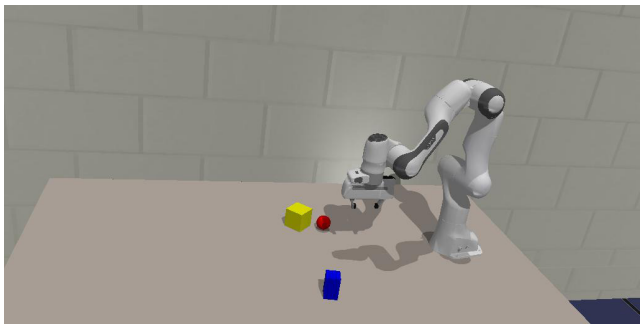


FIGURE 7 | An example scene of the robotic workspace in CoppeliaSim. Three random objects are spawned at arbitrary positions and rotations. This scene is used for the experiments in section 3.3.

the training process and does not limit performance on this use case, as the robot is still able to observe all objects placed on the workspace from a top view. During training, these observations are shuffled randomly, and a subset between 3 and 10 observations are selected and used as model inputs.

We design two cases for the active vision experiments in the robotic workspace. In the first case, we put an additional constraint on the height of the agent and only allow the agent to move in the x and y direction of the workspace, i.e. parallel with the table. We choose this to limit the potential viewpoints of the agent to observe the epistemic and instrumental behavior in more detail, with respect to the imagined views. In the second case, we allow the agent to also move along the z-axis. We can now evaluate the global behavior of the agent and observe that when it explores a new area, it will first prefer viewpoints from higher vantage points in which it can observe a large piece of the workspace, after which it will move down to acquire more detailed observations.

3.3.1. Active Vision With 2 Degrees of Freedom

This experiment considers the case where the artificial agent is limited to 2 degrees of freedom. We limit the degrees of freedom to make the analysis of the behavior more interpretable. The results of this experiment are shown in **Figure 8**.

Even though the generative model is capable of inferring the state and generating an imagined view for any viewpoint in a continuous space of the robotic working area, it would be computationally expensive to compute the expected free energy for all potential viewpoints. Instead, we sample a uniform grid of potential future viewpoints over the robotic workspace, and evaluate the expected epistemic value for these samples using the method described in section 2.

First, only the epistemic value is considered. We look at the behavior for an active vision agent for the scene visualized in **Figure 7**. For results on additional scenes, the reader is referred to **Supplementary Material**. The agent starts in an initial position in which it can not observe any of the objects that are lying on the table. Its current observation is shown in the first image of **Figure 8C**. The agent imagines the entire workspace to be empty without objects, this can be seen in the imagined observations for the potential viewpoints, shown in **Figure 8B**. The epistemic value is computed for all potential viewpoints, and is shown in **Figure 8A**. The largest epistemic values are located in the center of the table, as the agent believes that observations from these locations will allow it to learn more. After moving to the viewpoint with the highest epistemic value, the agent observes the yellow cube and the red ball. The generative model is then able to reconstruct these objects correctly at the potential viewpoints, which can be observed in the second plot of **Figure 8B**. We notice, that after observing these objects, the agent still prefers to look at these positions for a number of steps. The internal model of the environment is still being updated, which we can see in the sharper reconstructions in the first and second row of **Figure 8**. This can be attributed to the aggregation strategy for the approximate posterior. A single observation of the objects will not transform the distribution entirely, but a weighted mean and variance is computed. This results in a slower process for updating the state distribution, and it can result in the agent trying to observe the same area for a number of steps. Similar to the experiment in section 3.1, the observations can be seen to improve as the latent distribution improves. After a few steps, this distribution converges to a fixed value as can also be noted by the decreasing epistemic values shown in **Figure 8A**. Additionally, as the agent imagines no new objects at the other viewpoints, it does not believe they will influence its belief over $\mathbf{s}$. After the agent has refined its internal model, in step 7, the viewpoints it has not yet observed result in a higher epistemic value, after which the agent moves to this location. It finally observes the blue cube in the top which is then also reconstructed in the imagined views.

In a second experiment, we evaluate the behavior that emerges when the full expected free energy is used to drive viewpoint selection. Both the epistemic and instrumental values are computed and used to acquire the expected free energy for every potential viewpoint. The instrumental value is computed as the log likelihood of the expected observation under a desired goal prior distribution. We choose the distribution of this preferred observation as a multivariate Gaussian in which each pixel is an independent Gaussian with as mean value the target goal observation and a fixed variance of 0.65. We empirically determined this value for the goal variance which yields a good trade-off between the epistemic and instrumental behavior. In this case we use an observation of the blue cube as goal observation, namely the final observation from the epistemic exploration, and shown in **Figure 8C**. Please note that any observation could be used as a goal.

When we look at the behavior that emerges in **Figure 9**, we notice that initially the agent has no idea where it can observe it's preferred observation. This can be observed by the uniform instrumental value shown in **Figure 9B** at step 0. The epistemic value takes the upper hand, and the chosen viewpoint is again in the center of the table, similar as in the case when only the epistemic value was considered. At this new viewpoint, the agent observes the yellow cube and the red ball. Notice how the instrumental value becomes lower at these viewpoints in **Figure 9B**. The agent realizes that these viewpoints will not aid in the task to reach the blue object. However, as the epistemic value at this time step is larger than the range of the instrumental value

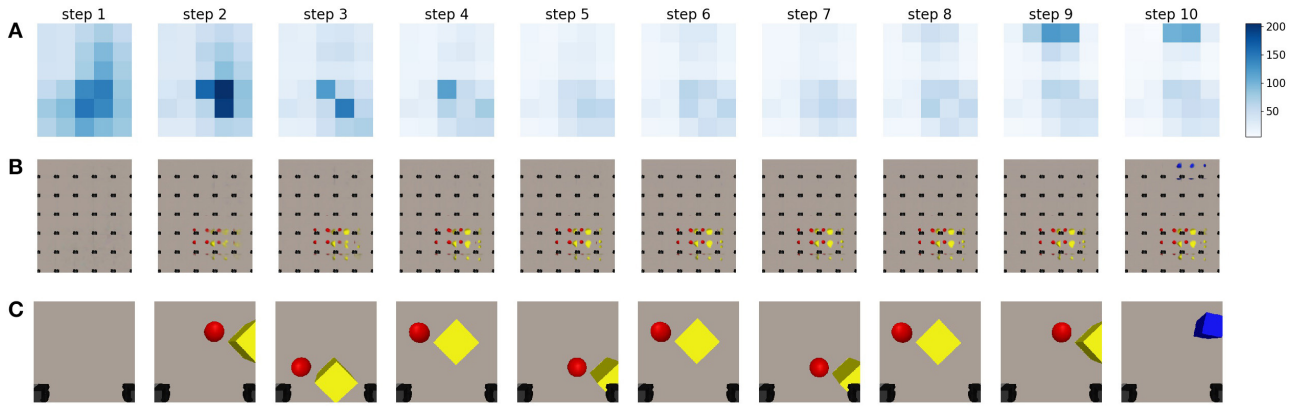


FIGURE 8 | A sequence of steps performed by an artificial agent driven through minimization of the epistemic term from the expected free energy. Information-seeking behavior can be observed. **(A)** A representation of the epistemic value for different potential future viewpoints in different steps of the sequence. The legend is provided on the right, darker values mean higher epistemic values, and are thus more likely to be chosen by the agent. **(B)** The imagined observations for the potential viewpoints at different steps in a sequence executed by an agent driven according to the active inference paradigm. **(C)** The last observation the agent has acquired from the previous step. The black squares in the bottom of each frame are the gripper handles.

at this viewpoint, they contradict each other and the epistemic value is still dominant. Please note that while the absolute value of the instrumental term is much higher than the epistemic term, these are relative to each other. The range of the instrumental term is in the same range as the epistemic value. After observing a few observations, the instrumental term finally takes the upper hand and the agent is driven away to further explore the area. It finally finds the blue cube in the top right in the 7th step. As the instrumental value is very high for this observation, it now takes the upper hand and the agent will naturally remain at this location. Notice how the agent has found the object in less steps than when it was only driven through epistemic value. Because the agent now prefers to search and reach its goal observation, it will avoid getting stuck at a specific location as long as this is not the preferred observation. It is therefore better at finding the target to reach, however it will not necessarily explore the entire workspace, as it would when only considering the epistemic term given enough steps. It is important to note that the instrumental value to the right of the target value is low in magnitude. The model believes it is unlikely that it will find the target observation here. This can be attributed to the pixel-wise log likelihood that is computed, even though the object is in view, because it is at different pixel locations, this will be a less likely observation than an area of the table that does not contain objects. To combat this characteristic, we sample the grid of potential viewpoints with a lot of overlap between the neighboring views.

3.3.2. Extending to Three Degrees of Freedom

Finally, we no longer constrain the movement along the z-axis for the robot manipulator. The orientation is still in a fixed downwards position. We still consider the same scene as in the previous experiments and start the robot gripper in the same initial position without any observations. We evaluate whether this third degree of freedom improves the speed at which the area can be uncovered, and whether the chosen actions matches the biological behavior encountered in for example an owl. The owl will fly to a high vantage point to search for its prey, and move down when it has localized it (Friston et al., 2016).

We task the robot to find the blue cube from the final observation in **Figure 8** again. The different achieved robot poses and their corresponding observations are shown in **Figure 10**. In the executed trajectory, we notice that the owl-like behavior emerges through the minimization of expected free energy. Initially, the agent has no knowledge about the workspace and moves its gripper and corresponding camera toward a higher vantage point from which it can observe the workspace. Initially, the agent only observes a red and a yellow object, after which it moves closer to inspect these objects. It has updated its internal model by observing the object from up close, and it is clear through the instrumental value that the desired observation is not at this location. In a similar manner as explained in the experiment with two degrees of freedom, the agent again moves to a higher vantage point, but more to the center of the table. It is now able to observe both the blue cuboid and the edges of the red and yellow objects. It has localized the target and moves toward its preferred state. The agent does not move in the subsequent steps, showing that it has reached the point that provides it with the lowest expected free energy. We also notice that the agent has found the object faster than in the previous experiment. The additional degree of freedom is immediately exploited by the free energy principle. For the acquired results on additional scenes, the reader is referred to **Supplementary Material**.

4. DISCUSSION

In the above experiments, we have shown that it is possible to use the active inference paradigm as a natural solution for active vision on complex tasks in which the distribution over the environment is not defined upfront. Similar to prior work on learning state space models for active inference (Çatal et al., 2020), we learn our generative model directly from data.

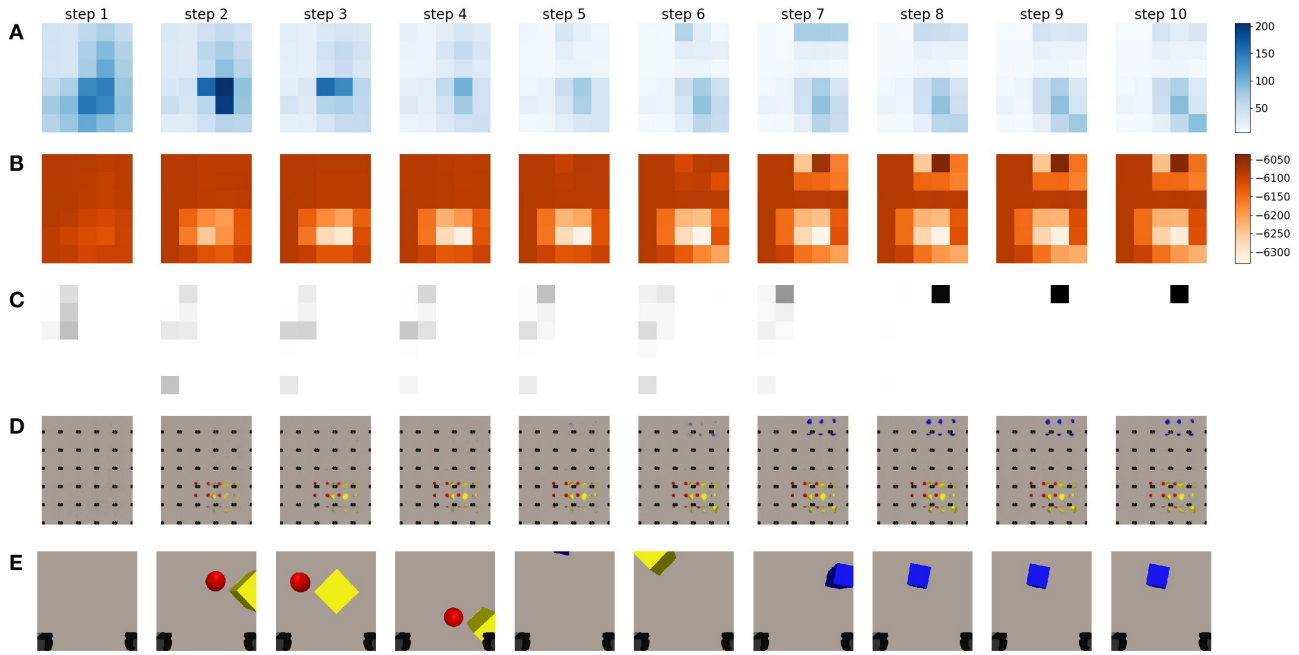


FIGURE 9 | This figure represents a sequence of movement when driving viewpoint selection through minimization of the expected free energy. The agent is now equipped with a preferred state. As a preferred state, we set the final observation from the epistemic exploration shown in **Figure 8C**, but in principle, any observation could be used. **(A)** A representation of the epistemic value for different potential future viewpoints in different steps of the sequence. The legend is provided on the right, darker values mean higher epistemic values, and are thus more likely to be chosen by the agent. **(B)** A representation of the instrumental value for different potential future viewpoints in different steps of a sequence. The legend is provided on the right, darker values mean higher instrumental values and are thus more likely to be chosen by the agent. **(C)** The chosen viewpoint at this step is shown by a black square. This is done by applying the softmax operation to the full expected free energy. **(D)** The imagined observations for the potential viewpoints for different steps in a sequence executed by an active inference driven agent. **(E)** The last observation the agent has acquired from the previous step. The black squares in the bottom of each frame are the gripper handles.

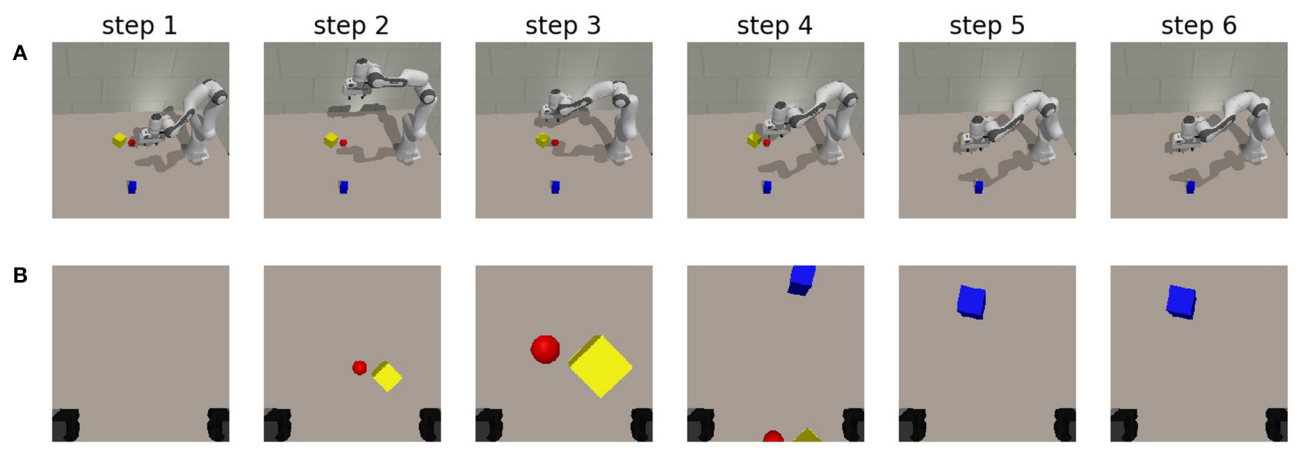


FIGURE 10 | The chosen trajectory for an artificial agent when viewpoints are chosen by the minimization of expected free energy. The preferred state is chosen randomly as the final observation from the epistemic exploration shown in **Figure 8C**. **(A)** Shows the pose of the robot at each step. Step 1 represents the initial state. **(B)** The corresponding observation from the gripper at that step.

We have observed that a sheer epistemic agent will explore the environment by moving to different viewpoints in the world. When we use the full expected free energy to drive viewpoint selection, we observe that epistemic foraging behavior emerges, and the agent will explore the environment with random saccades and will move toward a higher vantage point to observe a larger area at one time, similar to the behavior of an owl scavenging for prey.

For robots to solve complex tasks, one of the first steps is to perceive the world and understand the current situation. This

work shows that the learned generative model is capable of being used in a neurologically inspired solution for perception of the world. As this theoretical framework of active inference is already equipped to deal with actions that perturb the world, this solution can be extended with a more complex generative model that is able to estimate the changes the agent, or other autonomous beings can make in the world.

While our approach allows to learn the generative model purely from pixel data, this also has a couple of drawbacks. In our case for instance, the model is trained using a large amount of data in a simulation environment with a restricted number of object shapes and colors. To be applicable for real-world scenarios, probably an even larger model and dataset are required. Also, it is clear that the reconstructions are not sharp, and blurry objects are reconstructed. This is typical for a variational auto-encoder, and while many approaches exist to create sharper reconstructions (Makhzani et al., 2015; Heljakka et al., 2018, 2020; Huang et al., 2018), we argue that this is not necessary for our case. As long as the generated observations are spatially correlated and the object properties such as size and color are correctly reconstructed, the generative model will be capable of working within the active inference framework. This can be compared to someone trying to remember the fine details of a recently visited building. A person is able to draw the general structure of the building, but will find difficulty to draw each stone correctly with the correct shade. However, this is not necessary to find the door and navigate through the building. Nevertheless, by using the mean squared error in pixel space to train the likelihood model, small-sized objects will generate a small gradient signal, and will be difficult for the model to encode. To mitigate this, one could look at different loss functions, for example perceptual loss (Johnson et al., 2016) or contrastive loss (Hadsell et al., 2006).

Our approach evaluates the expected free energy for a number of considered potential viewpoints. The computational complexity of this algorithm scales linearly with the number of considered viewpoints. However, given enough GPU memory, this algorithm can easily be modified to compute the expected free energy for all potential viewpoints in parallel, making it an algorithm with constant time complexity. Provided that the neural network can be run on a GPU, it can be used for real-time control of physical robot manipulators.

In future work, we want to investigate more efficient methods for evaluating the free energy and planning in a complex state space. In this case, it was feasible to evaluate the expected free energy for each viewpoint as we sampled a limited grid of future viewpoints and only looked at one step in the future. The amount of expected free energy values to compute would increase exponentially, as more time steps ahead are considered. Additionally, in future work we would like to add object interaction, i.e., allowing the robot to move objects in a specific desired configuration. Moreover, this approach will be increasingly important in collaboratory settings. The robotic agent can encounter occlusions and limited field of view for multiple reasons such as other humans obstructing objects or placing things in front of the target object. It is in these situations essential to be able to reason about the scene and choosing the optimal next viewpoint. In follow-up work, the actions of human collaborators can be modeled through their own free energy minimization scheme and can be integrated in the active inference framework to select the next best view. Finally, the goal is to evaluate this method on a real-life robot.

Related Work

The related work falls in two categories, i.e., scene representation learning and related work in the area of active vision. There is a lot of research that considers the problem of scene representation learning and proposes different neural network architectures to aid the process of learning proper representation models of our neural network architecture. In the second part we consider the domain in active vision, this is an active research domain in traditional computer vision problems, but has also been applied to many reinforcement learning tasks.

Scene representation learning is a research field in which the goal is to learn a good representation of the environment. A vast amount of work exists that considers representation learning for separate objects. Multi-View CNN (MVCNN) uses views from multiple viewpoints to learn a representation for classification and segmentation (Su et al., 2015). DeepVoxels uses a geometric representation of the object, in which each voxel has a separate feature vector, which is then rendered through a neural renderer (Sitzmann et al., 2019a). In their follow-up work on Scene Representation Networks, this was extend to replace the voxelized representation by a neural network, estimated through a hypernetwork, that predicts a feature vector for any point in 3D space. These features are then rendered through a neural renderer (Sitzmann et al., 2019b).

Object-centric models have also gained a lot of attention lately. These models stem from the seminal work on Attend Infer Repeat (Eslami et al., 2016) in which a distinct latent code, which separately encodes the position and the type of object, is predicted per object in the scene. This is done through a recurrent neural network that is capable of estimating when all objects are found. In SQ-AIR, this work is extended to sequences of images, and a discovery and propagation mechanism was introduced to track objects through different frames (Kosiorek et al., 2018). These have been extended to better handle physical interactions (Kossen et al., 2020) or be more scalable (Crawford and Pineau, 2020; Jiang et al., 2020). These extensions have also been combined by Lin et al. (2020). 3D-RelNet is also an object-centric model that predicts a pose for each object and their relation to the other objects in the scene (Kulkarni et al., 2019). While these approaches seem promising, in their current implementation they only consider video data from a fixed camera viewpoint. These models do not lend themselves to an active vision system.

Implicit representation models learn the three dimensional properties of the world directly from observations with no intermediate representation. A single neural network is then created for each scene. Neural Radiance Fields (NeRF) learn to infer the color values for each three dimensional point through a differentiable ray tracer from a set of observations (Mildenhall et al., 2020). The follow-up work by Park et al. (2020) adapts the algorithm for a more robust optimization and the work by Xian

et al. (2020) extends this to deal with video sequences. SIRENs also belong to this category, however, this network is optimized directly from the three dimensional point cloud (Sitzmann et al., 2020). While these works often result in very sharp reconstructions with a large amount of detail present in the scenes, they are difficult to optimize due to the large training times and do not allow for new observations to be added on the fly.

The last category of methods encodes the scene in a latent vector that describes the scene in a black box approach. The latent vector does not enforce geometric constraints. The Generative Query Network does this by encoding all observations separately into a latent vector, which is then summed to acquire a global representation of the scene (Eslami et al., 2018). This latent vector can be sampled and decoded through an autoregressive decoder (Gregor et al., 2015), which is then optimized in an end-to-end fashion. This work considers full scenes in which the observer can navigate. This has also been extended with an attention mechanism to separately encode parts of each observation, in order to better capture the information (Burgess et al., 2019). Our model most resembles this GQN architecture, as this is a straightforward implementation that allows for arbitrary viewpoints and which could easily be extended with our Bayesian aggregation strategy. Other approaches result in sharper reconstructions, however they either optimize a neural network per scene, work with a fixed observer viewpoint, or only consider separate objects.

Active vision systems are called active since they can change the camera extrinsic parameters to improve the quality of the perception (Aloimonos et al., 1988). In most active vision research, the next best viewpoints are selected to improve the amount of observations need to scan an area, for exploration and mapping or for reconstruction of the world.

Most traditional methods use a frontier-based approach to select the next viewpoint (Yamauchi, 1997; Chen et al., 2011; Fraundorfer et al., 2012; Forster et al., 2014; Kriegel et al., 2015; Hepp et al., 2018). The frontier is defined as the boundary between the observed area and the unobserved area, and thus these models require an explicit geometric representation of the world. Typically these methods use a discretized map of the world, an occupancy grid in 2D (Yamauchi, 1997) or a voxelized rasterization in 3D (Fraundorfer et al., 2012). The points on the frontier are then evaluated through a utility function that scores the amount of information that will be gained. These utility functions are often hand-crafted and uncertainty or reconstruction based (Wenhardt et al., 2007; Dunn and Frahm, 2009; Forster et al., 2014; Kriegel et al., 2015; Isler et al., 2016; Delmerico et al., 2018; Hepp et al., 2018).

With the rise of deep learning, active vision problems has also been tackled through learning-based approaches. The problem has been cast as a set covering optimization problem in which a reinforcement learning agent has to select the least amount of views to observe the area (Devrim Kaba et al., 2017). This approach assumes that the area is known in advance, and that an agent can be trained on this. It does not allow for unseen environments. Other deep learning techniques have also been proposed. Hepp et al. (2018) learn a utility function using a data-driven approach that predicts the amount of new information gained from a given viewpoint. They learn this directly using supervision with oracle data. Instead of learning a utility function, deep neural networks that directly predict the next-best viewpoint have also been researched (Doumanoglou et al., 2016; Mendoza et al., 2020). These methods require a ground-truth "best" view, for which a dataset is created using the full scene or object information.

Biology has inspired work on active vision and perception as well. An active vision system for robotic manipulators was proposed that is inspired by the way primates deal with their visual inputs (Ognibene and Baldassare, 2014). Rasouli et al. (2019) propose a probabilistic bio-inspired attention-based visual search system for mobile robotics. Similar to our work, active inference has already been applied to different active vision settings. Mirza et al. (2016) show that the free energy principle can be used for visual foraging. They define a classification task, where the agent must acquire visual cues to correctly classify the scenario it is in. Follow-up work (Conor et al., 2020) considers a hierarchical scene in which decisions are made at multiple levels. Fovea-based attention to improve perception and recognition on image data has been performed through the free energy principle (Daucé, 2018). While these approaches show promising results, they all consider designed scenarios for which the state space can be carefully crafted in advance.

Our approach closely connects to traditional active vision systems in which a utility function is evaluated. The expected free energy formulation is used as a utility function in our work. However, in contrast to these traditional approaches, we use a deep neural network to encode the representation of the environment instead of using geometric representations or hand-crafting the distributions that are acquired. While active vision techniques that use neural networks typically use these models to predict the next best viewpoint directly, or predict a learned utility function. We reason that the expected free energy is a natural solution to this problem, as this is the utility function that determine the actions of living organisms (Friston, 2013). We use our neural networks to imagine future states, belief about the environment and, similar to the work Finn and Levine (2017), use these to plan the agent's actions.

5. CONCLUSION

In this paper we investigated whether the active inference paradigm could be used for a robotic searching and reaching task. As it is impossible for real-world scenarios to define the generative model upfront, we investigated the ability to use a learned generative model to this end. We showed that we were able to approximate a generative model using deep neural networks and that this can be learned directly from pixel observations by means free energy minimization. To this end we expanded the Generative Query Network by aggregating the latent distributions from each observation through a Gaussian multiplication. We conducted an ablation study and showed that this model had similar performance

as other aggregation methods when operating in the training range, and that the model outperformed other techniques when multiple observations were considered. In a second experiment we evaluated whether this model was capable of inferring information about a cup, namely its orientation and whether or not it has a handle. We showed that the agent actively samples the world from viewpoints that allow itself to reduce the uncertainty on its belief state distributions. In the third case, we show that an artificial agent with a robotic manipulator explores the environment until it has observed all objects in the workspace. We showed that if the viewpoints are chosen by minimization of the expected free energy when provided with a target goal, the agent explores the area in a biologically-inspired manner and navigates toward the goal viewpoint once it has acquired enough information to determine this specific viewpoint.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

TVa and TVe conceived and performed the experiments. TVa, OÇ, and TVe worked out the mathematical basis for the experiments. TVa, TVe, OÇ, CD, and BD contributed to the manuscript. BD supervised the experiments. All authors contributed to the article and approved the submitted version.

FUNDING

This research received funding from the Flemish Government (AI Research Program). OÇ was funded by a Ph.D. grant of the Flanders Research Foundation (FWO). Part of this work has been supported by Flanders Innovation & Entrepreneurship, by way of grant agreement HBC.2020.2347.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.642780/full#supplementary-material

REFERENCES

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *Int. J. Comput. Vis*. 1, 333–356. doi: 10.1007/BF00133571

Beal, M. J. (2003). *Variational algorithms for approximate Bayesian inference* (Ph.D. thesis). University College London, London, United Kingdom.

Billard, A., and Kragic, D. (2019). Trends and challenges in robot manipulation. *Science* 364:6446. doi: 10.1126/science.aat8414

Blender Online Community (2018). *Blender - a 3D Modelling and Rendering Package*. Amsterdam: Blender Foundation; Stichting Blender Foundation.

Burgess, C. P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., et al. (2019). Monet: Unsupervised scene decomposition and representation. *arXiv [Preprint]. arXiv:1901.11390*.

Çatal, O., Wauthier, S., De Boom, C., Verbelen, T., and Dhoedt, B. (2020). Learning generative state space models for active inference. *Front. Comput. Neurosci*. 14:103. doi: 10.3389/fncom.2020.574372

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., et al. (2015). *ShapeNet: An Information-Rich 3D Model Repository*. Technical Report, Stanford University; Princeton University; Toyota Technological Institute at Chicago.

Chen, S., Li, Y., and Kwok, N. M. (2011). Active vision in robotic systems: a survey of recent developments. *Int. J. Robot. Res*. 30, 1343–1377. doi: 10.1177/0278364911410755

Conor, H. R., Berk, M. M., Thomas, P., Karl, F., Igor, K., Arezoo, P. (2020). Deep active inference and scene construction. *Front. Artif. Intell*. 3:509354. doi: 10.3389/frai.2020.509354

Crawford, E., and Pineau, J. (2020). "Exploiting spatial invariance for scalable unsupervised object tracking," in *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020* (New York, NY: AAAI Press), 3684–3692. doi: 10.1609/aaai.v34i04.5777

Daucé, E. (2018). Active fovea-based vision through computationally-effective model-based prediction. *Front. Neurorobot*. 12:76. doi: 10.3389/fnbot.2018.00076

Delmerico, J., Isler, S., Sabzevari, R., and Scaramuzza, D. (2018). A comparison of volumetric information gain metrics for active 3d object reconstruction. *Auton. Robots* 42, 197–208. doi: 10.1007/s10514-017-9634-0

Devrim Kaba, M., Gokhan Uzunbas, M., and Nam Lim, S. (2017). "A reinforcement learning approach to the view planning problem," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 6933–6941. doi: 10.1109/CVPR.2017.541

Doumanoglou, A., Kouskouridas, R., Malassiotis, S., and Kim, T.-K. (2016). "Recovering 6d object pose and predicting next-best-view in the crowd," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Las Vegas, NV), 3583–3592. doi: 10.1109/CVPR.2016.390

Dunn, E., and Frahm, J.-M. (2009). "Next best view planning for active model improvement," in *Proceedings of the British Machine Vision Conference*, eds A. Cavallaro, S. Prince, and D. Alexander (BMVA Press). doi: 10.5244/C.23.53

Eslami, S. A., Heess, N., Weber, T., Tassa, Y., Szepesvari, D., Hinton, G. E., et al. (2016). "Attend, infer, repeat: fast scene understanding with generative models," in *Advances in Neural Information Processing Systems* (Barcelona), 3225–3233.

Eslami, S. M. A., Rezende, D. J., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., et al. (2018). Neural scene representation and rendering. *Science* 360, 1204–1210. doi: 10.1126/science.aar6170

Finn, C., and Levine, S. (2017). "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Singapore), 2786–2793. doi: 10.1109/ICRA.2017.7989324

Forster, C., Pizzoli, M., and Scaramuzza, D. (2014). "Appearance-based active, monocular, dense reconstruction for micro aerial vehicles," in *Conference: Robotics: Science and Systems (RSS)* (Berkely, CA) doi: 10.15607/RSS.2014.X.029

Fraundorfer, F., Heng, L., Honegger, D., Lee, G. H., Meier, L., Tanskanen, P., et al. (2012). "Vision-based autonomous mapping and exploration using a quadrotor MAV," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Vilamoura), 4557–4564. doi: 10.1109/IROS.2012.6385934

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci*. 11, 127–138. doi: 10.1038/nrn2787

Friston, K. (2013). Life as we know it. *J. R. Soc. Interface* 10:20130475. doi: 10.1098/rsif.2013.0475

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O'Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neurosci. Biobehav. Rev*. 68, 862–879. doi: 10.1016/j.neubiorev.2016.06.022

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., et al. (2018). Neural processes. *arXiv [Preprint]. arXiv:1807.01622*.

Gregor, K., Danihelka, I., Graves, A., Rezende, D., and Wierstra, D. (2015). "Draw: a recurrent neural network for image generation," in *International Conference on Machine Learning* (Lille: PMLR), 1462–1471.

Hadsell, R., Chopra, S., and Lecun, Y. (2006). "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* (New York, NY), 1735–1742. doi: 10.1109/CVPR.2006.100

Häni, N., Engin, S., Chao, J.-J., and Isler, V. (2020). "Continuous object representation networks: novel view synthesis without target view supervision," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, (Vancouver, BC).

Heljakka, A., Solin, A., and Kannala, J. (2018). "Pioneer networks: progressively growing generative autoencoder," in *Asian Conference on Computer Vision* (Perth: Springer), 22–38. doi: 10.1007/978-3-030-20887-5_2

Heljakka, A., Solin, A., and Kannala, J. (2020). "Towards photographic image manipulation with balanced growing of generative autoencoders," in *The IEEE Winter Conference on Applications of Computer Vision* (Snowmass Village, CO), 3120–3129. doi: 10.1109/WACV45572.2020.9093375

Hepp, B., Dey, D., Sinha, S. N., Kapoor, A., Joshi, N., and Hilliges, O. (2018). "Learn-to-score: efficient 3d scene exploration by predicting view utility," in *Proceedings of the European Conference on Computer Vision (ECCV)* (Munich), 437–452. doi: 10.1007/978-3-030-01267-0_27

Huang, H., He, R., Sun, Z., Tan, T., et al. (2018). "Introvae: Introspective variational autoencoders for photographic image synthesis," in *Advances in Neural Information Processing Systems* (Montreal, QC), 52–63.

Isler, S., Sabzevari, R., Delmerico, J., and Scaramuzza, D. (2016). "An information gain formulation for active volumetric 3d reconstruction," in *2016 IEEE International Conference on Robotics and Automation (ICRA)* (Stockholm), 3477–3484. doi: 10.1109/ICRA.2016.7487527

Jiang, J., Janghorbani, S., de Melo, G., and Ahn, S. (2020). "SCALOR: generative world models with scalable object representations," in *8th International Conference on Learning Representations, ICLR 2020* (Addis Ababa: OpenReview.net). Available online at: https://openreview.net/forum?id=SJxrKgStDH

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision – ECCV 2016*, eds B. Leibe, J. Matas, N. Sebe, and M. Welling (Cham: Springer International Publishing), 694–711.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *J. Basic Eng*. 82, 35–45. doi: 10.1115/1.3662552

Kingma, D. P., and Ba, J. (2015). "Adam: a method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, eds Y. Bengio and Y. LeCun (San Diego, CA).

Kingma, D. P., and Welling, M. (2014). "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014* (Banff, AB, Canada).

Kosiorek, A. R., Kim, H., Posner, I., and Teh, Y. W. (2018). "Sequential attend, infer, repeat: generative modelling of moving objects," in *Advances in Neural Information Processing Systems* (Montreal, QC).

Kossen, J., Stelzner, K., Hussing, M., Voelcker, C., and Kersting, K. (2020). "Structured object-aware physics prediction for video modeling and planning," in *International Conference on Learning Representations* (Addis Ababa). Available online at: https://openreview.net/forum?id=B1e-kxSKDH

Kriegel, S., Rink, C., Bodenmüller, T., and Suppa, M. (2015). Efficient next-best-scan planning for autonomous 3d surface reconstruction of unknown objects. *J. Real-Time Image Process*. 10, 611–631. doi: 10.1007/s11554-013-0386-6

Kulkarni, N., Misra, I., Tulsiani, S., and Gupta, A. (2019). "3D-relnet: joint object and relational network for 3d prediction," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2212–2221. doi: 10.1109/ICCV.2019.00230

Lin, Z., Wu, Y.-F., Peri, S., Fu, B., Jiang, J., and Ahn, S. (2020). Improving generative imagination in object-centric world models. *arXiv:2010.02054*.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. (2015). Adversarial autoencoders. *arXiv [Preprint]. arXiv:1511.05644*.

Matsumoto, T., and Tani, J. (2020). Goal-directed planning for habituated agents by active inference using a variational recurrent neural network. *Entropy* 22:564. doi: 10.3390/e22050564

Mendoza, M., Vasquez-Gomez, J. I., Taud, H., Sucar, L. E., and Reta, C. (2020). Supervised learning of the next-best-view for 3D object reconstruction. *Pattern Recogn. Lett*. 133, 224–231. doi: 10.1016/j.patrec.2020.02.024

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020) "Nerf: Representing scenes as neural radiance fields for view synthesis", in *Frahm Computer Vision? ECCV 2020. ECCV 2020. Lecture Notes in Computer Science*, Vol. 12346, eds A. Vedaldi, H. Bischof, T. Brox, and J. M. Frahm (Glasgow, UK; Cham: Springer) doi: 10.1007/978-3-030-58452-8_24

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE* 13:e190429. doi: 10.1371/journal.pone.0190429

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

Ognibene, D., and Baldassare, G. (2014). 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. Mental Dev*. 7, 3–25. doi: 10.1109/TAMD.2014.2341351

Park, K., Sinha, U., Barron, J. T., Bouaziz, S., Goldman, D. B., Seitz, S. M., et al. (2020). Deformable neural radiance fields. *arXiv [Preprint]. arXiv:2011.12948*.

Parr, T., and Friston, K. J. (2017). The active construction of the visual world. *Neuropsychologia* 104, 92–101. doi: 10.1016/j.neuropsychologia.2017.08.003

Perez, E., Strub, F., de Vries, H., Dumoulin, V., and Courville, A. (2018). "Film: visual reasoning with a general conditioning layer," in *Proceedings of the AAAI Conference on Artificial Intelligence* (New Orleans, LA), Vol. 32. Available online at: https://ojs.aaai.org/index.php/AAAI/article/view/11671

Rasouli, A., Lanillos, P., Cheng, G., and Tsotsos, J. K. (2019). Attention-based active visual search for mobile robots. *Auton. Robots* 44, 131–146. doi: 10.1007/s10514-019-09882-z

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). "Stochastic backpropagation and approximate inference in deep generative models," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014* (Beijing), 1278–1286.

Rezende, D. J., and Viola, F. (2018). Taming vaes. *CoRR, abs/1810.00597*.

Rohmer, E., Singh, S. P. N., and Freese, M. (2013). "Coppeliasim (formerly v-rep): a versatile and scalable robot simulation framework," in *Proc. of The International Conference on Intelligent Robots and Systems (IROS)* (Tokyo). doi: 10.1109/IROS.2013.6696520

Sitzmann, V., Martel, J. N. P., Bergman, A. W., Lindell, D. B., and Wetzstein, G. (2020). "Implicit neural representations with periodic activation functions," in *Proc. NeurIPS*.

Sitzmann, V., Thies, J., Heide, F., Nießner, M., Wetzstein, G., and Zollhöfer, M. (2019a). "Deepvoxels: Learning persistent 3d feature embeddings," in *Proc. Computer Vision and Pattern Recognition (CVPR)* (Long Beach, CA). doi: 10.1109/CVPR.2019.00254

Sitzmann, V., Zollhöfer, M., and Wetzstein, G. (2019b). "Scene representation networks: continuous 3d-structure-aware neural scene representations," in *Advances in Neural Information Processing Systems* (Vancouver, BC).

Srihasam, K., and Bullock, D. (2008). Target selection by the frontal cortex during coordinated saccadic and smooth pursuit eye movements. *J. Cogn. Neurosci*. 21, 1611–1627. doi: 10.1162/jocn.2009.21139

Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. G. (2015). "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Santiago), 945–953. doi: 10.1109/ICCV.2015.114

Wenhardt, S., Deutsch, B., Angelopoulou, E., and Niemann, H. (2007). "Active visual object reconstruction using d-, e-, and t-optimal next best views," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN), 1–7. doi: 10.1109/CVPR.2007.383363

Xian, W., Huang, J.-B., Kopf, J., and Kim, C. (2020). Space-time neural irradiance fields for free-viewpoint video. *arXiv [Preprint]. arXiv:2011.12950*.

Yamauchi, B. (1997). "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation'* (Monterey, CA), 146–151. doi: 10.1109/CIRA.1997.613851

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.

APPENDIX

TABLE A1 | Neural network architecture.

	Layer	Neurons/filters
Posterior (ϕ)	Convolutional (1 × 1)	64
	Convolutional (3 × 3)	16
	LeakyReLU	
	FiLM (conditioned on v_k)	16
	Convolutional (3 × 3)	32
	LeakyReLU	
	FiLM (conditioned on v_k)	32
	Convolutional (3 × 3)	64
	LeakyReLU	
	FiLM (conditioned on v_k)	64
	Convolutional (3 × 3)	128
	LeakyReLU	
	FiLM (conditioned on v_k)	128
	Linear	2 × latent size

The posterior model describes the encoder used in the neural network. The latent size varies from experiment to experiment. In the ShapeNet experiment, the latent size is 64, in the experiment of the cup, the latent size is 9. In the final case, for the robotic workspace, the latent size is 256. In the posterior model, each 3 × 3 convolution uses a stride of 2 to reduce the spatial resolution of the data. The 1 × 1 convolutions use a stride of 1.

TABLE A2 | Neural network architecture of the likelihood model.

	Layer	Neurons/filters
Likelihood (ψ)	Linear	4 × 4 × 3
	LeakyReLU	
	Convolutional (3 × 3)	128
	LeakyReLU	
	Convolutional (3 × 3)	128
	LeakyReLU	
	FiLM (conditioned on v_k and s)	128
	Convolutional (3 × 3)	64
	LeakyReLU	
	Convolutional (3 × 3)	64
	LeakyReLU	
	FiLM (conditioned on v_k and s)	64
	Convolutional (3 × 3)	32
	LeakyReLU	
	Convolutional (3 × 3)	32
	LeakyReLU	
	FiLM (conditioned on v_k and s)	32
	Convolutional (3 × 3)	16
	LeakyReLU	
	Convolutional (3 × 3)	16
	LeakyReLU	
	FiLM (conditioned on v_k and s)	16
	Convolutional (1 × 1)	3

This model estimates the pixel values of a potential viewpoint. Each 3 × 3 convolution is preceded by a linearly upsample step that doubles the image resolution. The 1 × 1 convolutions use a stride of 1.

frontiers in Neurorobotics

BRIEF RESEARCH REPORT
published: 30 March 2021
doi: 10.3389/fnbot.2021.639999

Gazing at Social Interactions Between Foraging and Decision Theory

Alessandro D'Amelio and Giuseppe Boccignone*

PHuSe Lab, Department of Computer Science, Universitá degli Studi di Milano, Milan, Italy

Finding the underlying principles of social attention in humans seems to be essential for the design of the interaction between natural and artificial agents. Here, we focus on the computational modeling of gaze dynamics as exhibited by humans when perceiving socially relevant multimodal information. The audio-visual landscape of social interactions is distilled into a number of multimodal patches that convey different social value, and we work under the general frame of foraging as a tradeoff between local patch exploitation and landscape exploration. We show that the spatio-temporal dynamics of gaze shifts can be parsimoniously described by Langevin-type stochastic differential equations triggering a decision equation over time. In particular, value-based patch choice and handling is reduced to a simple multi-alternative perceptual decision making that relies on a race-to-threshold between independent continuous-time perceptual evidence integrators, each integrator being associated with a patch.

Keywords: audio-visual attention, gaze models, social interaction, multimodal perception, drift-diffusion model, decision theory, perceptual decisions

OPEN ACCESS

Edited by:
Tom Foulsham,
University of Essex, United Kingdom

Reviewed by:
Giacinto Barresi,
Italian Institute of Technology (IIT), Italy
Roy S. Hessels,
Utrecht University, Netherlands

****Correspondence:***
Alessandro D'Amelio
alessandro.damelio@unimi.it

Received: *10 December 2020*
Accepted: *09 March 2021*
Published: *30 March 2021*

Citation:
D'Amelio A and Boccignone G (2021) Gazing at Social Interactions Between Foraging and Decision Theory. Front. Neurorobot. 15:639999. doi: 10.3389/fnbot.2021.639999

1. INTRODUCTION

The main concern of this work is modeling gaze dynamics as exhibited by humans when perceiving socially relevant multimodal information. Such dynamics accounts for gaze deployment as unfolding in time, depending on where observers look, how long and when. It is known that under certain circumstances humans spend the majority of time scrutinizing people, markedly their eyes and faces, and spotting persons that are talking (cfr., Foulsham et al., 2010, for framing this study, but see Hessels, 2020 for an in-depth discussion under general conditions and an up-to-date review). This is not surprising since social gazing abilities are likely to have played a significant role very early in the primate lineage (Shepherd and Platt, 2007).

Gaze, the act of directing the eyes toward a location in the visual world, is considered a good measure of overt attention (Kustov and Robinson, 1996). This makes the research problem addressed here relevant for many aspects, with promising applications in different fields, such as social robotics, social gaze analysis, and clinical studies (Hessels, 2020). Endowing artificial agents with the ability to gaze at social cues—a building block for many dyadic, triadic, and multiparty interactions (Hessels, 2020)- has been deemed essential since early attempts to build socially competent robots (Admoni and Scassellati, 2017; Wiese et al., 2017). A growing body of research is devoted to quantitatively assess how humans gather social information through gaze so to infer other persons' intentions, feelings, traits, expertise, or even expectations and to analyse group dynamics (Staab, 2014; Rubo and Gamer, 2018; Grossman et al., 2019; Guy et al., 2019; Jording et al., 2019). Over the years, a broad research spectrum has been established from traditional laboratory

studies of social attention or social gaze to interactive settings, unveiling the complexity of the problem (but see Hessels, 2020 for an enlightening and in-depth discussion). The conversational videos we are exploiting have the virtue of displaying real people embedded in a dynamic situation while being relatively controlled stimuli (Foulsham et al., 2010). In clinical research gaze is central to the investigation of attention mechanisms in groups of patients with atypical development in the appraisal of social cues, e.g., social anxiety disorder, autism spectrum disorder, schizophrenia (Klein et al., 2019). To such end, the analysis of social perception by employing contextually rich video stimuli poses little cognitive demands to the participants (Rubo and Gamer, 2018). Meanwhile, modeling gaze as a dynamical stochastic process that unfolds in space and time is gaining currency in clinical studies (e.g., Korda et al., 2016; Ioannou et al., 2020).

Surprisingly, limited research has addressed the computational modeling of eye guidance in a multimodal setting; only a handful of works have considered social cues in such setting (cfr. Tavakoli et al., 2020, and Boccignone et al., 2020, for a review). Yet, even when limiting to the unimodal case of visual stimuli, gaze dynamics has been by and large overlooked in computer vision in spite of the pioneering work of Aloimonos et al. (1988), Ballard (1991), and Bajcsy and Campos (1992). The current state of affairs is that effort is mostly spent to model salience (Borji and Itti, 2013; Borji, 2021) as a tool for predicting where/what to look at (for a critical discussion, see Tatler et al., 2011; Le Meur and Liu, 2015; Foulsham, 2019; Boccignone et al., 2020; Zhang et al., 2020).

Here we take a different stance and we focus on modeling gaze dynamics. To such end we build on foraging theory. Foraging is a general term that includes where animals search for food and which sorts of food they eat (Stephens, 1986; Bartumeus and Catalan, 2009). In brief, the animal strives for maximizing his intake of food in a "patchy" landscape: moment by moment it selects the most convenient patch, moves to the patch and starts foraging in that location. While exploiting the patch, the animal gains energy at a rate that decreases as the food becomes depleted: thus, at any time, he has to make a decision whether to stay or leave for the next patch (MacArthur and Pianka, 1966).

Foraging is an appealing and principled framework for dealing with gaze. The idea is simple: gaze deployment is the result of the foraging behavior of the observer. Consider **Figure 1**. The top-left image displays a video frame of a conversational clip overlaid with a number of computed audio-visual patches. The gaze trajectory of a perceiver, who is viewing and listening to the clip, unfolds such that local, within-patch exploitation alternates with long between-patch relocations (cfr. **Figure 1**, bottom-right image). Indeed, much like the foraging animal, the perceiver contends with two problems: *What* defines a patch as valuable to gaze at? *How* is gaze guided within and between patches?

The idea of exploiting the foraging framework has gained currency in the visual attention field and human cognition theories (e.g., Hills, 2006; Pirolli, 2007; Cain et al., 2012; Wolfe, 2013; Ehinger and Wolfe, 2016; Mirza et al., 2016), and it is deemed more than an informing metaphor (Wolfe, 2013). It has been argued that what was once foraging for tangible resources in a physical space became, over evolutionary time, foraging for information in cognitive space (Hills, 2006).

In this perspective, the selection of individual patches is not the most relevant issue (Wolfe, 2013; Ehinger and Wolfe, 2016). Of more interest is when does a forager leave one patch for the next one. Namely, the primary metric of concern in animal ecology studies is the patch giving-up time (GUT). The most influential account of average patch leaving behavior is Charnov's Marginal Value Theorem (MVT, Charnov, 1976). The MVT states that it is time to move when the rate of energy gain from the currently visited patch drops below the average rate. The latter, in turn, depends on the rate at which resources can be extracted from patches and on the time for relocating to the next patch. Accordingly, a poor patch yielding a low energy gain should be abandoned earlier.

Recently, a model has been proposed (Boccignone et al., 2020) that takes into account the above questions in order to reframe gaze deployment as the behavior of a stochastic forager while visiting audio-visual patches that convey different social value. Most relevant, the patch leaving time was obtained via the stochastic version of the MVT (McNamara, 1982). However, the advantage of having a general solution derived from first principles in the framework of optimal Bayesian foraging (Bartumeus and Catalan, 2009) is mitigated by a computational cost that might impact on possible application, such as social robotics (cfr., **Supplementary Figures 2**, **3**).

In this brief research report we investigate a patch handling model, which is alternative to that proposed in Boccignone et al. (2020). Here, the decision of relocating gaze from one patch to the other relies on simple multi-alternative perceptual decision making that embeds both patch leaving and choice. The latter takes stock of recent work that spells out animal foraging in terms of an evidence accumulation process (Davidson and El Hady, 2019). In our case evidence denotes the estimate of the relative value of scrutinizing a patch with respect to the others. We consider an integration-to-threshold mechanism, namely a race-to-threshold between continuous-time independent evidence integrators, each being associated with a patch. A snapshot of the process is displayed in the top-right panel of **Figure 1**, which shows the stochastic evolution of patch-related evidence. Meanwhile, in the same vein of Boccignone et al. (2020), the spatial displacement of gaze within and between patches is obtained via an Ornstein-Uhlenbeck (O-U) process that operates at two different spatial scales, local and global (bottom panels of **Figure 1**).

As a result, the gaze deployment problem can be parsimoniously formalized, both in time and in space, through the evolution of a set of Langevin-type stochastic differential equations. Then the question arises whether the model presented here retains the same basic response features obtained by Boccignone et al. (2020) while being computationally more efficient.

In the Methods section, the model is presented to bare essentials together with the experimental setup and the evaluation protocol. In the Results section the outcomes of the model are juxtaposed with those from the method introduced in Boccignone et al. (2020); comparison with other methods is

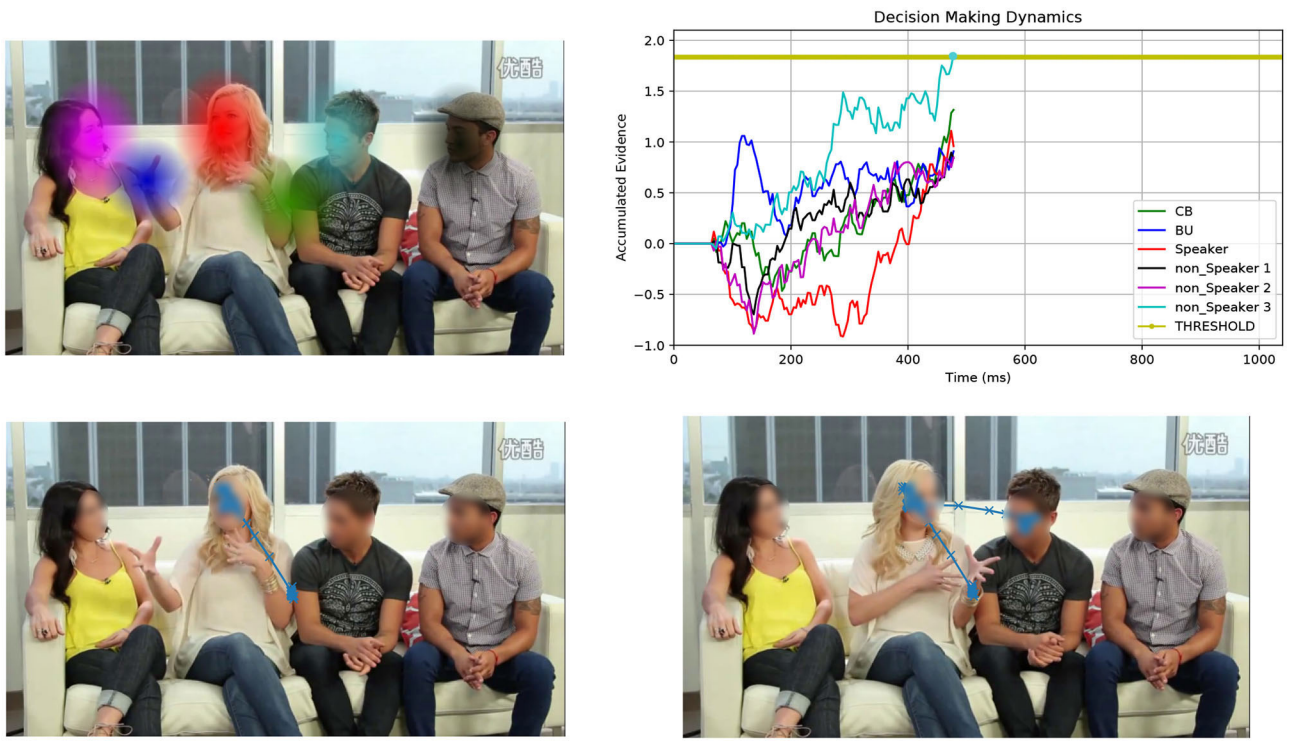


FIGURE 1 | Overall view of the patch cycle at the basis of the proposed model. (Top left) At any time t the perceiver captures the multimodal landscape of social interactions as a set of audio-visual patches that convey different social value (speakers, faces, gestures, etc.); patches are shown as colored Gaussian blobs that overlay the original video frame. (Bottom left) The simulated 2D spatial random walk (O-U process) is displayed starting from the frame center up to current gaze location within the red patch (speaker's face). (Top right) The decision making dynamics instantiated as the stochastic evolution (1D random walk with drift) of independent racers, one for each patch (patches and racers are coded by corresponding colors); the current patch (red blob) is scrutinized until one of the racers (winner) hits the threshold; the winner sets the next gaze attractor on the corresponding patch; in this case the light blue patch is the winner (non-speaking face); (Bottom right) The simulated gaze trajectory within the new chosen patch after between-patch relocation has been performed. See text for details.

available in the **Supplementary Material** section, too. It is shown that the simulated scan paths exhibit features that are statistically similar to those of eye movements of human observers that were eye tracked while watching and listening to conversational clips in a free-viewing condition. Notably, the performance attained is comparable, albeit relying on a simpler mechanism, and at a low computational cost. Eventually, we discuss the results so far achieved, highlighting the novelties of the method and its pitfalls, while addressing its implications in perspective.

2. METHODS

2.1. The Model

The input to the model at time t is the multimodal landscape, which we define as the time-varying ensemble of audio-visual patches $\mathcal{W}(t) = \{\mathcal{P}_p(t)\}_{p=1}^{N_P}$. These serve as regions of gaze attraction. Each patch is shaped as a 2-D Gaussian with localization parameter (mean) $\boldsymbol{\mu}_p$ and shape parameter (covariance matrix) $\boldsymbol{\Sigma}_p$. One example is provided in the top-left image of **Figure 1** displaying the set of computed patches $\mathcal{W}(t)$ as Gaussian blobs that overlay the original video frame; the patches correspond to the current speaker's face, the faces of the listeners, the speaker's hand gesture, and a center-bias patch. It is worth noting that the model needs not to rely upon any specific technique for deriving the pre-attentive representation $\mathcal{W}(t)$, as long as it captures relevant social multimodal information within the scene (persons, speakers, gestures, etc.).

Moment by moment, the perceiver, who is viewing and listening to the audio-visual clip, will (1) select one patch to gaze at, most likely the speaking face, (2) scrutinize it for a certain time, (3) move to a different patch, and so forth. Denote $\mathbf{r}_F(t) = (x_F(t), y_F(t))$ the vector of the spatial coordinates of gaze at time t.

The evolution over time of $\mathbf{r}_F(t)$ defines a trajectory, that is the spatiotemporal dynamics of gaze. Such trajectories are best described as the unfolding of local displacements within a patch followed by larger relocations between patches. Gaze allocation to one patch depends on the time-varying context of the scene and on the value V_p that each patch p is assigned within such context (e.g., the value of a patch including a face of a speaking person changes when the person becomes silent). In our setting, no specific external task or goal is given (free-viewing condition). Then, if the ultimate objective of an active perceiver is total reward maximization (Zhang et al., 2020), reward can be related to the "internal" value (Berridge and Robinson, 2003). The latter has different psychological facets including affect (implicit "liking" and conscious pleasure) and motivation

(implicit incentive salience, "wanting"). Indeed, social signals are expected to induce responses as other reward stimuli do, i.e., motivational approach as well as hedonic response (Vernetti et al., 2018).

Under such circumstances, the behavior of the perceiver can be formalized as that of a forager, gaze being the means to gather valuable information within the scene. At any time t, the forager is engaged either in local patch *exploitation* or in landscape *exploration* across patches. This entails solving the decision making problems of patch choice and patch giving up, together with setting the appropriate spatial dynamics for visiting the currently handled patch or relocating to a new one.

As to the decision problem, here, rather than resorting to optimal Bayesian foraging (Boccignone et al., 2020), we frame it in the physics of optimal decision making (Bogacz et al., 2006; Gold and Shadlen, 2007). Decision making depends upon the forager's estimate of the relative value of scrutinizing a patch with respect to the others, namely the evidence $q_p(t)$ assigned to patch p at time t (Bogacz et al., 2006; Gold and Shadlen, 2007). In turn, the evidence depends on both the patch value V_p and the overall dynamics of the patch ensemble (cfr. section 2.1.1 below). Evidence accumulation is computed by integrating a 1D Markov Gaussian process in the form of a Langevin-type drift-diffusion model. The decision making process is summarized via the evolution of state variables s_t and p_t^*. Both depend upon the evidence $q_p(t)$: the first one is a binary random variable accounting for the switching from within-patch exploitation ($s_t = 0$) to between-patch relocation ($s_t = 1$); the second variable indexes the patch chosen to be handled at time t, thus $p_t^* \in \{1, \cdots, N_P\}$.

Eventually, based on the forager's decisions, the stochastic evolution of gaze deployment, namely the spatiotemporal trajectory $\mathbf{r}_F(t)$, is generated by a 2D Markov Gaussian process. Precisely the latter is a 2D Ornstein-Uhlenbeck (O-U) process, which operates at two different spatial scales, within-patch and between patches, respectively. The O-U process is a mean reverting process where patches serve as trajectory attractors (cfr. section 2.1.2); the typical outcome of the O-U process is displayed in the bottom panels of **Figure 1**.

The model can be succinctly formalized as follows:

2.1.1. Decision Making Dynamics

We represent the perceptual decision making problem as a continuous-time race model in a multi-choice setting (Bogacz et al., 2006; Krajbich and Rangel, 2011), where the N_P patches compete one against the other to attract gaze. The response at time t is obtained by evolving over time, for each patch, the evidence accumulation process until a choice is made (top-right panel of **Figure 1**). Evidence in favor of each patch is accumulated at different rates depending on the patch value and on whether it is being gazed. For each patch p, the process has the form of the following stochastic differential equation (SDE):

$$dq_p(t) = I_p(t)dt + cdW(t), \qquad p = 1, \cdots, N_P. \tag{1}$$

The drift term $I_p(t)$ denotes the mean rate of incoming evidence; the second term cdW (W being a Wiener process) represents white noise, which is Gaussian distributed with mean 0 and variance c^2dt.

Equation (1) can be numerically integrated between 0 and t with initial condition $q_p(t) = 0$: (Lemons, 2002; Kloeden and Platen, 2013):

$$q_p(t') = q_p(t) + I_p(t)\delta t + c\sqrt{\delta t}z(t), \qquad p = 1, \cdots, N_P, \tag{2}$$

with $z(t) \sim \mathcal{N}(0, 1)$ and δt being the time increment $t' = t + \delta t$. We set $c = 1$; drift $I_p(t)$ is computed as follows:

Assume that the value V_p is available for each patch p on the basis of the patch type; this can be derived, for instance, from eye tracking data as the prior probability of gazing at speaking persons, non-speakers, etc., within the social scene. The drift rate $I_p(t)$ associated to the racer of the p-th patch at time t depends on whether or not patch p is being currently exploited, i.e., $p = p^*$ and $p \neq p^*$, respectively, and on the relative patch value ν_p:

$$I_p(t) = \Psi(p, p^*)\nu_p. \tag{3}$$

Define the gazing function Ψ as

$$\Psi(p, p^*) = \begin{cases} e^{-\frac{\phi}{V_p}t} & p = p^* \\ 1 & \text{otherwise} \end{cases}, \tag{4}$$

ϕ being a positive constant; the relative value ν_p is

$$\nu_p = \eta \frac{V_p}{V_{p^*}} \; e^{-\kappa \left\| \boldsymbol{\mu}_p - \boldsymbol{\mu}_{p^*} \right\|} \tag{5}$$

In Equation (5) the negative exponential $e^{-\kappa \left\| \boldsymbol{\mu}_p - \boldsymbol{\mu}_{p^*} \right\|}$, $\kappa > 0$ accounts for the visibility of the patch p from the current patch p^*. The visibility is weighted by the $\eta \frac{V_p}{V_{p^*}}$ term, $\eta > 0$, in order to scale the drift rates of all patches as a function of the prior value of the current one. As a consequence, the average accumulation rate is reduced when visiting valuable patches (hence producing higher residence times); it is increased when visiting poorer ones that will be given up earlier. Clearly, the exponential term implies higher drift rates for the currently visited patches since promoting the nearest sites, including the current one. This entails high probability for the current patch to be chosen again. Meanwhile, in order to avoid the process being stuck to the current patch, the function Ψ (Equation 4) decreases the drift rate of the visited patch exponentially in time. The drift rates of most valuable patches will be affected by a slower decrease, thus allowing for longer patch exploration.

Coming back to Equation (1), $q_p(t)$ grows at the rate $I_p(t)$ on average, but also diffuses due to the accumulation of noise. A decision is made as soon as the random walk of one among the $q_p(t)$ variables crosses a barrier a. This is accounted for by the decision equation

$$s_{p,t} = H(q_p(t) - a), \qquad p = 1, \cdots, N_P, \tag{6}$$

where H is the Heaviside function and $s_{p,t}$ denotes the response function related to patch p, clearly, a piece-wise constant function

admitting only the values 0 and 1. Race termination occurs as any $q_p(t)$ reaches the decision criterion, that is $s_{p,t} = 1$. Then, the choice of the motion regime or scale (i.e., local vs. global) accounted for by s_t, and that of the attractor indexed by p_t^* can be written

$$s_t = s_{p,t}, \qquad p_t^* = p. \tag{7}$$

When $p_t^* \neq p_{t-1}^*$, that is the chosen patch is different from the previous one, a between-patch relocation occurs, and $s_t = 1$ until the new patch is reached (bottom panels of **Figure 1**); otherwise, $(p_t^* = p_{t-1}^*)$, s_t is set to 0 and the exploration of the current patch is resumed.

2.1.2. Spatial Dynamics

Given the state (s_t, p_t^*), and following Boccignone et al. (2020), the spatial dynamics of gaze is obtained by evolving the FOA position $\mathbf{r}_F(t)$ over time through the state-dependent stochastic differential equation that defines the 2D O-U process

$$d\mathbf{r}_F(t) = \mathbf{B}_{p^*}^{(s_t)}[\boldsymbol{\mu}_{p^*}^{(s_t)} - \mathbf{r}_F(t)]dt + \mathbf{D}_{p^*}^{(s_t)}(\mathbf{r}_F(t))d\mathbf{W}^{(s_t)}(t). \tag{8}$$

This generates a mean reverting trajectory, $\boldsymbol{\mu}_{p^*}^{(s_t)}$ being the attractor location (center of mass of the selected patch). Clearly, when $s_t = 1$ the attractor serves as the target of a large scale gaze relocation; when $s_t = 0$, the attractor constrains local patch exploitation. Examples of the O-U outcome are displayed in the bottom panels of **Figure 1**.

In Equation (8), the 2×2 matrix $\mathbf{B}_{p^*}^{(s_t)}$ controls the strength of attraction (drift) of $\mathbf{r}_F$ toward the location $\boldsymbol{\mu}$; $\mathbf{D}_{p^*}^{(s_t)}$ is a 2×2 matrix representing the diffusion parameter of the 2D Brownian motion $\mathbf{W}(t)$. Precisely, for the 2D mean-reverting O-U process, $\mathbf{B}_{p^*}^{(s_t)} = (b_{x,p^*}^{(s_t)}, b_{y,p^*}^{(s_t)})^T$, $\mathbf{D}_{p^*}^{(s_t)} = (\sigma^{(s_t)})^2\mathbb{I}$, with $\mathbf{W} = (W_x, W_y)^T$ denoting independent Brownian processes. Equation (8) can be integrated so that the evolution in time of $\mathbf{r}_F(t) = (x_F(t), y_F(t))$ between 0 and t is computed by numerically advancing the gaze position through the update equation from t to $t' = t + \delta t$, i.e., δt time units later, and initial condition $x_0 = x_F(t)$:

$$\begin{aligned} x_F(t') &= x_F(t)e^{-b_{x,p^*}^{(s_t)}\delta t} + \mu_x(1 - e^{-b_{x,p^*}^{(s_t)}\delta t}) \\ &\quad + \sqrt{\gamma_x(1 - e^{-2b_{x,p^*}^{(s_t)}\delta t})}z(t) \\ y_F(t') &= y_F(t)e^{-b_{y,p^*}^{(s_t)}\delta t} + \mu_y(1 - e^{-b_{y,p^*}^{(s_t)}\delta t}) \\ &\quad + \sqrt{\gamma_y(1 - e^{-2b_{y,p^*}^{(s_t)}\delta t})}z(t) \end{aligned} \tag{9}$$

with $z \sim \mathcal{N}(0, 1)$. As to the O-U parameters, the drift terms $b_{x,p}^{(s_t)}$ and $b_{y,p}^{(s_t)}$ are set proportional to the width of the patch p if $s_t = 0$, or proportional to the distance from the target patch, otherwise. The diffusion terms are $\gamma_x^{(s_t)} = \frac{\sigma^{(s_t)}}{b_{x,p^*}^{(s_t)}}, \gamma_y^{(s_t)} = \frac{\sigma^{(s_t)}}{b_{y,p^*}^{(s_t)}}$ with $\sigma^{(s_t)}$ proportional to the average distance between patches if $s_t = 1$; equal to 1, otherwise.

2.2. Experimental Set-Up

Our experimental set-up can be recapped as follows:

As to stimuli and eye tracking data we use a large publicly available dataset (Xu et al., 2018), which is influential in current research on computational modeling of attention (Borji, 2021). We evaluate the proposed model (from now on, `Proposed`) by straightforward comparison to the `GazeDeploy` model (Boccignone et al., 2020). The main goal is the assessment of the effectiveness and the computational efficiency of the novel decision making procedure. For what concerns confronting with other models, only a few have been proposed that are experimentally at the ready for actual simulation of gaze deployment, i.e., with the capability of handling time-varying scenes and the availability of a software implementation (e.g., Boccignone and Ferraro, 2014; Zanca et al., 2020). For the sake of completeness, full evaluation with respect to these models and their variants is reported in the **Supplementary Table 1** and **Supplementary Figure 7**.

The evaluation protocol involves the simulation of both models to generate gaze trajectories. These are then quantitatively compared with data from human observers via the ScanMatch (Cristino et al., 2010) and the MultiMatch (Jarodzka et al., 2010; Dewhurst et al., 2012) metrics. Details are given in the sections below.

2.2.1. Stimuli and Eye Tracking Data

The adopted dataset (Xu et al., 2018) consists of 65 one-shot conversation scenes from YouTube and Youku, involving 1–27 different faces for each scene. The duration of the videos is cut down to be around 20 s, with a resolution of $1,280 \times 720$ pixels. The dataset includes eye tracking recordings from 39 different participants (26 males and 13 females, aging from 20 to 49, with either corrected or uncorrected normal eyesight), who were not aware of the purpose of the experiment. A 23-inch LCD screen was used to display the test videos at their original resolution. Eye tracking was carried out using a Tobii X2-60 eye tracker at 60 Hz. All subjects were required to sit on a comfortable chair with a viewing distance of about 60 cm from the LCD screen; no chin rest was used. Before viewing videos, each subject was required to perform a 9-point calibration for the eye tracker. The subjects were asked to free-view videos displayed at random order. The 65 test videos were divided into three sessions, and there was a 5-min rest after viewing each session to avoid eye fatigue. Moreover, a 10-s blank period with black screen was inserted between two successive videos for a short rest. Event classification into saccades and fixations with relative duration was performed via eye tracker embedded algorithms with default settings. Eventually, $1,011,647$ fixations in total were retained.

A caveat concerns the lack of full data quality reporting compliant with the criteria discussed by Holmqvist et al. (2012), considering the high level of noise (low precision) of the Tobii X2-60 eye tracker. On the other hand, this issue is in our case mitigated by the fact that when performing within-patch analysis, we are mostly interested in a phenomenological description of local gaze dynamics. Clearly, this would have been a serious impediment, if we had recursively applied our method to scrutinize specific items within the patch (e.g., the

eyes for gauging gaze direction, or other facial cues for expression recognition). In foraging terms (Stephens, 1986), such recursion would account for prey choice and handling. However, this goal was out of the scope of the present investigation.

2.2.2. Evaluation Protocol

We compare the scan paths simulated from a number of model-based, "artificial" observers with those recorded from human observers (the `Real` model). The rationale is to assess whether simulated behaviors are characterized by statistical properties that are significantly close to those featured by human subjects eye tracked while watching conversational videos. Put simply, any model can be considered adequate if model-generated scan paths mimic those generated by human observers (which we regard as samples of the `Real` model) while gazing at the same audio-visual stimuli.

As to the evaluation metrics, we adopt the ScanMatch (Cristino et al., 2010) and the MultiMatch (Jarodzka et al., 2010; Dewhurst et al., 2012) methods. ScanMatch (SM) is apt to provide an overall performance summary, whilst MultiMatch (MM) specifically addresses the many dimensions of gaze dynamics. The SM and MM metrics are computed on scan paths, that is a sequence of fixations and saccades. The `Proposed` and the `GazeDeploy` models generate continuous gaze trajectories that can be assimilated to *raw data* produced by eye trackers. Yet, the exploration and exploitation dynamics can be thought of as following a "saccade and fixate" strategy (Land, 2006). Further, the conversational stimuli we are using result in limited motion of patches, mostly due to head turning and hand gestures. Then, to classify fixation and saccade events in model generated trajectories we adopt, from a data analysis perspective, a functional definition of such events (Hessels et al., 2018). We consider a fixation as a period of time during which a static or a moderately displacing part of the visual stimulus (the patch) on the screen is gazed at and that in a human observer would be projected to a relatively constant location on the retina. This corresponds to local dynamics in the exploitation stage. Accordingly, saccades are the gaze shifts for redirecting the line of sight to a new patch of interest, as performed along the exploration stage. This is operationalized using the NSLR-HMM algorithm (Pekkanen and Lappi, 2017) with default settings; the original implementation is available from online repository (cfr., **Supplementary Material**, Computer Code). The algorithm classifies fixations, saccades, smooth pursuits, and post-saccadic oscillations. To serve our purposes, smooth pursuits were retained as fixations.

In detail, SM divides a scan path spatially and temporally into several bins and then codes it to form a sequence of letters. The frame width was divided into 14 bins, while the height was split in eight bins; the temporal bin size was set to 50 ms. Two scan paths are thus encoded to two strings that are compared by maximizing the similarity score. This metric indicates the joint spatial, temporal and sequential similarity between two scan paths, higher SM score denoting a better matching. Complementary, the MM metrics computes five distinct measures that capture the different scan path features: shape, direction, length, position, and duration. Higher score of each metric means better matching. The MM algorithm allows for scan paths sequences to be simplified in order to reduce their complexity. This is carried out by grouping together saccades of angular or amplitude differences below some predefined thresholds. Likewise, fixations are grouped if their duration is shorter than a duration threshold. In the adopted evaluation protocol no simplification was performed (i.e., no use of the direction, length, and duration thresholds), as even small differences in scan paths performed on a dynamic stimuli can correspond to major differences in the attended scene.

The evaluation protocol runs as follows: assume a number N_{obs} of human observers. Then, for each video in the test set: (1) compute SM and MM similarity scores for each possible pair of the N_{obs} observers (`Real` vs. `Real`); (2) for each model: (2.a) generate gaze trajectories from artificial observers; (2.b) parse/classify trajectories into scan paths (saccades and fixations with the relative duration) via the NSLR-HMM algorithm (Pekkanen and Lappi, 2017); (2.c) compute SM and MM scores for each possible pair of real and N_{obs} artificial scan paths (`Real` vs. `Model`). Eventually, (3) return the average SM and MM scores for `Real` vs. `Real` and `Real` vs. `Model` comparisons.

In what follows we consider each MM dimension to be a stand-alone score. Thus, the analysis uses six different scores: the five obtained from the MM dimensions of shape (MM_{Shape}), direction (MM_{Dir}), length (MM_{Len}), position (MM_{Pos}), and duration (MM_{Dur}), plus the SM score SM.

2.2.3. Simulation Details

The rationale of the simulations was to focus on the performance of the different gaze control strategies of the `Proposed` and of the `GazeDeploy` models. The input provided to either model was the same, namely the patch representation recapped in the **Supplementary Material**, Patch computation. The bottom layers of patch computation (face detection, speaker detection) rely on deep neural network modules that were independently optimized on a different dataset (Boccignone et al., 2019).

In addition, a baseline `Random` model was adopted. This simply generates random gaze shifts by sampling (x, y) fixation coordinates from an isotropic Gaussian distribution located at the center of the scene (center-bias). The Gaussian standard deviation is set proportional to the height of the video frames. The fixation duration is sampled from a uniform distribution ranging from 67 to 1,699 ms corresponding to the 0.01 and 0.99 quantiles of the empirical distribution of real fixations duration.

To optimize on model parameters, ten subjects were randomly sampled out of the 39 participants and their scan paths used to determine the free parameters of the proposed model via a grid search maximizing metric scores according to the procedure described in section 2.2.2. This yielded the optimal values $\phi = 0.18$, $\eta = 5$, $\kappa = 15$, and $a = 1.7$. The same procedure was performed to optimize `GazeDeploy` free parameters, as described in (Boccignone et al., 2020). The remaining 29 subjects were used for evaluation.

The code for the simulation of all models is available in online repositories (cfr., **Supplementary Material**, Computer Code).

3. RESULTS

A demonstration of the output obtained from model simulation is included in the **Supplementary Video 1**. The result is by and large representative of those obtained on the whole dataset.

Overall, the simulated model generates scan paths that mimic human scan paths in terms of spatiotemporal statistics (but see **Supplementary Figure 5** for a concrete example on a single video): the saccade amplitude distributions exhibit a multimodal shape, with short saccades preferred to long ones; fixation duration distributions from both real and simulated data reveal a right-skewed and heavy-tailed shape; *prima facie*, a high similarity can also be noticed between saccade direction distributions of real and simulated data. The same conclusions can be drawn by observing **Supplementary Figure 6**, which reports the same statistics and comparison on the whole dataset.

For visualization purposes, **Figure 2** depicts at a glance the estimated empirical densities of the similarity scores achieved by using the protocol introduced in section 2.2.2. Scores obtained from the `Real` vs. `Real` comparison represent the gold standard. A preliminary, qualitative inspection shows that the `Proposed` model, much like the `GazeDeploy` model, gives rise to empirical densities that are close to those yielded by real subjects. This holds for all dimensions, with the exception of the direction score MM_{Dir}.

For what concerns the efficiency of the two methods, all things being equal as regards the input provided (the ensemble of audio-visual patches $\mathcal{W}(t)$ and the O-U spatial dynamics), the computational cost of the decision making procedures of `Proposed` and `GazeDeploy` amounts to the 0.2 and the 44.6%, on average, of the total computation time, respectively, at frame rate. A summary of the cost profiling is reported in **Supplementary Figure 4**.

As to the quantitative evaluation of the effectiveness of the methods, in the following we adopt well-established statistical tests in order to assess whether or not each model generates scan paths that significantly differ from those of human observers and to gauge the size of such difference (effect size).

3.1. Statistical Analyses

In a nutshell, we are interested in performing a statistical comparison of the performance between multiple models over each video of the adopted dataset. This is the typical repeated measure analysis between multiple groups, for which standard ANOVA is usually performed. The ANOVA test requires populations distributions to be normal and homoscedastic (with the same finite variance). If either normality or homoscedasticity cannot be ensured, non-parametric statistical tests (like the Friedman test) should be employed. In the analyses that follow, the SM metric and each dimension of the MM metric are treated as separate scores. Significance level of all statistical tests is $\alpha = 0.05$.

As to scores MM_{Shape} and MM_{Len}, the Shapiro-Wilk test with Bonferroni correction rejected the null hypothesis of normality as opposed to the SM, MM_{Dir}, MM_{Pos}, and MM_{Dur} scores.

For all scores the null hypothesis of homoscedasticity of distributions was rejected by either Bartlett (in case of normality of distributions) or Levene (non-Gaussian distributions) tests. Hence, the Friedman test with Nemenyi *post-hoc* analysis was performed. The results for each score are depicted in **Figure 3** via the corresponding Critical Difference (CD) diagrams. These provide quantitative support for the preliminary observations offered by the empirical densities in **Figure 2**.

Notably, according to the SM metric and the adopted assessment strategy, the scan paths simulated from the `GazeDeploy` and `Proposed` procedures cannot be distinguished from those of `Real` subjects (this is further demonstrated by the fact that these two models achieve *small* or *negligible* effect sizes, as reported in **Supplementary Table 1**).

The SM score can be conceived as an overall summary of the performance of the considered models. A deeper analysis can be performed by inspecting the individual dimensions provided by the MM metric. One important result is delivered by the MM_{Dur} dimension, summarizing the similarity of fixation duration between aligned scan paths: again, the `Proposed` and `GazeDeploy` models cannot be distinguished from the gold standard (`Real`), exhibiting *negligible* and *small* effect sizes, respectively (see **Supplementary Table 1**).

A similar conduct is exhibited by the MM_{Shape}, MM_{Len}, and MM_{Pos} scores.

The MM_{Dir} score is worthy of mention: in this case, `GazeDeploy` and the `Proposed` procedures perform comparably with the `Random` model. This is probably due to the fact that saccade direction modeling is not addressed by both models, but just absorbed into the gaze shift policy at hand.

4. DISCUSSION

We set out to investigate the modeling of gaze dynamics as exhibited by a perceiver who scrutinizes socially relevant multimodal information. This effort was developed under the framework of foraging behavior.

The work presented here builds upon previous one (Boccignone et al., 2020). However, in that case the cogent problems of patch choice and leave were framed within an optimal Bayesian setting (Bartumeus and Catalan, 2009). Here, in a different vein, we considered a simple multi-alternative perceptual decision making approach. This relies on a race-to-threshold between independent integrators, each integrator being associated with a patch (Bogacz et al., 2006; Ditterich, 2010; Krajbich and Rangel, 2011). In consequence, the eye guidance problem can be parsimoniously formalized in terms of the evolution of the stochastic differential equations (1) and (8) together with the decision equation (6).

The gain in simplicity and computational efficiency does not come to the cost of performance as it might have been expected. The results so far achieved, when inspected under the lens of statistics, show that the proposed method is comparable to the `GazeDeploy` method in terms of either the overall performance, as measured by the SM score, and the specific scores gauged by MM. In particular, the remarkable

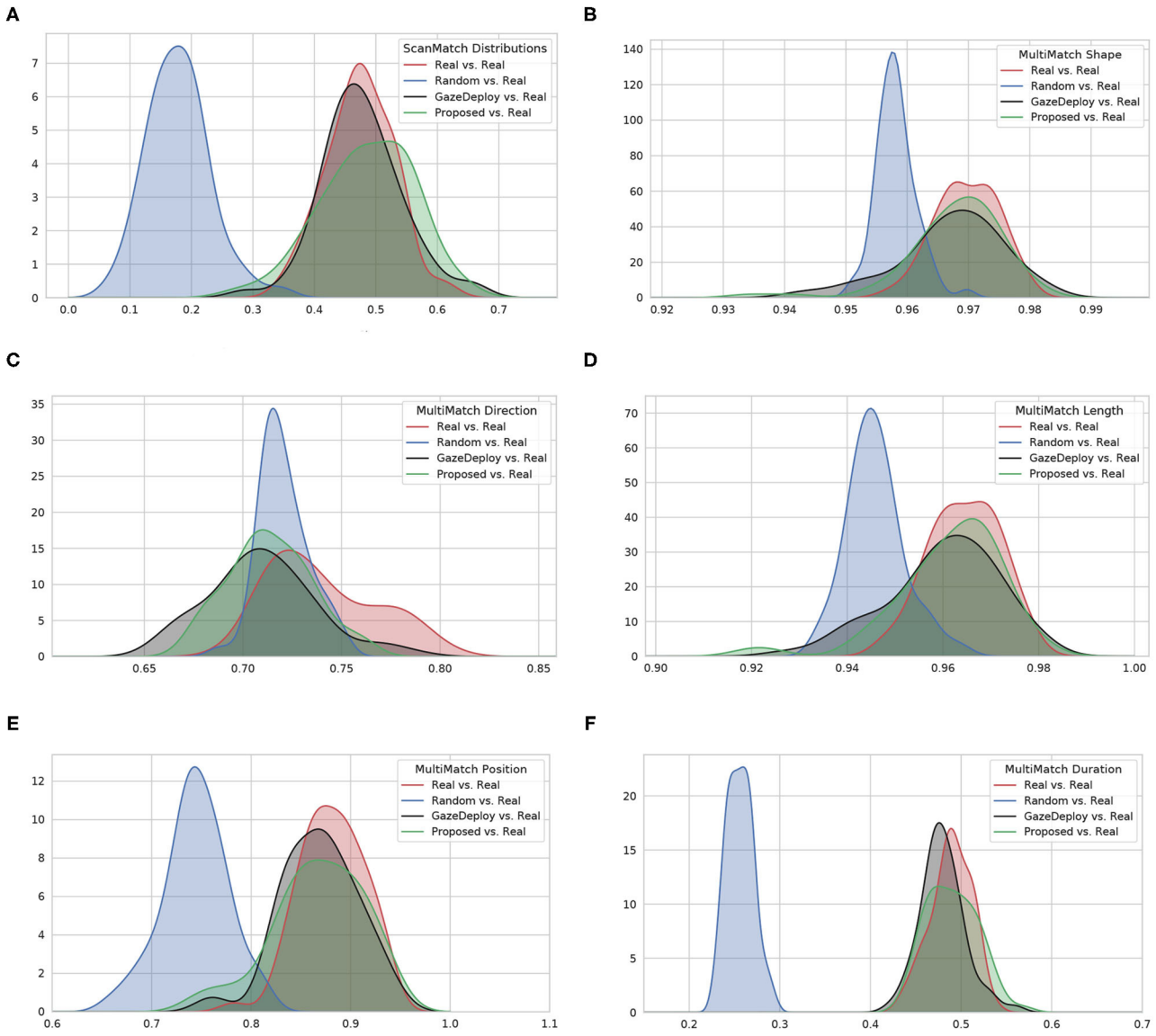


FIGURE 2 | The estimated empirical densities *f*(*score*) for the considered models (via Kernel Density estimation). **(A)** Shows the distributions for the ScanMatch score; **(B–F)** show the distributions related to the five MultiMatch dimensions.

result obtained by `GazeDeploy` for what concerns fixation duration—which in that case was related to the MVT modeling of the giving-up—is also replicated by this simpler method. Thus, a question arises in regard to the relations, if any, between the two models. A thorough discussion of this point would carry us deep into establishing formal connections between the methods, that is out of the scope of this brief research report. A few considerations must here suffice.

Optimal foraging theory, markedly the MVT and its stochastic extension, provides general rules for when an animal should leave a patch. This lays the theoretical foundation for assessing optimal decision-making, though lacking mechanistic explanation. It has been shown under appropriate conditions (Davidson and El Hady, 2019) that optimal foraging relying on patch-leaving decisions can be connected to a stochastic evidence accumulation model of foraging, namely a drift-diffusion model (DDM, Ratcliff et al., 2016). This describes the process through which an animal gathers information to make decisions. The DDM can be solved for conditions where foraging decisions are optimal and equivalent to the MVT (Davidson and El Hady, 2019). Notably, the DDM can be extended to a multi-alternative DDM (Bogacz et al., 2006). The latter, for instance, has been applied to eye tracking experiments involving multiple choice in value-based decision (Krajbich and Rangel, 2011). The continuous-time independent race integrators that we used here, should be considered as a theoretically sub-optimal solution; yet,

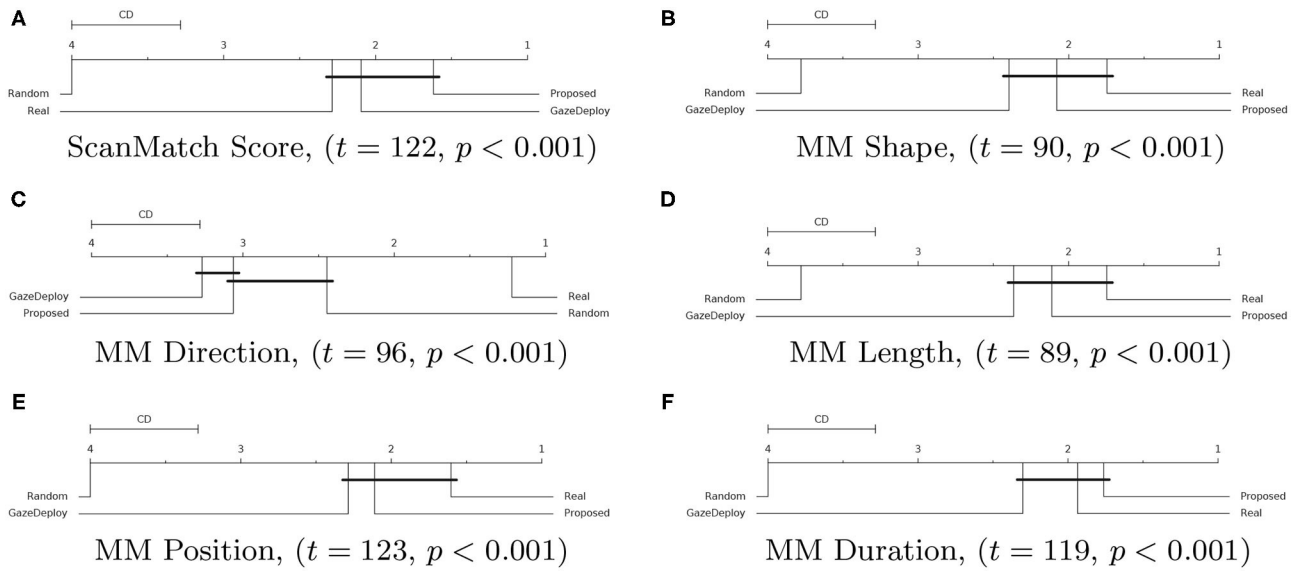


FIGURE 3 | Critical Difference (CD) diagrams of the *post-hoc* Nemenyi test ($\alpha = 0.05$) for the ScanMatch **(A)** and MultiMatch scores **(B–F)** when comparing the proposed model with the GazeDeploy procedure, the gold standard and a baseline random model. Diagrams can be read as follows: the difference between two models is significant if the difference in their ranks is larger than the CD. Models that are not significantly different from one another are connected by a black CD line. Friedman's test statistic (t) and p-value (p) are reported in brackets.

according to results we gathered so far, it qualifies as a viable solution for trading down complexity of the full MVT approach. Overall, differently from optimal foraging theory, DDMs and generalizations (Bogacz et al., 2006; Ratcliff et al., 2016) provide a mechanistic framework suitable to unravel behavioral and neural underpinnings of value-based decision making. Interestingly enough, stochastic race accumulator have been proposed to model neural activity for action selection in the pre-motor areas (Ognibene et al., 2006). Also, from a neurobiological standpoint, a body of evidence suggests the firing properties of neurons that are likely to drive decisions in the LIP and the FEF are well-described by stochastic accumulator models (Gold and Shadlen, 2007).

The Langevin-type equation formalizing evidence accumulation is entangled with the 2D spatial Langevin-type equation (O-U process) accounting for the two different scales of landscape exploration and of local patch exploitation. On the one hand this succinctly permits the use of one and only dynamics of oculomotor behavior in the vein of current literature suggesting that visual fixation is functionally equivalent to visual exploration on a spatially focused scale (the functional continuum hypothesis, Otero-Millan et al., 2013). On the other hand, the strict interplay between the evidence accumulation equation and the 2-D multiscale gaze shift equation puts forward the present study for having a special bearing on current proposals in computational models that address the focal and ambient dichotomy and the relation between saccade amplitude and fixation duration (Le Meur and Fons, 2020). This issue was well-known in the eye tracking literature (Unema et al., 2007) but overlooked in the computational modeling of visual attention.

Beyond the merit of the above theoretical aspects, the model bears on potential applications for researchers interested in social gaze. Our approach allows for operationalizing the effect of social information on gaze allocation in terms of both decision making and value attributed to different kinds of gaze attractors. Meanwhile, it takes into account spatial tendencies in the unfolding of gaze trajectories. The basic foraging dimensions of value-based patch selection and patch handling over time pave the way for analysing in a principled framework social gaze as related to persons' intentions, feelings, traits, and expertise by exploiting semantically rich multimodal dynamic scenes. Video stimuli are clearly advantageous when investigating social attention compared to static stimuli (Risko et al., 2012). Complex, dynamic and contextually rich video clips elicit more natural and representative viewing behavior in participants, even though it might deviate from that found in everyday situations (Risko et al., 2012; Hessels, 2020). In a sense, this experimental arrangement should provide a better approximation to a "real world" social dynamic context, thus bearing higher ecological validity. However, the latter is a problematic claim (one good place to look for further reflection on these matters is Holleman et al., 2020). In what follows, we shall limit our discussion to particular contexts of social robotics. Yet, in general, our model and set-up can be useful for investigating social attention under a variety of circumstances, such as in clinical populations as discussed in the Introduction.

The computational efficiency of the method shows promise for application in robotics, markedly in social robotics, where active vision plays an important role and where social robot's sensitivity to environmental information and the ability to localize the people around itself is crucial (Admoni and

Scassellati, 2017; Wiese et al., 2017; Zhang et al., 2020). Social robots need to gather information about their human fellows to facilitate mutual understanding and coordination (Zhang et al., 2020). Designing robot gaze itself is challenging and difficult to standardize due to the variations in physical robots and human participants, while burdened with architectural constraints. Early research efforts (Breazeal et al., 2001) relied on simple saliency-based schemes (Itti et al., 1998) inherited from computer vision (Shic and Scassellati, 2007; Ferreira and Dias, 2014); in the last decade these have been reshaped in the form of deep neural nets, such as convolutional networks (Zhang et al., 2020). Yet, the aptness of accounting for task, value and context in the visuo-motor loop is crucial. In this perspective, it is acknowledged that socially interactive robots would greatly benefit from the development of probabilistic real-time frameworks that implement automatic attention mechanisms (Ferreira and Dias, 2014). For instance, in a recent work (Rasouli et al., 2020), active visual behavior has been grounded in the probability of gazing at a location that accounts for an empirical exploitation/exploration trade-off; here, the same issue is set but in a principled framework. Also, the stochasticity, which is inherent to our approach, has proved to be strategic. It has been reported (Martinez et al., 2008) that a stochastic gaze control mechanism enables the i-Cub robot to explore its environment up to three times faster compared to the standard winner-take-all mechanism (Itti et al., 1998). Indeed, stochasticity makes the robot sensitive to new signals and flexibly change its attention. This, in turn, enables efficient exploration of the environment as the basis for action learning along interactive tasks (Nagai, 2009a,b). Further, the proposed method is suitable to be implemented in both overt and covert gaze action selection and generation (Rea et al., 2014). Results achieved here in a multimodal conversational setting are likely to be relevant in everyday multimodal settings where the robot is requested to gaze at people around (Zibafar et al., 2019). Clearly, in a real world context the bottom layer of patch computation should efficiently embed suitable methods that have been applied for speaker localization in the field of humanoid robotics (e.g., Zibafar et al., 2019; Rea et al., 2020).

This study has several caveats. For instance, statistical analyses have highlighted problems in gaze direction modeling. This is a difficult hurdle to face. Some contextual rules have been proposed in the computer vision field (Torralba et al., 2006) and in the psychological literature (Tatler and Vincent, 2008). However, these might be put into question out of the lab and in dynamic environments. One solution could be that of a data-driven strategy (Le Meur and Coutrot, 2016; Hu et al., 2020), albeit raising in turn the problem of generalizability. Further, the accumulator model lacks of a detailed account for the actual handling of within-patch items (i.e., what would be considered "prey handling" in the animal ecology field). One example is the processing of components of facial expression and gaze of people involved in the interaction. Here, the bare phenomenological account that we have presented forgoes processing details. Nevertheless, different policies of deploying gaze to specific items in facial expressions might also affect emotional responses (Schomaker et al., 2017; Rubo and Gamer, 2018). These aspects need to be further investigated.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

AD'A: study and model design, software implementation, experiments, statistical analyses, and manuscript writing. GB: study and model design, statistical analyses, and manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported in part by University of Milano under Grant PSR 2019, Sensing and foraging of information for resource allocation in spatial and social contexts: statistical models and optimization.

ACKNOWLEDGMENTS

AD'A thanks Prof. Tom Foulsham for enlightening discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.639999/full#supplementary-material

Supplementary Video 1 | The supplementary video shows a simulation of the proposed model on a video clip. A visual demonstration of both the decision making dynamics and spatial dynamics is provided.

Supplementary Table 1 | Central tendencies for each score and model computed as mean (M) or median (MED) with associated dispersion metrics (standard deviation, SD or median absolute deviation, MAD. Effect sizes are computed as the Cohen's d or the Cliff's between the given model and real subjects.

Supplementary Figure 1 | A sketch of the patch computation procedure from the audio/visual input.

Supplementary Figure 2 | The prediction by MVT is that a poor patch should be abandoned earlier than a rich patch. The time axis starts with a travel time with no energy gain after which the forager finds a patch. The shapes of the red and black gain curves, arising from resource exploitation, represent the cumulative rewards of a "rich" and a "poor" patch, respectively. For each curve, the osculation point of the tangent defines the optimal patch residence time (adapted from Boccignone et al., 2020).

Supplementary Figure 3 | Overall description of the switching behavior. The first block depicts the typical trend of the instantaneous reward rate for two types of patches (rich and poor). These can be conceived as Giving Up Time (GUT) functions; as time goes by, the GUT function approaches the quality threshold Q, the run being faster for poorer patches. At any time step the decision stay/go is taken by sampling a Bernoulli RV (third block) whose parameter is given by the distance between the GUT function and the quality threshold at that time (opportunely scaled by a logistic function, c.f.r. second block).

Supplementary Figure 4 | Time profiling. **(a)** Time required (seconds) by the modules composing the Proposed and GazeDeploy method for the analysis and simulation on a single video frame (results reported on a logarithmic scale). **(b)** Percentage of computation time required by Pre-attentive modules (Face/Speaker Detection, Spatio-Temporal Saliency and patch computation) and actual Gaze Deployment (Decision Making and Spatial Dynamics) for the GazeDeploy procedure. **(c)** Comparison of time requirements between the GazeDeploy and Proposed procedures in relation to those of the Pre-attentive modules. **(d)** Percentage of computation time required by Pre-attentive modules (Face/Speaker Detection, Spatio- Temporal Saliency and patch computation) and actual Gaze Deployment (Decision Making and Spatial Dynamics) for the Proposed procedure.

Supplementary Figure 5 | (a) Frame of video 008 with overlaid heatmap of real fixations. **(b)** Frame of video 008 with overlaid heatmap of generated fixations. **(c)** Real (red) and Generated (blue) saccades amplitude distribution. **(d)** Real (red) and Generated (blue) fixations duration distribution. **(e)** Real saccades direction distribution. **(f)** Generated saccades direction distribution.

Supplementary Figure 6 | (a) Heatmap of real fixations on the whole dataset **(b)** Heatmap of generated fixations on the whole dataset **(c)** Saccades amplitude distribution on the whole dataset for Real (red) and Generated (blue) scanpaths **(d)** Fixations duration distribution on the whole dataset for Real (red) and Generated (blue) scanpaths **(e)** Real saccades direction distribution on the whole dataset **(f)** Generated saccades direction distribution on the whole dataset.

Supplementary Figure 7 | Critical Difference (CD) diagrams of the *post-hoc* Nemenyi test ($\alpha = 0{:}05$) for the ScanMatch and MultiMatch scores when comparing different models proposed in literature plus the gold standard and a baseline random model. Friedman's test statistic (t) and p-value (p) are reported in brackets.

REFERENCES

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

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. (1988). Active vision. *Int. J. Comput. Vis*. 1, 333–356. doi: 10.1007/BF00133571

Bajcsy, R., and Campos, M. (1992). Active and exploratory perception. *CVGIP Image Understand*. 56, 31–40. doi: 10.1016/1049-9660(92)90083-F

Ballard, D. (1991). Animate vision. *Artif. Intell*. 48, 57–86. doi: 10.1016/0004-3702(91)90080-4

Bartumeus, F., and Catalan, J. (2009). Optimal search behavior and classic foraging theory. *J. Phys. A Math. Theor*. 42:434002. doi: 10.1088/1751-8113/42/43/434002

Berridge, K. C., and Robinson, T. E. (2003). Parsing reward. *Trends Neurosci*. 26, 507–513. doi: 10.1016/S0166-2236(03)00233-9

Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2019). "Give ear to my face: modelling multimodal attention to social interactions," in *Computer Vision-ECCV 2018 Workshops*, eds L. Leal-Taixé and S. Roth (Cham: Springer International Publishing), 331–345. doi: 10.1007/978-3-030-11012-3_27

Boccignone, G., Cuculo, V., D'Amelio, A., Grossi, G., and Lanzarotti, R. (2020). On gaze deployment to audio-visual cues of social interactions. *IEEE Access* 8, 161630–161654. doi: 10.1109/ACCESS.2020.3021211

Boccignone, G., and Ferraro, M. (2014). Ecological sampling of gaze shifts. *IEEE Trans. Cybernet*. 44, 266–279. doi: 10.1109/TCYB.2013.2253460

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., and Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev*. 113:700. doi: 10.1037/0033-295X.113.4.700

Borji, A. (2021). Saliency prediction in the deep learning era: successes and limitations. *IEEE Trans. Pattern Anal. Mach. Intell*. 43, 679–700. doi: 10.1109/TPAMI.2019.2935715

Borji, A., and Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell*. 35, 185–207. doi: 10.1109/TPAMI.2012.89

Breazeal, C., Edsinger, A., Fitzpatrick, P., and Scassellati, B. (2001). Active vision for sociable robots. *IEEE Trans. Syst. Man Cybernet. A Syst. Hum*. 31, 443–453. doi: 10.1109/3468.952718

Cain, M. S., Vul, E., Clark, K., and Mitroff, S. R. (2012). A bayesian optimal foraging model of human visual search. *Psychol. Sci*. 23, 1047–1054. doi: 10.1177/0956797612440460

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theor. Popul. Biol*. 9, 129–136. doi: 10.1016/0040-5809(76)90040-X

Cristino, F., Mathôt, S., Theeuwes, J., and Gilchrist, I. D. (2010). Scanmatch: a novel method for comparing fixation sequences. *Behav. Res. Methods* 42, 692–700. doi: 10.3758/BRM.42.3.692

Davidson, J. D., and El Hady, A. (2019). Foraging as an evidence accumulation process. *PLoS Comput. Biol*. 15:e1007060. doi: 10.1371/journal.pcbi.1007060

Dewhurst, R., Nyström, M., Jarodzka, H., Foulsham, T., Johansson, R., and Holmqvist, K. (2012). It depends on how you look at it: scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behav. Res. Methods* 44, 1079–1100. doi: 10.3758/s13428-012-0212-2

Ditterich, J. (2010). A comparison between mechanisms of multi-alternative perceptual decision making: ability to explain human behavior, predictions for neurophysiology, and relationship with decision theory. *Front. Neurosci*. 4:184. doi: 10.3389/fnins.2010.00184

Ehinger, K. A., and Wolfe, J. M. (2016). When is it time to move to the next map? Optimal foraging in guided visual search. *Attent. Percept. Psychophys*. 78, 2135–2151. doi: 10.3758/s13414-016-1128-1

Ferreira, J. F., and Dias, J. (2014). Attentional mechanisms for socially interactive robots-a survey. *IEEE Trans. Auton. Ment. Dev*. 6, 110–125. doi: 10.1109/TAMD.2014.2303072

Foulsham, T. (2019). "Scenes, saliency maps and scanpaths," in *Eye Movement Research*, eds C. Klein and U. Ettinger (Cham: Springer), 197–238. doi: 10.1007/978-3-030-20085-5_6

Foulsham, T., Cheng, J. T., Tracy, J. L., Henrich, J., and Kingstone, A. (2010). Gaze allocation in a dynamic situation: effects of social status and speaking. *Cognition* 117, 319–331. doi: 10.1016/j.cognition.2010.09.003

Gold, J. I., and Shadlen, M. N. (2007). The neural basis of decision making. *Annu. Rev. Neurosci*. 30, 535–574. doi: 10.1146/annurev.neuro.29.051605.113038

Grossman, R. B., Mertens, J., and Zane, E. (2019). Perceptions of self and other: social judgments and gaze patterns to videos of adolescents with and without autism spectrum disorder. *Autism* 23, 846–857. doi: 10.1177/1362361318788071

Guy, N., Azulay, H., Kardosh, R., Weiss, Y., Hassin, R. R., Israel, S., et al. (2019). A novel perceptual trait: gaze predilection for faces during visual exploration. *Sci. Rep*. 9:10714. doi: 10.1038/s41598-019-47110-x

Hessels, R. S. (2020). How does gaze to faces support face-to-face interaction? A review and perspective. *Psychon. Bull. Rev*. 27, 856–881. doi: 10.3758/s13423-020-01715-w

Hessels, R. S., Niehorster, D. C., Nyström, M., Andersson, R., and Hooge, I. T. C. (2018). Is the eye-movement field confused about fixations and saccades? A survey among 124 researchers. *R. Soc. Open Sci*. 5:180502. doi: 10.1098/rsos.180502

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cogn. Sci*. 30, 3–41. doi: 10.1207/s15516709cog0000_50

Holleman, G. A., Hooge, I. T., Kemner, C., and Hessels, R. S. (2020). The 'real-world approach' and its problems: a critique of the term ecological validity. *Front. Psychol*. 11:721. doi: 10.3389/fpsyg.2020.00721

Holmqvist, K., Nyström, M., and Mulvey, F. (2012). "Eye tracker data quality: what it is and how to measure it," in *Proceedings of the Symposium on Eye Tracking Research and Applications* (New York, NY), 45–52. doi: 10.1145/2168556.2168563

Hu, Z., Li, S., Zhang, C., Yi, K., Wang, G., and Manocha, D. (2020). DGaze: CNN-based gaze prediction in dynamic scenes. *IEEE Trans. Vis. Comput. Graph*. 26, 1902–1911. doi: 10.1109/TVCG.2020.2973473

Ioannou, C., Seernani, D., Stefanou, M. E., Biscaldi-Schaefer, M., Van Elst, L. T., Fleischhaker, C., et al. (2020). Social visual perception under the eye of bayesian theories in autism spectrum disorder using advanced modeling of spatial and temporal parameters. *Front. Psychiatry* 11:585149. doi: 10.3389/fpsyt.2020.585149

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259. doi: 10.1109/34.730558

Jarodzka, H., Holmqvist, K., and Nyström, M. (2010). "A vector-based, multidimensional scanpath similarity measure," in *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)* (New York, NY: ACM), 211–218. doi: 10.1145/1743666.1743718

Jording, M., Engemann, D., Eckert, H., Bente, G., and Vogeley, K. (2019). Distinguishing social from private intentions through the passive observation of gaze cues. *Front. Hum. Neurosci.* 13:442. doi: 10.3389/fnhum.2019.00442

Klein, C., Seernani, D., Ioannou, C., Schulz-Zhecheva, Y., Biscaldi, M., and Kavšek, M. (2019). "Typical and atypical development of eye movements," in *Eye Movement Research*, eds C. Klein and U. Ettinger (Cham: Springer), 635–701. doi: 10.1007/978-3-030-20085-5_15

Kloeden, P. E., and Platen, E. (2013). *Numerical Solution of Stochastic Differential Equations*, Vol. 23. Berlin: Springer Science & Business Media.

Korda, A. I., Koliaraki, M., Asvestas, P. A., Matsopoulos, G. K., Ventouras, E. M., Ktonas, P. Y., et al. (2016). Discrete states of attention during active visual fixation revealed by markovian analysis of the time series of intrusive saccades. *Neuroscience* 339, 385–395. doi: 10.1016/j.neuroscience.2016.10.012

Krajbich, I., and Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proc. Natl. Acad. Sci. U.S.A.* 108, 13852–13857. doi: 10.1073/pnas.1101328108

Kustov, A., and Robinson, D. (1996). Shared neural control of attentional shifts and eye movements. *Nature* 384:74. doi: 10.1038/384074a0

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

Le Meur, O., and Coutrot, A. (2016). Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision Res.* 121, 72–84. doi: 10.1016/j.visres.2016.01.005

Le Meur, O., and Fons, P.-A. (2020). "Predicting image influence on visual saliency distribution: the focal and ambient dichotomy," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Short Papers* (New York, NY: Association for Computing Machinery). doi: 10.1145/3379156.3391362

Le Meur, O., and Liu, Z. (2015). Saccadic model of eye movements for free-viewing condition. *Vision Res.* 116, 152–164. doi: 10.1016/j.visres.2014.12.026

Lemons, D. S. (2002). *An Introduction to Stochastic Processes in Physics*. Baltimore, MD: JHU Press.

MacArthur, R. H., and Pianka, E. R. (1966). On optimal use of a patchy environment. *Am. Nat.* 100, 603–609. doi: 10.1086/282454

Martinez, H., Lungarella, M., and Pfeifer, R. (2008). *Stochastic Extension to the Attention-Selection System for the iCub*. University of Zurich, Technical Report.

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theor. Popul. Biol.* 21, 269–288. doi: 10.1016/0040-5809(82)90018-1

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

Nagai, Y. (2009a). "From bottom-up visual attention to robot action learning," in *Proceedings of 8 IEEE International Conference on Development and Learning* (Los Alamitos, CA: IEEE Press), 1–6. doi: 10.1109/DEVLRN.2009.5175517

Nagai, Y. (2009b). "Stability and sensitivity of bottom-up visual attention for dynamic scene analysis," in *Proceedings of the 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems* (St. Louis, MO: IEEE Press), 5198–5203. doi: 10.1109/IROS.2009.5354466

Ognibene, D., Mannella, F., Pezzulo, G., and Baldassarre, G. (2006). "Integrating reinforcement-learning, accumulator models, and motor-primitives to study action selection and reaching in monkeys," in *Proceedings of the 7th International Conference on Cognitive Modelling-ICCM06* (Trieste), 214–219.

Otero-Millan, J., Macknik, S. L., Langston, R. E., and Martinez-Conde, S. (2013). An oculomotor continuum from exploration to fixation. *Proc. Natl. Acad. Sci. U.S.A.* 110, 6175–6180. doi: 10.1073/pnas.1222715110

Pekkanen, J., and Lappi, O. (2017). A new and general approach to signal denoising and eye movement classification based on segmented linear regression. *Sci. Rep.* 7:17726. doi: 10.1038/s41598-017-17983-x

Pirolli, P. (2007). *Information Foraging Theory: Adaptive Interaction With Information*. New York, NY: Oxford University Press.

Rasouli, A., Lanillos, P., Cheng, G., and Tsotsos, J. K. (2020). Attention-based active visual search for mobile robots. *Auton. Robots* 44, 131–146. doi: 10.1007/s10514-019-09882-z

Ratcliff, R., Smith, P. L., Brown, S. D., and McKoon, G. (2016). Diffusion decision model: current issues and history. *Trends Cogn. Sci.* 20, 260–281. doi: 10.1016/j.tics.2016.01.007

Rea, F., Kothig, A., Grasse, L., and Tata, M. (2020). Speech envelope dynamics for noise-robust auditory scene analysis in robotics. *Int. J. Hum. Robot* (Madrid). 17:2050023. doi: 10.1142/S0219843620500231

Rea, F., Sandini, G., and Metta, G. (2014). "Motor biases in visual attention for a humanoid robot," in *2014 IEEE-RAS International Conference on Humanoid Robots* (IEEE), 779–786. doi: 10.1109/HUMANOIDS.2014.7041452

Risko, E., Laidlaw, K., Freeth, M., Foulsham, T., and Kingstone, A. (2012). Social attention with real versus reel stimuli: toward an empirical approach to concerns about ecological validity. *Front. Hum. Neurosci.* 6:143. doi: 10.3389/fnhum.2012.00143

Rubo, M., and Gamer, M. (2018). Social content and emotional valence modulate gaze fixations in dynamic scenes. *Sci. Rep.* 8:3804. doi: 10.1038/s41598-018-22127-w

Schomaker, J., Walper, D., Wittmann, B. C., and Einhäuser, W. (2017). Attention in natural scenes: affective-motivational factors guide gaze independently of visual salience. *Vision Res.* 133, 161–175. doi: 10.1016/j.visres.2017.02.003

Shepherd, S. V., and Platt, M. L. (2007). Spontaneous social orienting and gaze following in ringtailed lemurs (lemur catta). *Anim. Cogn.* 11:13. doi: 10.1007/s10071-007-0083-6

Shic, F., and Scassellati, B. (2007). A behavioral analysis of computational models of visual attention. *Int. J. Comput. Vis.* 73, 159–177. doi: 10.1007/s11263-006-9784-6

Staab, J. P. (2014). The influence of anxiety on ocular motor control and gaze. *Curr. Opin. Neurol.* 27, 118–124. doi: 10.1097/WCO.0000000000000055

Stephens, D. W. (1986). *Foraging Theory*. Princeton, NJ: Princeton University Press.

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

Tatler, B., and Vincent, B. (2008). Systematic tendencies in scene viewing. *J. Eye Mov. Res.* 2, 1–18. doi: 10.16910/jemr.2.2.5

Tavakoli, H. R., Borji, A., Kannala, J., and Rahtu, E. (2020). "Deep audio-visual saliency: baseline model and data," in *ACM Symposium on Eye Tracking Research and Applications, ETRA '20 Short Papers* (Stuttgart: ACM), 1–5. doi: 10.1145/3379156.3391337

Torralba, A., Oliva, A., Castelhano, M., and Henderson, J. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychol. Rev.* 113:766. doi: 10.1037/0033-295X.113.4.766

Unema, P., Pannasch, S., Joos, M., and Velichkovsky, B. (2007). Time course of information processing during scene perception: the relationship between saccade amplitude and fixation duration. *Visual Cogn.* 12, 473–494. doi: 10.1080/13506280444000409

Vernetti, A., Senju, A., Charman, T., Johnson, M. H., and Gliga, T. (2018). Simulating interaction: using gaze-contingent eye-tracking to measure the reward value of social signals in toddlers with and without autism. *Dev. Cogn. Neurosci.* 29, 21–29. doi: 10.1016/j.dcn.2017.08.004

Wiese, E., Metta, G., and Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Front. Psychol.* 8:1663. doi: 10.3389/fpsyg.2017.01663

Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? Foraging rules in human visual search. *J. Vis.* 13:10. doi: 10.1167/13.3.10

Xu, M., Liu, Y., Hu, R., and He, F. (2018). Find who to look at: turning from action to saliency. *IEEE Trans. Image Process.* 27, 4529–4544. doi: 10.1109/TIP.2018.2837106

Zanca, D., Melacci, S., and Gori, M. (2020). Gravitational laws of focus of attention. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 2983–2995. doi: 10.1109/TPAMI.2019.2920636

Zhang, R., Saran, A., Liu, B., Zhu, Y., Guo, S., Niekum, S., et al. (2020). "Human gaze assisted artificial intelligence: a review," in *IJCAI: Proceedings of the Conference*, Vol. 2020 (Yokohama: NIH Public Access), 4951. doi: 10.24963/ijcai.2020/689

Zibafar, A., Saffari, E., Alemi, M., Meghdari, A., Faryan, L., Pour, A. G., et al. (2019). State-of-the-art visual merchandising using a fashionable social robot: Roma. *Int. J. Soc. Robot.* 11, 1–15. doi: 10.1007/s12369-019-00566-3

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.

The handling Editor declared a past collaboration with one of the authors AD'A.

frontiers in Neurorobotics

HYPOTHESIS AND THEORY
published: 13 April 2021
doi: 10.3389/fnbot.2021.651432

Generative Models for Active Vision

Thomas Parr[1]*, *Noor Sajid*[1], *Lancelot Da Costa*[1,2], *M. Berk Mirza*[3] and *Karl J. Friston*[1]

[1] Wellcome Centre for Human Neuroimaging, Queen Square Institute of Neurology, London, United Kingdom, [2] Department of Mathematics, Imperial College London, London, United Kingdom, [3] Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology & Neuroscience, King's College London, London, United Kingdom

The active visual system comprises the visual cortices, cerebral attention networks, and oculomotor system. While fascinating in its own right, it is also an important model for sensorimotor networks in general. A prominent approach to studying this system is active inference—which assumes the brain makes use of an internal (generative) model to predict proprioceptive and visual input. This approach treats action as ensuring sensations conform to predictions (i.e., by moving the eyes) and posits that visual percepts are the consequence of updating predictions to conform to sensations. Under active inference, the challenge is to identify the form of the generative model that makes these predictions—and thus directs behavior. In this paper, we provide an overview of the generative models that the brain must employ to engage in active vision. This means specifying the processes that explain retinal cell activity and proprioceptive information from oculomotor muscle fibers. In addition to the mechanics of the eyes and retina, these processes include our choices about where to move our eyes. These decisions rest upon beliefs about salient locations, or the potential for information gain and belief-updating. A key theme of this paper is the relationship between "looking" and "seeing" under the brain's implicit generative model of the visual world.

Keywords: active vision, generative model, inference, Bayesian, oculomotion, attention

OPEN ACCESS

Edited by:
Dimitri Ognibene,
University of Milano-Bicocca, Italy

Reviewed by:
Emmanuel Dauce,
Centrale Marseille, France
Giuseppe Boccignone,
University of Milan, Italy

***Correspondence:**
Thomas Parr
thomas.parr.12@ucl.ac.uk

Received: *09 January 2021*
Accepted: *15 March 2021*
Published: *13 April 2021*

Citation:
Parr T, Sajid N, Da Costa L, Mirza MB and Friston KJ (2021) Generative Models for Active Vision. Front. Neurorobot. 15:651432. doi: 10.3389/fnbot.2021.651432

INTRODUCTION

This paper reviews visual perception, but in the opposite direction to most accounts. Normally, accounts of vision start from photons hitting the retina and follow a sequence of neurons from photoreceptor to visual cortex (and beyond) (Goodale and Milner, 1992; Wallis and Rolls, 1997; Riesenhuber and Poggio, 1999; Serre et al., 2007; DiCarlo et al., 2012). At each step, we are told about the successive transformation of these data to detect edges, contours, objects, and so on, starting from a 2-dimensional retinal image and ending with a representation of the outside world (Marr, 1982/2010; Perrett and Oram, 1993; Carandini et al., 2005). In this paper, we reverse this account and ask what we would need to know to generate a retinal image. Our aim is to formalize the inference problem the brain must solve to explain visual data. By framing perceptual inference or synthesis in terms of a forward or generative model, we arrive at the space of hypothetical explanations the brain could call upon to account for what is happening on the retina (Helmholtz, 1878 (1971); MacKay, 1956; Neisser, 1967; Gregory, 1968, 1980; Yuille and Kersten, 2006).

The motivation for this perspective comes from formalisations of brain function in terms of (active) inference (Friston et al., 2017; Da Costa et al., 2020). The idea is that the brain makes use of an implicit model of how sensory data are generated. Perception is then the inversion of this model to find the causes of our sensations (Von Helmholtz, 1867; Gregory, 1980; Doya, 2007). Here, the term 'inversion' refers to the use of (approximate) Bayesian inference to compute posterior

probabilities that represent (Bayesian) beliefs about the world. This is an inversion in the sense that we start with a model of how the world generates sensations and ask what the sensations we obtain tell us about the world. Central to this is the bidirectionality inherent in inference. It is this bidirectionality that manifests in neurobiology (Friston et al., 2017a; Parr and Friston, 2018b), where messages are passed reciprocally between neural populations

In a sense, everything we have said so far only brings us to the point that vision is not just 'bottom-up' but that it has an important "top-down" element to it—which is uncontroversial (Zeki and Shipp, 1988; Lee and Mumford, 2003; Spratling, 2017). However, we take this one step further and argue that if the messages passed up visual hierarchies are the inversion of a (top-down) generative model, then all we need to do is understand this model, and the ascending pathways should emerge naturally, under some neuronally plausible message passing scheme. For this reason, we will focus upon the problem that the visual brain must solve and will not concern ourselves with the details of its solution, reserving this for a future paper.

Perceptual inference is just one part of the story (Ferro et al., 2010; Andreopoulos and Tsotsos, 2013; Zimmermann and Lappe, 2016; Pezzulo et al., 2017). We only sample a small portion of our sensory environment at any one time. In the context of vision, this depends upon where our retina is pointing. This tells us that, to generate a retinal image, we need to take account of how we choose where to look (Ognibene and Baldassarre, 2014). The problem of deciding where to look, and of influencing the biophysical processes required to implement these decisions, are also inference problems. The first relies upon the notion of planning as inference (Botvinick and Toussaint, 2012). Here, we treat alternative action sequences as a set of hypotheses. To select among these, we must weigh prior beliefs about the best course of action against the evidence sensory data afford to each plan. Under active inference, the priors are assumed to favor those plans for which there is a high expected information gain (Lindley, 1956; Itti and Koch, 2000; Itti and Baldi, 2006; Friston et al., 2015; Yang et al., 2016). In short, we have to plan our visual palpation of the world in a way that allows us to construct a scene in our heads that best predicts "what would happen if I looked over there" (Hassabis and Maguire, 2007; Schmidhuber, 2010; Zeidman et al., 2015; Mirza et al., 2016).

The process of implementing these plans is also an inference problem but cast in a slightly different way. In its variational form, approximate Bayesian inference can be framed as optimisation. The inference is deemed optimal when a lower bound on the Bayesian model evidence—the probability of data given a model—is maximized (Beal, 2003; Winn and Bishop, 2005; Dauwels, 2007). While this lower bound can be maximized by closing the gap between the bound and the evidence, it can also be maximized by selecting data that cohere with the model, increasing the evidence itself (c.f., self-evidencing (Hohwy, 2016)). The implication is that we can use action to change the data generating process to fit the world to the model, in addition to fitting the model to the world. For active vision (Wurtz et al., 2011), this means predicting the proprioceptive data we might expect from the oculomotor muscles if a given eye movement is made. Maximizing the evidence then means changing—through contraction or relaxation—muscle lengths until the predicted input is achieved. This can be regarded as a formalization of the equilibrium point hypothesis for motor control (Feldman and Levin, 2009), which posits that all we need do is specify some desired setpoint that can be fulfilled through brainstem (or spinal) reflexes.

To address these issues, we divide this paper into two main sections. First, we deal with the 'seeing' problem. Here, we start from a given environment (e.g., a room we might find ourselves in) and our location in it and ask what pattern of retinal cell activity we would predict. This depends upon the contents of that environment (e.g., the furniture in the room) and the location and geometry of those contents. In addition, it depends upon where we are in the environment, which way we are facing, and the orientation of our eyes relative to our head. We then turn to the 'looking' problem, and its constituents: where to look and how to look there. By formulating looking and seeing as generative models, we reduce the problems to a series of conditional dependencies. As our interest here is in active vision as implemented by the brain, we keep in mind the anatomical manifestations of these conditional dependencies as connections between neural populations.

SEEING

In this section, our aim is to generate a retinal image. **Figure 1** provides an overview of the generative model in (Forney) factor graph format (Loeliger, 2004; Loeliger et al., 2007; Forney and Vontobel, 2011; Laar and Vries, 2016; de Vries and Friston, 2017; van de Laar and de Vries, 2019). As we will appeal to this formalism throughout, we will briefly describe the conventions. As the name suggests, this graphical notation depends upon factorizing the problem into a series of smaller problems. If we assume a set of latent (or hidden) variables x that generate our retinal image y, we can write down a probability distribution that can be decomposed according to the conditional dependencies in the generative model. For example:

$$P\left(y, x^1, x^2, x^3, \ldots\right) = P\left(y|x^1\right) P\left(x^1|x^2, x^3\right) P\left(x^2|x^4\right) \ldots \quad (1)$$

To construct a factor graph of Equation (1), we would take each factor on the right-hand side and draw a square. We then draw a line coming out of this square for every variable that appears inside the factor. If that variable appears in another factor, we connect the line to the square representing the other factor. For those used to looking at Bayesian networks—where edges denote factors—it is worth emphasizing that edges in a factor graph denote random variables. This may seem a little abstract. However, we will go through the components of the factor graph in **Figure 1** in detail over the next few sections. The important thing to begin with is that the upper left of the factor graph relates to scene and object identity. In contrast, the upper right deals with locations and directions. The separation of these explanatory variables offers our first point of connection with neuroanatomy, as this closely resembles the "what" (ventral) and "where" (dorsal) visual streams that support object and spatial

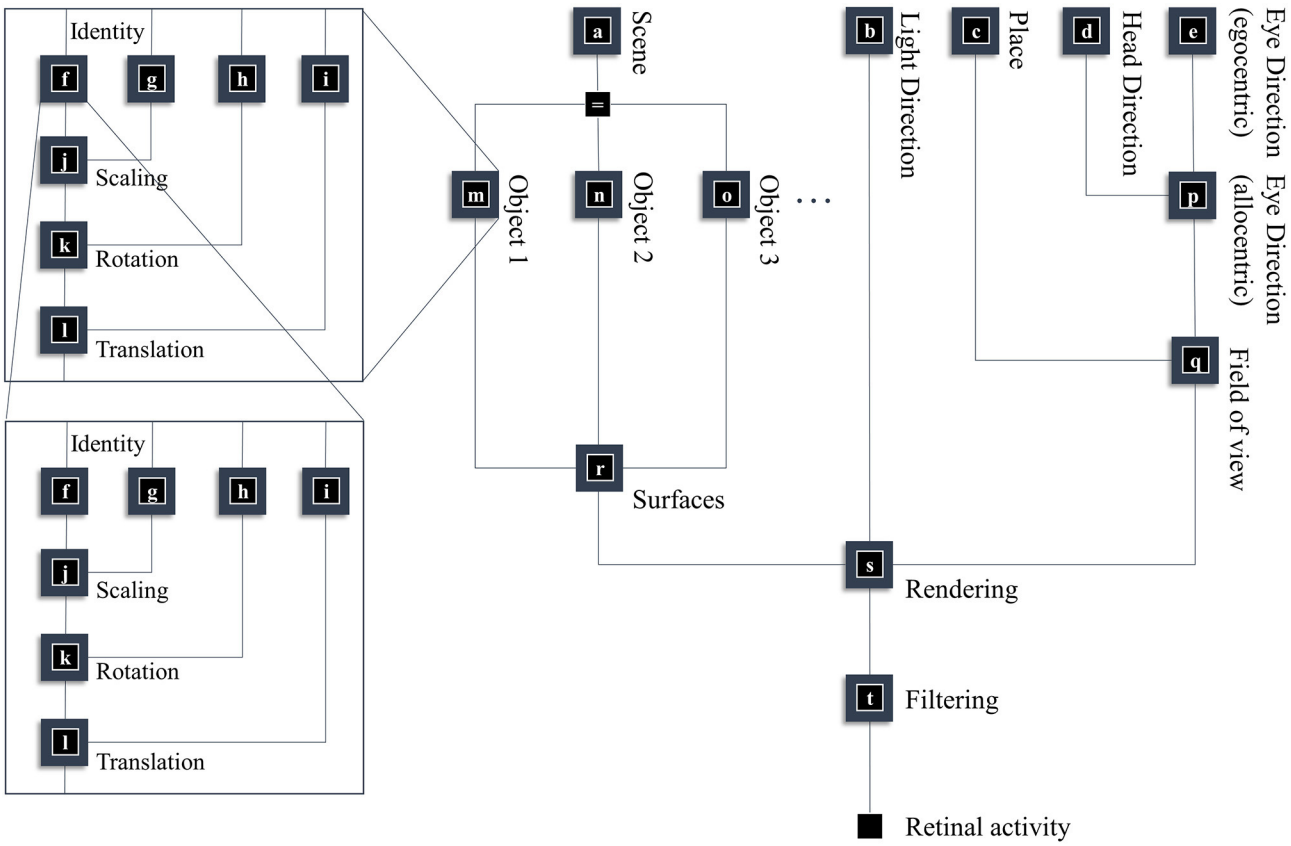


FIGURE 1 | A generative model for seeing. This figure offers a summary of a model that generates a retinal image. Starting from the scene in which we find ourselves, we can predict the objects we expect to encounter. These objects may be recursively defined, through identifying their constituent parts and performing a series of geometric (affine) transformations that result in the configuration of parts, the scaling and rotation of this overall configuration, and the placement of these objects in an allocentric reference frame. The nested parts of the graphical structure (on left hand side) indicate this recursive aspect, e.g., an object is defined by an identity, and scaling, rotation and translation parameters; furthermore, it may be composed of various sub-objects each of which has these attributes. To render a retinal image, we also need to know where the retina is and which way it is pointing (i.e., the line of sight available to it). This depends upon where we are in the environment, our heading direction, and where we choose to look. Subsequent figures unpack the parts of this model in greater detail.

vision, respectively (Mishkin et al., 1983). The sections on The Ventral Stream and The (Extended) Dorsal Stream deal with these pathways, and the section on The Retinocortical Pathway deals with their convergence.

The Ventral Stream

This section focuses upon the identity and shape of the things causing our visual sensations. From a neurobiological perspective, the structures involved in object and scene identification are distributed between the occipital and temporal lobes (Kravitz et al., 2013). The occipitotemporal visual areas are referred to as the 'what' pathway or the ventral visual stream. The occipital portion of the pathway includes cells with receptive fields responsive to concentric circles (Hubel and Wiesel, 1959) and gratings (Hegdé and Van Essen, 2007). The temporal portion contains cells with more abstract response properties, relying upon more specific feature configurations that are invariant to size, view, or location (Deco and Rolls, 2004; DiCarlo et al., 2012). We will start from the more abstract (temporal) end of this pathway and work our way toward the simpler features at the occipital end.

The first thing we need to know, to generate an image, is the environment in which we find ourselves. A schematic of a simple environment is shown in **Figure 2**, which shows three possible rooms—each of which contains two objects that can appear in different locations. If we knew which of these rooms we were in, we could predict which objects were present. This is approximately the same structure as used in previous accounts of scene construction in a 2-dimensional world (Mirza et al., 2016). It has neurobiological validity as evidenced by the proximity of the inferotemporal cortex, associated with object recognition (Logothetis and Sheinberg, 1996; Tanaka, 1996), to the parahippocampal gyrus, associated with recognition of places (Epstein et al., 1999), hinting at how the brain might represent dependencies between scenes and their constituent objects.

Once we know which objects we expect to be present, we can associate them with their 3-dimensional geometry. To generate these objects, we assume they are constructed from simpler structures—for the purposes of illustration, spheres.

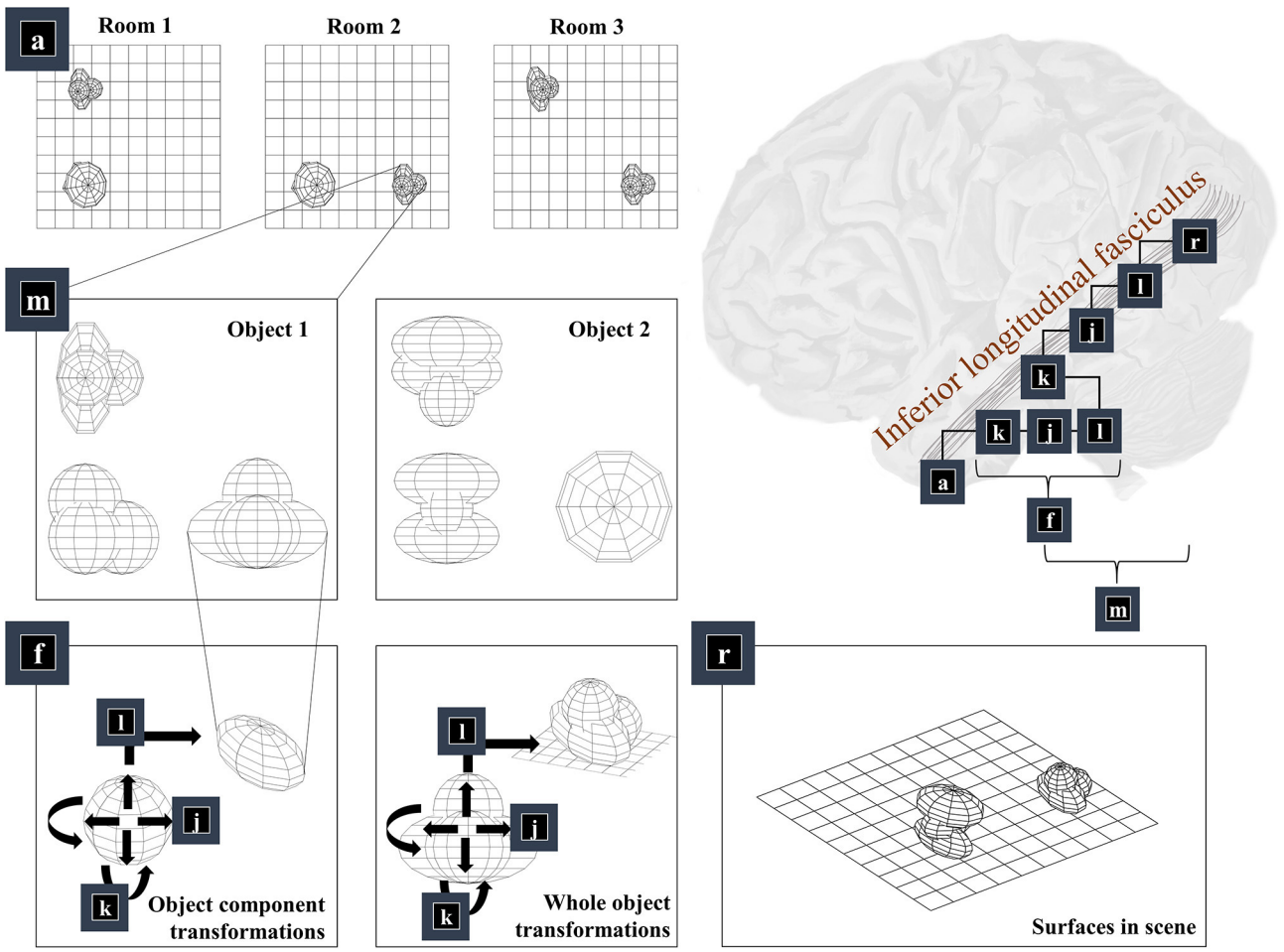


FIGURE 2 | The "what" pathway. This figure focuses upon part of the factor graph in **Figure 1**. In the upper right, this factor graph is reproduced as it might be implemented neuroanatomically. Here, the factors are arranged along the occipitotemporal "what" pathway, which loosely follows those cortical areas superficial to a white matter tract called the inferior longitudinal fasciculus. The panels on the left show the sequence of steps implemented by these factors. Factor **a** is the distribution over alternative rooms (or scenes) we could find ourselves in. This implies a categorical distribution assigning a probability to each of the rooms provided in the example graphics. Conditioned upon being in a room, we may be able to predict which objects are present. Two example objects are shown (from three orthogonal views). The conditional probability distribution for the geometry of the objects given the room is given by the **m** factor, which is here broken down into several constituent factors. First, we need to know the identity of the objects in the room (**f**), which we operationalise in terms of the configuration of the parts of that object. Specifically, we decompose **f** into a series of transformations (**k**-**l**) applied to a set of spheres, which represent the constituents of the object. In principle, we could have used other objects in place of spheres, or could have applied this procedure recursively, such that the constituents of an object can themselves be decomposed into their constituents. Once we have our object, we can apply the same transformations (**k-l**) to the whole to position it our room. We do this for each object in the room, eventually coming to a representation of all the surfaces in the scene (**r**), shown in the lower right panel.

Figure 1 illustrates the recursive aspect to this, where the object factor (**m**) is decomposed into a series of geometric (affine) transformations applied to a structure as identified by the object identity factor (**f**), which itself can be decomposed into a series of transformations of simpler features. In other words, an object's geometry depends upon the configuration of its features (e.g., the legs and surface of a table), but these features can themselves depend upon configurations of simpler features (Biederman, 1987). Implicit in this perspective is that the scene itself is simply the highest level of the recursion, comprising features (objects) that themselves comprise simpler features. **Figure 2** illustrates this idea graphically.

Taking a step back, we need to be able to represent the shape of a feature before we can start applying transformations to it. One way of doing this is to construct a mesh. Meshes specify the vertices of the surfaces that comprise an object (Baumgart, 1975), effectively setting out where we would expect to find surfaces. This is the form shown in the graphics of **Figure 2**—where we have omitted occluded surfaces for visual clarity. Note that we have taken a subtle but important step here. We have moved from discussing categorical variables like scene or object identity and have started working in a continuous domain. At this point, we can apply geometric transforms to our objects. The first is the scaling of an object (factor **j**), which is a simple

linear transform using a matrix (S) whose diagonal elements are positive scaling coefficients along each dimension. This is applied to each coordinate vector of our mesh. Expressing this as a factor of a probability distribution, we have:

$$P\left(x^j|x^f,x^g\right) = \delta\left(S(x^g)x^f - x^j\right)$$
$$S\left([\alpha,\beta,\gamma]\right) = \begin{bmatrix} e^{\alpha} & & \\ & e^{\beta} & \\ & & e^{\gamma} \end{bmatrix} \quad (2)$$

The x variables represent the edges in the graph of **Figure 1**. The superscripts indicate the factor from which the edge originates (i.e., the square node above the edge). The x^f variable includes the coordinates of the vertices of each surface of the object. This is transformed based upon the scaling in each dimension (in the x^g variable) to give the scaled coordinates x^j. The scaling variables are treated as log scale parameters. This means we can specify factor **g** to be a Gaussian distribution without fear of negative scaling. However, we could relax this constraint and allow for negative scaling (i.e., reflection). In addition, we could include off-diagonal elements to account for shear transforms. In Equation (2), δ is the Dirac delta function—a limiting case of the (zero-centered) normal distribution when variance tends to zero. It ensures there is non-zero probability density only when its argument is zero. This is a way of expressing an equality as a probability density. We could have used a normal distribution here, but for very large objects, with many surfaces, the associated covariance matrices could become unwieldy. It is simpler to absorb the uncertainty into the priors over the (log) scaling parameters.

Our next step is to apply rotations to the object. Here, we use a rotation matrix (R) that has the form:

$$P\left(x^k|x^j,x^h\right) = \delta\left(R(x^h)x^j - x^k\right)$$
$$R([\theta,\phi,\varphi]) = \begin{bmatrix} \cos(\phi)\cos(\varphi) & -\cos(\phi)\cos(\varphi) & \sin(\phi) \\ \cos(\theta)\sin(\varphi)+\sin(\theta)\sin(\phi)\cos(\varphi) & \cos(\theta)\sin(\varphi)-\sin(\theta)\sin(\phi)\sin(\varphi) & -\sin(\theta)\cos(\phi) \\ \sin(\theta)\sin(\varphi)-\cos(\theta)\sin(\phi)\cos(\varphi) & \sin(\theta)\cos(\varphi)+\cos(\theta)\sin(\phi)\sin(\varphi) & \cos(\theta)\cos(\phi) \end{bmatrix} \quad (3)$$

As in Equation (2), we use the Dirac delta distribution such that the rotated coordinates can only plausibly be the original coordinates, rotated. This defines the **k** factor.

Finally, we translate the objects (factor **l**). This is simply a matter of adding the same vector to all vertices of the mesh and centring a Dirac delta distribution for x^l on this value. **Figure 2** shows two applications of these three operations that give us the components of object 1 (lower left panel) and that place object 1 in a particular place in our scene (lower middle panel). The factor **r** simply concatenates the surfaces from all objects such that x^r is simply a list of surfaces.

Is there any validity to the idea that the brain might generate objects with a series of geometrical transforms of this sort? Evidence in favor of this comes from two lines of research. One is in psychological experiments which show that, during object recognition, reaction times scale with the angle of rotation that would have to be performed to bring that object into a familiar configuration (Cooper and Shepard, 1973; Tarr and Pinker, 1989)—suggesting a form of implicit mental rotation. This is consistent with the idea that the brain optimizes its model through updating beliefs about the degree of rotation until it best fits the data at hand.

The second line of evidence is from neurophysiological studies into invariance of neural responses to different properties. To understand the relevance of invariant representations, note that the transforms we have described do not commute with one another. To see this, consider what would happen if we were to rotate the sphere before rescaling it. The implication is that, if there are objects whose identity is preserved with changes in its geometry, we should expect to see different sorts of invariance emerge at different stages along the visual hierarchy. At the highest levels, we might expect neural responses to be consistent for an object, no matter how it is oriented, scaled, or translated. As we descend toward the occipital lobe, we might anticipate these invariances being lost, in sequence. This is exactly what happens (Rust and DiCarlo, 2010; Grill-Spector and Weiner, 2014; Tacchetti et al., 2018), with inferotemporal cortical cells responding to specific objects, regardless of their size, position (Ito et al., 1995), or the angle from which they are viewed (Ratan Murty and Arun, 2015). As we move toward the occipital cortex, neurons become more sensitive to the rotation of an object (Gauthier et al., 2002; Andresen et al., 2009). On reaching areas V2-V4 of the early visual cortex, the receptive fields of neurons are many times smaller than those in inferotemporal cortex (Kravitz et al., 2013). This means they respond only when a stimulus is in a specific region of space, implying loss of translation invariance. Evidence that the brain inverts a model of this sort comes from studies illustrating that the activity of (feedforward) convolutional neural networks trained on visual data—which implicitly account for the requisite transforms—aligns with gamma-band activity in visual cortices (Kuzovkin et al., 2018). This frequency band is crucial in ascending neural message passing (Bastos et al., 2015) associated with model inversion (Friston, 2019).

While we chose affine transforms for simplicity, it is worth emphasizing that the generative model is highly non-linear. This is most striking for the recursive part of the ventral stream model, which alternates between linear operations (affine transformations of the shapes) and non-linear operations (selection between shapes). To invert this kind of model, one would employ a linear operation to undo the affine transformations for each component of an object. On finding the log likelihood of the inverted shape for each component, one could compute a posterior by adding the log prior for each component and taking a non-linear softmax transform. This is then repeated for the next level of the recursion, eventually returning a categorical distribution over plausible objects that could be causing visual data. The alternation between linear

and non-linear operations—in the inversion of this model—could explain why deep learning architectures, that alternate in this way, have been so successful in machine vision. Non-affine transformations could be incorporated through using a spatial basis set to deform the objects or their components—analogous to the models employed for spatial normalization in image analysis (Arad et al., 1994; Ashburner and Friston, 1999; Shusharina and Sharp, 2012). This would involve adding additional factors into the ventral stream model that represent these deformations but would not change the overall anatomy of the model.

In summary, we have gone from prior beliefs about the room we occupy to beliefs about the objects in that room. These are decomposed into their constituent parts, and the surfaces that define these parts. At the occipital end of the pathway, we have a set of surfaces. Taken individually, these surfaces could belong to any object. Each occupies a smaller portion of space than the complete objects. This means that, in the process of generating the geometric structures we will need for vision, we have traversed the ventral visual pathway from the large, abstract receptive fields of the inferior temporal cortices to the smaller, simpler receptive fields of the occipital lobe.

A final consideration for this section is the consequence of damage to the brain structures implementing this generative model. Ventral visual stream lesions give rise to an interesting category of neuropsychological syndromes, broadly referred to as agnosia (Adler, 1944; Benson and Greenberg, 1969; Greene, 2005). There are many variants of agnosia, but common to all is a failure to recognize something. Visual agnosia is an inability to recognize objects, sometimes restricted to specific categories. For example, prosopagnosia is a form of visual agnosia specific to faces (Sacks, 2014). Generative modeling offers a useful perspective on agnosia, as any lesions to the ventral stream impair the capacity of a model to predict the visual data that would be anticipated if a given object were present. If we assumed that a given lesion removed all neurons involved in representing object 1 from **Figure 2** or cut the connections that predicted the surfaces anticipated when object 1 is present, we could generate as many images as we wanted by sampling from the generative model without ever generating one characteristic of object 1. Without this hypothesis available to the brain, it is unable to invert the data-generating process to arrive at the conclusion that object 1 is present. Despite this, it might still be possible to identify its constituent parts, particularly if these parts are like those found in other objects.

The (Extended) Dorsal Stream

Now that we know the positions and orientations of the surfaces in our scene, we need to know the same for our retina. To know where our retina is, the first thing we need to know is where our head is in allocentric space. In other words, where we are in our environment. The part of the brain most associated with this is outside of the classical visual brain. It is the hippocampal formation that famously contains place (and grid) cells, which increase their firing rate when an animal is in specific places (or at repeating intervals) in an environment (Moser et al., 2008).

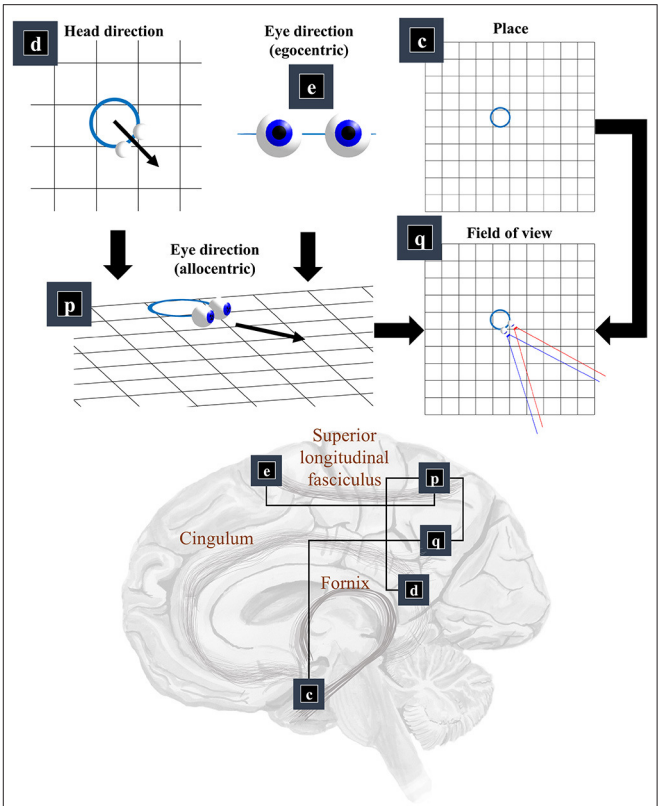


FIGURE 3 | The "where" pathway. This figure shows the factors that conspire to generate a field of view. This shows how the allocentric head and egocentric eye-directions (factors **d** and **e**) can be combined to compute an allocentric eye-direction (factor **p**). When this vector is placed so that it originates from the place (factor **c**) we find ourselves in, we have our field of view (factor **q**). The graphic in the lower part of this figure maps the associated factors onto the brain structures thought to be involved in representing these variables. The frontal eye-fields (factor **e**) and the retrosplenial cortex (factor **d**), both project to the parietal cortex (factor **p**), which includes regions sensitive to allocentric eye-directions. This communicates with temporoparietal regions (factor **q**), which are also accessible to hippocampal outputs (factor **c**) via a pathway comprising the fornix, mammillothalamic, and cingular white matter tracts.

Figure 3 illustrates this by placing factor **c**—prior beliefs about place—in the medial temporal lobe.

We need more than the location of the head to be able to locate the retina. First, we need to know which way the head is facing. Head-direction cells, which fire maximally when an animal is oriented along a given direction in its environment, are found distributed throughout the brain (Taube et al., 1990; Taube, 1995; Blair et al., 1998). Specifically, they are found in the constituents of the Papez circuit (Papez, 1995), originally thought to mediate emotional responses. Together, the place and head-direction tell us where the eyes are, but they do not pinpoint the retinal location. For this, we also need to know the direction in which the eyes are pointing. Combining the head-direction (factor **d**) with the egocentric eye-direction (factor **e**), we can compute the allocentric eye-direction (factor **p**). With information about place, this gives us our field of view (factor **q**).

Expressed as a probability distribution, factor **p** is:

$$P\left(x^p|x^d, x^e\right) = \delta\left(x^d + x^e - x^p\right) \tag{3}$$

This ensures the allocentric eye-direction is given by the angle of the head plus the angle of the eyes relative to the head. We can augment this for each eye, to allow for their convergence—i.e., that the directions of the left and right eye are not parallel to one another. Factor **q** is a little more complicated but involves constructing arrays representing locations of retinal cells or, more simply, locations in front of the lens that, if light were to pass through the location and reach the lens, would refract to a given retinal photoreceptor (or group of photoreceptors). We generate one array for each eye. We make a simplification here in that we assume we are dealing with a small foveal area such that we can ignore the global topography of the retina. As such, we treat the array of cells as uniformly spaced. A more complete retinal model would take account of the log-polar organization (Javier Traver and Bernardino, 2010), in which the density of photoreceptors decreases with retinal eccentricity—i.e., distance from the fovea. This array, along with the location of the lens, gives us our field of view. Taking the outermost cells from each array, we simply project from the lens, through that location. This generates the blue and red lines in the **q** panel of **Figure 3**. The x^q variables are tuples, for each element of the retinal array, containing the location and a unit vector representing its preferred angle of incidence.

The classical 'where' pathway involves the occipitoparietal cortices. **Figure 3** shows how the factors needed to compute a field of view could converge upon the parietal lobe, assuming we assign factor **q** to the temporoparietal cortices. Interestingly, these regions have been associated with the ability to take another point of view in several different senses. Electrical stimulation of these regions on the right side of the brain can induce out of body experiences (Blanke et al., 2002), where people feel as if they are observing the world from a vantage point outside of their body. We also talk informally about seeing things from another person's point of view. This relates to theory of mind, and the ability to infer another's perspective at a more abstract level. These functions are also associated with the temporoparietal cortices (Abu-Akel and Shamay-Tsoory, 2011; Santiesteban et al., 2012). The implication is that the same machinery may be involved in taking a viewpoint, both in the literal and metaphorical sense, and that this machinery is housed in the temporoparietal region. Some have argued that this representation of viewpoint is central to the first-person perspective that underwrites conscious experience (Seth, 2009; Williford et al., 2018).

The retrosplenial cortex is a good candidate for factor **d**, given its role in relating visual 'where' data with head-direction (Marchette et al., 2014; Shine et al., 2016). Specifically, it is responsive to where we have to look to find stable, unambiguous, landmarks (Auger et al., 2012). Lesions to this region impair the representation of head-direction in other parts of the brain—notably the anterior thalamus—even in the presence of clear visual landmarks (Clark et al., 2010). Neuropsychological evidence supports this assignment, as lesions to the retrosplenial cortex can cause a form of topographical disorientation, where patients lose their sense of direction (Aguirre and D'Esposito, 1999).

The translation from head-centered eye-direction to a world-centered reference frame (i.e., factors **e** and **p**) is consistent with the connections from the frontal eye fields to the parietal lobe. These connections are underwritten by a white matter tract known as the superior longitudinal fasciculus (Makris et al., 2005; Thiebaut de Schotten et al., 2011). The parts of the brain connected by this tract are referred to as the attention networks (Corbetta and Shulman, 2002; Szczepanski et al., 2013)—identified through their recruitment in attentional tasks during neuroimaging studies. The frontal eye fields (Bruce et al., 1985) and intraparietal sulcus (Pertzov et al., 2011) both contain neurons sensitive to eye position, in different coordinate systems.

In summary, the generation of a line of sight depends upon the head location and direction, and the position of the eyes relative to the head. These are represented in the medial temporal lobe, the frontal lobe, and medial parietal structures. The convergence of axonal projections from these regions to the lateral parietal lobe provides the dorsal visual stream with key information, which can be reciprocally exchanged with the occipital cortices. While we have adopted the rhetoric of "what" and "where" streams, it is interesting to note that the controllable aspects of the generative model all relate to the "where" stream. This provides a useful point of connection to a complementary framing of the two visual streams. Under this alternative perspective (Goodale and Milner, 1992), the ventral stream is thought to support perception, while the primary role of the dorsal stream is to inform action. This view is informed by neuropsychological findings (Goodale et al., 1991), including the ability of those with dorsal stream lesions to see objects they cannot grasp, and the ability of those with lesions to other parts of the visual cortices grasp objects they could not see.

The Retinocortical Pathway

So far, we have generated a set of surfaces, and a field of view. The final challenge of our 'seeing' generative model is to convert these to a pair of retinal images. This is analogous to the process of rendering in computer graphics (Shum and Kang, 2000). There are many ways to implement sophisticated rendering schemes, and a review of these is outside the scope of this paper. We will outline one way in which a simple form of rendering may be implemented and consider whether this has neurobiological correlates.

For any given retinal photoreceptor, we can trace an imaginary line out through the lens of the eye and ask which surface it will first encounter. If it does not pass through any surface, this means there is nothing that can reflect light in the direction of that cell, and the receptor will not be activated. However, if it does encounter a surface, we must determine the intensity of light that surface reflects in the direction opposite to our imaginary line. This is similar to the ray tracing method in computer graphics (Whitted, 1980), and depends upon the rendering equation (Kajiya, 1986):

$$P\left(x^s|x^r, x^q, x^b\right) = \delta\left(\Lambda\left(x^q, x^r, x^b\right) - x^s\right)$$

$$\Lambda\left(u,v,z\right) = \eta\left(u,v\right) \times \left(\underbrace{\alpha(u,v)}_{\text{Ambient}} + \int_S \underbrace{\Lambda\left(v,w,z\right)\beta\left(u,v,w\right)dw}_{\text{Reflected}}\right) \quad (4)$$

The variables in the conditioning set are the light direction (x^b), as a unit vector, and tuples containing information about the surfaces of objects (x^r) and the retinal cells (x^q). The η function acts as an indicator as to whether a line passing through the lens, that would refract light to a specific retinal cell (u), intersects with a point on a surface (v) before reaching any other surface. It is one if so, and zero otherwise. The α function plays the role of ambient lighting, and we assume this is a constant for all surfaces, for simplicity. The β function determines the proportion of light reaching a surface from other sources (w)—e.g., reflected off other surfaces (S)—that is reflected toward u. The recursive structure of the integral part of this expression resembles the recursive marginalization that underwrites belief-propagation schemes (Frey and MacKay, 1998; Yedidia et al., 2005). Recursive expressions of this sort can usually be solved either analytically—e.g., through re-expression in terms of an underlying differential equation—or numerically. In principle, we could construct a factor graph like that of **Figure 1**, using the β functions as our factors, determining the dependencies between the level of illumination of each surface. The integral includes all surfaces S that could reflect light to surface v. To simplify, we ignore the dependencies between surfaces, and assume a single level of recursion (i.e., surfaces reflect light to the retina, but the light incident on a surface originates directly from the light source). This means we choose $S = z$, so that Equation (4) simplifies to:

$$\Lambda\left(u,v,z\right) = \eta\left(u,v\right)\left(\alpha(u,v) + \eta\left(v,z\right)\alpha\left(v,z\right)\beta\left(u,v,z\right)\right) \quad (5)$$

The key differences between different approaches to generating images rest upon the choice of β. We follow the approach outlined in (Blinn, 1977):

$$\beta\left(u,v,z\right) = \underbrace{c_1 \max\left(0, v_n \cdot z\right)}_{\text{Diffuse}} + \underbrace{c_2\left(v_n \cdot \frac{u_n + z}{\sqrt{(u_n+z)\cdot(u_n+z)}}\right)^{c_3}}_{\text{Specular}} \quad (6)$$

Equation (6) uses the subscript n to indicate (normalized) unit vectors drawn from the u and v tuples (which also include the coordinates of the origins of these vectors). For u_n, this vector is parallel to the line from the lens outwards—in the opposite direction to the light that would be refracted to a specific group of cells on the retina. For v_n it is the normal unit vector to the surface in question[1]. Equation (6) includes a diffuse term, which allows for light to be reflected equally in all directions, where the amount reflected depends upon the angle of incidence. In **Figure 4**, we see how this lighting component catches some surfaces but not others, and the way in which it induces shadows (via multiplication with the η function). The specular component accounts for the relationship between the angle of incidence and the angle of reflectance from a surface (Phong, 1975). To gain some intuition for this term, imagine shining a torch into a mirror. The reflection will appear maximally bright when the angle between the torch and the normal to the mirror is equal to the angle between your eye and the normal to the mirror and will rapidly decay on moving either eye or torch.

A simplification made in the above is to treat the lens as a point, neglecting the fact that there are a range of angles of light that could be focused upon a given cell in the retina. In reality, neighboring photoreceptors may encounter photons reflected from the same point on a surface. To account for the artificial high frequency components introduced during this discretisation of space, we apply a blurring effect (factor **t**) This is based upon a discrete cosine transform followed by attenuation of those coefficients corresponding to these high frequencies followed by the inverse transform. Specifically, we multiply the coefficients by a Gaussian function centered on the low frequency components. An interesting consequence of this relates to the inversion of this model. Undoing this process would mean replacing the high frequency components. This enhancement might give the appearance of edge detection—a common role afforded to cells in the early visual pathway with center-surround receptive fields (Crick et al., 1980; Marr et al., 1980). In addition, it could account for the sensitivity of early visual neurons to specific spatial frequencies, and the widespread use of grating stimuli and Gabor patches in experiments designed to interrogate these cells (Mahon and De Valois, 2001).

An important feature of this generative model is the fact that surfaces on the left of the head (in egocentric space) are projected to the right side of both retinas. Similarly, surfaces on the right of the head are projected to the left side of both retinas. This is interesting in the sense that there are two sorts of deficit we could induce. As shown on the left of **Figure 4**, we could disconnect one retina, precluding surfaces from either side of space from generating an image on this side. This generates images consistent with monocular blindness. Alternatively, by precluding any surface on one side of space from causing retinal cell activation, we lose activity on the same side of both retinas—i.e., a homonymous hemianopia. This maps to the deficits found on lesions to the retinocortical pathway before and after the optic chiasm, respectively (Lueck, 2010; Wong and Plant, 2015). This highlights the inevitability of these visual field defects following lesions to the visual pathway, under the assumption that the brain uses a model that represents the same surfaces as causes of data on both retinas.

The generative model ultimately must generate the data it seeks to explain. For our purposes, these data are the signals sent from the retina to the visual cortex. However, it is possible to take this further and to specify the kinds of generative model used within the retina itself. Attempts to do this have focused upon a prior belief about the smoothness of input across the retina and have provided useful accounts of efficient retinal processing as predictive coding (Srinivasan et al., 1982; Hosoya et al., 2005).

[1]If v contains the four vectors corresponding to the vertices of a quadrilateral surface, then v_n is obtained (with appropriate normalization) as $v_n \propto (v_1 - v_2) \times (v_4 - v_2)$.

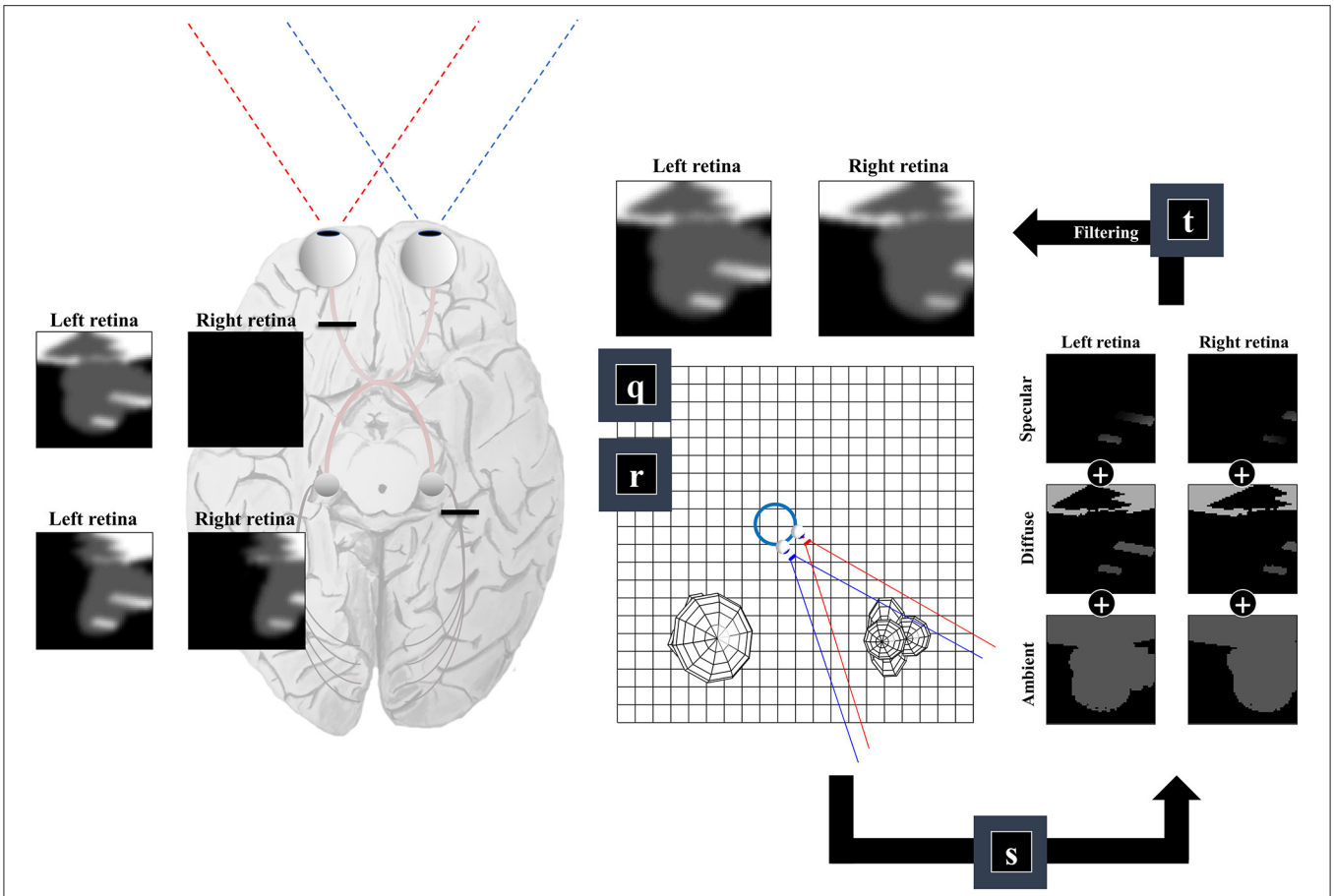


FIGURE 4 | The retinocortical pathway. This figure takes the results from **Figures 2**, **3** and combines these to arrive at images in each retina. From factors **r** and **q** we have our field of view and the surfaces it captures. We can then project from each retinal cell (shown as pixels in the retinal images) to see whether any surface is encountered. For the first surface we reach, we combine the ambient, diffuse, and specular lighting components (factor **s**). This depends upon factor **b** from **Figure 1**, which provides a lighting direction. Once we have the sum of these lighting components, we apply a blurring (factor **t**) to the image to compensate for the artificial high frequency components introduced by our simplifications. Practically, this is implemented by finding the coefficients of a 2-dimensional discrete cosine transform, multiplying this by a Gaussian function centered on the low frequency coefficients, and then performing the inverse transform. Note that the final image is inverted across the horizontal and vertical planes. This is due to the light reflecting off surfaces on the temporal visual fields being propagated to the retina on the nasal side, and vice versa (with the same inversion in the superior and inferior axis). The fact that the same surface can cause activation of both the right and left retina implies a divergence in the predictions made by parts of the brain dealing in surfaces (e.g., striate cortices) about retinal input. The graphic on the left illustrates the two sorts of visual field defect resulting from this divergence—either interrupting the influence of any surface on one retina (upper image) or interrupting the influence of a subset (e.g., the right half) of all surfaces on either retina (lower image).

LOOKING

As alluded to above, retinal data depends not just upon what is "out there" in our environment, but upon where we direct our gaze. **Figure 5** takes factors **d** and **e** from **Figure 1**, and conditions these upon a policy variable. This accounts for the fact that our choices determine where our eyes and our head are facing. In addition, **Figure 5** shows some of the non-visual sensory modalities that result from these explanatory variables. These depend upon dynamical systems, as the motion of the head and eyes cause changes in vestibular and proprioceptive modalities. This is of particular importance when thinking about movement as the solution to an inference problem. When acting so as to minimize any discrepancy between predicted and realized sensations, thereby maximizing the evidence for a model, the predicted consequences of action become central to the performance of that action. The section on The Brainstem unpacks the generation of proprioceptive data from the oculomotor muscles and the relationship to the oculomotor brainstem. The section on The Basal Ganglia then focuses upon formulation of prior beliefs about the policy—and its neurobiological manifestation in the oculomotor loops of the basal ganglia. Together, these can be seen in the spirit of agenda-driven perspectives (Ballard and Zhang, 2020) on action, where we unpack a selected policy into the set of processes that must be initiated at lower levels of a model to execute or realize that policy.

The Brainstem

This section focuses upon the biophysics of oculomotion that underwrites implementations of saccadic eye movements. Modeling the eyes is relatively straightforward. They tend to

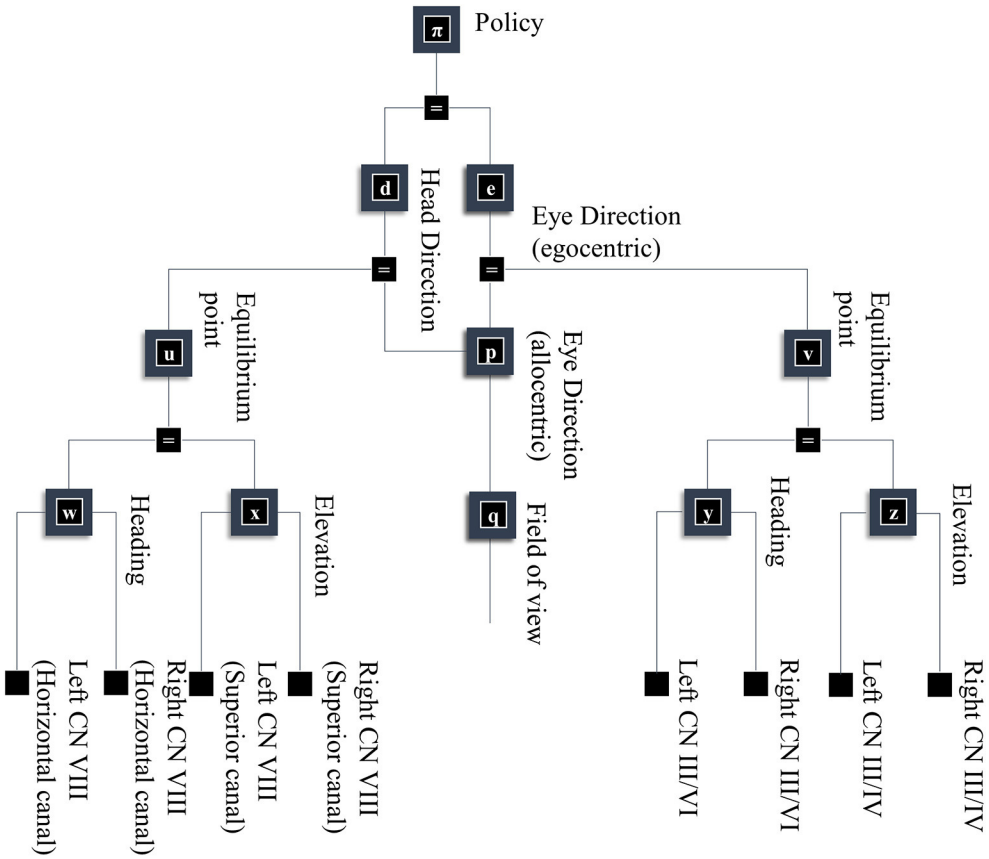


FIGURE 5 | A generative model for looking. This figure builds upon part of the model shown in **Figure 1**. Specifically, it unpacks some of the other sources of data resulting from the eye and head-direction factors and includes a policy variable that determines priors over these variables. These depend upon dynamical systems. This means predicting an equilibrium point (or attractor) that the eyes or head are drawn toward. These dynamics may be divided into changes in the elevation or heading angles. For head movements, the velocity of the head causes changes in the semi-circular canals in the inner ear, communicated to the brain by cranial nerve (CN) VIII. For eye movements, the position and velocity of the eyes give rise to proprioceptive signals due to stretch of the oculomotor muscle tendons, communicated to the brain by CN III, IV, and VI.

move together[2] and can be described using Newton's second law applied to rotational forces (McSpadden, 1998). This describes the relationship between a torque τ applied at radius r to a point mass m and an angle θ:

$$\begin{aligned} \tau &= mr^2\ddot{\theta} \\ \Rightarrow \int_0^\infty \tau(r)dr &= \ddot{\theta}\int_0^\infty m(r)r^2dr \end{aligned} \tag{7}$$

The second line of this equation relates the first to a solid object, where the torque and the density ($m(r)$) of the object can vary with the radius. The oculomotor muscles that generate torques insert into the surface of the eyeballs, meaning we can simplify Equation (7) as follows:

$$\begin{aligned} \tau(r) &= \tau\delta(r - r_{\max}) \Rightarrow \\ \tau &= J\ddot{\theta} \end{aligned}$$

$$J \triangleq \int_0^\infty m(r)r^2dr \tag{8}$$

The term J in the final line is a constant known as the "moment of inertia." Equation (8) implies the following equations of motion:

$$\begin{aligned} \theta &\triangleq \begin{bmatrix} \theta \\ \dot{\theta} \end{bmatrix} \\ \dot{\theta} &= f(\phi,\theta) \triangleq \begin{bmatrix} \dot{\theta} \\ J^{-1}\tau(\phi) \end{bmatrix} \end{aligned} \tag{9}$$

All that is left is to provide a functional form for the torque. We can choose this such that the eyes come to rest at an angle ϕ:

$$\tau(\phi,\theta,\dot{\theta}) = \phi - \theta - \kappa\dot{\theta} \tag{10}$$

This is analogous to the torque associated with a swinging pendulum. The constant κ determines the damping, which precludes large oscillations around ϕ. We can interpret ϕ as a target or setpoint, in the spirit of the equilibrium point hypothesis

[2]Unless you are a chameleon: Katz et al. (2015).

of motor control (Feldman and Levin, 2009). Now that we have the equations of motion of the eye—noting that we have a single equation for both eyes to enforce conjugacy[3]. (Parr and Friston, 2018a)—we must detail the sensory consequences of these movements. These are given as follows:

$$g(\theta, \omega) \triangleq \begin{bmatrix} \theta - \frac{1}{2}\omega \\ \dot{\theta} \\ \theta + \frac{1}{2}\omega \\ \dot{\theta} \end{bmatrix} \tag{11}$$

Here, ω represents the convergence of the eyes, accommodating the fact that the angle between the two can vary. The first two rows relate to the left eye, and the last two to the right. Equation (11) assumes a direct mapping from the angular positions and velocities of each eye to the proprioceptive input from the oculomotor muscles, consistent with the role of II and Ia sensory afferents (Cooper and Daniel, 1949; Cooper et al., 1951; Ruskell, 1989; Lukas et al., 1994), respectively.

Converting Equations (9–11) to factors of a probability distribution, we have:

$$\begin{aligned} P\left(\dot{x}^{v} | x^{v}, x^{e}\right) &= \mathcal{N}\left(f\left(x^{v}, x^{e}\right), \Pi_{f}\right) \\ P\left(y^{y}, y^{z} | x^{v}\right) &= \mathcal{N}\left(g\left(x^{v}, \omega\right), \Pi_{g}\right) \end{aligned} \tag{12}$$

The superscripts here refer to the factors determining the prior densities of each variable in the graph of **Figure 5**. The precision matrices Π stand for inverse covariances. Each of these factors can itself be factorized (assuming diagonal precision matrices) into elevation and heading angles and into left and right eyes. The oculomotor brainstem is well-suited to implementing this part of the forward model (and its inversion). The superior colliculus[4] projects to the raphe interpositus nucleus (Gandhi and Keller, 1997; Yoshida et al., 2001), and via this structure to two nuclei that represent the first (elevation and heading) factorization. The paramedian pontine reticular formation mediates horizontal saccades (Strassman et al., 1986), while the rostral interstitial nucleus of the medial longitudinal fasciculus mediates vertical saccades (Büttner-Ennever and Büttner, 1978). These nuclei then project to the cranial nerve nuclei that communicate directly with oculomotor muscles. The cranial nerve nuclei on the right of the midbrain connect to the muscles of the right eye, and those on the left connect to the left eye. This represents the second factorization into left and right eyes. **Figure 6** shows how this factorization may manifest anatomically and illustrates the proprioceptive data we would anticipate on simulating the dynamics outlined above.

This just leaves the question as to where the equilibrium point (x^{v}) comes from. As we have said, the superior colliculus—a midbrain structure—is an important junction in the descending pathway to the oculomotor brainstem. Via factor **v**, the dynamics depend upon factor **e**, which is the same variable that appears in our frontal eye fields in **Figure 3**. The frontal eye fields project to the superior colliculus (Künzle and Akert, 1977; Hanes and Wurtz, 2001), as shown in **Figure 6**. However, factor **e** is conditioned upon the policy, implying we may have several alternative equilibrium points available to the superior colliculus. To adjudicate between these, we need another input to the colliculus that selects between policies. We have previously argued that the output nuclei of the basal ganglia could fulfill this role (Parr and Friston, 2018c). This is consistent with the projections from the substantia nigra pars reticulata to the superior colliculus (Hikosaka and Wurtz, 1983). The selection between alternative policies is the focus of section The basal ganglia. A similar analysis could be made of head movements and the vestibular data they generate. We omit this here to avoid duplication of the concepts outlined above. More generally, selecting a series of attracting points, as we have for saccadic eye movements, offers a useful way of representing environmental dynamics, including those that are out of our control. For instance, by replacing the static prior over object location with a series of transition probabilities, we could predict the next location given the current location. This converts the static elements of the model into a hidden Markov model. By associating each possible location with an attracting point, we can predict the continuous trajectories of the object as it is drawn from one location to the next (Huerta and Rabinovich, 2004; Friston et al., 2011). This style of dynamical modeling for active inference has been exploited in the context of a 2-dimensional visual search task (Friston et al., 2017a), and in control of arm movements in 3-dimensions (Parr et al., 2021).

The Basal Ganglia

In thinking about the problem of where to look, we must consider a set of subcortical nuclei known to play an important role in planning (Jahanshahi et al., 2015). The basal ganglia receive input from much of the cerebral cortex and provide output to the superior colliculus, among other structures. This means they are well-positioned to evaluate alternative action plans based upon the beliefs represented by the cortex, and to modulate the cortical projections to the colliculus to bring about the most likely eye movements. As such, these nuclei have frequently been associated with inferences about what to do in the process theories associated with active inference (Friston et al., 2017a,b; Parr and Friston, 2018b).

What makes one eye-movement better than another? One way to think about this is to frame the problem as one of experimental design (Itti and Koch, 2000; Friston et al., 2012). The best experiments (or eye movements) are those

[3]This assumption of conjugacy may underwrite internuclear ophthalmoplegia. This is a syndrome—caused by brainstem demyelination or stroke—in which the predictions required for one eye to move towards the nose (while the other moves away from it) are interrupted. This violation of the conjugacy assumption has consequences for the contralateral eye, which exhibits a pathological oscillatory nystagmus.

[4]The superior colliculus exhibits a log-polar retinotopy which implies the x^{v} variable might be represented in this coordinate system. The functional relevance of this is that the probability density for x^{v}, when translated into polar or Cartesian coordinates, will assign higher variance to more eccentric values. This has been proposed as an explanation for the increased variance of saccadic endpoints for more eccentric locations in a Cartesian frame, despite unform variance in log-polar reference frames (Daucé and Perrinet, 2020).

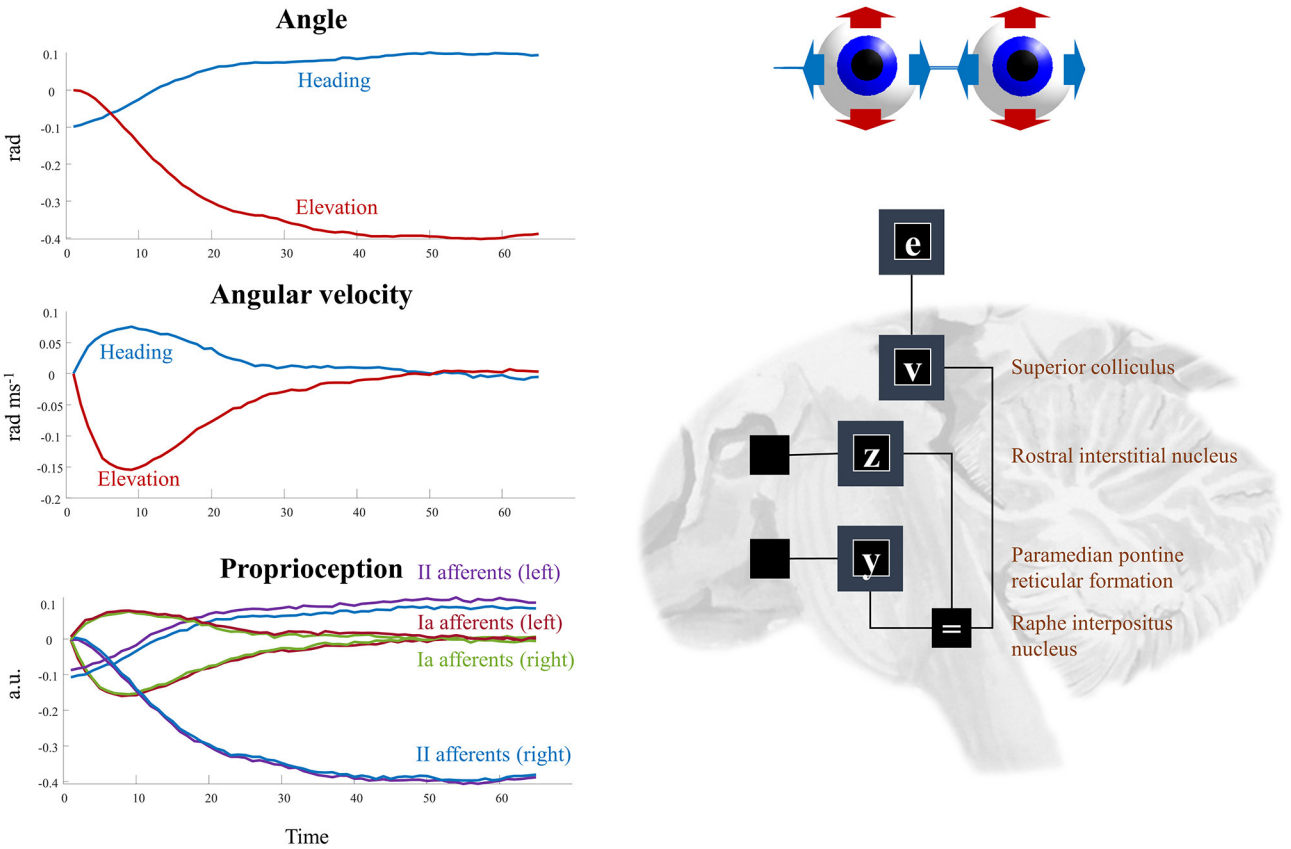


FIGURE 6 | Oculomotion. The plots on the left of this figure show an example of the kinds of dynamics that result from Equations (9–12). These illustrate a single saccade toward some equilibrium point determined by factor **v**. The first two plots detail the hidden states, comprising the heading angle of the eyes, their elevation, and the rates of change of each of these. In addition, the third plot illustrates the proprioceptive data we might expect these dynamics to generate. These are divided into sensory neurons that report instantaneous muscle tendon stretch (II afferents) and those that report changes in this (Ia afferents) for the right and left eye. Note that these differ only in the heading angle—as eye movements are congruent. The constant discrepancy in the heading angle results from the angle of convergence of the eyes. On the right, the factors are arranged to be consistent with the brainstem structures that deal with saccades in the vertical (factor **z**) and horizontal (factor **y**) directions. The proprioceptive signal is expressed in arbitrary units (a.u.) which could be converted to firing rates with the appropriate (e.g., sigmoidal) transforms.

that maximize expected information gain[5]—i.e., the mutual information (Lindley, 1956) between data (y) and hypotheses or causes (x) under some design or policy (π):

$$\begin{aligned} \mathbb{I}[X, Y|\pi] &= D_{KL}\left[P(x, y|\pi)||P(x|\pi)P(y|\pi)\right] \\ &= \mathbb{E}_{P(y|\pi)}\underbrace{\left[D_{KL}\left[P(x|y, \pi)||P(x|\pi)\right]\right]}_{\text{Information gain}} \\ &= \underbrace{H\left[P(y|\pi)\right]}_{\text{Predictive Entropy}} - \underbrace{\mathbb{E}_{P(x|\pi)}\left[H\left[P(y|x, \pi)\right]\right]}_{\text{Expected Ambiguity}} \end{aligned} \tag{13}$$

Equation (13) shows three different expressions of the mutual information, incorporating KL-Divergences—quantifying how different two distributions are from one another—and entropies. An entropy (H) is a measure of the dispersion or uncertainty associated with a probability distribution. The first line says that the expected information gain is greatest when the joint distribution of data and their causes, under a given policy, is very different from the product of the two marginal distributions. The second line expresses this in terms of the expected update from prior to posterior—i.e., the information gain. The third line breaks this down into two components. These are easiest to understand when thinking about what makes a good experiment. The first thing is that it should tell us something we do not already know. An experiment for which we can already confidently predict our measurements is a poor experiment. Such experiments are penalized by the predictive entropy term, which favors those experiments for which the predicted measurements are maximally uncertain, i.e., not known beforehand.

Figure 7 illustrates the relevance of the predictive entropy in adjudicating between alternative fields of view. This shows two (of many) possible head-directions and the visual input this generates in each of the three rooms shown in **Figure 2**. Imagine we are uncertain about the room we occupy, but relatively

[5]From the perspective of active inference, this is normally augmented with an additional distribution that ascribes greater probability to preferred datapoints, turning the mutual information into an expected free energy. However, we focus upon information seeking specifically, under the assumption that eye movements are primarily exploratory (i.e., preferences over visual data are uniform). This is a special case of an expected free energy.

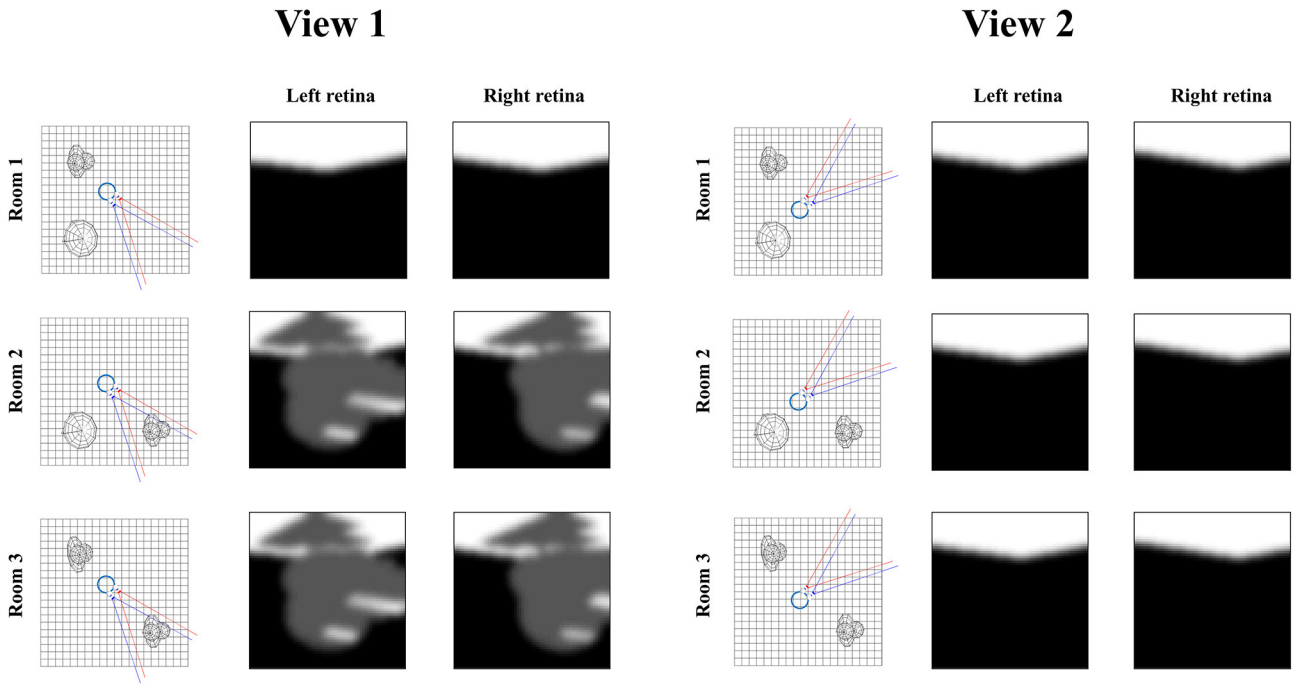


FIGURE 7 | Expected information gain. This figure highlights the role of predictive entropies in adjudicating between salient actions. This shows two alternative fields of view we could choose between, through making eye or head movements. If we were not sure which room we were in, view 1 (toward the southeast corner) would be associated with a high predictive entropy, and is consequently useful in resolving uncertainty. In contrast, view 2 (toward the northeast corner) has zero predictive entropy, and does not help distinguish between rooms.

confident about everything else. View 1 could give rise to a view with no object, or with object 1. We can be confident that view 2 will always lead to a view with no object, as none of the three rooms have an object in this location. Any actions leading to view 1 (by moving eyes or head) will be associated with a higher predictive entropy than actions leading to view 2 (zero entropy). Intuitively this is sensible, as we will be able to tell from the consequences of view 1 whether we are in room 1, or in room 2 or 3. We will gain no information about the room from view 2. Once we have seen object 1 in view 1, we know we are in room 2 or 3, and there is no added information available in this view. We would always anticipate seeing the same thing here. At this point, the southwest or northwest corners of the room may become more salient, allowing disambiguation between the rooms that are still plausible.

The expected ambiguity term in Equation (13) expresses the fact that, even if sensory input is unpredictable, it is not necessarily useful. Everything else being equal, expected ambiguity underwrites the imperative to sample precise and unambiguous visual sensations. Perhaps the simplest example is keeping our eyes open. When our eyes are closed (or the lights are off), the probability of every retinal cell firing is roughly the same, which corresponds to a maximally ambiguous state of affairs.

The basal ganglia appear to be key in quantifying information gain (Sheth et al., 2011; White et al., 2019). However, they are part of a broader network of regions involved in making these decisions. This is important, in the sense that information gain is a functional (function of a function) of beliefs. As such, the broad range of inputs to the basal ganglia from the cortex and elsewhere may give them access to these beliefs across different modalities. This is evidenced by disorders of salience attribution, like sensory neglect syndromes (Husain et al., 2001; Fruhmann Berger et al., 2008; Parr and Friston, 2017a)—which occur with lesions to the superior longitudinal fasciculus (c.f., **Figure 3**) (Bartolomeo et al., 2007, 2012) in addition to basal ganglia structures (Karnath et al., 2002). In the context of active vision, at least, the basal ganglia appear to be the point at which the most epistemically valuable saccadic movements are determined, given the direct influence of this subcortical network over the superior colliculus (Hikosaka and Wurtz, 1983).

RELATED WORK

While we have focused upon the sort of generative model the brain could employ, we have neglected the question as to how a model of this sort might develop in the first place. Prominent approaches to learning of such models from machine vision include capsule networks (Sabour et al., 2017) and the Generative Query Network (GQN) (Eslami et al., 2018). The former is a supervised learning technique in which capsules, groups of neurons representing attributes of an entity causing visual data, optimize their connections between multiple convolutional layers to associate images with their labels. The latter is an unsupervised learning approach—reminiscent of a variational autoencoder (Kingma and Welling, 2013; An and Cho, 2015)—that learns two functions. The first is a function

from observations to a representation of a scene and the second is a generative function that predicts observations, in a viewpoint-dependent manner, under the current scene representation. The two are jointly optimized based upon the fidelity with which observations are predicted given the scene representation. While unsupervised in the sense that no labeled training data are used, this approach could be viewed as supervised learning of a function from viewpoint to visual data.

There are important shared features between the generative model presented in this paper and those that emerge from training capsule networks or the GQN. Perhaps the most striking is the importance of factorization. In capsule networks, factors are an integral part of the network. Each neuron within a capsule represents distinct features in relation to other neurons. This allows a capsule—representing a given object—to represent that object in multiple orientations, or colors. In the GQN, factorization emerges from training on environments in which different attributes can vary independently. For instance, training on views of red cubes, red triangles, and blue spheres enables reconstruction of, previously unobserved, red spheres. In this paper, we have highlighted the factorization of different explanatory variables (i.e., latent causes) that manifest in different visual streams—for instance, changing our viewpoint does not change object identity, and vice versa.

A second shared feature is the increase in the spatial scale of receptive fields, as we move from observations to their causes. In capsule networks, this arises from their convolutional architecture. In our generative model, the convergence of high dimensional pixel spaces through to hidden layers with fewer and fewer units is represented, in reverse, by the generation of objects from scenes, surfaces from objects, and pixel intensities from surfaces.

Given that there are successful machine learning approaches available—that effectively learn the structure of a generative model for visual rendering—it would be reasonable to ask what is added by the approach pursued here. In short, the benefit is transparency, in the sense of both explainability and interpretability (Marcinkevičs and Vogt, 2020). The benefits of approaches based upon deep learning are that they scale well, and that the models they learn emerge from the statistical regularities in the data on which they are trained. However, the interpretability of the resulting models is not always straightforward. In contrast, specifying an explicit generative model affords an explicit interpretation of the ensuing inferences. This may not matter when developing new approaches to visual rendering but is crucial in advancing hypotheses as to how the brain (and other sentient artifacts) solves active vision problems. The account advanced in this paper is not designed to replace machine learning but offers an example of the kind of generative model they might implicitly learn.

DISCUSSION

In this paper, we set out a generative model capable of generating simple retinal images. Our aim was to determine the set of explanatory variables the brain could call upon to explain these visual data, the dependencies between these variables, and the anatomical connectivity that could support the requisite neuronal message passing. In other words, we sought to identify the problem the visual brain must solve. From a neurobiological perspective, one conclusion we could draw from this analysis is that few parts of the brain are not involved in active vision.

We have seen how beliefs about scenes, and the objects in those scenes, thought to be represented in the temporal lobe, are combined with beliefs about the retinal location. The latter depend upon the parietal cortices and their relationship with medial temporal and frontal lobe structures. If we know the retinal location and the set of surfaces in a scene, we can compute which surfaces lie within our field of view and determine (for a given light source) the influence of those surfaces on retinal cells. This is the retinocortical pathway in reverse. Explanations of visual data afforded by a model of this sort are highly sensitive to where the retina is. This means part of the explanation must always include our choices about where we position our retina. Central to this is the computation of expected information gain, which implicates the oculomotor loops of the basal ganglia. In addition, the process of acting to change our eye (or head) position—when viewed as an inference problem—requires that we predict all of the sensory consequences of the action we hope to execute. We detailed how this could play out in the oculomotor brainstem, predicting the proprioceptive data we hope to realize.

Clearly, there are limitations to the model presented here, and many aspects of vision that are not accounted for. It is useful to consider how these could be incorporated in this generative model. First, there are other ways, in addition to moving our eyes, in which we can influence our visual environment. For instance, we could move our hands in our field of view (Limanowski and Friston, 2020). We could go further and move objects around in the environment or assume that other agents can do so. This means unfolding the prior beliefs from **Figure 1** in time, such that they factorize into a series of policy-dependent transition probabilities. Time-dependence adds an interesting twist to the expected information gain, as it means that the posterior predictive entropy grows over time for unobserved locations. The reason for this is simple. The longer the time since looking in each location, the greater the probability that something has changed. This is consistent with Jaynes' maximum entropy principle (Jaynes, 1957). The result is a form of inhibition of return (Posner et al., 1985), the duration of which varies with the precision of probabilistic transitions over time (Parr and Friston, 2017b). The duration of this inhibition of return is one of the crucial differences between static and dynamic environments: reflecting the possibility that things have changed since each location was last fixated. This engenders loss of confidence about state of affairs at that location—and an epistemic affordance of return that increases with time. This relates to other visual phenomena, even in the absence of overt eye movements. Periodic redirection of covert attention—a form of mental action (Rizzolatti et al., 1987; Hohwy, 2012; Limanowski and Friston, 2018)—based upon the accumulated uncertainty of unattended features reproduces binocular rivalry phenomena (Parr et al., 2019), in which perception alternates between different images presented to each eye (Leopold and Logothetis, 1999; Hohwy et al., 2008).

We have omitted interesting questions about texture and color vision. Textured surfaces could be modeled through varying the constants (c_1, c_2, c_3) from Equation (6) and the ambient lighting (α) as functions of their location on a surface. Color vision could be incorporated simply by repeating section The Retinocortical Pathway for several different wavelengths of light—specifically, the red, green, and blue wavelengths detected by different cone photoreceptors (Nathans et al., 1986). This would aid in disambiguating the roles of magnocellular and parvocellular streams, involved in dissociable aspects of trichromatic and monochromatic vision (Masri et al., 2020). The magnocellular stream also seems to have a key role in detecting motion (Merigan et al., 1991) – something that is highly relevant in the context of active event recognition (Ognibene and Demiris, 2013).

From a computational perspective, there are important outstanding questions about the role of precision (i.e., neuromodulation) which may involve second order thalamic nuclei, like the pulvinar (Kanai et al., 2015), and the cholinergic basal nucleus of Meynert (Moran et al., 2013). These could be accommodated in this model through including prior beliefs about the precision or variance associated with regions of the visual field. This may be particularly relevant in understanding how subcortical structures participate in visual perception. For instance, the role of the amygdala in enhancing the perception of fearful faces (Pessoa et al., 2006; Adolphs, 2008) could be formulated as inferences about the precision of visual features consistent with this emotional state. Another important computational feature was omitted in our discussion of models of oculomotion. We neglected to mention the role of generalized coordinates of motion (acceleration, jerk and higher order temporal derivatives) (Friston et al., 2010), which offer a local approximation to the trajectory of dynamical variables, as opposed to an instantaneous value. This has important implications for things like sensorimotor delays (Perrinet et al., 2014), accounting for small discrepancies in the time the brainstem receives a proprioceptive signal compared to the time an oculomotor muscle contracted. In brief, representations of the local trajectory enable projections into the immediate past or future. To see how generalized coordinates of motion can be incorporated into a factor graph, see (Friston et al., 2017a).

Why is it useful to formulate a generative model of active vision? There are several answers to this question. The first is that having a forward model is the first step in designing an inference scheme that inverts the model. This is a matter of undoing everything that was done to generate visual data, so that their causes can be revealed. There have been promising advances in practical, scalable, model inversion for active vision from a robotics perspective, that use deep neural networks to learn a generative model that predicts camera images (Çatal et al., 2020), leading to Bayes optimal behavior in a real environment. Similar approaches have been developed both in the visual domain (Fountas et al., 2020; van der Himst and Lanillos, 2020), and in a generic (non-visual) control setting, which may also have applications for high-dimensional visual data (Tschantz et al., 2020). By treating vision as active, we can design agents that actively sample the environment to resolve their uncertainty, in high-dimensional, incongruent settings. This takes us beyond static deep learning models which, although apt at simple classification tasks (LeCun and Bengio, 1995; Jin et al., 2017), are unable to handle the complexity involved in human active vision.

The second is that this model generates behavior (i.e., saccades). As we highlighted in section The Basal Ganglia, the saccades performed depend upon prior beliefs. This means measured eye movements could be used to draw inferences about the parameters of prior beliefs in the model used by an experimental participant, or clinical patient (Mirza et al., 2018; Cullen et al., 2020). Virtual reality technologies offer a useful way to investigate this, with tight control over the visual environment combined with eye-tracking (Limanowski et al., 2017; Harris et al., 2020a,b). In principle, we could present visual data consistent with the generative model set out here and use this to test hypotheses about the structure of the generative model used by the brain, or about the parameters of each factor. One such hypothesis as to the anatomical implementation has been set out in the figures. However, it is important to recognize that this is one of many hypotheses that could have been advanced. Crucially, a generative model for visual data allows us to generate stimuli that vary according to specific hidden causes. This would allow for alternative anatomical hypotheses to be evaluated through neuroimaging, as we would anticipate variation in a given hidden state should lead to variation in beliefs about this state, and changes in neural activity—i.e., belief updating—in those regions representing these beliefs.

The third utility of forward models of this sort is that understanding the conditional dependencies in a model, and by implication the structure of the neuronal message passing that solves the model, we have an opportunity to frame questions about classical disconnection syndromes (Geschwind, 1965a,b) in functional (computational) terms (Sajid et al., 2020). We have briefly touched upon some of these syndromes, including visual field defects, agnosia, and neglect. Generative models of active vision let us express the mechanisms that underwrite these syndromes in the same formal language—that of aberrant prior beliefs. This approach is commonly used to characterize inferential pathologies in computational psychiatry (Adams et al., 2015).

CONCLUSION

Under modern approaches to theoretical neurobiology—including active inference—brain function is understood in terms of the problems it solves. Its biology recapitulates the structure of this problem. In this paper, we have attempted to define the problem faced by the active visual system. This is framed as explaining visual input, where good explanations involve not just the external environment, but how we choose to position our sensors (i.e., retinas) in that environment. This explanation takes the form of a predictive model comprising factors that determine the geometry of objects expected in a given room, the placement of the retina in that room, and the combination of these variables in generating a retinal image. The factors involved in determining the placement of the retina can be further unpacked in terms of their causes—i.e., the most

epistemically rich saccades—and their consequences for the dynamics of, and proprioceptive inputs from, the eyes. We hope that this paper provides a useful reference that brings together the probabilistic models required for aspects of biological active vision.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://github.com/tejparr/Generative-Models-Active-Vision.

AUTHOR CONTRIBUTIONS

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

FUNDING

LD was supported by the Fonds National de la Recherche, Luxembourg (Project code: 13568875). NS was supported by the Medical Research Council (MR/S502522/1). KF is a Wellcome Principal Research Fellow (Ref: 088130/Z/09/Z).

REFERENCES

Abu-Akel, A., and Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia* 49, 2971–2984. doi: 10.1016/j.neuropsychologia.2011.07.012

Adams, R. A., Huys, Q. J., and Roiser, J. P. (2015). Computational Psychiatry: towards a mathematically informed understanding of mental illness. *J. Neurol. Neurosurg. Psychiatry* 87, 53–63. doi: 10.1136/jnnp-2015-310737

Adler, A. (1944). Disintegration and restoration of optic recognition in visual agnosia: analysis of a case. *Arch. Neurol. Psychiatry* 51, 243–259. doi: 10.1001/archneurpsyc.1944.02290270032004

Adolphs, R. (2008). Fear, faces, and the human amygdala. *Curr. Opin. Neurobiol.* 18, 166–172. doi: 10.1016/j.conb.2008.06.006

Aguirre, G. K., and D'Esposito, M. (1999). Topographical disorientation: a synthesis and taxonomy. *Brain* 122, 1613–1628. doi: 10.1093/brain/122.9.1613

An, J., and Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. *Special Lect. IE* 2, 1–18. Available online at: http://dm.snu.ac.kr/static/docs/TR/SNUDM-TR-2015-03.pdf

Andreopoulos, A., and Tsotsos, J. K. (2013). A computational learning theory of active object recognition under uncertainty. *Int. J. Comput. Vis.* 101, 95–142. doi: 10.1007/s11263-012-0551-6

Andresen, D. R., Vinberg, J., and Grill-Spector, K. (2009). The representation of object viewpoint in human visual cortex. *Neuroimage* 45, 522–536. doi: 10.1016/j.neuroimage.2008.11.009

Arad, N., Dyn, N., Reisfeld, D., and Yeshurun, Y. (1994). Image warping by radial basis functions: application to facial expressions. *CVGIP Graphical Models Image Process.* 56, 161–172. doi: 10.1006/cgip.1994.1015

Ashburner, J., and Friston, K. J. (1999). Nonlinear spatial normalization using basis functions. *Hum. Brain Mapp.* 7, 254–266. doi: 10.1002/(SICI)1097-0193(1999)7:4<254::AID-HBM4>3.0.CO;2-G

Auger, S. D., Mullally, S. L., and Maguire, E. A. (2012). Retrosplenial cortex codes for permanent landmarks. *PLoS ONE* 7:e43620. doi: 10.1371/journal.pone.0043620

Ballard, D. H., and Zhang, R. (2020). The hierarchical evolution in human vision modeling. *Trends Cogn. Sci.* Available online at: https://www.cs.utexas.edu/~zharucs/publications/2020_TiCS_hier.pdf

Bartolomeo, P. M., Thiebaut de Schotten, M., and Chica, A. B. (2012). Brain networks of visuospatial attention and their disruption in visual neglect. *Front. Hum. Neurosci.* 6:110. doi: 10.3389/fnhum.2012.00110

Bartolomeo, P. M., Thiebaut de Schotten, M., and Doricchi, F. (2007). Left unilateral neglect as a disconnection syndrome. *Cereb. Cortex* 17, 2479–2490. doi: 10.1093/cercor/bhl181

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, M., Oostenveld, R., Dowdall, J. R., et al. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron* 85, 390–401. doi: 10.1016/j.neuron.2014.12.018

Baumgart, B. G. (1975). "A polyhedron representation for computer vision," in *Proceedings of the May 19-22, 1975, National Computer Conference and Exposition* (New York, NY), 589–596. doi: 10.1145/1499949.1500071

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. London: University of London.

Benson, D. F., and Greenberg, J. P. (1969). Visual form agnosia: a specific defect in visual discrimination. *Arch. Neurol.* 20, 82–89. doi: 10.1001/archneur.1969.00480070092010

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychol. Rev.* 94:115. doi: 10.1037/0033-295X.94.2.115

Blair, H. T., Cho, J., and Sharp, P. E. (1998). Role of the lateral mammillary nucleus in the rat head direction circuit: a combined single unit recording and lesion study. *Neuron* 21, 1387–1397. doi: 10.1016/S0896-6273(00)80657-1

Blanke, O., Ortigue, S., Landis, T., and Seeck, M. (2002). Stimulating illusory own-body perceptions. *Nature* 419, 269–270. doi: 10.1038/419269a

Blinn, J. F. (1977). "Models of light reflection for computer synthesized pictures," in *Proceedings of the 4th Annual Conference on Computer Graphics and Interactive Techniques* (San Jose, CA: Association for Computing Machinery), 192–198. doi: 10.1145/563858.563893

Botvinick, M., and Toussaint, M. (2012). Planning as inference. *Trends Cogn. Sci.* 16, 485–488. doi: 10.1016/j.tics.2012.08.006

Bruce, C. J., Goldberg, M. E., Bushnell, M. C., and Stanton, G. B. (1985). Primate frontal eye fields. II Physiological and anatomical correlates of electrically evoked eye movements. *J. Neurophysiol.* 54, 714–734. doi: 10.1152/jn.1985.54.3.714

Büttner-Ennever, J., and Büttner, U. (1978). A cell group associated with vertical eye movements in the rostral mesencephalic reticular formation of the monkey. *Brain Res.* 151, 31–47. doi: 10.1016/0006-8993(78)90948-4

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., et al. (2005). Do we know what the early visual system does? *J. Neurosci.* 25, 10577–10597. doi: 10.1523/JNEUROSCI.3726-05.2005

Çatal, O., Wauthier, S., Verbelen, T., De Boom, C., and Dhoedt, B. (2020). Deep active inference for autonomous robot navigation. *arXiv [Preprint] arXiv*:2003.03220.

Clark, B. J., Bassett, J. P., Wang, S. S., and Taube, J. S. (2010). Impaired head direction cell representation in the anterodorsal thalamus after lesions of the retrosplenial cortex. *J. Neurosci.* 30:5289. doi: 10.1523/JNEUROSCI.3380-09.2010

Cooper, L. A., and Shepard, R. N. (1973). "Chronometric studies of the rotation of mental images," in *Proceedings of the Eighth Annual Carnegie Symposium on Cognition* (Pittsburgh, PA: Carnegie-Mellon University), 75–176. doi: 10.1016/B978-0-12-170150-5.50009-3

Cooper, S., Daniel, P., and Whitteridge, D. (1951). Afferent impulses in the oculomotor nerve, from the extrinsic eye muscles. *J. Physiol.* 113, 463. doi: 10.1113/jphysiol.1951.sp004588

Cooper, S., and Daniel, P. M. (1949). Muscle spindles in human extrinsic eye muscles. *Brain* 72, 1–24. doi: 10.1093/brain/72.1.1

Corbetta, M., and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nat. Rev. Neurosci.* 3, 201–215. doi: 10.1038/nrn755

Crick, F. H., Marr, D. C., and Poggio, T. (1980). *An Information Processing Approach to Understanding the Visual Cortex*. MIT Libraries.

Cullen, M., Monney, J., Mirza, M. B., and Moran, R. (2020). A meta-bayesian model of intentional visual search. *arXiv [Preprint] arXiv:* 2006.03531.

Da Costa, L., Parr, T., Sajid, N., Veselic, S., Neacsu, V., and Friston, K. (2020). Active inference on discrete state-spaces: a synthesis. *J. Math. Psychol.* 99:102447. doi: 10.1016/j.jmp.2020.102447

Daucé, E., and Perrinet, L. (2020). "Visual search as active inference," in *International Workshop on Active Inference* (Ghent: Springer). doi: 10.1007/978-3-030-64919-7_17

Dauwels, J. (2007). "On variational message passing on factor graphs," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on*, *IEEE* (Nice). doi: 10.1109/ISIT.2007.4557602

de Vries, B., and Friston, K. J. (2017). A factor graph description of deep temporal active inference. *Front. Comput. Neurosci.* 11:95. doi: 10.3389/fncom.2017.00095

Deco, G., and Rolls, E. T. (2004). A Neurodynamical cortical model of visual attention and invariant object recognition. *Vis. Res.* 44, 621–642. doi: 10.1016/j.visres.2003.09.037

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron* 73, 415–434. doi: 10.1016/j.neuron.2012.01.010

Doya, K. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. Cambridge, MA: MIT press. doi: 10.7551/mitpress/9780262042383.001.0001

Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area: recognition, navigation, or encoding? *Neuron* 23, 115–125. doi: 10.1016/S0896-6273(00)80758-8

Eslami, S. M. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., et al. (2018). Neural scene representation and rendering. *Science* 360:1204. doi: 10.1126/science.aar6170

Feldman, A. G., and Levin, M. F. (2009). "The equilibrium-point hypothesis – past, present and future," in *Progress in Motor Control: A Multidisciplinary Perspective*, ed D. Sternad (Boston, MA: Springer), 699–726. doi: 10.1007/978-0-387-77064-2_38

Ferro, M., Ognibene, D., Pezzulo, G., and Pirrelli, V. (2010). Reading as active sensing: a computational model of gaze planning during word recognition. *Front. Neurorobot.* 4:6. doi: 10.3389/fnbot.2010.00006

Forney, G. D. Jr., and Vontobel, P. O. (2011). Partition functions of normal factor graphs. *arXiv [Preprint] arXiv:1102.0316*.

Fountas, Z., Sajid, N., Mediano, P. A., and Friston, K. (2020). Deep active inference agents using Monte-Carlo methods. *arXiv [Preprint] arXiv:2006.04176*.

Frey, B. J., and MacKay, D. J. C. (1998). "A revolution: belief propagation in graphs with cycles," in *Proceedings of the 1997 conference on Advances in Neural Information Processing Systems 10* (Denver, CO: MIT Press), p. 479–485.

Friston, K., Adams, R. A., Perrinet, L., and Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Front. Psychol.* 3:151. doi: 10.3389/fpsyg.2012.00151

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., and Pezzulo, G. (2017). Active inference: a process theory. *Neural Comput.* 29, 1–49. doi: 10.1162/NECO_a_00912

Friston, K., Mattout, J., and Kilner, J. (2011). Action understanding and active inference. *Biol. Cybern.* 104, 137–160. doi: 10.1007/s00422-011-0424-z

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., Stephan, K., Li, B., and Daunizeau, J. (2010). Generalised filtering. *Math. Probl. Eng.* 2010:621670. doi: 10.1155/2010/621670

Friston, K. J. (2019). Waves of prediction. *PLoS Biol.* 17:e3000426. doi: 10.1371/journal.pbio.3000426

Friston, K. J., Parr, T., and de Vries, B. (2017a). The graphical brain: belief propagation and active inference. *Netw. Neurosci.* 1, 381–414. doi: 10.1162/NETN_a_00018

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017b). Deep temporal models and active inference. *Neurosci. Biobehav. Rev.* 77, 388–402. doi: 10.1016/j.neubiorev.2017.04.009

Fruhmann Berger, M., Johannsen, L., and, Karnath, H.-O. (2008). Time course of eye and head deviation in spatial neglect. *Neuropsychology* 22, 697–702. doi: 10.1037/a0013351

Gandhi, N. J., and Keller, E. L. (1997). Spatial distribution and discharge characteristics of superior colliculus neurons antidromically activated from the omnipause region in monkey. *J. Neurophysiol.* 78, 2221–2225. doi: 10.1152/jn.1997.78.4.2221

Gauthier, I., Hayward, W. G., Tarr, M. J., Anderson, A. W., Skudlarski, P., and Gore, J. C. (2002). BOLD activity during mental rotation and viewpoint-dependent object recognition. *Neuron* 34, 161–171. doi: 10.1016/S0896-6273(02)00622-0

Geschwind, N. (1965a). Disconnexion syndromes in animals and man. I. *Brain* 88, 237–237. doi: 10.1093/brain/88.2.237

Geschwind, N. (1965b). Disconnexion syndromes in animals and man. II. *Brain* 88:585. doi: 10.1093/brain/88.3.585

Goodale, M. A., and Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends Neurosci.* 15, 20–25. doi: 10.1016/0166-2236(92)90344-8

Goodale, M. A., Milner, A. D., Jakobson, L. S., and Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature* 349, 154–156. doi: 10.1038/349154a0

Greene, J. D. W. (2005). Apraxia, agnosias, and higher visual function abnormalities. *J. Neurol. Neurosur. Psychiatry* 76(Suppl. 5):v25. doi: 10.1136/jnnp.2005.081885

Gregory, R. L. (1968). Perceptual illusions and brain models. *Proc. R Soc. Lond B.* 171:179–196. doi: 10.1098/rspb.1968.0071

Gregory, R. L. (1980). Perceptions as hypotheses. *Phil. Trans. R. Soc. Lond. B.* 290, 181–197. doi: 10.1098/rstb.1980.0090

Grill-Spector, K., and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nat. Rev. Neurosci.* 15, 536–548. doi: 10.1038/nrn3747

Hanes, D. P., and Wurtz, R. H. (2001). Interaction of the frontal eye field and superior colliculus for saccade generation. *J. Neurophysiol.* 85, 804–815. doi: 10.1152/jn.2001.85.2.804

Harris, D. J., Buckingham, G., Wilson, M. R., Brookes, J., Mushtaq, F., Mon-Williams, M., et al. (2020a). The effect of a virtual reality environment on gaze behaviour and motor skill learning. *Psychol. Sport Exerc.* 50:101721. doi: 10.1016/j.psychsport.2020.101721

Harris, D. J., Buckingham, G., Wilson, M. R., Brookes, J., Mushtaq, F., Mon-Williams, M., et al. (2020b). Exploring sensorimotor performance and user experience within a virtual reality golf putting simulator. *Virtual Real.* doi: 10.1007/s10055-020-00480-4

Hassabis, D., and Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends Cogn. Sci.* 11, 299–306. doi: 10.1016/j.tics.2007.05.001

Hegdé, J., and Van Essen, D. C. (2007). A comparative study of shape representation in macaque visual areas V2 and V4. *Cereb. Cortex* 17, 1100–1116. doi: 10.1093/cercor/bhl020

Helmholtz, H. (1878 (1971)). *The Facts of Perception. The Selected Writings of Hermann von Helmholtz*. R. K. Middletown. Connecticut: Wesleyan University Press. 384.

Hikosaka, O., and Wurtz, R. H. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. IV. Relation of substantia nigra to superior colliculus. *J. Neurophysiol.* 49, 1285–1301. doi: 10.1152/jn.1983.49.5.1285

Hohwy, J. (2012). Attention and conscious perception in the hypothesis testing brain. *Front. Psychol.* 3:96. doi: 10.3389/fpsyg.2012.00096

Hohwy, J. (2016). The self-evidencing brain. *Noûs* 50, 259–285. doi: 10.1111/nous.12062

Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition* 108, 687–701. doi: 10.1016/j.cognition.2008.05.010

Hosoya, T., Baccus, S. A., and Meister, M. (2005). Dynamic predictive coding by the retina. *Nature* 436, 71–77. doi: 10.1038/nature03689

Hubel, D. H., and Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* 148, 574–591. doi: 10.1113/jphysiol.1959.sp006308

Huerta, R., and Rabinovich, M. (2004). Reproducible sequence generation in random neural ensembles. *Phys. Rev. Lett.* 93:238104. doi: 10.1103/PhysRevLett.93.238104

Husain, M., Mannan, S., Hodgson, T., Wojciulik, E., Driver, J., and Kennard, C. (2001). Impaired spatial working memory across saccades contributes to abnormal search in parietal neglect. *Brain* 124, 941–952. doi: 10.1093/brain/124.5.941

Ito, M., Tamura, H., Fujita, I., and Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophysiol.* 73, 218–226. doi: 10.1152/jn.1995.73.1.218

Itti, L., and Baldi, P. (2006). Bayesian surprise attracts human attention. *Adv. Neural Inf. Process. Syst.* 18:547. Available online at: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.4948&rep=rep1&type=pdf

Itti, L., and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vis. Res.* 40, 1489–1506. doi: 10.1016/S0042-6989(99)00163-7

Jahanshahi, M., Obeso, I., Rothwell, J. C., and Obeso, J. A. (2015). A fronto-striato-subthalamic-pallidal network for goal-directed and habitual inhibition. *Nat. Rev. Neurosci.* 16, 719–732. doi: 10.1038/nrn4038

Javier Traver, V., and Bernardino, A. (2010). A review of log-polar imaging for visual perception in robotics. *Rob. Auton. Syst.* 58, 378–398. doi: 10.1016/j.robot.2009.10.002

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev. II* 106, 620–630. doi: 10.1103/PhysRev.106.620

Jin, K. H., McCann, M. T., Froustey, E., and Unser, M. (2017). Deep convolutional neural network for inverse problems in imaging. *IEEE Trans. Image Process.* 26, 4509–4522. doi: 10.1109/TIP.2017.2713099

Kajiya, J. T. (1986). The rendering equation. *SIGGRAPH Comput. Graph.* 20, 143–150. doi: 10.1145/15886.15902

Kanai, R., Komura, Y., Shipp, S., and Friston, K. (2015). Cerebral hierarchies: predictive processing, precision and the pulvinar. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140169. doi: 10.1098/rstb.2014.0169

Karnath, H. O., Himmelbach, M., and Rorden, C. (2002). The subcortical anatomy of human spatial neglect: putamen, caudate nucleus and pulvinar. *Brain* 125, 350–360. doi: 10.1093/brain/awf032

Katz, H. K., Lustig, A., Lev-Ari, T., Nov, Y., Rivlin, E., and Katzir, G. (2015). Eye movements in chameleons are not truly independent – evidence from simultaneous monocular tracking of two targets. *J. Exp. Biol.* 218:2097. doi: 10.1242/jeb.113084

Kingma, D. P., and Welling, M. (2013). Auto-encoding variational bayes. *arXiv [Preprint] arXiv:1312.6114*.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* 17, 26–49. doi: 10.1016/j.tics.2012.10.011

Künzle, H., and Akert, K. (1977). Efferent connections of cortical, area 8 (frontal eye field) in Macaca fascicularis. A reinvestigation using the autoradiographic technique. *J. Comp. Neurol.* 173, 147–164. doi: 10.1002/cne.901730108

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J.-P., Baciu, M., Kahane, P., et al. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. *Commun. Biol.* 1:107. doi: 10.1038/s42003-018-0110-y

Laar, T. V. D., and Vries, B. D. (2016). A probabilistic modeling approach to hearing loss compensation. *IEEE/ACM Trans. Audio Speech Lang. Process.* 24, 2200–2213. doi: 10.1109/TASLP.2016.2599275

LeCun, Y., and Bengio, Y. (1995). "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, ed M. A. Arbib (Cambridge, MA: MIT press) 255–258.

Lee, T. S., and Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *JOSA A* 20, 1434–1448. doi: 10.1364/JOSAA.20.001434

Leopold, D. A., and Logothetis, N. K. (1999). Multistable phenomena: changing views in perception. *Trends Cogn. Sci.* 3, 254–264. doi: 10.1016/S1364-6613(99)01332-7

Limanowski, J., and Friston, K. (2018). 'Seeing the dark': grounding phenomenal transparency and opacity in precision estimation for active inference. *Front. Psychol.* 9:643. doi: 10.3389/fpsyg.2018.00643

Limanowski, J., and Friston, K. (2020). Active inference under visuo-proprioceptive conflict: simulation and empirical results. *Sci. Rep.* 10:4010. doi: 10.1038/s41598-020-61097-w

Limanowski, J., Kirilina, E., and Blankenburg, F. (2017). Neuronal correlates of continuous manual tracking under varying visual movement feedback in a virtual reality environment. *Neuroimage* 146, 81–89. doi: 10.1016/j.neuroimage.2016.11.009

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* 27, 986–1005. doi: 10.1214/aoms/1177728069

Loeliger, H. (2004). An introduction to factor graphs. *IEEE Signal Process. Mag.* 21, 28–41. doi: 10.1109/MSP.2004.1267047

Loeliger, H. A., Dauwels, J., Hu, J., Korl, S., Ping, L., and Kschischang, F. R. (2007). The factor graph approach to model-based signal processing. *Proc. IEEE* 95, 1295–1322. doi: 10.1109/JPROC.2007.896497

Logothetis, N. K., and Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci.* 19, 577–621. doi: 10.1146/annurev.ne.19.030196.003045

Lueck, C. J. (2010). Loss of vision. *Pract. Neurol.* 10:315. doi: 10.1136/jnnp.2010.223677

Lukas, J. R., Aigner, M., Blumer, R., Heinzl, H., and Mayr, R. (1994). Number and distribution of neuromuscular spindles in human extraocular muscles. *Invest. Ophthalmol. Vis. Sci.* 35, 4317–4327.

MacKay, D. M. C. (1956). "The epistemological problem for automata," in *Automata Studies*, eds C. Shannon, and J. McCarthy (Princeton, NJ, Princeton University Press), 235–251. doi: 10.1515/9781400882618-012

Mahon, L. E., and De Valois, R. L. (2001). Cartesian and non-Cartesian responses in LGN, V1, and V2 cells. *Vis. Neurosci.* 18, 973–981. doi: 10.1017/S0952523801186141

Makris, N., Kennedy, D. N., McInerney, S., Sorensen, A. G., Wang, R., Caviness, V. S. Jr., et al. (2005). Segmentation of subcomponents within the superior longitudinal fascicle in humans: a quantitative, *in vivo*, DT-MRI study. *Cereb. Cortex* 15, 854–869. doi: 10.1093/cercor/bhh186

Marchette, S. A., Vass, L. K., Ryan, J., and Epstein, R. A. (2014). Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe. *Nat. Neurosci.* 17, 1598–1606. doi: 10.1038/nn.3834

Marcinkevičs, R., and Vogt, J. E. (2020). Interpretability and explainability: a machine learning zoo mini-tour. *arXiv [Preprint] arXiv:2012.01805*.

Marr, D. (1982/2010). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT press. doi: 10.7551/mitpress/9780262514620.001.0001

Marr, D., Hildreth, E., and Brenner, S. (1980). Theory of edge detection. *Proc. R. Soc. Lond. B. Biol. Sci.* 207, 187–217. doi: 10.1098/rspb.1980.0020

Masri, R. A., Grünert, U., and Martin, P. R. (2020). Analysis of parvocellular and magnocellular visual pathways in human retina. *J. Neurosci.* 40, 8132–8148. doi: 10.1523/JNEUROSCI.1671-20.2020

McSpadden, A. (1998). *A Mathematical Model of Human Saccadic Eye Movement*. Lubbock, TX: Texas Tech University.

Merigan, W. H., Byrne, C. E., and Maunsell, J. H. (1991). Does primate motion perception depend on the magnocellular pathway? *J. Neurosci.* 11, 3422–3429. doi: 10.1523/JNEUROSCI.11-11-03422.1991

Mirza, M. B., Adams, R. A., Mathys, C., and Friston, K. J. (2018). Human visual exploration reduces uncertainty about the sensed world. *PLoS ONE* 13:e0190429. doi: 10.1371/journal.pone.0190429

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

Mishkin, M., Ungerleider, L. G., and Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* 6, 414–417. doi: 10.1016/0166-2236(83)90190-X

Moran, R. J., Campo, P., Symmonds, M., Stephan, K. E., Dolan, R. J., and Friston, K. J. (2013). Free energy, precision and learning: the role of cholinergic neuromodulation. *J. Neurosci.* 33, 8227–8236. doi: 10.1523/JNEUROSCI.4255-12.2013

Moser, E. I., Kropff, E., and Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.* 31, 69–89. doi: 10.1146/annurev.neuro.31.061307.090723

Nathans, J., Thomas, D., and Hogness, D. S. (1986). Molecular genetics of human color vision: the genes encoding blue, green, and red pigments. *Science* 232, 193–202. doi: 10.1126/science.2937147

Neisser, U. (1967). *Cognitive Psychology*. New York, NY: Appleton-Century-Crofts.

Ognibene, D., and Baldassarre, G. (2014). Ecological active vision: four bio-inspired principles to integrate bottom-up and adaptive top-down attention tested with a simple camera-arm robot. *IEEE Trans. Auton. Ment. Dev.* 7, 3–25. doi: 10.1109/TAMD.2014.2341351

Ognibene, D., and Demiris, Y. (2013). *Towards Active Event Recognition*. New York, NY: IJCAI. https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=7058320

Papez, J. W. (1995). A proposed mechanism of emotion. 1937. *J. Neuropsychiatry Clin. Neurosci.* 7, 103–112. doi: 10.1176/jnp.7.1.103

Parr, T., Corcoran, A. W., Friston, K. J., and Hohwy, J. (2019). Perceptual awareness and active inference. *Neurosci. Consciousness* 2019:niz012. doi: 10.1093/nc/niz012

Parr, T., and Friston, K. J. (2017a). The computational anatomy of visual neglect. *Cereb. Cortex* 28, 777–790. doi: 10.1093/cercor/bhx316

Parr, T., and Friston, K. J. (2017b). Uncertainty, epistemics and active inference. *J. R. Soc. Interface* 14:20170376. doi: 10.1098/rsif.2017.0376

Parr, T., and Friston, K. J. (2018a). Active inference and the anatomy of oculomotion. *Neuropsychologia* 111, 334–343. doi: 10.1016/j.neuropsychologia.2018.01.041

Parr, T., and Friston, K. J. (2018b). The anatomy of inference: generative models and brain structure. *Front. Comput. Neurosci.* 12:90. doi: 10.3389/fncom.2018.00090

Parr, T., and Friston, K. J. (2018c). The discrete and continuous brain: from decisions to movement—and back again. *Neural Comput.* 30, 2319–2347. doi: 10.1162/neco_a_01102

Parr, T., Limanowski, J., Rawji, V., and Friston, K. (2021). The computational neurology of movement under active inference. *Brain*. doi: 10.1093/brain/awab085. [Epub ahead of print].

Perrett, D. I., and Oram, M. W. (1993). Neurophysiology of shape processing. *Image Vis. Comput.* 11, 317–333. doi: 10.1016/0262-8856(93)90011-5

Perrinet, L. U., Adams, R. A., and Friston, K. J. (2014). Active inference, eye movements and oculomotor delays. *Biol. Cybern.* 108, 777–801. doi: 10.1007/s00422-014-0620-8

Pertzov, Y., Avidan, G., and Zohary, E. (2011). Multiple reference frames for saccadic planning in the human parietal cortex. *J. Neurosci.* 31, 1059–1068. doi: 10.1523/JNEUROSCI.3721-10.2011

Pessoa, L., Japee, S., Sturman, D., and Ungerleider, L. G. (2006). Target visibility and visual awareness modulate amygdala responses to fearful faces. *Cereb. Cortex* 16, 366–375. doi: 10.1093/cercor/bhi115

Pezzulo, G., Donnarumma, F., Iodice, P., Maisto, D., and Stoianov, I. (2017). Model-based approaches to active perception and control. *Entropy* 19:266. doi: 10.3390/e19060266

Phong, B. T. (1975). Illumination for computer generated pictures. *Commun. ACM* 18, 311–317. doi: 10.1145/360825.360839

Posner, M. I., Rafal, R. D., Choate, L. S., and Vaughan, J. (1985). Inhibition of return: neural basis and function. *Cogn. Neuropsychol.* 2, 211–228. doi: 10.1080/02643298508252866

Ratan Murty, N. A., and Arun, S. P. (2015). Dynamics of 3D view invariance in monkey inferotemporal cortex. *J. Neurophysiol.* 113, 2180–2194. doi: 10.1152/jn.00810.2014

Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025. doi: 10.1038/14819

Rizzolatti, G., Riggio, L., Dascola, I., and Umiltá, C. (1987). Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention. *Neuropsychologia* 25, 31–40. doi: 10.1016/0028-3932(87)90041-8

Ruskell, G. (1989). The fine structure of human extraocular muscle spindles and their potential proprioceptive capacity. *J. Anat.* 167:199.

Rust, N. C., and DiCarlo, J. J. (2010). Selectivity and tolerance (invariance) both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* 30:12978. doi: 10.1523/JNEUROSCI.0179-10.2010

Sabour, S., Frosst, N., and Hinton, G. E. (2017). Dynamic routing between capsules. *arXiv [Preprint] arXiv:1710.09829.*

Sacks, O. (2014). *The Man Who Mistook His Wife for a Hat*. New York, NY: Pan Macmillan.

Sajid, N., Parr, T., Gajardo-Vidal, A., Price, C. J., and Friston, K. J. (2020). Paradoxical lesions, plasticity and active inference. *Brain Commun.* 2:fcaa164. doi: 10.1093/braincomms/fcaa164

Santiesteban, I., Michael Banissy, J., Catmur, C., and Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Curr. Biol.* 22, 2274–2277. doi: 10.1016/j.cub.2012.10.018

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990-2010). *IEEE Trans. Auton. Ment. Dev.* 2, 230–247. doi: 10.1109/TAMD.2010.2056368

Serre, T., Oliva, A., and Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci. U.S.A.* 104, 6424–6429. doi: 10.1073/pnas.0700622104

Seth, A. (2009). Explanatory correlates of consciousness: theoretical and computational challenges. *Cognit. Comput.* 1, 50–63. doi: 10.1007/s12559-009-9007-x

Sheth, S. A., Abuelem, T., Gale, J. T., and Eskandar, E. N. (2011). Basal ganglia neurons dynamically facilitate exploration during associative learning. *J. Neurosci.* 31, 4878–4885. doi: 10.1523/JNEUROSCI.3658-10.2011

Shine, J. P., Valdés-Herrera, J. P., Hegarty, M., and Wolbers, T. (2016). The human retrosplenial cortex and thalamus code head direction in a global reference frame. *J. Neurosci.* 36, 6371–6381. doi: 10.1523/JNEUROSCI.1268-15.2016

Shum, H., and Kang, S. B. (2000). "Review of image-based rendering techniques," in *Visual Communications and Image Processing 2000, International Society for Optics and Photonics* (Perth, WA). doi: 10.1117/12.386541

Shusharina, N., and Sharp, G. (2012). Image registration using radial basis functions with adaptive radius. *Med. Phys.* 39, 6542–6549. doi: 10.1118/1.4756932

Spratling, M. W. (2017). A hierarchical predictive coding model of object recognition in natural images. *Cognit. Comput.* 9, 151–167. doi: 10.1007/s12559-016-9445-1

Srinivasan, M. V., Laughlin, S. B., Dubs, A., and Horridge, G. A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B. Biol. Sci.* 216, 427–459. doi: 10.1098/rspb.1982.0085

Strassman, A., Highstein, S., and McCrea, R. (1986). Anatomy and physiology of saccadic burst neurons in the alert squirrel monkey. I. Excitatory burst neurons. *J. Comp. Neurol.* 249, 337–357. doi: 10.1002/cne.902490303

Szczepanski, S. M., Pinsk, M. A., Douglas, M. M., Kastner, S., and Saalmann, Y. B. (2013). Functional and structural architecture of the human dorsal frontoparietal attention network. *Proc. Natl. Acad. Sci. U.S.A.* 110, 15806–15811. doi: 10.1073/pnas.1313903110

Tacchetti, A., Isik, L., and Poggio, T. A. (2018). Invariant recognition shapes neural representations of visual input. *Ann. Rev. Vis. Sci.* 4, 403–422. doi: 10.1146/annurev-vision-091517-034103

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annu. Rev. Neurosci.* 19, 109–139. doi: 10.1146/annurev.ne.19.030196.000545

Tarr, M. J., and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cogn. Psychol.* 21, 233–282. doi: 10.1016/0010-0285(89)90009-1

Taube, J. S. (1995). Head direction cells recorded in the anterior thalamic nuclei of freely moving rats. *J. Neurosci.* 15, 70–86. doi: 10.1523/JNEUROSCI.15-01-00070.1995

Taube, J. S., Muller, R. U., and Ranck, J. B. (1990). Head-direction cells recorded from the postsubiculum in freely moving rats. Description, I., and quantitative analysis. *J. Neurosci.* 10, 420–435. doi: 10.1523/JNEUROSCI.10-02-00420.1990

Thiebaut de Schotten, M., Dell'Acqua, F., Forkel, S. J., Simmons, A., Vergani, F., Murphy, D. G., et al. (2011). A lateralized brain network for visuospatial attention. *Nat. Neurosci.* 14, 1245–1246. doi: 10.1038/nn.2905

Tschantz, A., Baltieri, M., Seth, A. K., and Buckley, C. L. (2020). "Scaling active inference," in *2020 International Joint Conference on Neural Networks (IJCNN)* (Glasgow: IEEE). doi: 10.1109/IJCNN48605.2020.9207382

van de Laar, T. W., and de Vries, B. (2019). Simulating active inference processes by message passing. *Front. Robot. AI* 6:20. doi: 10.3389/frobt.2019.00020

van der Himst, O., and Lanillos, P. (2020). *Deep Active Inference for Partially Observable MDPs. Active Inference.* Cham: Springer International Publishing. doi: 10.1007/978-3-030-64919-7_8

Von Helmholtz, H. (1867). *Handbuch der Physiologischen Optik*. Leipzig: Voss.

Wallis, G., and Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Prog. Neurobiol.* 51, 167–194. doi: 10.1016/S0301-0082(96)00054-8

White, J. K., Bromberg-Martin, E. S., Heilbronner, S. R., Zhang, K., Pai, J., Haber, S. N., et al. (2019). A neural network for information seeking. *Nat. Commun.* 10:5168. doi: 10.1038/s41467-019-13135-z

Whitted, T. (1980). An improved illumination model for shaded display. *Commun. ACM* 23, 343–349. doi: 10.1145/358876.358882

Williford, K., Bennequin, D., Friston, K., and Rudrauf, D. (2018). The projective consciousness model and phenomenal selfhood. *Front. Psychol.* 9:2571. doi: 10.3389/fpsyg.2018.02571

Winn, J., and Bishop, C. M. (2005). Variational message passing. *J. Machine Learn. Res.* 6, 661–694. Available online at: https://www.jmlr.org/papers/volume6/winn05a/winn05a

Wong, S. H., and Plant, G. T. (2015). How to interpret visual fields. *Pract. Neurol.* 15:374. doi: 10.1136/practneurol-2015-001155

Wurtz, R. H., McAlonan, K., Cavanaugh, J., and Berman, R. A. (2011). Thalamic pathways for active vision. *Trends Cogn. Sci.* 5, 177–184. doi: 10.1016/j.tics.2011.02.004

Yang, S. C.-H., Lengyel, M., and Wolpert, D. M. (2016). Active sensing in the categorization of visual patterns. *eLife* 5:e12215. doi: 10.7554/eLife.12215

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inf. Theory* 51, 2282–2312. doi: 10.1109/TIT.2005.850085

Yoshida, K., Iwamoto, Y., Chimoto, S., and Shimazu, H. (2001). Disynaptic inhibition of omnipause neurons following electrical stimulation of the superior colliculus in alert cats. *J. Neurophysiol.* 85, 2639–2642. doi: 10.1152/jn.2001.85.6.2639

Yuille, A., and Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10, 301–308. doi: 10.1016/j.tics.2006.05.002

Zeidman, P., Lutti, A., and Maguire, E. A. (2015). Investigating the functions of subregions within anterior hippocampus. *Cortex* 73, 240–256. doi: 10.1016/j.cortex.2015.09.002

Zeki, S., and Shipp, S. (1988). The functional logic of cortical connections. *Nature* 335, 311–317. doi: 10.1038/335311a0

Zimmermann, E., and Lappe, M. (2016). Visual space constructed by saccade motor maps. *Front. Hum. Neurosci.* 10:225. doi: 10.3389/fnhum.2016.00225

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 16 April 2021
doi: 10.3389/fnbot.2021.647930

Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks

Stefan Fuchs[†] *and Anna Belardinelli*[*]

Honda Research Institute Europe, Offenbach, Germany

Shared autonomy aims at combining robotic and human control in the execution of remote, teleoperated tasks. This cooperative interaction cannot be brought about without the robot first recognizing the current human intention in a fast and reliable way so that a suitable assisting plan can be quickly instantiated and executed. Eye movements have long been known to be highly predictive of the cognitive agenda unfolding during manual tasks and constitute, hence, the earliest and most reliable behavioral cues for intention estimation. In this study, we present an experiment aimed at analyzing human behavior in simple teleoperated pick-and-place tasks in a simulated scenario and at devising a suitable model for early estimation of the current proximal intention. We show that scan paths are, as expected, heavily shaped by the current intention and that two types of Gaussian Hidden Markov Models, one more scene-specific and one more action-specific, achieve a very good prediction performance, while also generalizing to new users and spatial arrangements. We finally discuss how behavioral and model results suggest that eye movements reflect to some extent the invariance and generality of higher-level planning across object configurations, which can be leveraged by cooperative robotic systems.

Keywords: intention recognition, shared autonomy, eye tracking, teleoperation, eye-hand coordination, Hidden Markov Models, human-robot interaction

OPEN ACCESS

Edited by:
Dimitri Ognibene,
University of Milano-Bicocca, Italy

Reviewed by:
Xiaoyu Wang,
UCLA Department of Mechanical and Aerospace Engineering, United States
Tom Foulsham,
University of Essex, United Kingdom

***Correspondence:**
Anna Belardinelli
anna.belardinelli@honda-ri.de

[†]***Present address:***
Stefan Fuchs,
Siemens AG, Berlin, Germany

Received: *30 December 2020*
Accepted: *12 March 2021*
Published: *16 April 2021*

Citation:
Fuchs S and Belardinelli A (2021) Gaze-Based Intention Estimation for Shared Autonomy in Pick-and-Place Tasks. Front. Neurorobot. 15:647930. doi: 10.3389/fnbot.2021.647930

1. INTRODUCTION

Shared autonomy has recently emerged as an ideal trade-off between full autonomy and complete teleoperation in the execution of remote tasks. The benefits of this approach rely on assigning to each party the aspects of the task for which they are better suited. The lower kinematic aspects of action execution are usually left to the robot while higher-level cognitive skills, like task planning and handling unexpected events, are typically concurrently exercised by the human, in a blend that can entail different degrees of autonomy for the robotic part (Goodrich et al., 2013; Beer et al., 2014; Schilling et al., 2016). Considering the often large asymmetry in terms of degrees of freedom or kinematic capabilities between the user input controller (e.g., joysticks) and the robotic effector, shared autonomy eases the operator cognitive load and speeds up execution improving motion fluency and precision. Since the user is setting the goals and the ways to achieve them, this collaborative effort relies on the robotic partner to first recognize the current human intention (*intent recognition*) and only afterwards to decide how much to assist with the execution (*arbitration*). Intention recognition should thus happen as early and as naturally as possible for the user to be relieved of explicitly directing the robot and for the robot to timely initiate the assisting

action. To this end, although several approaches have been proposed that rely on intent recognition from the user control input driving the robotic movement (Yu et al., 2005; Aarno and Kragic, 2008; Hauser, 2013; Javdani et al., 2015; Tanwani and Calinon, 2017; Yang et al., 2017), the most natural and timely way to predict intention both in assistive technologies and remote manipulation is certainly to use gaze cues, as reviewed in the next section. In light of the need to cope with sensorimotor delays (Miall and Reckess, 2002), gaze control itself in task-based scenarios can be considered as inherently predictive of a number of action-relevant aspects. Indeed, in moving our eyes we make use of knowledge- and sensorimotor-based experience (Belardinelli et al., 2016; Hayhoe, 2017; Henderson, 2017; Fiehler et al., 2019) to quickly retrieve the information needed to plan limb motion.

In this study, we focus on gaze-based intention prediction in teleoperating a robotic gripper in a simulated scenario, to investigate human eye-hand coordination under these conditions and to devise an intention estimation model to be later transferred to a real-world shared autonomy scenario. As a first setup for object manipulation, we concentrate on basic pick-and-place tasks as common in this kind of architectures (Javdani et al., 2015; Li et al., 2017; Jain and Argall, 2018, 2019; Shafti et al., 2019). Presented contributions are a behavioral assessment of eye-hand coordination in such scenarios and the design of two Gaussian Hidden Markov Model schemes trained on collected data, showing good generalizability across users and task configurations.

In the next sections, related work on gaze-based intention recognition is first reviewed; the experimental methods used in our setup and the devised models are further presented, followed by results obtained from behavioral analysis and model testing. We conclude by discussing emerged implications and future perspectives.

2. RELATED WORK

That the task shapes the way we look at the world has long been known, as shown by Yarbus (1967). In that study, it was shown that depending on the question the viewer was trying to answer different scanning patterns were produced on the very same image. A number of studies have replicated and confirmed Yarbus' experiment and managed to invert the process and estimate the task from eye data above chance level (e.g., Borji and Itti, 2014; Haji-Abolhassani and Clark, 2014; Kanan et al., 2014). The most popular and effective techniques to compute the probability of a given task given eye movements and possibly their sequence entail Naive Bayes classifier, Hidden Markov Models, SVM, multivariate pattern analysis, and random forests (see Boisvert and Bruce, 2016, for a more complete review). The largely increased diffusion of wearable cameras and eye-trackers in recent years has triggered research on daily activities recognition as observed from an egocentric perspective (Yi and Ballard, 2009; Fathi et al., 2012; Ogaki et al., 2012), hence relying on eye, hand, head, and possibly body coordination (see Nguyen et al., 2016, for a full review).

Yet the approaches above are concerned either with passive information-seeking or with general activity recognition rather than with simple action or proximal intention recognition. Indeed, two basic types of intention have been postulated (Bratman, 1987): a mental state concerning intention for the future (*distal intention*), not necessarily situated in a specific spatial and temporal context, and intentionality for an immediate action (*proximal intention*). From a temporal perspective, a proximal intention is very close to the executed action. Thus, the boundary between proximal intention recognition and action recognition is at times rather blurry. The later an intention is recognized the more advanced the execution of the corresponding action might be.

In a recent study considering object aligning tasks in Virtual Reality (Keshava et al., 2020), it was shown how already simple features, such as the proportion of Points-Of-Regard (POR) on distinct Areas-of-Interest (AoIs) within the objects could constitute a sufficient oculomotor signature to discriminate between four different tasks, which could be classified well above chance. In human-robot collaboration often the robot partner is aware of the activity context and for effective cooperation, it just needs to detect the current action intention of the human partner to help them with it. Huang and Mutlu (2016) have proposed a method for anticipatory control which allows a robot to predict the intent of the human user and plan ahead of the explicit command. In the task considered, a robotic arm prepares a smoothie by picking the ingredients selected vocally by a human user looking at an illustrated list. By means of eye tracking the robot infers the user intention before they utter it and anticipates picking the intended ingredient: an SVM was fed a feature vector of gaze features for each ingredient, such as the number of glances, duration of the first glance, total duration and whether it was the most recently glanced item as predictors of the currently intended ingredient. Although such an approach seems simple and effective in this case the human user was carrying out no parallel visuomotor control task that could yield spurious fixations.

Within shared autonomy approaches, as a first attempt at integrating gaze input from the user, Admoni and Srinivasa (2016) put forward a proposal relying on Javdani's framework (Javdani et al., 2015), where the probability distribution over the goals (hidden states) is updated by considering both user's eye movements and joystick commands as observations in a Partially Observable Markov Decision Process (POMDP), using hindsight optimization to solve it in real-time.

In a further study (Aronson et al., 2018), the authors present an eye tracking experiment aimed at comparing user behavior within-subjects in different teleoperation modalities, namely with more or less autonomy. In the scenario of an assistive robot arm spearing food bits from a plate to feed an impaired user, by looking at partly manually annotated gaze behavior, two patterns of fixations emerged: monitoring glances, meant to check the translational behavior of the arm approaching the intended food morsel, and planning glances, which select the target morsel before starting the arm actuation, as in natural eye-hand coordination (Johansson et al., 2001; Hayhoe et al., 2003). Haji Fathaliyan et al. (2018) proposed a method to

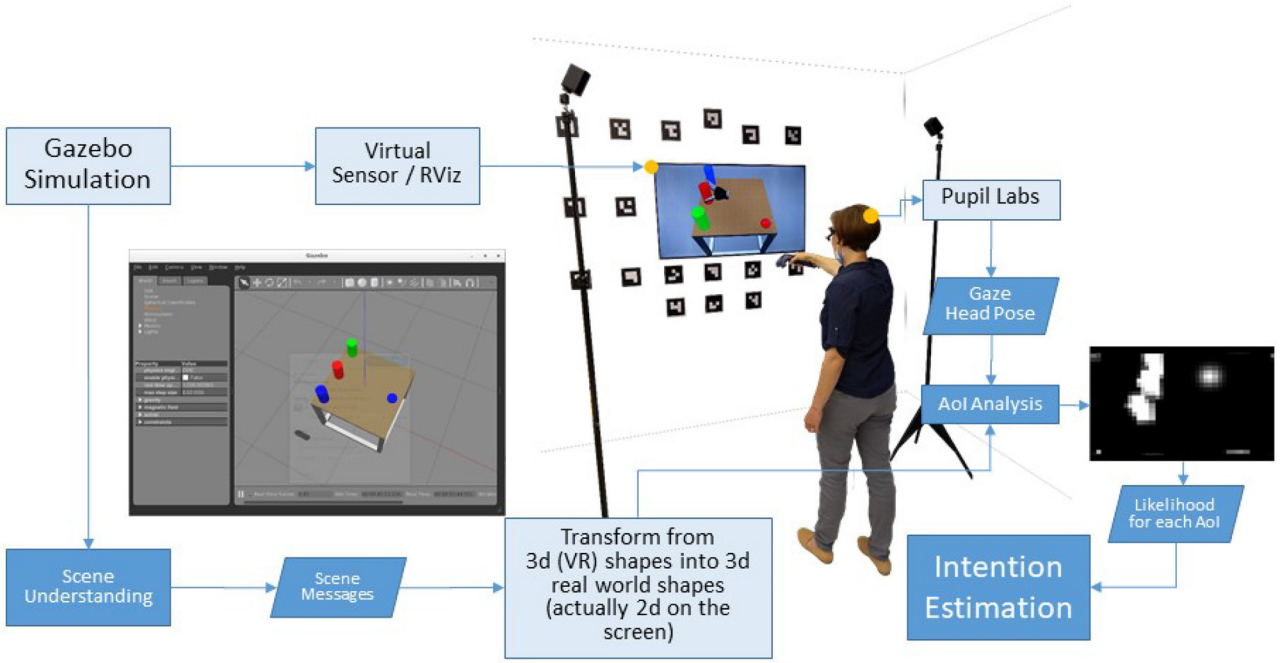


FIGURE 1 | The experiment was carried out in a simulated scenario. The Gazebo simulator features physics and a virtual sensor, with its measurements rendered on the screen. The user controls the Pick-and-Place-Task with a HTC Vive controller, tracked by the Vive Lighthouse system. At the same time, the Vive controller and a Pupil Labs eye-tracking system are used to estimate the human sensorimotor state.

localize gaze on 3D objects by projecting the gaze vector on point cloud representations of the objects manipulated by a person preparing a powdered drink. Using Dynamic Time Warping barycentric averaging, sequences of gazed objects were obtained encapsulating the typical temporal patterns of object interaction that could be used for action recognition. Very recently, the same group used features extracted by this method to recognize action primitives in different activities (Wang et al., 2020). However, data were collected using natural eye-hand coordination, with participants executing the task themselves, which represents a different situation from a teleoperation scenario both on a perceptual and action control level. In the context of assistive robotics, a number of other studies have also considered gaze information (at times combined with multimodal interfaces, such as BCI and haptic feedback) to operate robotic limbs and wheelchairs (Schettino and Demiris, 2019; Zeng et al., 2020). Often in these cases, the gaze is used to implicitly but actively point the system to the object the impaired user wants the robot to interact with (Li and Zhang, 2017; Wang et al., 2018; Shafti et al., 2019).

Our study follows similar motivations as Aronson et al. (2018) and Wang et al. (2020) and complements those results, while not being aimed specifically at assistive applications, but rather trying to leverage human dexterity and eye-hand coordination to improve performance in teleoperated manipulation tasks. To investigate human oculomotor behavior during teleoperation in a more controlled scenario and with a more natural input interface, we designed an experiment in simulation, where the participant would control the remote robot arm by means of their own arm movements via motion tracking. We reasoned that this would produce more natural scanpaths and reaching behavior, without the cognitive overload of a controller with few DOFs, but still showing how the user copes on a sensorimotor level with the task of controlling a remote arm. These behavioral cues were further collected to train a proof-of-concept model able to predict the current intention in pick-and-place tasks in similar teleoperated scenarios, to be later deployed in a real-world setup[1]. Since many teleoperation scenarios relay visual input through a camera, we displayed the scene on a screen and used eye tracking glasses to retrieve the (POR) on the 2D display.

In our approach, we plan to work with multiple objects and to recognize different sequential sub-tasks, hence we chose to model scanpaths via Hidden Markov Models (HMM), which present the benefit of considering the temporal dimension of the gaze shifts and can better deal with spurious fixations and gaze samples and varying eye tracking frequency (Belardinelli et al., 2007; Coutrot et al., 2018; Boccignone, 2019). Our experimental setting and the intention estimation model are detailed in the next sections.

[1]To avoid confusion with terms sometimes used interchangeably, sometimes meaning different things, we here refer to task as the overarching ongoing activity, e.g., pick and place, while intention implies the commitment to perform the current proximal action/sub-task, e.g., reaching to grasp.

3. BEHAVIORAL EXPERIMENT METHODS AND ANALYSIS

3.1. Participants

This study has been conducted after the outbreak of the COVID-19 pandemic. Hence, a number of participants suitable for this kind of study could not be recruited and to minimize infection risks only associates of the Honda Research Institute participating in this project were asked to take part in data collection on a voluntary basis (N = 4, including the authors). We complied with the measures of the *Occupational Safety and Health Standard* emanated by the German Federal Ministry of Labour and Social Affairs by keeping a safe distance and wearing face masks. The study was approved by the Bioethics Committee of Honda.

Participants had normal or corrected-to-normal vision, were all right-handed, and gave informed consent to participate in the study.

3.2. Experimental Setup and Procedure

The experiment was carried out in a simulated scenario created with the Gazebo Simulator[2] (see **Figure 1**). The scene was captured with a virtual sensor and displayed on a widescreen (1.21 × 0.68 m) with HD resolution in front of the participant, who was standing at a distance of about 1.5 m. Participants wore a binocular Pupil Core eye-tracker by Pupil Labs, working at 100 Hz with a reported accuracy of 0.6°. They also held in the right hand the HTC Vive controller, tracked by the Vive Lighthouse system for input control in the teleoperation task. All physical devices and surfaces were sanitized after each use. After instructions, participants were required to wear the eye tracking glasses, to adjust the eye and scene cameras according to the experimenter's directions, and to perform a 5-point calibration.

The experimental stimuli consisted of three cylinders presented in two configurations (in different blocks): either aligned on the left side of a table (numbered as follows: 0 for the top, 1 for the middle, 2 for the bottom of the table) or at the vertices of a virtual triangle (0 for the top vertex, 1 for the bottom right, 2 for the bottom left; see **Figure 2**). Colors were permuted anew in each trial. Along with the cylinders, a disk would appear on the right side of the table, at one of three positions (denoted as: 0 top, 1 middle, 2 bottom). The disk specified the current pick-and-place task: the color indicated which cylinder to pick up and the position of the disk where the cylinder had to be placed down on the table. The task would be executed by a robotic gripper in the virtual scene, operated by the participant's movements. The position and orientation of the Vive controller grasped by a user's hand were tracked and mapped onto the gripper. Just the robot hand was visible and could be controlled by the participant as the own hand. No robotic arm kinematics was simulated in the mapping of the movement. Participants were required to reach with the controller in their hand toward the target and to grab it by pressing the button on the controller under the index finger. They had then to move the cylinder to the other side and release it on the place position, in so ending the trial. Between trials, a resting time of 5 s was given, followed by a fixation cross and indications on how to move the controller back to the rest position. As soon as the controller reached the starting position, the next trial started.

The cylinders were 20 cm high and with a radius of 5 cm. In the lined-up configuration, they were placed 30 cm apart, while in the triangular configuration cylinder 1 and 2 were about 21 cm apart and both were 22.6 cm apart from the cylinder in position 0. The robotic gripper was about 16 cm long from the wrist to the midpoint between the fingers.

3.3. Design and Data Processing

We designed two different arrangements of the cylinders since we hypothesized that the positions of the objects would require different movement trajectories and oculomotor strategies. In this way, we could investigate the impact of the spatial arrangement of the items on the gaze behavior.

Each trial consisted of a reach-and-grasp phase and a transport-to-place phase to the placing target position. The two phases are separated by the gripper grasping the picking target. In this sense, in the following, picking times are considered as the time from the start of the trial to the grasp event detected via button press. The transport phase spans the time from the grasp event to the end of the trial, i.e., when the gripper button was released and the cylinder in hand was within 10 cm of the placing disk. The tasks are defined by the positions of the respective targets, e.g., *pick_0* for picking at the pick position 0 or *place_1* for placing at the place position 1. In each trial the target pick and place positions were randomly generated. This has led to an uneven number of pick-place target combinations for each participant. Lined-up and triangle arrangements were probed in separate blocks. Specifically, the final dataset consisted of sequences containing for the lined-up configuration 63 examples of *pick_0* and *pick_1*, 54 of *pick_2*, 60 of *place_0*, 55 of *place_1* and 65 of *place_2*. For the triangular configuration the dataset contained 35 examples of *pick_0* and *pick_1*, 27 of *pick_2*, 26 of *place_0*, 30 of *place_1*, and 41 of *place_2*.

Instead of working with relative eye coordinates, we used the fiducial markers and the scene camera of the Pupil Labs device to localize the eye-tracking-glasses in the scene w.r.t. the world and screen, respectively. Fixations represent a very popular cue in eye-movement data analysis and might seem an obvious choice in this intention estimation application. The parameterization of a fixation identification method, however, might be very arbitrary. Usually, thresholds are chosen to determine when exactly fixations start and when they end. Thus, the parameters of a fixation identification algorithm can have a dramatic impact on our higher-level analyses (Salvucci and Goldberg, 2000). Further, the system will be required to work online eventually and online fixation recognition is not always accurate while further increasing the computational load. The temporal information related to dwelling time in the AoIs (the objects of interest in the scene) during fixations is still learned and considered by the HMM all along.

For these reasons, we decided to work with gaze samples that were mapped on the scene according to the following approach (depicted in **Figure 3**). A heatmap with a discrete resolution

[2]http://gazebosim.org/

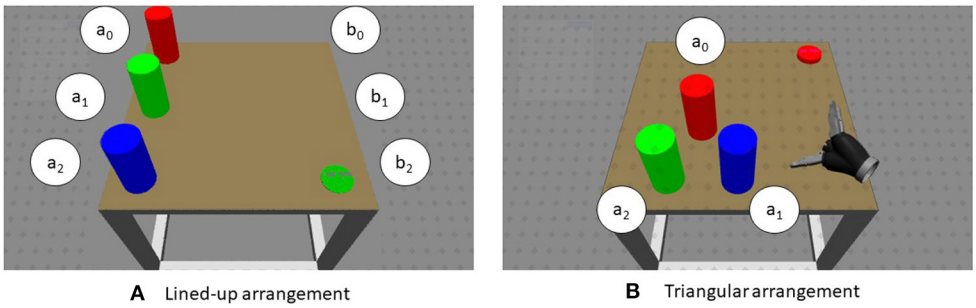


FIGURE 2 | Example of scenes used in each trial. The objects to pick up were displayed lined up on the left side (or in a triangular arrangement) in three different colors while the disc on the right could similarly appear at each of three positions on the right side. The color of the disk signified which cylinder was to pick up, the position of the disk denoted the position for the placing down. The white disks are just shown here to label the picking and placing positions. **(A)** Lined-up arrangement. **(B)** Triangular arrangement.

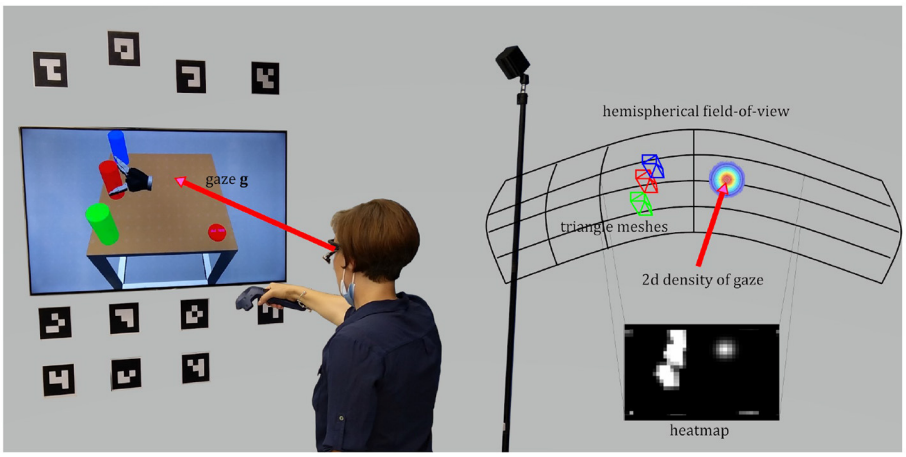


FIGURE 3 | The user's field-of-view is approximated by a hemispherical heatmap. The density of a 2d normal distribution centered on the point-of-regard represents the gaze and its uncertainty. The surface integral over the triangles of a certain object is the likelihood of this AoI.

represents the hemispherical field-of-view of the participant. In this case a sampling of 1° is used and the heatmap comes with a resolution of 180 by 90 px. The user's eye gaze **g** is represented by a two-dimensional normal distribution and the density is plotted onto the heatmap with gaze uncertainty σ and location centered on μ[3]. The choice of the size of σ might depend on the accuracy and precision of the eye tracking measurements. Here, we set $\sigma = 2^\circ$ which is in accordance with the size of the human fovea. All potential scene objects are represented as triangle meshes with a bounding box made of at least 12 triangles. The pose of the objects is delivered by a scene understanding module and given the localization of the eye tracking glasses, the object poses can be transformed into the head coordinate system. Mesh triangles, that are visible to the user (i.e., normal of triangle directed toward the user), are plotted onto the heatmap. The surface integral of the density function over these triangles represents the likelihood that this area is regarded by the user. The complete likelihood (of each object to be regarded by the user) is the sum of all visible triangles the object is made of. In order to not overemphasize large objects, all likelihoods are normalized by their visible areas. For each object an Area-of-Interest was defined, for a total of seven AoIs: for the picking objects the areas $\{a_0, a_1, a_2\}$, for the placing positions the areas $\{b_0, b_1, b_2\}$, plus an area R for the robotic gripper.

As a result, this so-called *Area-of-Interest-analysis* provides for every gaze sample **g** a feature vector **F** entailing the likelihood computed for each of these AoIs:

$$\mathbf{F}_t = \{\mathrm{P}(\mathrm{AoI} = a_0|\mathbf{g}_t), \mathrm{P}(\mathrm{AoI} = a_1|\mathbf{g}_t), ..., \mathrm{P}(\mathrm{AoI} = R|\mathbf{g}_t)\} \quad . \tag{1}$$

These were logged along with the current hand position and robot gripper position and with the current grasping state (defined as the binary state of the grasping button). Trial samples were further labeled with a Boolean feature to state if the trial was successful. Indeed, if the grasp failed for any reason multiple

[3]The gaze was mapped in this way since in a later stage we plan to move the simulation into a virtual reality headset with embedded eye tracking and the gaze mapping on the scene can stay unaltered.

grasp attempts could be observed or none at all if the cylinder was toppled down and fell off the table.

3.4. Behavioral Analysis

To get a better picture of the gaze behavior during the presented task, we looked at some behavioral measures, seeking confirmation of some of the patterns described in Aronson et al. (2018). Due to the low number of participants thus far, we could not perform an inferential statistics analysis within subjects to test any hypothesis, hence for the most part we report a descriptive analysis computed over the whole dataset depending on the different tasks.

Two exemplary trajectories for different pick-and-place tasks and object configurations are depicted in **Figure 4**. At any time, the AoI collecting the highest likelihood is considered the one currently looked at. Upon motion onset the AoI corresponding to the place target (whose color determines also the picking target) is glanced. This is a planning glance, as defined in Aronson et al. (2018). Right afterward the gaze moves to the picking target. It must be noted that this glance at the place location is due to the way the task is designed. Possibly, it would not be observed if the picking target was communicated to the user in another way, e.g., verbally, with the placing target displayed as a gray disk, for instance. During the transport phase, the gaze targets the placing target. During both the reach-to-grasp phase and the transport-to-place phase the robot AoI (gray) is checked in a monitoring pattern, to make sure the gripper is moving in the intended direction.

For each trial, we consider two intentions/(sub)tasks, one picking and one placing intention, separated by the keypress triggering the grasping. To get a more complete overview of the time the gaze spent in different AoIs across tasks, the relative time distribution of gaze on each AoI was computed and is presented in **Figure 5**. To make the picture easier to interpret, we considered that for each intention there are actually just five semantic entities that are relevant to describe the gaze behavior, namely: the pick target (e.g., a_0 for *pick_0*), the pick distractors (e.g., a_1, a_2 for *pick_0*), the place target (e.g., any of the b_i AoIs depending on the current task), the place distractors (e.g., any of the b_i AoIs that are not the target) and the robot hand. Analogously for the place intention, the place target would be the specific AoI related to that task, while the pick target could be any of the picking positions and the distractors are the pick and place AoIs that are not the current pick nor place target of the trial.

As can be noted from **Figure 5**, for either task of pick or place, the distributions of the gaze time share a common pattern on a semantic level, i.e., the target of the task is longer dwelled on. Thus, these tasks can be distinguished as each task corresponds to a different semantic target (picking or placing objects). Yet the distributions are also distinctive within each task, considering that each action target is the AoIs related to the task target, i.e., the corresponding a_i position in the pick tasks and the corresponding b_i position in the place tasks. In the pick tasks the place target is briefly looked at to learn the pick target, while in the place tasks the pick target receives also some attention since, after pressing the button for the grasp and in absence of any haptic feedback, the gaze checks that the object is correctly grasped. This evolution in time can indeed be appreciated better in **Figure 6**, where trials across intentions were averaged on a normalized time axis between the start of the trial and the grasp for the pick trials and between the grasp event and the end of the trial for the place trials. In these latter, independently of the place target, it can be noted how for about the initial 30% of the placing task the pick target is still looked at, to visually check whether the object is lifted up with the gripper, hence confirming the grasp was successful. Interestingly, in both configurations and tasks also the robot effector receives a discrete amount of gaze time and in the pick trials shares a lot of gaze distribution with the pick target. Of course toward the end of the pick and place trials the gripper is close to the pick/place targets and the gaze can have both within the fovea or in the parafoveal space and monitor them at the same time. Still, in natural eye-hand coordination, the hand instead is rarely looked at (Johansson et al., 2001) because proprioceptive information and peripheral vision usually suffice to monitor it. This suggests that in this teleoperation scenario the unusual sensorimotor mapping from the arm and controller to the three-fingered robotic gripper, especially considering the grasp pose, and possibly some delay in the tracking makes the user uncertain about the effector movements and current pose. Participants, thus, produced multiple monitoring glances (Aronson et al., 2018) during the movements to visually adjust the effector trajectory and pose. However, in general, the distributions looked rather distinctive across tasks, suggesting that it could be possible to reliably discriminate among them, while they looked rather similar across picking configurations, hinting to the possibility to generalize from one to the other. The pick distractors are looked at especially during picking, since the gaze checks the neighboring cylinders in order to decide the best grasp and in order not to collide with them. This is especially the case in picking at position 1 in the lined-up case and overall in the triangle configuration since the cylinders are all close to one another. The place distractors do not receive any attention since in each task only the target position was made visible with a disk in this experiment (see **Figure 2**).

To gain further insight into the difficulty of the task, we looked into the number of failed trials across picking tasks. Error rates were computed for the three pick tasks in the two configurations. The picking action in the lined-up configuration was successful in the 71.4% of *pick_0* cases, 88.9% for *pick_1* trials, and 79.6% for *pick_2*. In the triangular configuration, the grasp was successful in 68.6, 88.6, and 85.2% of cases, for the same picking cases, respectively. The users could accomplish the task in the vast majority of the cases, but a significant number of failed grasps occurred when picking at position 0 in both configurations.

This could be the case for different reasons. In the lined-up configuration, the 0 position is the rearmost and the one requiring to stretch the arm until the furthest edge of the table. However, 3D depth on a 2D plane is badly estimated, especially in a virtual scene where size cues are more difficult to gauge and the own body could not be used as reference either. In the triangle configuration, the 0 position is closer to the user, yet the other two objects are placed in front of it, requiring to pick the cylinder from above or—for a right-handed user—trying to avoid the cylinder in position 1 going around it. The

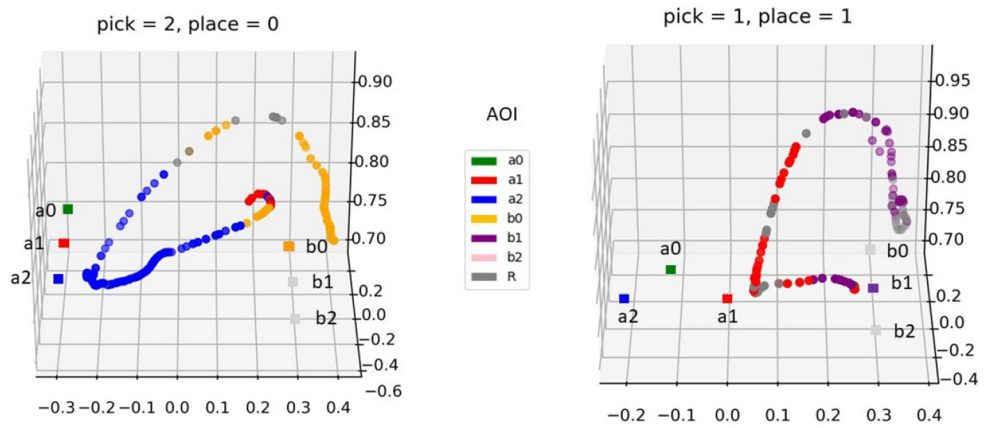


FIGURE 4 | Two exemplary trajectories of the hand during the pick-and-place tasks (left: pick in position 2 and place in position 0; right: pick in position 1 and place in position 1). The movement samples are colored with the currently gazed AoI (see legend). The square markers denote the picking and placing positions: the former are here denominated and colored as the respective AoIs, for the latter only the current placing target is presented in color (the other targets are in light gray since they were not visible but their position is shown for reference).

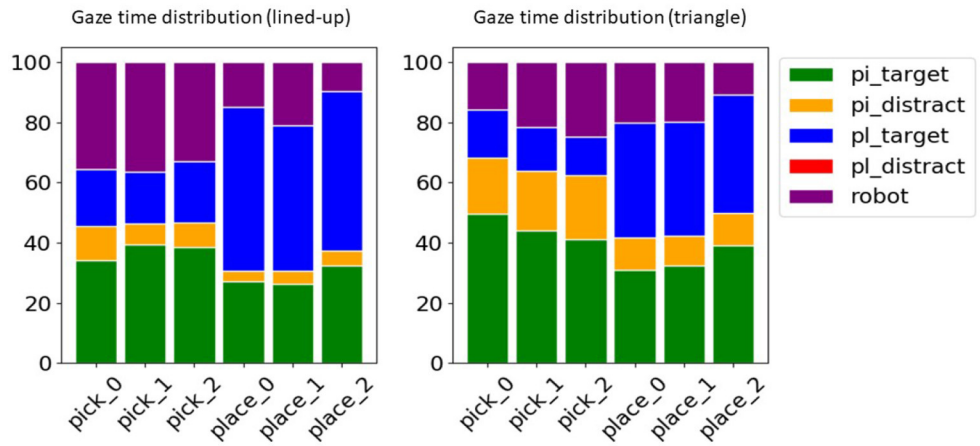


FIGURE 5 | Relative distribution of the time the gaze spent on semantic AoIs across tasks for the lined-up (left) and the triangle arrangement (right). In the pick tasks the respective picking AoIs (pi_target) are more looked at, in the place tasks the respective placing AoIs (pl_target). The other AOIs were summarized in the pick distractors (pi_distract, i.e., the cylinders not to be picked up), place distractors (pl_distract, i.e., possible place-down locations other than the target), and the robot hand (robot).

depth estimation difficulty could yet be ameliorated in a virtual reality set-up.

A similar pattern emerges also looking at picking times. In this case, we considered only successful trials since in a failed trial no grasp or more than one grasp could occur. Looking at **Figure 7**, it can be noted again that the rearmost position requires the longest reaching time. The difference is significant between position 0 and position 2 [Bonferroni corrected Welch's t-test, $t(67.11) = 3.56, p = 0.002$] and between position 1 and 2 [$t(88.34) = 2.94, p = 0.012$]. In the case of the triangular configuration, also items in position 2 require a more careful movement, since a right-handed person needs to mind avoiding the cylinder in position 1 when approaching the cylinder in position 2 with the open gripper [position 0–1: $t(27.24) = 4.02, p = 0.001$, position 1-2: $t(40.54) = -4.15, p < 0.001$].

4. COMPUTATIONAL MODELING AND RESULTS

4.1. Modeling Intentions With Gaussian HMMs

Our approach aims at predicting the proximal intention, i.e., the current action and the involved object. Gaze not only comes with a specific pattern during action execution but also provides early cues that indicate parameters of a pick and place task, such as which object to pick or where to place it down. These parameters are defined by the proximal intention (Bratman, 1987; Pacherie, 2008). The temporal gaze pattern can be represented with a Gaussian Hidden Markov Model (see **Figure 8**). The hidden states $\mathbf{X}(t)$ describe the internal intention process and might relate to *looking at the target object* or *looking at the placing*

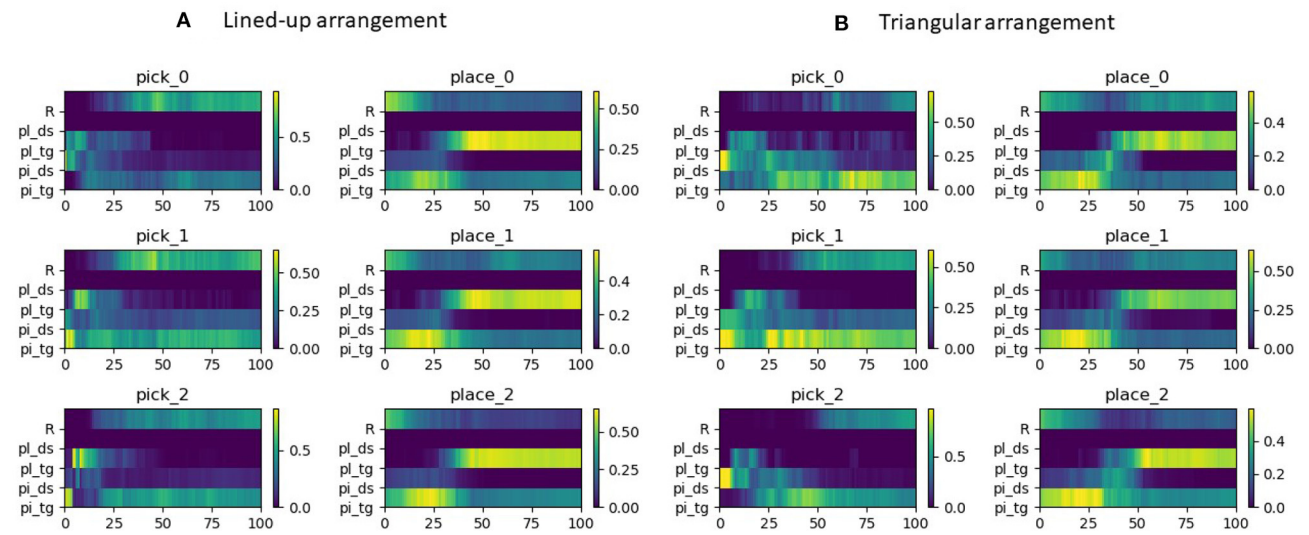


FIGURE 6 | Gaze distribution on a normalized time axis across semantic AoIs. Time = 0 signifies starting of the trial and time = 100 the closing of the gripper on the object for the pick sequences. This last event represents the start of the normalized time axis for the place sequences, with 100 indicating the end of the trial. **(A)** Lined-up arrangement. **(B)** Triangular arrangement.

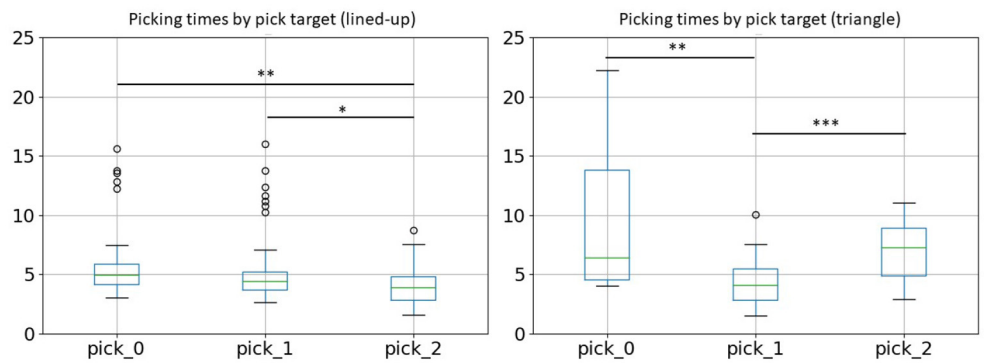


FIGURE 7 | Picking times across picking position for the two object configurations. The rearmost position requires a longer reaching time in both configurations, also due to difficult depth estimation. In the triangle configuration the forefront position on the left (pick_2), besides the rearmost position (pick_0), requires a longer picking time.

position. However, this is just an assumption, while the hidden Markov process drives an observable gaze sequence $\mathbf{Y}(t)$. The gaze sequence is described by the sequence of AoI likelihoods as derived from the multivariate Gaussian distribution (see section 3). The distribution of these AoI likelihoods at a particular time is governed by the emission probabilities of the hidden Markov process given the state of the hidden variable at that time. This approach is independent of the gaze sequence length, i.e., observation sampling and execution velocity, as long as the sequences are scaled linearly.

We defined six intentions to be recognized: three pick-up intentions (for each of the three cylinders) and three place intentions (for each of the three placing positions). Hence, six HMMs have been configured with five internal states. The observation vector of an HMM comprises eight components: the AoI likelihoods of the three cylinders, the AoI likelihoods of the three possible placing positions, the AoI likelihood of the robot, and the trigger button state of the Vive controller. The transition and emission parameters were learned by each HMM, which was fed the respective training sequences (between 19 and 31 observations sequences for each model for a total of 160). These sequences were all performed by two users. The training was done offline with data only from the lined-up arrangement and only successful pick-and-place tasks (no multiple grasp attempts, no toppling or dropping of the cylinders).

Figure 9 sketches the online intention recognition approach. At every time step t the observations from the last Δt s are used to compute the log-probability of these observations under each of the trained HMMs. The HMM with the highest log-probability exceeding a given threshold ($\kappa = 0$) is taken as prediction of the

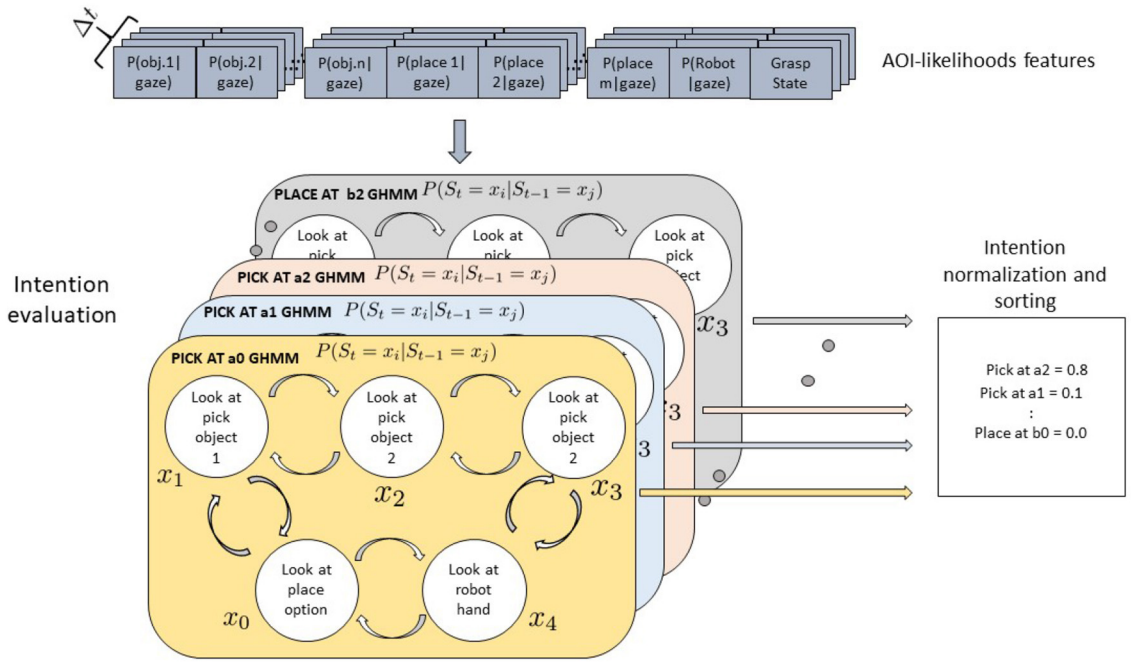


FIGURE 8 | Gaussian Hidden Markov Models for a Pick-and-Place-Task. Five hidden states **X** are shown which might represent the perceptual state of the user looking at the object to be picked, at the robot, at the placing position target, or at the teleoperated robotic hand. Arrows between states represent transition probabilities: for the sake of legibility here only transitions between adjacent states are shown but actually all states are fully connected. A model is defined for each pick-object and place-at-location intention. Each model receives as input a sequence of vectors of AoI likelihoods, representing the probability of the object under the gaze distribution at different time steps. The emissions probabilities defining the probability of each state to emit the observed features are learned from the data and assumed Gaussian. Each model outputs a likelihood of the corresponding intention after observing the current sequence of features.

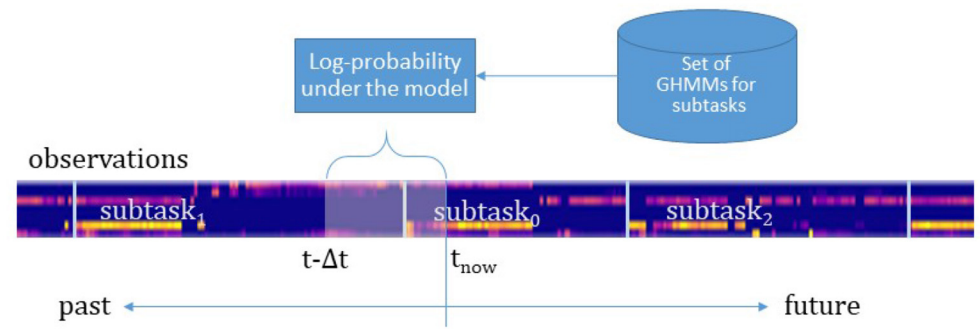


FIGURE 9 | The observations made in the last Δt s are used to compute the log-probability of these observations under each of the trained GHMMs. Here, an example sequence of subtasks with the observations is shown. The feature vectors are color-coded, vertically plotted, and concatenated (generating the bluish bar). The length Δt of the time window decides on the accuracy and the earliness of the intention predictions.

respective intention. If no model scores over the threshold, no intention is confidently recognized. The offline training and the online recognition are implemented in Python with the help of the *hmmlearn*-library[4].

The performance of this approach is tested on data from four users (between 17 and 28 observation sequences for each intention, respectively, for a total of 128 sequences). The testing data comprised unseen sequences in the lined-up arrangement from the two users used for training plus sequences from two additional users. Moreover, testing was done also on sequences from blocks with triangular arrangement (between 19 and 33 sequences for each intention, for a total of 156).

4.2. Intention Recognition Results

To evaluate the intention recognition performance, we looked on the one hand at how accurate was the prediction whenever a

[4]https://hmmlearn.readthedocs.io

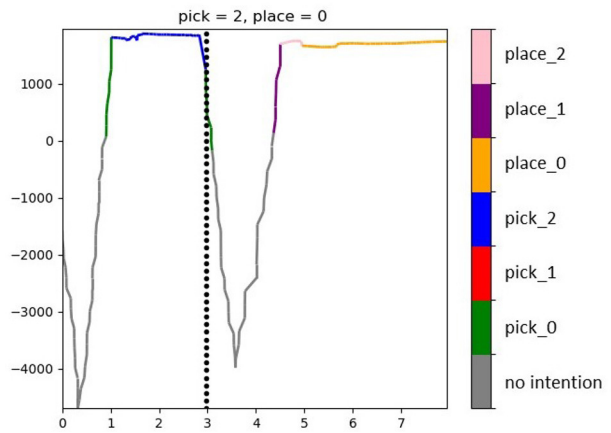


FIGURE 10 | Likelihood of the predicted intention for a trial picking in position 2 and placing in position 0. The black dashed vertical line shows the time of the grasp event. At any time the likelihood of the currently highest model score is depicted in the color associated with the respective intention (see legend). If the highest score is below the threshold $\kappa = 0$, no intention is accepted (gray line).

prediction was indeed available (that is, the proportion of correct predictions over the overall number of delivered predictions). On the other hand, as stated in Ellis et al. (2013) and Wang et al. (2020), there is a trade-off between accuracy and observational latency. Indeed, the more evidence is accumulated before making a prediction, the more accurate the prediction is going to be. Yet, in the case of systems that should act on that prediction, the earlier this comes the better. To this end, maximizing accuracy can be at odds with minimizing latency. We looked also at this kind of latency and call it *predictability*, because of the way it is operationalized. Predictability refers to the fraction of action execution time where an intention is confidently recognized (regardless of whether right or wrong), defined as the ratio between the number of action samples for which an intention estimation is over the threshold κ and the overall number of samples in the action. At the beginning of a new trial when the gaze is still wandering between the placing target to check its color and the pick target, perhaps also checking the pick distractors, it is most likely that the models cannot deliver a confident enough prediction. Similarly, in the transport phase, after checking the successful grasp, the gaze quickly moves from the pick target to the place target. In this case, the observed time window might contain both samples related to the grasped object and to the place destination, hence even the highest-scoring model might deliver a very low likelihood score (under the threshold). This can be appreciated in the example in **Figure 10**: in the beginning of the trial no model reaches a confident enough likelihood score, but as soon as evidence is accumulated in favor of a picking action, the winning likelihood oversteps the threshold. At first, the wrong picking intention is predicted while later the correct model reaches the highest likelihood. A similar course is displayed after the grasp event, with the likelihood going down and then rising again in favor of a placing action.

Figure 11A shows the accuracy and predictability of the intention recognition when using a time window of 0.9 s for the lined-up arrangement. On average, the HMM with the best log-probability being above the given threshold ($\kappa = 0$) indicates the correct intention in 78% of cases (chance level = 16.7%).

Figure 12A highlights the relationship between the time window Δt, accuracy, and predictability. With a longer time window both the prediction accuracy and the predictability decrease. A longer time window has the effect to include more observation samples belonging to a previous action rather than the current intention. This is sketched in **Figure 9**. As a result, either the log-probability threshold is not exceeded or an incorrect intention is recognized. There is a maximum accuracy at a time window of 0.9 s with a predictability of 77%. That is, after at least 23% of the action execution the right action is predicted in 78% of cases. Given this earliness, we can speak of intention recognition.

Figure 12B plots a similar relationship between time window and performance for the triangle-shaped arrangement. The optimal time window size here is 1.3 s with an accuracy of 75% (chance level = 16.7%) and a predictability of 78%. The accuracy curve seems to be flattened because the action execution times in this setup come with a larger spread. Especially, picking up the cylinders at positions 0 and 1 is more challenging and causes a longer execution time compared to the other sub-tasks in this triangle setup. This issue is apparent also in **Figure 11B** with more distant whiskers and extended boxes for *pick_0* and *pick_1*.

Furthermore, the plots in **Figures 11**, **12** confirm the observations described in section 3.4. The gaze behavior seems to be independent of the spatial arrangement of the objects in the scene. This fact is very well-represented by the HMMs, which have been trained only on lined-up arrangement data, but perform almost as well on the triangle arrangement data.

Moreover, **Figure 13** shows the confusion matrices for the two tested spatial arrangements. It can be appreciated that when the model delivers a wrong prediction it usually mistakes neighboring picking or placing locations, but still correctly identifies the task.

4.3. An Alternative Model: The Semantic GHMM

Our hypothesis in designing the behavioral experiment and the model presented above was that the object positions and configuration would have an effect on the observed sensorimotor behavior. Yet, both the results of the behavioral analysis (cfr. **Figures 5**, **6**) and the results of the modeling of separate action-object intentions show how gaze patterns are pretty similar across picking positions and configurations and how the models even generalize well to a new configuration with different picking positions. This suggests that rather the current motor primitive (pick or place), represented at a symbolic, semantic level, determines a prototypical sensorimotor pattern, which gets further specified by the motor system depending on the current situation (motor intentions as put forward by Pacherie, 2008). Yet, these further adjustments are at an intra-class level, preserving the general inter-primitive discriminability. This is

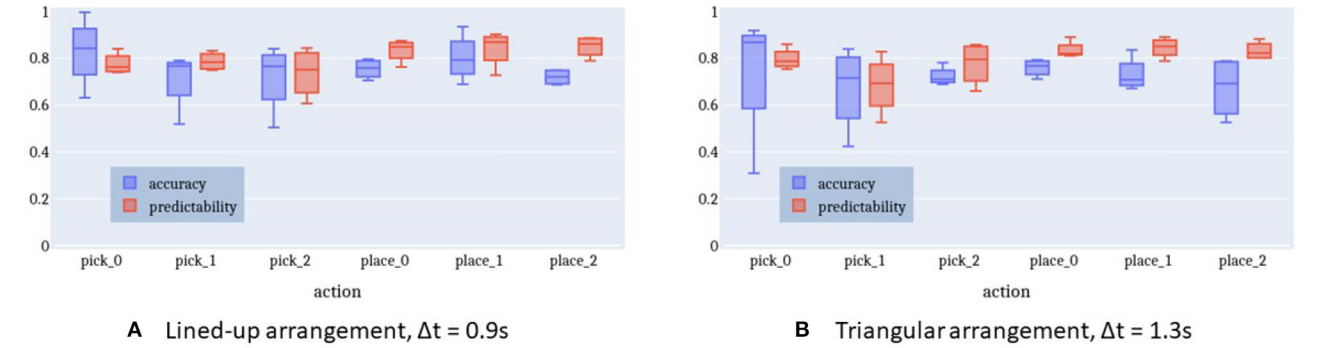


FIGURE 11 | Accuracy and predictability of online intention recognition for lined-up arrangement and triangle shape arrangement. **(A)** Lined-up arrangement, Δt = 0.9 s. **(B)** Triangular arrangement, Δt = 1.3 s.

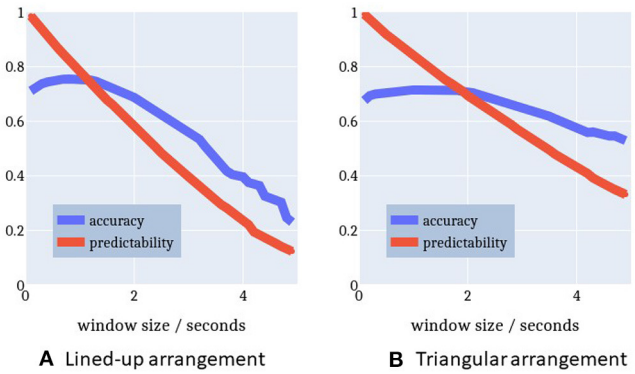


FIGURE 12 | Relationship between time window (min = 0.1 s, max = 5 s) and performance measures. Time window and performance are inversely proportional to each other. A maximum accuracy for a reasonable predictability is reached at time window Δt = 0.9 s for the lined-up arrangement and Δt = 1.3 s for the triangle arrangement, respectively. **(A)** Lined-up arrangement. **(B)** Triangular arrangement.

probably specifically the case in the simplified context we have worked with here, where no real grasping is executed but still, the gripper needs to be placed correctly on the cylinder to allow a firm grasp or to achieve a stable placing down. The model proposed above has further the limitation of scalability: if more pick and place targets were added to the scene, possibly even dynamically appear or disappear, new models would need to be instantiated and trained, while also the feature vector to the models being correspondingly adapted every time. The same would of course occur if a further action would be added to the mix, with every possible combination of object and action being explicitly modeled and trained.

For these reasons, we also devised an alternative, semantic model to be tested against the first model. In this case, just two models are instantiated and trained, one for the pick and one for the place action. The same observations as used in the previous approach have been translated into a flexible object- and action-based arrangement of the feature vector fed to each GHMM model. Thus, the models receive always the same number of features, regardless of the number of objects in the scene. Assuming that n objects and m placing options are present in the scene and that the corresponding AoIs come labeled as either "pickable" or "placeable" candidates, to get the likelihood that object i (or at position i) is currently the pick target, the following vector is fed the *pick* GHMM:

$$\begin{aligned}\mathbf{F}_t = \{&\mathrm{P}(\mathrm{AoI} = a_i|\mathbf{g}_t), \sum_{j\neq i} \mathrm{P}(\mathrm{AoI} = a_j|\mathbf{g}_t), \sum_{k=1}^{m} \mathrm{P}(\mathrm{AoI} \\ &= b_k|\mathbf{g}_t), \mathrm{P}(\mathrm{AoI} = Robot|\mathbf{g}_t), grasping\_state_t)\} \quad . \quad (2)\end{aligned}$$

and to get instead the likelihood that coaster i (or position i) is currently the place target, the following vector is fed the *place* GHMM:

$$\begin{aligned}\mathbf{F}_t = \{&\mathrm{P}(\mathrm{AoI} = b_i|\mathbf{g}_t), \sum_{j\neq i} \mathrm{P}(\mathrm{AoI} = b_j|\mathbf{g}_t), \sum_{k=1}^{n} \mathrm{P}(\mathrm{AoI} \\ &= a_k|\mathbf{g}_t), \mathrm{P}(\mathrm{AoI} = Robot|\mathbf{g}_t), grasping\_state_t)\} \quad . \quad (3)\end{aligned}$$

The two models were instantiated with four hidden states each (representing *looking at a pick* or *place target*, wandering with the gaze on any *distractor for pick* or *for place* or *looking at the robot hand*) and trained and tested with the same data as the first model. Any time a new sample is available from the AoI analysis, the two models are submitted n and m differently arranged feature vectors, respectively, and produce as many likelihood scores, with the highest-ranking taken as the estimated intention. This process is exemplified in **Figure 14**.

Results show that with the semantic models, the accuracy increases reaching a mean value of 88.0% and of 89.7% for the lined-up and triangular configuration, respectively (see **Figure 15**). On the one hand, the lower number of states (4) used in the semantic model might have contributed to the increase of the recognition accuracy. Indeed, although the naïve model uses one more state than the semantic model and this might fit the

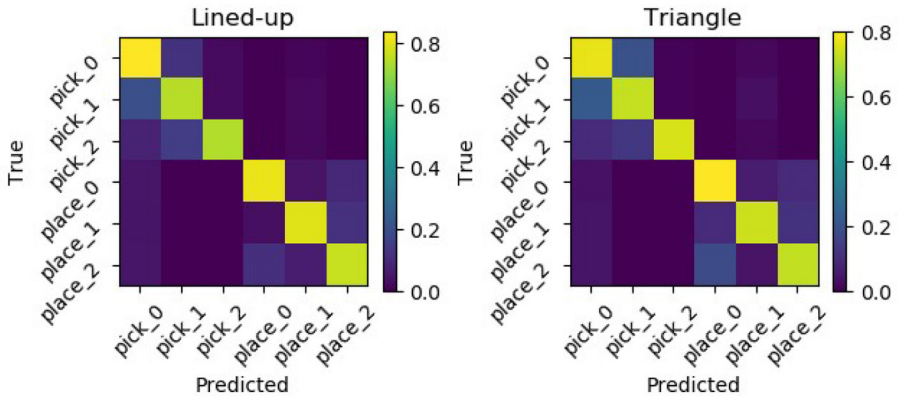


FIGURE 13 | Normalized confusion matrices for the two picking arrangements. Errors are mostly made mistaking neighboring locations but still classifying the task correctly.

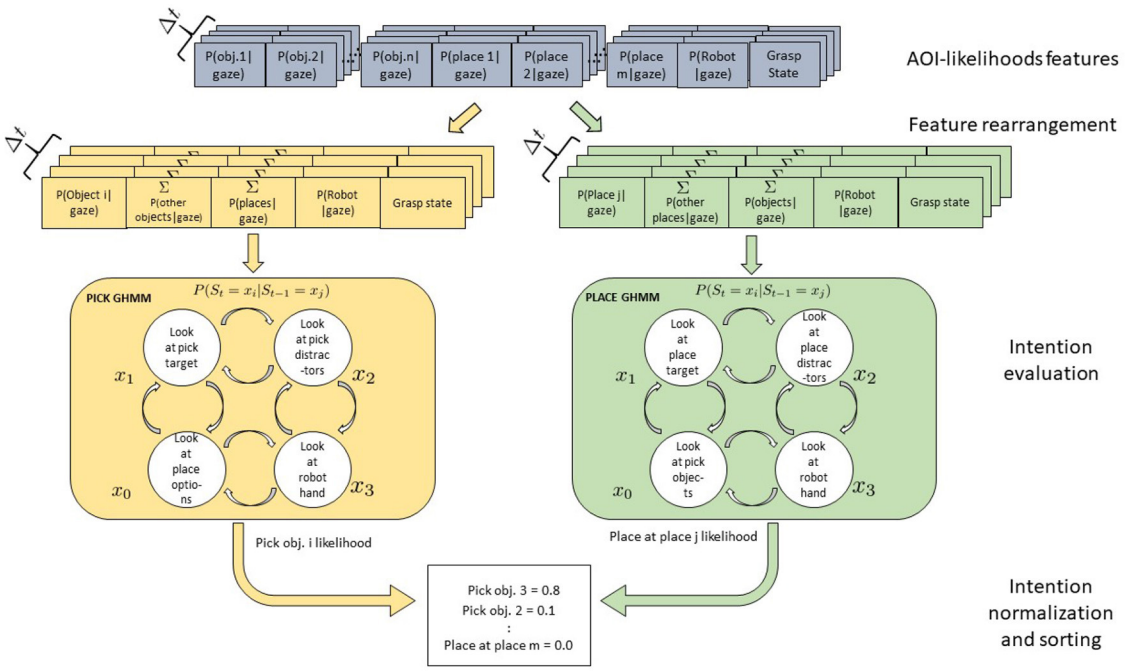


FIGURE 14 | Gaussian Hidden Markov Models for a Pick-and-Place-Task. Four hidden states *x* are shown which might represent the perceptual state of the user looking at the object to be picked, at the robot, at the placing position target, or at the teleoperated robotic hand. Arrows between states represent transition probabilities: for the sake of legibility here only transitions between adjacent states are shown but actually all states are fully connected. Two models are defined for the pick and place intentions. Each model receives as input a vector of rearranged AoI likelihoods, with the first element representing the object to be tested for pick (place), while the second sums up the AoI likelihood of the other pick (place) distractors. The third element sums the features of the other objects relevant for the other action, while the robot and the grasping state features stay the same.

training data better, a higher number of states can unnecessarily overcomplicate the model and produce overfitting. On the other hand, the two semantic models have access to more training data w.r.t the six naive models: in general they can abstract better as to what defines a pick or a place action across the different targets. Considering the normalized confusion matrices depicted in **Figure 16**, in this case, no mistake is made between the two actions: the semantic models seem to be able to better learn the importance of the grasping state feature in discriminating the two actions, as further shown in the next subsection.

4.4. Effect of Grasping State on Performance

Assessing the model performance against a 16.7% chance level could be misleading since the grasping state constitutes a powerful binary cue to tell the two actions apart and hence a

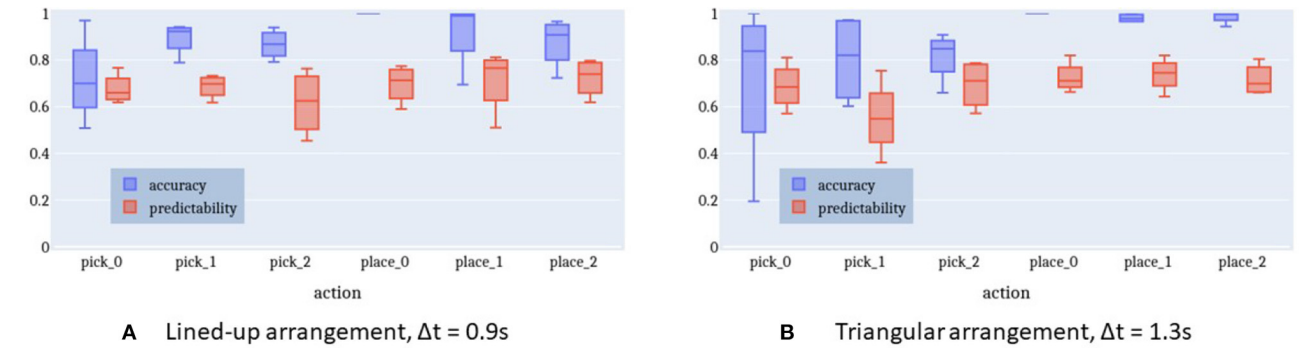


FIGURE 15 | Accuracy and predictability of online intention recognition for lined-up arrangement and triangle shape arrangement in the semantic model. **(A)** Lined-up arrangement, Δt = 0.9 s. **(B)** Triangular arrangement, Δt = 1.3 s.

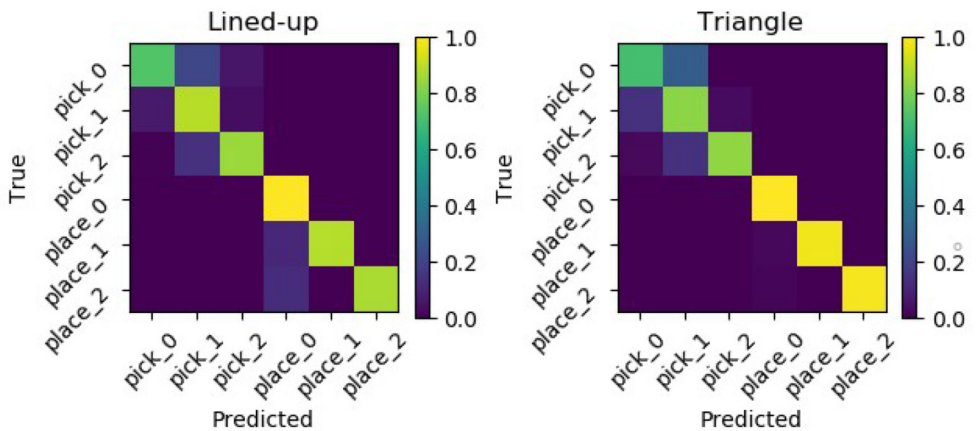


FIGURE 16 | Normalized confusion matrices from the semantic GHMM for the two used spatial arrangements.

33.3% chance level would be a fairer assessment. For this reason, both models were trained and tested again without the grasping state feature. Results with and without this feature are reported in **Tables 1**, **2**. The accuracy substantially decreases for both models, but still remains on an above-chance level. The predictability rises almost at ceiling levels, probably because the models consider at any time the current fixation as indicative enough of the current intention, irrespective of its compatibility with the current grasp state, and outputs the corresponding intention. While on an overall level the naive and the semantic models achieve a similar accuracy of above 50%, it can be seen in the confusion matrices in **Figure 17** that the naive model somehow manages to differentiate between the two different actions while the semantic model basically classifies most intentions as place intentions. A possible explanation could be that without the grasping cue or any other common-sense prior knowledge about picking and placing, the semantic models can only generally infer an intention to interact with an object. In this case, the "place" model, which relies on clearer, longer fixations on one target (see **Figure 6**), is the most confident about its predictions, while the "pick" model sees the gaze likelihood distributed among more objects. The naive models, on the other hand, which were separately trained on the single object/locations, manage to retrieve some of the regular patterns of each single intention. Still, it is reasonable to expect that with a larger dataset, both models could better learn the scanpath differences evident in **Figure 6** and better discriminate between the two actions just by means of gaze features.

4.5. Comparison to Active Fixation-Based Approaches

The specific nature of our setup makes it difficult to compare our system to other approaches presented in the literature, since often either natural eye-hand coordination (Haji Fathaliyan et al., 2018; Wang et al., 2020) or no eye-hand coordination at all (Huang and Mutlu, 2016) is used for intention recognition (see section 2). In teleoperation, especially in the context of assistive technologies, the user is often required to actively fixate the object of interest for a certain amount of time in order to trigger an associated action (Wang et al., 2018; Cio et al., 2019; Shafti et al., 2019). We compare here our system to such approaches, to verify the advantage of a probabilistic framework over a deterministic, sensory-driven one. To this end, we computed the classification performance when considering a fixation as a time window Δt where the same AoI had consistently the highest likelihood and

TABLE 1 | Accuracy for the naive and semantic GHMM models with and without the grasping state feature.

Naive model	With GS (%)	Without GS (%)	Semantic model	With GS (%)	Without GS (%)
Lined-up	78.3	53.5	Lined-up	88.0	54.6
Triangular	75.3	51.5	Triangular	89.7	58.8

TABLE 2 | Predictability for the naive and semantic GHMM models with and without the grasping state feature.

Naive model	With GS (%)	Without GS (%)	Semantic model	With GS (%)	Without GS (%)
Lined-up	77.1	98.5	Lined-up	70.5	98.5
Triangular	78.2	99.0	Triangular	71.9	99.2

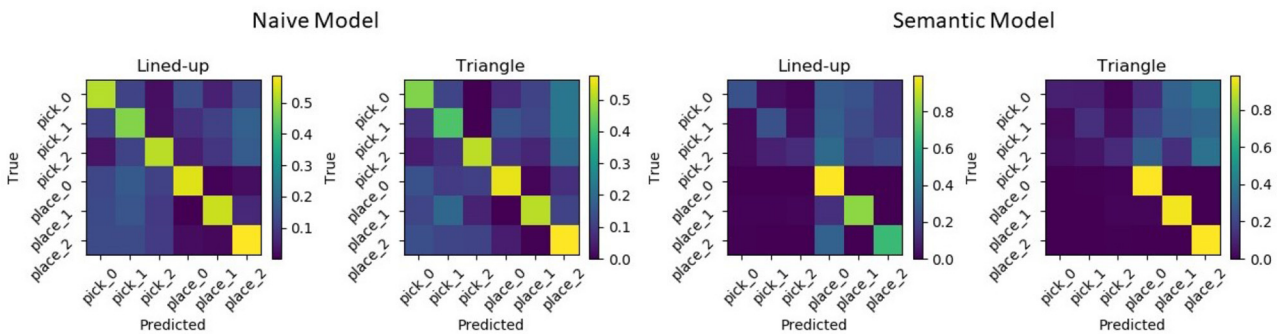


FIGURE 17 | Normalized confusion matrices for the naive (left) and semantic (right) models for both configurations when training and testing the model without the grasping state feature. The semantic model achieves a slightly higher overall accuracy but fundamentally recognizes highly accurately just the placing intentions.

took as prediction the corresponding intention (e.g., if a_0 was fixated for Δt s the prediction would be "pick at a_0"). Note that this method assumes that each object is associated with only one action (either pick or place), thus it would not mistake the two actions. We used Δt of different sizes to compare the fixation performance to our model ($\Delta t = 0.9$ for the lined-up, $\Delta t = 1.3$ for the triangular configuration), to the approach by Shafti et al. (2019) ($\Delta t = 1.5$ s), and to the approach by Wang et al. (2018) ($\Delta t = 2$ s). It must be stressed that in our case the users were not actively fixating the objects to make their intentions legible: when fixating naturally, it is rarely the case that fixations this long occur, hence we further tested a shorter window $\Delta t = 0.5$. Accuracy and predictability results are reported in **Table 3**. Even if this system reaches an accuracy at times comparable with that of our model without grasping state feature, still the predictability is considerably lower. That is, only when a fixation on an object is ongoing for sufficient time a prediction is available, while most of the time no intention is predicted. Indeed, considering a shorter $\Delta t = 0.5$ produces the best accuracy and predictability scores. In contrast, our model emits a more reliable prediction earlier in time and maintains it even when the gaze is not on the target object since transitions are accounted for.

5. DISCUSSION AND CONCLUSIONS

We presented a study aimed at investigating eye-hand coordination and gaze-based intention recognition during teleoperated pick-and-place tasks. The ultimate goal is to transfer such intention recognition into a shared autonomy architecture. To this end, in this first study on the one hand data was collected and analyzed in order to have a baseline characterization of user behavior in a fully teleoperated modality. On the other hand, collected data was used to train a model flexible enough to work with different users and in possibly different settings.

In teleoperation contexts natural eye-hand coordination is somehow disrupted since action is mediated by an input controller and executed by a robotic system. This arrangement upends the internal forward and inverse model predictions and places a further monitoring load on the visual system. Hence, as first studies besides this have shown (Aronson and Admoni, 2018; Aronson et al., 2018), investigating eye-hand coordination during teleoperation can shed light on the user's specific sensorimotor behavior and needs in such setting, prompting better design and models for intention recognition in such systems. Still, in contrast to those studies aimed at assistive applications, we strove for a more natural control input based on motion tracking. In this way, we aim to elicit and exploit patterns of eye-hand coordination similar to those used in real grasping and acting. The analysis of eye and hand behavior has revealed that, although participants in most cases managed to successfully operate the gripper in the pick-and-place task, still some positions required more grasp attempts and longer reaching times. This is in part due to the impaired depth estimation on the screen, however, the difficulty in aligning the gripper with the cylinder in the furthest

TABLE 3 | Accuracy and predictability of the AoI features considering different fixation times for comparison with the fixation times of 1.5 s used by Shafti et al. (2019) and of 2 s used by Wang et al. (2018).

Configuration (fixation time)	Lined-up (Δt = 0.5)	Triangle (Δt = 0.5)	Lined-up (Δt = 1.5)	Triangle (Δt = 1.5)	Lined-up (Δt = 2.0)	Triangle (Δt = 2.0)	Lined-up (Δt = 0.9)	Triangle (Δt = 1.3)
Accuracy	66.0%	62.4%	52.1%	54.3%	42.6%	48.8%	61.0%	56.3%
Predictability	47.0%	53.8%	24.1%	35.0%	16.1%	28.4%	37.0%	37.9%

Since in general shorter fixation times occurred in the trials, also a shorter 0.5 s interval was tested. A further comparison is done using a fixation time equal to the window times used in our model ($\Delta t = \{0.9, 1.3s\}$).

position or in avoiding bumping into cylinder 1 to grasp in position 2 in the triangular arrangement required extra care and slowed down the movement. Moreover, while the gaze behavior showed some similarities with natural eye-hand coordination, e.g., locating and guiding the hand to the target of the next proximal intention (Land et al., 1999), we found that both in the reaching and in the transport phase the robot gripper was looked at for quite some time, differently from what happens when grasping with the own hand (Johansson et al., 2001). This represents an indicator that the participant preferred to visually monitor the gripper movement in the absence of the usual proprioceptive coordination and tactile feedback. Furthermore, the object held in hand was looked at also after the grasping was triggered, again something that does not happen in natural eye-hand coordination, since tactile feedback confirms the expected contact event and successful grasping (Johansson and Flanagan, 2009). In this teleoperation scenario instead, the grasp had to be confirmed visually, hence the gaze lingered on the picked object and only after seeing the object moving along with the hand, moved on to the next distal intention (i.e., the placing position). Still, this kind of measures offers an insight into the user experience of the teleoperation task: as long as the uncertainty about the task execution is high, the gaze is less anticipative and lingers there where further information needs to be acquired to carry out the task. Although some of these issues could be mitigated with longer training, allowing the user to master the new visuomotor mapping and task (Sailer et al., 2005), an intention recognition model embedded in a shared autonomy architecture that could adjust the robot movement and grasping pose to reliably produce the intended grasp would shorten these training times. This would allow a more natural eye-hand coordination and relieve the gaze system of monitoring every sub-task unfolding and transition with extra care. That is, an effective shared autonomy system would be validated by shorter execution times, fewer failed grasp attempts, and more anticipative gaze behavior with less time spent monitoring the grasped object and the robot gripper. This would confirm that the user trusts the robotic partner to correctly infer and assist with the current intention but that their sense of agency is preserved since they anticipate the next subtask in their plan (see on this the discussion in Haji Fathaliyan et al., 2018).

Apart from these considerations, as shown for a different task (Keshava et al., 2020), we also found that the gaze behavior still was reliably different across tasks and could be hence learned and predicted effectively. To this end, a Hidden Markov Model was first devised for each of the intentions to be recognized. The normalized likelihoods of the gaze (represented as a Gaussian distribution) to be on each of the objects in the scene along with the grasping state were considered as emissions of the HMM. The system was trained on pick-and-place tasks from two users and then tested on similar unseen sequences from the two users plus two other users. Considering a time window of 0.9 s where emissions are accumulated and then scored by the six GHMMs, the system achieves a well above chance accuracy across all tasks, returning a prediction as early as after seeing 22% of the current action, on average. Here, the concept of predictability, indicating the portion of the task for which an estimate is available, relates to that of observational latency. As pointed out in Wang et al. (2020), even a very accurate prediction is of very low utility if it is not delivered in time for the system to plan and execute a supporting operation before the user has carried out the action themselves. The generalizability of the system was further tested on a different geometrical configuration of the pick task, delivering comparable accuracy and predictability. Even more accurate results were obtained by a second intention recognition scheme, which modeled the two basic actions (pick and place) and scored the likelihood of each picking or placing target by appropriately arranging the features representing the likelihood of the gaze on the different AoIs. Also, in this case, generalization was higher with new users and configurations, while practically no confusion between the two classes was observed. This kind of model offers also the possibility of scaling up the system to new picking objects and support surfaces, not previously seen during training: the dimension of the feature vector fed to the GHMMs stays in fact constant and the arrangement of the features determines which object is evaluated as pick/place target. The amount of gaze distribution captured by other objects of the same category (which should still be comparatively low compared to the real target) and by all those of the other category is indeed considered as two collective features, independent of the number of present objects or support surfaces.

A test without the binary grasping state feature, yet, showed a less consistent performance of the semantic model with respect to the naive model: the semantic model perfectly recognized the place intentions but mostly misclassified the pick intentions as place intentions. This might be due to the fact that the semantic model relies more strongly on the grasping state to determine the action and uses the gaze data to infer the object of interest, while the six naive GHMMs better learned the specificity of the scanpaths in the different conditions.

In any case, considering the similar semantic distributions of gaze time within equivalent sub-tasks and across spatial

configurations, these results suggest that there is a certain invariance in the gaze patterns. These are mainly shaped by the general sub-task at a higher level. At least in simple manipulation tasks and object configurations, sequences of gaze glances at objects are more heavily determined and constrained by the current subtask structure (pick vs. place), once the target is specified, rather than by the contingent spatial setup. That is, also the oculomotor plan subserving and directing the motor plan seems to reflect the syntactic structure of action (Pastra and Aloimonos, 2012).

These are promising results for the further development of our intention recognition system and its embedding in a real-world shared autonomy scenario. Current and future work is going to expand both the training and testing sets with multiple participants as well as considering more and different objects and tasks. A richer dataset with data from naive participants would indeed provide a better characterization not only of the users' sensorimotor behavior in itself, but it would indeed allow testing for learning effects within each participant and individual differences between participants. This would also help to investigate the co-adaptation process between human and robotic systems (Gallina et al., 2015). On the one hand, indeed, visuomotor adaptation to the new environment produces effective motor learning enabling the user to better handle the initially unfamiliar sensorimotor mapping. This effect could override some of the features learned by the intention recognition framework and should hence be accounted for. On the other hand, it should be investigated how different users cope on a sensorimotor level with the same task. This could both help to understand the generalization limits across users of a pre-trained model and to identify possibilities for user customization. Experimenting with a more complex scenario in terms of number and configuration of objects and support surfaces would very likely affect the high accuracy observed in this study, yet would also offer insight into meaningful ways to effectively assist the user and on ways to tackle the trade-off between accuracy and observational latency also downstream, at the behavior control level.

DATA AVAILABILITY STATEMENT

The dataset presented in this article is not readily available: considering the small number of participants, the dataset was not intended for public dissemination. Requests to access the dataset can be directed to Anna Belardinelli (anna.belardinelli@honda-ri.de).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Honda Bioethics Committee. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

SF and AB designed and conducted the study. SF devised the computational models. SF and AB analyzed the results. AB performed the behavioral data analysis and interpretation and wrote the main manuscript.

REFERENCES

Aarno, D., and Kragic, D. (2008). Motion intention recognition in robot assisted applications. *Robot. Auton. Syst.* 56, 692–705. doi: 10.1016/j.robot.2007.11.005

Admoni, H., and Srinivasa, S. (2016). "Predicting user intent through eye gaze for shared autonomy," in *2016 AAAI Fall Symposium Series* (Arlington, VA).

Aronson, R. M., and Admoni, H. (2018). "Gaze for error detection during human-robot shared manipulation," in *Fundamentals of Joint Action Workshop, Robotics: Science and Systems* (Pittsburgh, PA).

Aronson, R. M., Santini, T., Kübler, T. C., Kasneci, E., Srinivasa, S., and Admoni, H. (2018). "Eye-hand behavior in human-robot shared manipulation," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (Chicago, IL), 4–13. doi: 10.1145/3171221.3171287

Beer, J. M., Fisk, A. D., and Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *J. Hum. Robot Interact.* 3, 74–99. doi: 10.5898/JHRI.3.2.Beer

Belardinelli, A., Barabas, M., Himmelbach, M., and Butz, M. V. (2016). Anticipatory eye fixations reveal tool knowledge for tool interaction. *Exp. Brain Res.* 234, 2415–2431. doi: 10.1007/s00221-016-4646-0

Belardinelli, A., Pirri, F., and Carbone, A. (2007). Bottom-up gaze shifts and fixations learning by imitation. *IEEE Trans. Syst. Man Cybernet. B Cybernet.* 37, 256–271. doi: 10.1109/TSMCB.2006.886950

Boccignone, G. (2019). "Advanced statistical methods for eye movement analysis and modelling: a gentle introduction," in *Eye Movement Research*, eds C. Klein and U. Ettinger (Cham: Springer), 309–405. doi: 10.1007/978-3-030-20085-5_9

Boisvert, J. F., and Bruce, N. D. (2016). Predicting task from eye movements: on the importance of spatial distribution, dynamics, and image features. *Neurocomputing* 207, 653–668. doi: 10.1016/j.neucom.2016.05.047

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

Bratman, M. (1987). *Intention, Plans, and Practical Reason*, Vol. 10. Cambridge, MA: Harvard University Press.

Cio, Y. S. L. K., Raison, M., Leblond Ménard, C., and Achiche, S. (2019). Proof of concept of an assistive robotic arm control using artificial stereovision and eye-tracking. *IEEE Trans. Neural Syst. Rehabil. Eng.* 27, 2344–2352. doi: 10.1109/TNSRE.2019.2950619

Coutrot, A., Hsiao, J. H., and Chan, A. B. (2018). Scanpath modeling and classification with hidden markov models. *Behav. Res. Methods* 50, 362–379. doi: 10.3758/s13428-017-0876-8

Ellis, C., Masood, S. Z., Tappen, M. F., LaViola, J. J., and Sukthankar, R. (2013). Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vis.* 101, 420–436. doi: 10.1007/s11263-012-0550-7

Fathi, A., Li, Y., and Rehg, J. M. (2012). "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision* (Berlin: Springer), 314–327. doi: 10.1007/978-3-642-33718-5_23

Fiehler, K., Brenner, E., and Spering, M. (2019). Prediction in goal-directed action. *J. Vis.* 19, 10–10. doi: 10.1167/19.9.10

Gallina, P., Bellotto, N., and Di Luca, M. (2015). "Progressive co-adaptation in human-machine interaction," in *2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, Vol. 2 (Colmar: IEEE), 362–368. doi: 10.5220/0005561003620368

Goodrich, M. A., Crandall, J. W., and Barakova, E. (2013). Teleoperation and beyond for assistive humanoid robots. *Rev. Hum. Factors Ergon.* 9, 175–226. doi: 10.1177/1557234X13502463

Haji Fathaliyan, A., Wang, X., and Santos, V. J. (2018). Exploiting three-dimensional gaze tracking for action recognition during bimanual

manipulation to enhance human-robot collaboration. *Front. Robot. AI* 5:25. doi: 10.3389/frobt.2018.00025

Haji-Abolhassani, A., and Clark, J. J. (2014). An inverse yarbus process: predicting observer's task from eye movement patterns. *Vis. Res.* 103, 127–142. doi: 10.1016/j.visres.2014.08.014

Hauser, K. (2013). Recognition, prediction, and planning for assisted teleoperation of freeform tasks. *Auton. Robots* 35, 241–254. doi: 10.1007/s10514-013-9350-3

Hayhoe, M. M. (2017). Vision and action. *Annu. Rev. Vis. Sci.* 3, 389–413. doi: 10.1146/annurev-vision-102016-061437

Hayhoe, M. M., Shrivastava, A., Mruczek, R., and Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *J. Vis.* 3, 6–6. doi: 10.1167/3.1.6

Henderson, J. M. (2017). Gaze control as prediction. *Trends Cogn. Sci.* 21, 15–23. doi: 10.1016/j.tics.2016.11.003

Huang, C. M., and Mutlu, B. (2016). "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Christchurch: IEEE), 83–90. doi: 10.1109/HRI.2016.7451737

Jain, S., and Argall, B. (2018). "Recursive bayesian human intent recognition in shared-control robotics," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid: IEEE), 3905–3912. doi: 10.1109/IROS.2018.8593766

Jain, S., and Argall, B. (2019). Probabilistic human intent recognition for shared autonomy in assistive robotics. *ACM Trans. Hum. Robot Interact.* 9, 1–23. doi: 10.1145/3359614

Javdani, S., Srinivasa, S. S., and Bagnell, J. A. (2015). "Shared autonomy via hindsight optimization," in *Robotics Science and Systems: Online Proceedings*, Vol. 2015 (Rome: NIH Public Access). doi: 10.15607/RSS.2015.XI.032

Johansson, R. S., and Flanagan, J. R. (2009). Coding and use of tactile signals from the fingertips in object manipulation tasks. *Nat. Rev. Neurosci.* 10, 345–359. doi: 10.1038/nrn2621

Johansson, R. S., Westling, G., Bäckström, A., and Flanagan, J. R. (2001). Eye-hand coordination in object manipulation. *J. Neurosci.* 21, 6917–6932. doi: 10.1523/JNEUROSCI.21-17-06917.2001

Kanan, C., Ray, N. A., Bseiso, D. N., Hsiao, J. H., and Cottrell, G. W. (2014). "Predicting an observer's task using multi-fixation pattern analysis," in *Proceedings of the Symposium on Eye Tracking Research and Applications* (Safety Harbor, FL), 287–290. doi: 10.1145/2578153.2578208

Keshava, A., Aumeistere, A., Izdebski, K., and Konig, P. (2020). "Decoding task from oculomotor behavior in virtual reality," in *Symposium on Eye Tracking Research and Applications* (Stuttgart), 1–5. doi: 10.1145/3379156.3391338

Land, M., Mennie, N., and Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. *Perception* 28, 1311–1328. doi: 10.1068/p2935

Li, S., and Zhang, X. (2017). Implicit intention communication in human–robot interaction through visual behavior studies. *IEEE Trans. Hum. Mach. Syst.* 47, 437–448. doi: 10.1109/THMS.2017.2647882

Li, S., Zhang, X., and Webb, J. D. (2017). 3-D-gaze-based robotic grasping through mimicking human visuomotor function for people with motion impairments. *IEEE Trans. Biomed. Eng.* 64, 2824–2835. doi: 10.1109/TBME.2017.2677902

Miall, R., and Reckess, G. (2002). The cerebellum and the timing of coordinated eye and hand tracking. *Brain Cogn.* 48, 212–226. doi: 10.1006/brcg.2001.1314

Nguyen, T. H. C., Nebel, J. C., and Florez-Revuelta, F. (2016). Recognition of activities of daily living with egocentric vision: a review. *Sensors* 16:72. doi: 10.3390/s16010072

Ogaki, K., Kitani, K. M., Sugano, Y., and Sato, Y. (2012). "Coupling eye-motion and ego-motion features for first-person activity recognition," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (IEEE)* (Providence, RI), 1–7. doi: 10.1109/CVPRW.2012.6239188

Pacherie, E. (2008). The phenomenology of action: a conceptual framework. *Cognition* 107, 179–217. doi: 10.1016/j.cognition.2007.09.003

Pastra, K., and Aloimonos, Y. (2012). The minimalist grammar of action. *Philos. Trans. R. Soc. B Biol. Sci.* 367, 103–117. doi: 10.1098/rstb.2011.0123

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

Salvucci, D., and Goldberg, J. (2000). "Identifying fixations and saccades in eye-tracking protocols," in *ETRA '00: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications* (Palm Beach Garden, FL), 71–78. doi: 10.1145/355017.355028

Schettino, V., and Demiris, Y. (2019). "Inference of user-intention in remote robot wheelchair assistance using multimodal interfaces," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Macau), 4600–4606. doi: 10.1109/IROS40897.2019.8968203

Schilling, M., Kopp, S., Wachsmuth, S., Wrede, B., Ritter, H., Brox, T., et al. (2016). "Towards a multidimensional perspective on shared autonomy," in *2016 AAAI Fall Symposium Series* (Arlington, VA).

Shafti, A., Orlov, P., and Faisal, A. A. (2019). "Gaze-based, context-aware robotic system for assisted reaching and grasping," in *2019 International Conference on Robotics and Automation (ICRA)* (Montreal, QC: IEEE), 863–869. doi: 10.1109/ICRA.2019.8793804

Tanwani, A. K., and Calinon, S. (2017). "A generative model for intention recognition and manipulation assistance in teleoperation," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC: IEEE), 43–50. doi: 10.1109/IROS.2017.8202136

Wang, M., Kogkas, A. A., Darzi, A., and Mylonas, G. P. (2018). "Free-view, 3D gaze-guided, assistive robotic system for activities of daily living," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Madrid), 2355–2361. doi: 10.1109/IROS.2018.8594045

Wang, X., Haji Fathaliyan, A., and Santos, V. J. (2020). Toward shared autonomy control schemes for human-robot systems: action primitive recognition using eye gaze features. *Front. Neurorob.* 14:66. doi: 10.3389/fnbot.2020.567571

Yang, T., Huang, W., Chui, C. K., Jiang, Z., and Jiang, L. (2017). "Stacked hidden markov model for motion intention recognition," in *2017 IEEE 2nd International Conference on Signal and Image Processing (ICSIP)* (Singapore), 266–271. doi: 10.1109/SIPROCESS.2017.8124546

Yarbus, A. L. (1967). "Eye movements during perception of complex objects," in *Eye Movements and Vision* (Boston, MA: Springer), 171–211. doi: 10.1007/978-1-4899-5379-7_8

Yi, W., and Ballard, D. (2009). Recognizing behavior in hand-eye coordination patterns. *Int. J. Hum. Robot.* 6, 337–359. doi: 10.1142/S0219843609001863

Yu, W., Alqasemi, R., Dubey, R., and Pernalete, N. (2005). "Telemanipulation assistance based on motion intention recognition," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation (IEEE)* (Barcelona), 1121–1126.

Zeng, H., Shen, Y., Hu, X., Song, A., Xu, B., Li, H., et al. (2020). Semi-autonomous robotic arm reaching with hybrid gaze-brain machine interface. *Front. Neurorobot.* 13:111. doi: 10.3389/fnbot.2019.00111

Conflict of Interest: When conducting the study both authors were employed by the Honda Research Institute Europe GmbH. During the writing of the manuscript, SF moved to Siemens AG.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 09 June 2021
doi: 10.3389/fnbot.2021.648527

A Dynamical Generative Model of Social Interactions

*Alessandro Salatiello†, Mohammad Hovaidi-Ardestani† and Martin A. Giese**

Section for Computational Sensomotorics, Department of Cognitive Neurology, Centre for Integrative Neuroscience, Hertie Institute for Clinical Brain Research, University Clinic Tübingen, Tübingen, Germany

The ability to make accurate social inferences makes humans able to navigate and act in their social environment effortlessly. Converging evidence shows that motion is one of the most informative cues in shaping the perception of social interactions. However, the scarcity of parameterized generative models for the generation of highly-controlled stimuli has slowed down both the identification of the most critical motion features and the understanding of the computational mechanisms underlying their extraction and processing from rich visual inputs. In this work, we introduce a novel generative model for the automatic generation of an arbitrarily large number of videos of socially interacting agents for comprehensive studies of social perception. The proposed framework, validated with three psychophysical experiments, allows generating as many as 15 distinct interaction classes. The model builds on classical dynamical system models of biological navigation and is able to generate visual stimuli that are parametrically controlled and representative of a heterogeneous set of social interaction classes. The proposed method represents thus an important tool for experiments aimed at unveiling the computational mechanisms mediating the perception of social interactions. The ability to generate highly-controlled stimuli makes the model valuable not only to conduct behavioral and neuroimaging studies, but also to develop and validate neural models of social inference, and machine vision systems for the automatic recognition of social interactions. In fact, contrasting human and model responses to a heterogeneous set of highly-controlled stimuli can help to identify critical computational steps in the processing of social interaction stimuli.

Keywords: social interactions, generative model, motion cues, social perception, social inference

OPEN ACCESS

Edited by:
Letizia Marchegiani, Aalborg University, Denmark

Reviewed by:
Ashley Liddiard, Ford Motor Company, United States
Bin Zhi Li, Chongqing Institute of Green and Intelligent Technology (CAS), China

***Correspondence:**
Martin A. Giese
martin.giese@uni-tuebingen.de

†These authors have contributed equally to this work

Received: *31 December 2020*
Accepted: *23 April 2021*
Published: *09 June 2021*

Citation:
Salatiello A, Hovaidi-Ardestani M and Giese MA (2021) A Dynamical Generative Model of Social Interactions. Front. Neurorobot. 15:648527. doi: 10.3389/fnbot.2021.648527

1. INTRODUCTION

Human and non-human primates are able to recognize the social interactions taking place in their environment quickly and effortlessly: with a few glances out of the window, we can easily understand whether two people are following each other, avoiding each other, fighting, or are engaging in some other form of social behavior. Notably, such interactive behaviors can be recognized even when the available visual information is poor: for example, when the scene we are watching is unfolding behind the leaves of a tree, at a considerable distance from us, or in a low-resolution video. In some of these situations, critical visual cues such as facial expressions might be completely occluded, yet our ability to make social inference is largely unaffected. Such perceptual ability is instrumental in allowing us to move in our social environment and flexibly interact with it, while abiding by the social norms (Troje et al., 2013). Therefore, it constitutes an important social skill that is worth characterizing and modeling also for the development of social robots.

Understanding the neural mechanisms underlying the inference of animacy and social interactions from visual inputs is a long-standing research challenge (Heider and Simmel, 1944; Michotte, 1946; Scholl and Tremoulet, 2000; Troje et al., 2013). Recent work has started identifying some of the responsible neural circuits (Castelli et al., 2000; Isik et al., 2017; Sliwa and Freiwald, 2017; Walbrin et al., 2018; Freiwald, 2020). Even though the detailed computational mechanisms mediating the formation of social percepts from visual inputs remain largely unknown, converging evidence has shown that the observation of biological motion alone is enough for humans to make accurate social inferences (e.g., Heider and Simmel, 1944; Tremoulet and Feldman, 2000; McAleer and Pollick, 2008; Roether et al., 2009). For example, Heider and Simmel (1944) demonstrated that humans can reliably decode animacy and social interactions from strongly impoverished stimuli consisting of simple geometrical figures moving around in the two-dimensional plane. Remarkably, despite their highly abstract nature, the visual stimuli used in this study were perceived as *alive* and sometimes even *anthropomorphic*: the agents were often considered as endowed with intentions, emotions, and even personality traits.

Several subsequent studies (e.g., Oatley and Yuill, 1985; Rimé et al., 1985; Springer et al., 1996; Castelli et al., 2000, 2002) replicated these findings using similar stimuli and showed that the inference of social interactions from impoverished stimuli is a cross-cultural phenomenon (Rimé et al., 1985) that is present even in 5-year-old preschoolers (Springer et al., 1996). Taken together, these findings support the view that the perception of animacy and social interactions might rely on some innate and automatic processing of low-level kinematic features present in the visual inputs, rather than on higher-level cognitive processing (Scholl and Gao, 2013).

The identification of the most critical visual features that shape these social percepts has also received great attention (Tremoulet and Feldman, 2000, 2006). For example, influential work suggested that these percepts are mediated by the detection of apparent violations of the principle of conservation of energy (Dittrich and Lea, 1994; Gelman et al., 1995; Csibra, 2008; Kaduk et al., 2013). Later research proved that also agent's orientation, velocity, and acceleration play a major role (Szego and Rutherford, 2008; Träuble et al., 2014). At the same time, neuroimaging work has shed light on some of the brain regions mediating these phenomena: the right posterior superior temporal sulcus (pSTS—Isik et al., 2017; Walbrin et al., 2018), the medial prefrontal cortex (mPFC—Castelli et al., 2000; Sliwa and Freiwald, 2017), and the right temporoparietal junction (TPJ—Castelli et al., 2000; Saxe and Kanwisher, 2003) are among the brain regions most frequently reported as being involved in the perception of social interaction. Interestingly, Schultz and Bülthoff (2019), recently identified another region—the right intraparietal sulcus (IPS)—that seems to be exclusively engaged during the perception of animacy.

Clearly, the success of both behavioral and neuroimaging social perception studies is tightly linked to the ability to finely control the visual stimuli that participants are exposed to. Specifically, such stimuli should ideally be generated through a process that allows complete parametric control, the creation of a high number of replicates with sufficient variety, and the gradual reduction of complexity. *Parametric control* (e.g., over agents' speed) facilitates the identification of brain regions and individual neurons whose activation covaries with the kinematic features of agents' behavior. *Variety* in classes of social interaction allows the characterization of the class-specific and general response properties of such brain regions. *Numerosity* allows averaging out response properties that are independent of social interaction processing. Finally, the ability to control stimulus complexity allows the generation of *impoverished stimuli* that are fundamental to minimize the impact of confounding factors, inevitably present, for example, in real videos. Similarly, such properties are also desirable when designing and validating neural and mechanistic models of human social perceptions: contrasting human and model responses to a variety of highly controlled stimuli can help discriminate between the computational mechanisms that the models capture well from those that need further refinement. This is especially critical for state-of-the-art deep learning models (e.g., Yamins et al., 2014), which can easily have millions of parameters and be prone to over-fitting.

Currently, no well-established method can generate visual stimuli for the analysis of social perception that satisfy all of the above conditions. Because of this, researchers often have to resort to time-consuming and class-specific, heuristic procedures. A creative approach to this problem has been the one adopted by Gordon and Roemmele (2014), where the task of generating videos was assigned to a set of participants—who were asked to create their own videos of socially interacting geometrical shapes, and to label them accordingly. However, typically, researches use visual stimuli where agents' trajectories are hand-crafted or hard-coded (e.g., Heider and Simmel, 1944; Oatley and Yuill, 1985; Rimé et al., 1985; Springer et al., 1996; Castelli et al., 2000, 2002; Baker et al., 2009; Gao et al., 2009, 2010; Kaduk et al., 2013; Träuble et al., 2014; Isik et al., 2017; van Buren et al., 2017; Walbrin et al., 2018), based on rules (e.g., Kerr and Cohen, 2010; Pantelis et al., 2014), or derived from real videos (e.g., McAleer and Pollick, 2008; McAleer et al., 2011; Thurman and Lu, 2014; Sliwa and Freiwald, 2017; Shu et al., 2018). All of these approaches suffer from significant limitations. Hand-crafted trajectories need to be generated *de novo* for each experimental condition and are not easily amenable to parametric control. Likewise, the extraction of trajectories from real videos also comes with its burdens: real videos need to be recorded, labeled, and heavily processed to remove unwanted background information. Rule-based approaches offer an interesting alternative. However, it is generally difficult to define natural classes of social interactions using rules akin to those used in Kerr and Cohen (2010) and Pantelis et al. (2014). Recent work (Schultz and Bülthoff, 2019; Shu et al., 2019, 2020) has generated visual stimuli using model-based methods; however, these models can only generate limited and generic classes of social interaction (namely, cooperative and obstructive behaviors). Finally, specialized literature on the collective behavior of humans and animals has produced a wealth of influential models (Blackwell, 1997; Paris et al., 2007; Luo et al., 2008; Russell et al., 2017); however, such models can

also typically account only for simple behaviors (e.g., feeding, resting, and traveling) and for basic interactions (e.g., avoidance and following).

To overcome the limitations of the above methods, in this work, we introduce a dynamical generative model of social interactions. In stark contrast to previous work, our model is able to automatically generate an arbitrary number of parameterized motion trajectories to animate virtual agents with 15 distinct interactive motion styles; the modeled trajectories include the six fundamental interaction categories frequently used in psychophysical experiments (i.e., *Chasing*, *Fighting*, *Flirting*, *Following*, *Guarding*, and *Playing*—Blythe et al. 1999; Barrett et al. 2005; McAleer and Pollick 2008) and nine relevant others. The model controls *speed*, and *motion direction*, arguably the two most critical determinants of social interaction perception (Tremoulet and Feldman, 2000; Szego and Rutherford, 2008; Träuble et al., 2014). Finally, we validated the model with three psychophysical experiments, which demonstrate that participants are able to consistently attribute the intended interaction classes to the animations generated with our model.

The rest of the paper is organized as follows. In section 2, we describe the generative model and the experiments we conducted to validate it. Next, in section 3, we summarize the experimental results. Finally, in section 4, we (1) explain how our results validate the developed model, (2) explain how the model compares to related work, and (3) discuss the main limitations of our model and future directions.

2. METHODS

2.1. Related Modeling Work

The generative model we introduce in this work builds on classical models of biological and robotic navigation. In the classical work by Reichardt and Poggio (1976), the authors proposed a dynamical model to describe the navigation behavior of flies intent on chasing moving targets as part of their mating behavior. The core idea was to consider the moving targets as *attractors* of the dynamical system describing the flies' trajectories. Subsequently, Schöner and Dose (1992) and Schöner et al. (1995) used a similar approach to develop a biomimetic control system for the navigation of autonomous robots. Critically, such a system was also able to deal with the presence of obstacles in the environment, which were modeled as *repellors*. Extending this system, Fajen and Warren (2003) built a model of human navigation that was able to closely capture the trajectories described by their participants as they walked naturally toward targets while avoiding obstacles on their way. Specifically, this model was able to describe the dynamics of the participants' average heading direction very accurately; however, their speed was roughly approximated as constant.

Alternative approaches can characterize richer navigation behaviors by jointly modeling both heading direction and speed dynamics. This idea was successfully used to control the motion of both autonomous vehicles (Bicho and Schöner, 1997; Bicho et al., 2000) and robotic arms (Reimann et al., 2011). Similar approaches have also been used in computer graphics to model the navigation of articulated agents (Mukovskiy et al., 2013).

2.2. The Generative Model

To model the interactive behavior of two virtual agents, we define, for each agent i, a dynamical system of two nonlinear differential equations. Specifically, the equations describe the dynamics of the agent's heading direction $\phi_i(t)$ and instantaneous propagation speed $s_i(t)$.

The heading direction dynamics, derived from Fajen and Warren (2003), are defined by:

$$\ddot{\phi}_i(t) = -b\dot{\phi}_i(t) + A(\phi_i(t), \psi_i^g(t)) + R(\phi_i(t), \boldsymbol{\psi}_i^o(t)) \tag{1}$$

In this equation, $A(\phi_i(t), \psi_i^g(t))$ defines the *attraction* of agent i to the goal g located along the direction $\psi_i^g(t)$, at a distance $d_i^g(t)$ from it. Similarly, $R(\phi_i(t), \boldsymbol{\psi}_i^o(t))$ defines the *repulsion* of agent i for the obstacles $\boldsymbol{o} = [o_1, o_2, ..., o_{N_{obst}}]^T$ located along the directions $\boldsymbol{\psi}_i^o(t)$, at a distance $\boldsymbol{d}_i^o(t)$ from it. These two functions are given by:

$$\begin{aligned} A(\phi_i(t), \psi_i^g(t)) &= -k^g(\phi_i(t) - \psi_i^g(t))(e^{-c_1 d_i^g(t)} + c_2) \\ R(\phi_i(t), \boldsymbol{\psi}_i^o(t)) &= k^o \sum_{n=1}^{N_{obst}} r^{o_n}(\phi_i(t)) \end{aligned} \tag{2}$$

The contributions of the individual obstacles to the repulsion function are given by:

$$r^{o_n}(\phi_i(t)) = (\phi_i(t) - \psi_i^{o_n}(t))(e^{-c_3|\phi_i(t) - \psi_i^{o_n}(t)|})(e^{-c_4 d_i^{o_n}(t)}) \tag{3}$$

In these equations, k^j and c_j are constants; o_n indicates the *nth* obstacle. Note that, in general, $\psi_i^{o_n}(t)$, which is the direction of the *nth* obstacle of the *ith* agent is time-dependent; for example, depending on the specific social interaction class it might be a function of the instantaneous heading direction of other agents.

The propagation speed dynamics are specified by the following stochastic differential equation:

$$\tau \dot{s}_i(t) = -s_i(t) + F_i(d_i^g(t)) + k_i^\epsilon \epsilon_i(t) \tag{4}$$

where $\epsilon_i(t)$ is Gaussian white noise. The nonlinear function F_i specifies how the agent's speed changes as a function of the distance from its goal:

$$F_i(d) = \frac{c_5}{1 + e^{-c_6^i(d - c_7^i)}} - c_8^i e^{-k_i d} + c_9^i \tag{5}$$

Critically, we choose this specific functional form because it provides us with enough flexibility to reproduce several relevant interaction classes, including the six fundamental interaction categories traditionally studied in psychophysical experiments (Blythe et al., 1999; Barrett et al., 2005; McAleer and Pollick, 2008): *Chasing*, *Fighting*, *Flirting*, *Following*, *Guarding*, and *Playing*.

To generate the trajectories, we first randomly sample a series of goal points for the first agent from a two-dimensional uniform distribution over the 2D plane of action. Such goal points are commonly referred to as *via points*. We then use the instantaneous position of the first agent as goal position for the

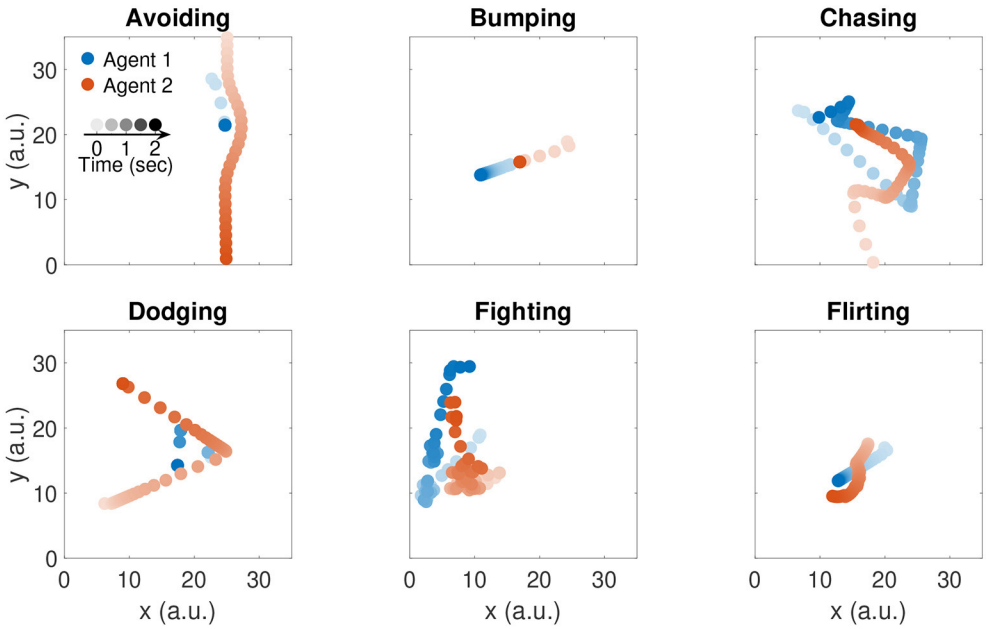


FIGURE 1 | Trajectories of six example social interactions. Color indicates agent identity: agent 1 is represented in blue; agent 2 is represented in red. Color saturation indicates time: darker colors indicate recent time samples.

Algorithm 1: Pseudocode for trajectory generation

Input: Class-specific parameters θ_c
Output: Agents' direction Φ and speed S
for each timestep t **do**
  for each agent i **do**
    compute goal direction $\psi_i^g(t)$
    compute distance from goal $d_i^g(t)$
    for each obstacle o_n **do**
      compute obstacle direction $\psi_i^{o_n}(t)$
      compute distance from obstacle $d_i^{o_n}(t)$
    end
    compute $\phi_i(t)$ integrating Equation (1)
    compute $s_i(t)$ integrating Equation (4)
  end
end
/*Note: $\psi_i^g(t)$ and $\psi_i^{o_n}(t)$ are either specified a priori or computed dynamically depending on the agent and social interaction class. For example, for simple behaviors (e.g., chasing) $\psi_1^g(t)$ and $\psi_1^{o_n}(t)$ are specified a priori, while $\psi_2^g(t) = \phi_1(t)$*/

second agent. Samples that are too close to the current agent's position are rejected. Further details about the implementation of the generative model are provided in the Algorithm 1 box. Representative trajectories of six example social interactions are illustrated in **Figure 1**. Note that the speed control dynamics are not influenced by the presence of obstacles, since their effect was not needed to realistically capture the social interactive behaviors we chose to model.

2.3. Model Validation

To assess whether our model is able to generate perceptually valid socially interactive behaviors, we carried out three behavioral experiments. In these experiments, we asked participants to categorize videos of interacting agents generated with our model in a free-choice task (Experiment 1), and in a forced-choice task (Experiment 2). Finally, we analyzed the semantic similarities between the labels chosen by the participants (Experiment 3).

2.3.1. Dataset Generation

To validate our approach, we chose to model the six fundamental interaction classes (i.e., *Chasing*, *Fighting*, *Flirting*, *Following*, *Guarding*, and *Playing*; Blythe et al. 1999; Barrett et al. 2005; McAleer and Pollick 2008), and nine other relevant ones (i.e., *Avoiding*, *Bumping*, *Dodging*, *Frightening*, *Meeting*, *Pulling Pushing*, *Tug of War*, and *Walking*) resulting in a total of 15 interaction classes. To generate the trajectories corresponding to these classes, we simulated the model with 15 distinct parameter sets, which we identified through a simulation-based heuristic procedure. A list of the most critical parameters is presented in **Table 1**. The complete dataset we generated for our experiments included five random realizations of each interaction class, for a total of 75 videos. Each random realization is defined by different via points and noise realizations.

2.3.2. Participants

A total of 39 participants with normal or corrected vision took part in the experiments: 13 in Experiment 1 (9 females, 4 males), ten in Experiment 2 (5 females, 5 males), and 16 in Experiment 3 (9 females, 7 males). All participants were college students attending the University of Tübingen and provided written

TABLE 1 | Main model parameters.

Interaction class	Agent 1							Agent 2						
	k	k^{ϵ}	c_5	c_6	c_7	c_8	c_9	k	k^{ϵ}	c_5	c_6	c_7	c_8	c_9
Avoiding	0	0	1	1	5	3	0	0	0	0.4	1	0	2.7	0
Bumping	0	0.9	1	0.8	0	0	0	0	1	0.8	10	0	1	0
Chasing	0	0	1	10	7	0	0	0	0	1	1	7	0	0
Dodging	0	0	1	0.5	7	5	0	0	0	3	1	0	0	0
Fighting	0.1	0	1	1	3	1	0	0.1	1	1	1	3	1	0
Flirting	0	0	1	1	5	0	0	0.5	1	0.6	1	2	1	0
Following	0	0	1	10	7	0	0	0	0	1	4	4	0	0
Frightening	0	0	1	1	5	0	0	0	0	1	1	5	0	0.5
Guarding	0	0	1	1	5	0	0	0	0	1	1	3	0	0.5
Meeting	0	0.2	1	2	0	6	0	0.5	1	0.22	3	0	6	0
Playing	0	0	1	1	5	0	0	0	1	1	1	10	0	0.5
Pulling	0	0	1	10	0	2.6	0	0	0	0.9	5	0	2.6	0
Pushing	0	0	1	10	0	2.5	0	0	0	0.1	1	0	0	2.5
Tug of War	0	0.2	1	10	0	6	0	0	0.5	0.9	5	0	0	0.5
Walking	0	0.2	1	10	0	1	0	0	0	0.22	10	0	0	0

informed consent before the experiments. All experiments were in full compliance with the Declaration of Helsinki. Participants were naïve to the purpose of the study and were financially compensated for their participation.

2.3.3. Experiment Setup

In Experiment 1 and Experiment 2, participants sat in a dimly lit room in front of an LCD monitor (resolution: 1,920 × 1,080, refresh rate: 60Hz), at a distance of 60cm from it. To ensure that all participants would observe the stimuli with the same view parameters and the same distance from the screen, they were asked to place their heads in a chin-and-forehead rest during the experimental sessions. The experiments started with a short familiarization session during which the participants learned to use the computer interface. Subsequently, the participants were shown the videos generated with our model. Their task was to describe the videos by using their own words (Experiment 1) or by selecting labels among those provided to them (Experiment 2), and to provide animacy ratings through a standard 0–10 Likert scale. To increase the confidence in their answers, we gave participants the opportunity to re-watch each video up to three times. The videos were presented in pseudo-randomized order over five blocks. Five-minute rest breaks were given after each block. The animated videos always showed two agents moving in a 2D plane following speed and direction dynamics generated offline with our model. Critically, unlike in previous work (Blythe et al., 1999; Barrett et al., 2005), our agents were very simple geometrical shapes, namely a blue circle and a red rectangle (as in Tremoulet and Feldman, 2000); this choice ensured that participants' perception would not be biased by additional visual cues beyond the agents' motion and relative positions. In Experiment 3, subjects were asked to fill out a questionnaire to rate the semantic similarity between social interaction classes (0–10 Likert scale).

2.3.4. Experiment 1

The first experiment was aimed at assessing whether subjects would perceive the motion of virtual agents generated with our model as a social interaction. The second goal of this experiment was the identification of unequivocal labels for the interaction classes generated with our model. To this end, we asked participants to watch all the videos in our stimulus set (section 2.3.1). After watching the videos, subjects were asked to provide their own interpretations by summarizing what they had perceived with a few sentences or keywords. Importantly, in this experiment, to make sure we would not bias the participants' perceptions, we did not provide them with any labels or other cues: they had to come up with their own words. In addition, subjects were asked to provide an animacy rating for each agent. The most commonly reported keywords were used as *ground-truth* interaction labels for the remaining experiments.

To test whether participants assigned different animacy ratings depending on agent identity and social interaction class, we fitted a linear mixed-effect model to the animacy ratings, with Agent and Social Interaction as fixed effects, and Subject as random effect:

$$\begin{aligned}\text{Animacy}_{sl} = \alpha_0 &+ \sum_{i=1}^{N_a} \beta_i \cdot \text{Agent}(i,l) \\ &+ \sum_{i=1}^{N_c} \gamma_i \cdot \text{SocialInteraction}(i,l) + b_{0s} + \epsilon_{sl} \quad (6)\end{aligned}$$

In this model, Animacy_{sl} is the lth animacy rating reported by subject s, with $s = 1, 2, ..., N_s$ and $l = 1, 2, ..., N_a N_c$; N_a, N_c, and N_s are the number of agents, social interaction classes, and subjects, respectively. Moreover, $\text{Agent}(i,l)$ is a dummy variable that is equal to 1 when the rating l is for agent i, and 0 otherwise. Similarly, $\text{SocialInteraction}(i,l)$ is a dummy variable that is equal

to 1 when the rating l is for social interaction i, and 0 otherwise. Finally, b_{0s} is the subject-specific random effect [$b_{0s} \sim N(0, \sigma_b^2)$] and ϵ_{sl} are the residual error terms [$\epsilon_{sl} \sim N(0, \sigma^2)$]. Notably, the model was fitted with a sum-to-zero constrain, that is $\sum_{i=1}^{N_a} \beta_i = 0$ and $\sum_{i=1}^{N_c} \gamma_i = 0$; therefore, in this model, α_0 represents the overall average animacy rating. All the analyses described in this and in the next sections were performed in MATLAB R2020a (The MathWorks, Natick, MA).

2.3.5. Experiment 2

The second experiment was aimed at further studying the social interaction classes perceived by the participants while watching our animated videos. To this end, new subjects were exposed to a subset of the videos in our original dataset. Specifically, for this experiment we excluded the videos corresponding to the classes *Following*, *Guarding*, and *Playing*, as these tended either to be often confused with other classes, or to be labeled with a broad variety of related terms. Critically, unlike in Experiment 1, after watching the videos, participants were asked to describe the videos by choosing up to three labels, among those selected in Experiment 1.

To assess the classification performance, we computed the confusion matrix M. In this matrix, each element $m_{i,j}$ is the number of times participants assigned the class j to a video from class i. Starting from M, we computed, for each social interaction class, Recall, Precision, and F_1 score. Recall measures the fraction of videos of class i that are correctly classified, and is defined as $Recall_i = m_{i=j} / \sum_{j=1}^{N_c} m_{i,j}$. Precision measures the fraction of times participants correctly assigned the class j to a video, and is defined as $Precision_j = m_{i=j} / \sum_{i=1}^{N_c} m_{i,j}$. Finally, the F_1 score is the harmonic mean of Precision and Recall; it measures the overall classification accuracy and is defined as $F_1 = 2 \cdot Precision \cdot Recall / (Precision + Recall)$.

To evaluate whether some classes were more likely to be confused with each other, we computed, for each pair of classes (i, j), with $i \neq j$, the empirical pairwise mislabeling probability, defined as $P_{MS}(i, j) = (m_{i,j} + m_{j,i}) / (\sum_{k=1}^{N_c} \sum_{l \neq k} m_{k,l})$.

To assess whether participants improved their classification performance during the experiment, we computed the average Precision, Recall, and F_1 score across social interaction class, as a function of experimental block; we then fitted linear models to test whether experimental block explained a significant fraction of variation in the performance measures defined above.

2.3.6. Experiment 3

The third and last experiment was aimed at assessing whether there are interpretable semantic similarities among the labels provided in Experiment 2. Some interaction classes were misclassified by the participants in Experiment 2. This suggests that either the generated animated videos are not distinctive enough or that the classes semantically overlap with each other. To disambiguate between the two options, we ran a semantic survey test with a new set of participants. Participants in this experiment did not watch any video. After providing them with precise definitions for each social interaction class, we asked them to indicate the level of semantic similarity for each pair of classes, by providing rates ranging from 0 to 10. Specifically, using this scoring system, participants were asked to assign 0 to pairs of classes perceived as not sharing any semantic similarity, and 10 to those perceived as equivalent classes.

To assess the geometry of the semantic similarity space, we first transformed all the similarity ratings s into distance ratings d by computing their complement (i.e., $d = 10 - s$), and then rescaled them between 0 and 1. All the resulting semantic distances collected from participant i were then stored in a matrix D^i. In this matrix, $D^i_{j,k} = 0$ if the classes j and k were considered as semantically equivalent by subject i; $D^i_{j,k} = 1$ if the classes j and k were considered as semantically unrelated. We then used non-metric multidimensional scaling (MDS; Shepard, 1962a,b) to visualize in a 2D space the underlying relational structure contained in the distance matrix.

To determine whether some groups of classes were consistently considered as semantically similar, we performed agglomerative hierarchical clustering on the distance matrix D using the Ward's linkage method (Ward, 1963), which minimizes the within-cluster variance. Clusters were then identified using a simple cut-off method, using as a threshold $\tau = 0.7 \cdot M_{WD}$, where M_{WD} is the maximum observed Ward's distance.

Finally, to estimate whether the semantic similarity between pairs of classes explained the mislabelings observed in Experiment 2, we computed the Pearson's correlation coefficient (ρ) between the empirical mislabeling probability $P_{MS}(j, k)$ measured in Experiment 2 and the semantic distance $D(j, k)$.

3. RESULTS

3.1. Experiment 1

As mentioned above, participants in this experiment were completely free to provide interpretations about the videos through either labels or short sentences. For each video class, we pooled together all the definitions and labels, and we considered the most used term as the *ground-truth* class label. **Figure 2** summarizes the reported labels for six example social interaction classes. The pie charts show that some classes such as *Avoiding* and *Fighting* tended to be consistently described with very few labels (i.e., 2 − 3). Other classes such as *Dodging* were instead described with more labels (i.e., 6). Regardless of the number of labels used to describe a social interaction class, these were generally semantically similar. For example, some classes were named interchangeably depending on the perspective from which subjects reported their interpretation about the videos. A typical example of this issue is the ambiguity between the classes *Pulling* and *Pushing*. On the other hand, some other classes (for instance *Bumping* and *Pushing*) were sometimes misclassified regardless of the perspective from which subjects might have observed the videos.

Average animacy ratings are reported in **Figure 3A**, with classes sorted in ascending order of average across-agent animacy. Agents, were consistently perceived as animate [$\alpha_0 = 53.27\%, t_{(299)} = 11.72, p = 2.3 \cdot 10^{-26}$]. This is consistent with the fact that self-propulsion (Csibra, 2008), goal directedness (van Buren et al., 2016), being reactive to social contingencies (Dittrich

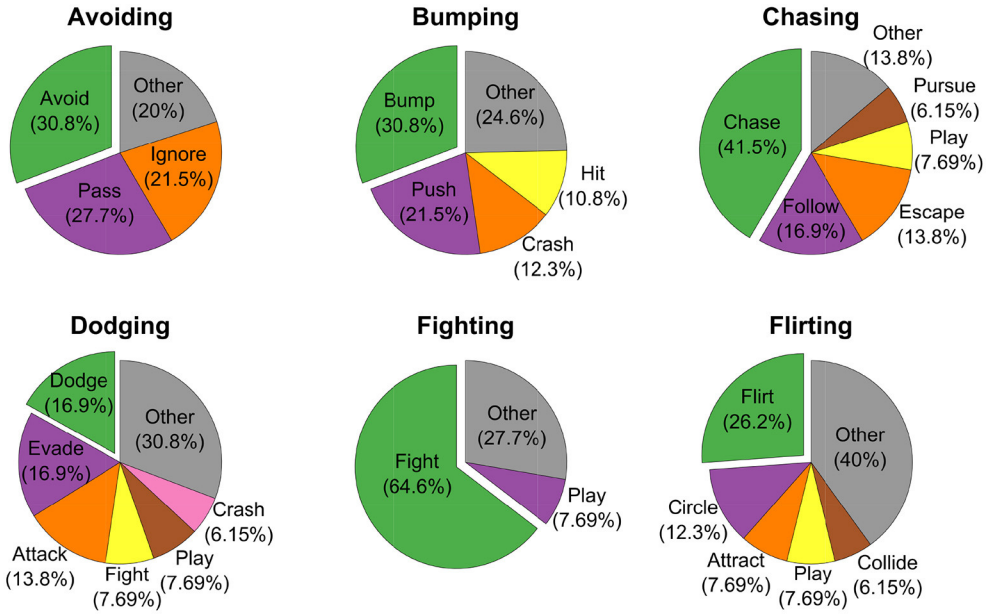


FIGURE 2 | Distribution of reported keywords for six example social interactions. Pie charts' titles indicate the true classes. Individual slices are assigned to all the keywords reported in Experiment 1 occurring with a frequency >5%. Keywords reported with a frequency <5% are pooled together in the slice *Other* (in gray). Offset slices (in green) represent the most frequently reported keywords.

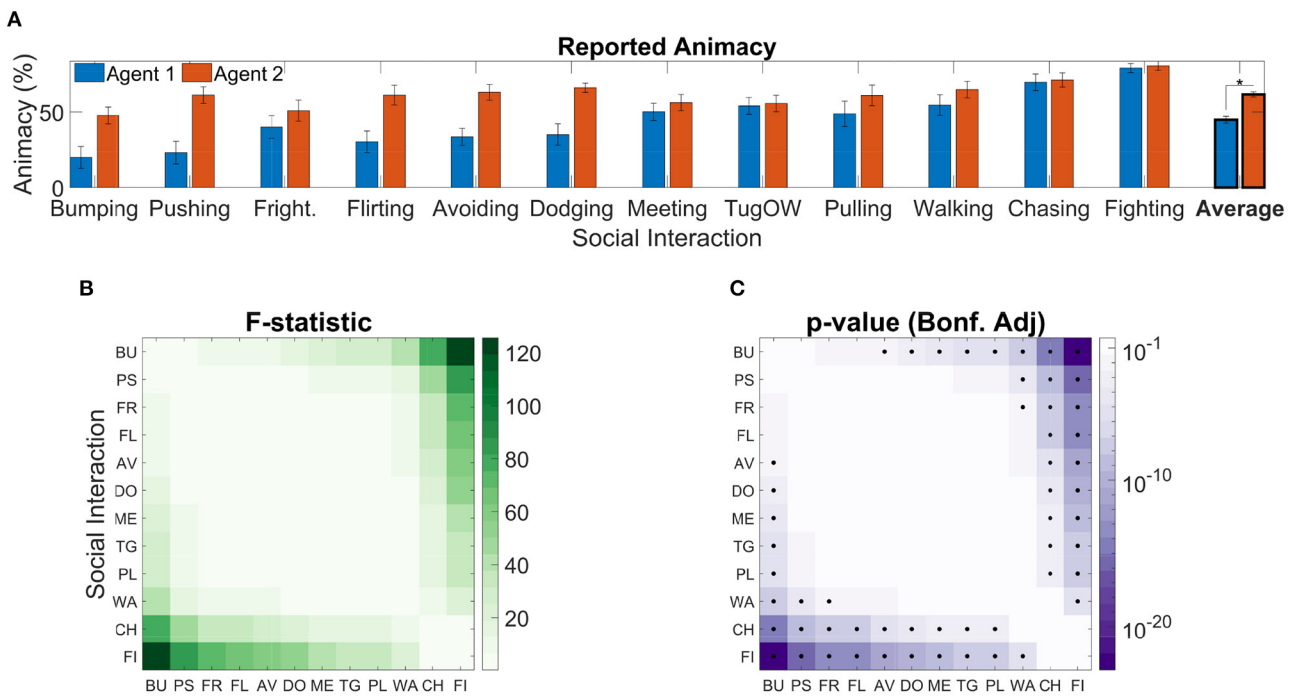


FIGURE 3 | Reported agent animacy. **(A)** Mean animacy ratings obtained in Experiment 1; error bars represent standard errors; results are rescaled between 0 and 100. Classes are sorted in ascending order by average across-agent animacy rating. The asterisk denotes a significant effect ($p < 0.05$) of Agent on Animacy [$F_{(1,299)} = 99.98$, $p = 1.74 \cdot 10^{-20}$]. **(B)** F-statistics of *post-hoc* tests to assess the difference in animacy ratings between social interaction classes [i.e., $F_{(1,299)}$]. **(C)** Bonferroni adjusted *p*-values corresponding to the F-statistics reported in **(B)**; black dots represent significant pairwise differences.

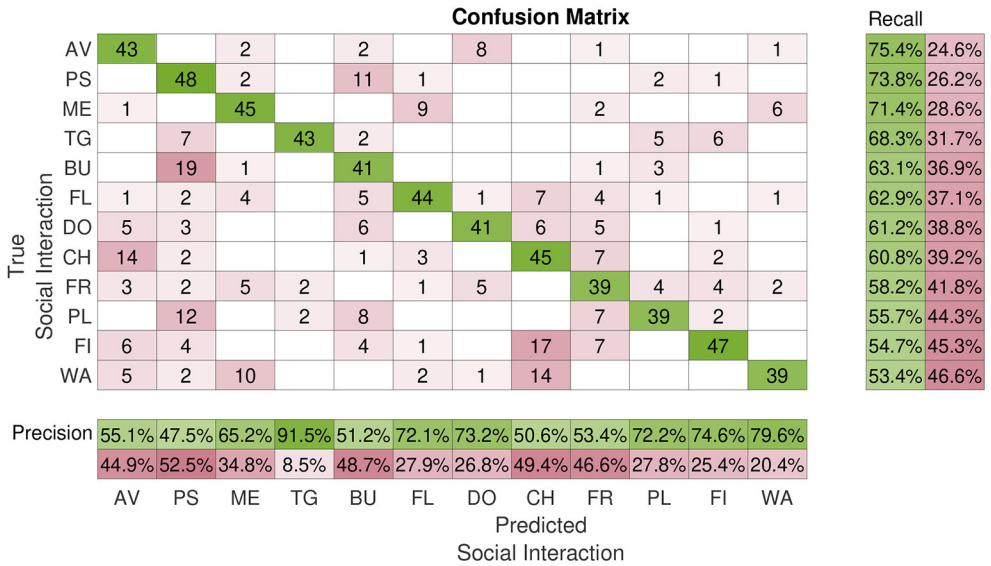


FIGURE 4 | Average classification performance. This figure shows the confusion matrix of the classification experiment (Experiment 2). Rows represent the true interaction class; columns the interaction class reported by the participants in Experiment 2. Matrix entries $m_{i,j}$ report the number of times participants assigned the class j to a video from class i. Rows and columns are sorted by decreasing Recall. AV, avoiding; BU, bumping; CH, chasing; DO, dodging; FI, fighting; FL, flirting; FR, frightening; ME, meeting; PL, pulling; PS, pushing; TG, tug of war; WA, walking.

and Lea, 1994), acceleration (Tremoulet and Feldman, 2000), and speed (Szego and Rutherford, 2008) are the most prominent cues for perceived animacy in psychophysical experiments. Moreover, the blue circle was consistently rated as less animate than the red rectangle [$\beta_1 = -\beta_2 = -8.37\%$, $t_{(299)} = -10$, $p = 1.74 \cdot 10^{-20}$], consistently with the finding that geometrical figures with a body axis are perceived as more animate than those without one, such as circles (Tremoulet and Feldman, 2000).

We further found a significant effect of social interaction on animacy [$F_{(11,299)} = 18.3$, $p = 8.29 \cdot 10^{-28}$]; this suggests that certain classes of social interactions tended to elicit stronger animacy percepts than others. To assess which specific pairs of classes were assigned significantly different animacy rating, we performed *post-hoc F*-tests. This analysis revealed that some classes consistently received higher average animacy ratings: for example, *Fighting* received higher animacy ratings than all other classes [$F_{(1,299)} \geq 24.04$, $p_{adj} \leq 1.03 \cdot 10^{-4}$], with the exception of *Chasing*, which was rated similarly [$F_{(1,299)} = 5.25$, $p_{adj} = 1$]. Analogously, *Bumping* tended to receive lower animacy ratings than all other classes [$F_{(1,299)} \geq 12.44$, $p_{adj} \leq 0.03$], with the exception of *Pushing*, *Frightening*, and *Flirting*, which were rated similarly [$F_{(1,299)} = 8.42$, $p_{adj} \geq 0.26$]. We report in **Figure 3B** all the *post-hoc* F-statistics, and in **Figure 3C** all the corresponding Bonferroni adjusted *p*-values.

3.2. Experiment 2

Figure 4 shows the total confusion matrix M of the classification task. Rows and columns are sorted by decreasing Recall. *Avoiding* was the most accurately classified class by our participants (*Recall* = 75.4%). However, even the hardest class was classified with largely above-chance accuracy (*Walking*: *Recall* = 53.4%; chance level: 8.3%). Nonetheless, there are obviously some misclassifications, especially between *Bumping* and *Pushing* ($m_{BU,PS} = 19$, $m_{PS,BU} = 11$), and between *Fighting* and *Chasing* ($m_{FI,CH} = 17$, $m_{CH,FI} = 2$). These two kinds of mislabeling alone accounted for a large fraction of the total number of mislabelings [$P_{MS}(BU, PS) = 9.8\%$, $P_{MS}(FI, CH) = 6.2\%$].

One possible reason for this misclassification could be the fact that these labels are semantically intrinsically similar and even real videos of these types of social interactions could be mislabeled. This line of reasoning is supported by the fact that in Experiment 1, *Pushing* was the second preferred keyword used to label videos of class *Bumping* (see **Figure 2**). Interestingly, both Precision and Recall (and thus F_1 score) significantly improved across experimental blocks [Precision: $t_{(3)} = 19.5$, $p = 2.93 \cdot 10^{-4}$; Recall: $t_{(3)} = 10.8$, $p = 1.68 \cdot 10^{-3}$; see **Figure 5**]. This indicates a latent learning of the categorization of the classes, which is remarkable since no external feedback about the correctness of the class assignments was provided during the experiment. Such a learning was particularly evident for the following often-confused pairs: *Tug of War* vs. *Pulling*, *Frightening* vs. *Avoiding*, and *Fighting* vs. *Pushing* (not shown).

3.3. Experiment 3

The pairwise semantic distance matrix D is plotted in **Figure 6A**: light shades of green indicate semantically close social interaction classes, while darker shades indicate semantically distant classes. The two pairs associated with the highest mislabeling probability in Experiment 2, *Bumping-Pushing*, and *Fighting-Chasing* [$P_{MS}(BU, PS) = 9.8\%$, $P_{MS}(FI, CH) = 6.2\%$] were generally considered as semantically similar [$d(BU, PU) = 0.49$, $d(FI, CH) = 0.65$]; however, they were not the most similar

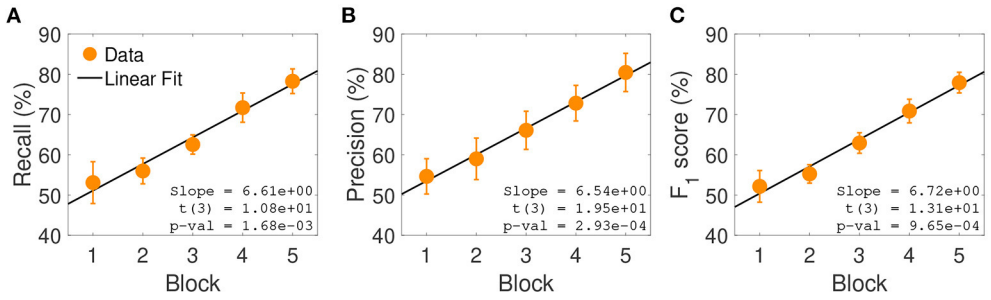


FIGURE 5 | Classification performance across experimental blocks. **(A)** Average block-wise recall. Results are averaged across subjects and social interactions; error bars represent standard errors. Insets show the slope of the estimated linear model, the corresponding t-statistic and p-value. **(B)** Average block-wise precision. **(C)** Average block-wise F_1 score.

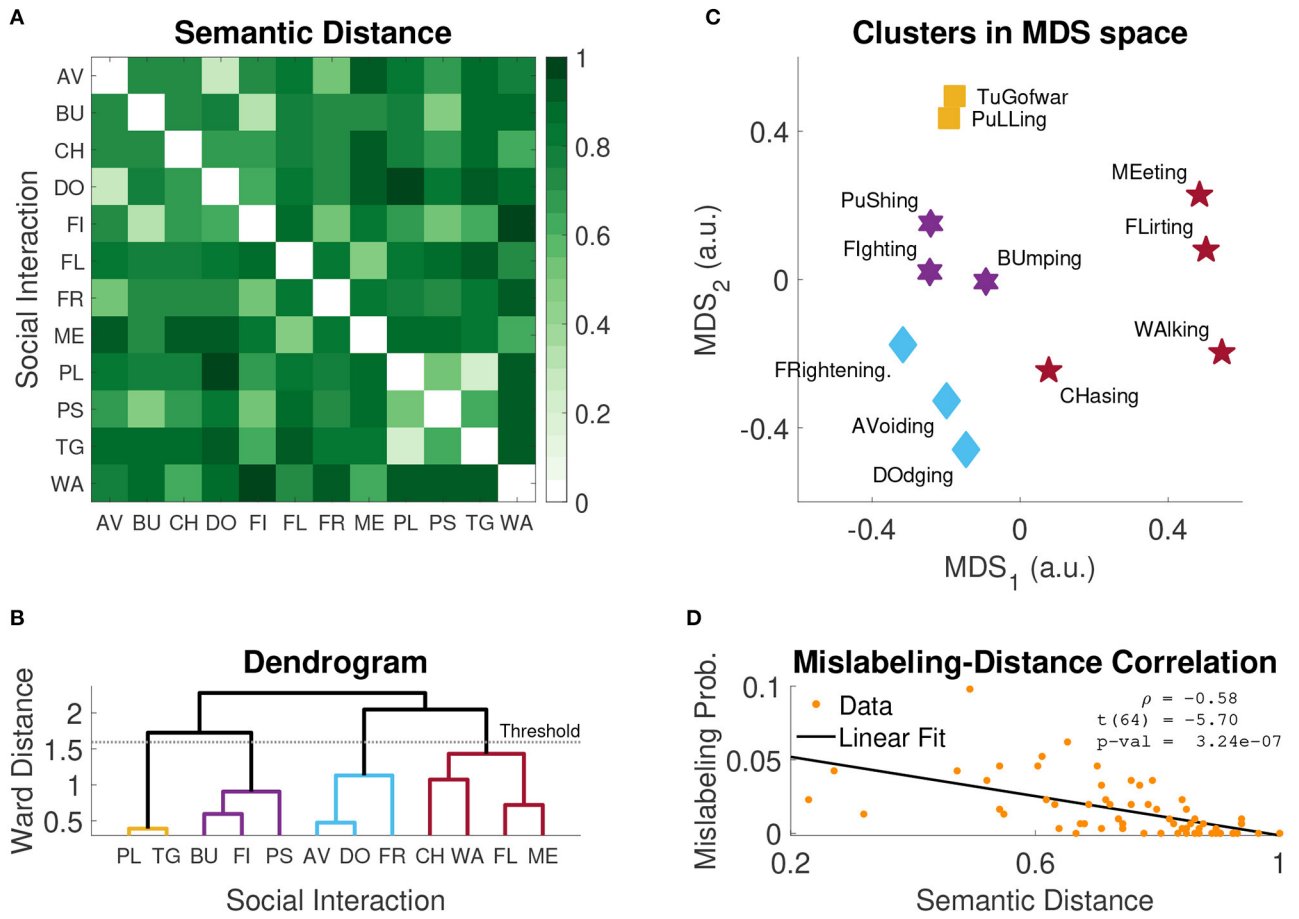


FIGURE 6 | Cluster analysis results. **(A)** Average semantic distances obtained in Experiment 3. **(B)** Dendrogram of hierachical clustering; the horizontal line represents the cut-off threshold used to identify the clusters (i.e., $0.7 * M_{WD}$, where M_{WD} is the maximum Ward distance). **(C)** Clusters of social interactions plotted in low dimensional distance-preserving 2D space identified with Multidimensional Scaling (MDS). **(D)** Average mislabeling probability (Experiment 2) as a function of semantic distance (Experiment 3); inset reports the Pearson's correlation coefficient, the corresponding t-statistic and p-value. AV, avoiding; BU, bumping; CH, chasing; DO, dodging; FI, fighting; FL, flirting; FR, frightening; ME, meeting; PL, pulling; PS, pushing; TG, tug of war; WA, walking.

pairs. Rather, the three most semantically similar pairs were *Pulling-Tug of War*, *Avoiding-Dodging*, and *Bumping-Fighting* [$d(PU, TG) = 0.23$, $d(AV, DO) = 0.27$, $d(BU, FI) = 0.32$]. Nevertheless, regardless of this apparent discrepancy for these few extreme examples, mislabeling probability $P_{MS}(i, j)$ and semantic distance $d(i, j)$ were significantly anti-correlated [$\rho =$

-0.58, $t_{(64)} = -5.7$, $p = 3.24 \cdot 10^{-7}$; **Figure 6D**]; this suggests that the more semantically similar two social interaction classes are, the more likely they are of being confused in a video labeling task.

Multidimensional scaling (MDS) provides a compact 2D visualization of the semantic similarity space (**Figure 6C**). Since MDS is inherently spatial, items that were rated as being highly similar are spatially close to each other in the final map. The map effectively shows which classes of social interactions are semantically similar and which are not. For example, let us consider the hypothetical groups G_1 ={Tug of War, Pulling} and G_2 ={Frightening, Avoiding, Dodging}. Participants recognized that *Tug of War* and *Pulling* involve similar interactions between the agents, and that these interactions are different from those occurring in the classes *Frightening*, *Avoiding*, and *Dodging*. For this reason, participants tended to assign high pairwise similarity scores to intra-group pairs, and low to inter-group pairs. This pattern of scoring is captured by MDS and evident in the resulting map (**Figure 6C**).

The agglomerative hierarchical cluster analysis on the distance matrix D (**Figure 6B**) confirms this intuition and identifies four distinct semantic clusters; such clusters are visualized in the MDS map with four different symbols (**Figure 6C**). This analysis supports the conclusion that misclassified labels tend to belong to the same semantic cluster. While not all misclassifications can be explained by semantic similarity, many confusions can be accounted for by this factor. For example, *Pushing* vs. *Bumping*, *Walking* vs. *Meeting*, *Avoiding* vs. *Dodging*.

To summarize, our analysis of semantic similarity shows that many of the labeling confusions observed in Experiment 2 can be explained by the semantic similarity of the class labels.

4. DISCUSSION

In this work, we introduced a novel framework for the automatic generation of videos of socially interacting virtual agents. The underlying model is a nonlinear dynamical system that specifies heading direction and forward speed of the agents. Our model is able to generate as many as 15 different interaction classes, defined by different parameter sets. We validated our model with three different behavioral experiments, in which participants were able to consistently identify the intended interaction classes. Our model is thus suitable for the automatic generation of animations of socially interacting agents. Furthermore, the generation process is also amenable to full parametric control. This feature allows the creation of highly-controlled and arbitrarily-large datasets for in-depth psychophysical and electrophysiological characterization of the perception of social interactions. The model thus overcomes the major limitations that come with hand-crafted, hard-coded, rule-based, and real-video-based approaches (1) to visual stimuli generation. Importantly, the generative nature of the model, makes it a valuable tool also for the development of mechanistic and neural *decoder* models of social perception: model responses to the heterogeneous set of highly-controlled social stimuli here introduced can be rigorously tested for the development of more accurate and brain-like decoder models that replicate human behavioral and neural responses. Recent work (Shu et al., 2018, 2019, 2020), aimed at building a mechanistic model of social inference, used a similar approach.

Shu et al. (2019, 2020) also proposed generative models of social interactions. Unlike the ones proposed in these studies, the generative model introduced in this work does not directly lend itself to the study of the interactions between intuitive physics and social inferences (Battaglia et al., 2013). However, substantial evidence suggests that physical and social judgments are mediated by different brain regions (Isik et al., 2017; Sliwa and Freiwald, 2017). More importantly, our model is not limited to describing cooperative and obstructive behaviors and thus seems better suited to study more general social interaction classes.

The identification of suitable parameters for the classes modeled in this work was not automatic: it was conducted using a simulation-based heuristic procedure. This is an obvious limitation of our work. Nevertheless, once the parameters are available, they can be used to automatically generate arbitrary numbers of coupled trajectories for each interaction class (by randomly sampling initial conditions, via-points, and noise). With this procedure, we were able to find suitable parameters for only 15 specific interaction classes. However, to the best of our knowledge, no other method is able to automatically generate more than a handful of individual or socially-interactive behaviors (Blackwell, 1997; Paris et al., 2007; Luo et al., 2008; Russell et al., 2017; Shu et al., 2019, 2020). Future work will extend the range of modeled classes by using system identification methods (e.g., Schön et al., 2011; Gao et al., 2018; Gonçalves et al., 2020) to automatically extract model parameters from preexisting trajectories—extracted, for example, from real videos.

Another possible limitation of our work is that all our participants were recruited from a German university; while this might, in theory, represent a biased sample, previous studies (Rimé et al., 1985) suggest that the perception of social interactions from impoverished stimuli is a phenomenon that is highly stable across cultures. Specifically, these authors showed that African, European, and Northern American participants provided similar interpretations to animated videos of geometrical shapes. This suggests that our findings would not have significantly changed if we had recruited a more heterogeneous sample.

In this work, we used the trajectories generated by our model to animate simple geometrical figures. The resulting abstract visual stimuli can be directly applied to characterize the kinematic features underlying the inference of social interactions. However, the trajectories can also be used as a basis for richer visual stimuli. For example, in ongoing work, we have been developing methods to link the speed and direction dynamics generated by the model to articulating movements of three-dimensional animal models. This approach allows the generation of highly controlled and realistic videos of interacting animals, which can be used to study social interaction perception in the corresponding animal models with ecologically valid stimuli. Furthermore,

contrasting the neural responses to impoverished and realistic visual stimuli can help identify the brain regions and neural computations mediating the extraction of the relevant kinematic features and the subsequent construction of social percepts.

Finally, even though the proposed model is mainly aimed to provide a tool to facilitate the design of in-depth psychophysical and electrophysiological studies of social interaction perception, we speculate that it can also be helpful in the development of machine vision systems for the automatic detection of social interactions. Specifically, the development of effective modern machine vision systems tends to be heavily dependent on the availability of large numbers of appropriately-labeled videos of social interactions (Rodríguez-Moreno et al., 2019; Stergiou and Poppe, 2019). A popular approach to this problem is to use clips extracted from already existing (YouTube) videos and movies. However, one of the reasons why feature-based (e.g. Kumar and John, 2016; Sehgal, 2018) and especially deep-neural-network-based (e.g., Karpathy et al., 2014; Carreira and Zisserman, 2017; Gupta et al., 2018) vision systems require *big data* is that they need to learn to ignore irrelevant information that is inevitably present in real videos. Therefore, we hypothesize that pre-training such systems with stylized videos of socially interacting agents—such as the very same generated by our model or appropriate avatar-based extensions—might greatly reduce their training time and possibly improve their performance. Future work will test this hypothesis.

To sum up, this work introduced a novel generative model of social interactions. The results of our psychophysical experiments suggest that the model is suitable for the automatic generation of arbitrarily-numerous and highly-controlled videos of socially interacting agents for comprehensive studies of animacy and social interaction perception. Our model can also be potentially used to create large, noise-free, and annotated datasets that can facilitate the development of mechanistic and neural models of social perception, as well as the design of machine vision systems for the automatic recognition of human interactions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the ethics board of the University of Tübingen. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was supported by the German Federal Ministry of Education and Research (BMBF FKZ 01GQ1704), the Human Frontiers Science Program (HFSP RGP0036/2016), the German Research Foundation (DFG GZ: KA 1258/15-1), and the European Research Council (ERC 2019-SyG-RELEVANCE-856495).

ACKNOWLEDGMENTS

The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting AS. The authors would also like to thank the participants who took part in the study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.648527/full#supplementary-material

REFERENCES

Baker, C. L., Saxe, R., and Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition* 113, 329–349. doi: 10.1016/j.cognition.2009.07.005

Barrett, H. C., Todd, P. M., Miller, G. F., and Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: a cross-cultural study. *Evol. Hum. Behav*. 26, 313–331. doi: 10.1016/j.evolhumbehav.2004.08.015

Battaglia, P. W., Hamrick, J. B., and Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U.S.A*. 110, 18327–18332. doi: 10.1073/pnas.1306572110

Bicho, E., Mallet, P., and Schöner, G. (2000). Target representation on an autonomous vehicle with low-level sensors. *Int. J. Robot. Res*. 19, 424–447. doi: 10.1177/02783640022066950

Bicho, E., and Schöner, G. (1997). The dynamic approach to autonomous robotics demonstrated on a low-level vehicle platform. *Robot. Auton. Syst*. 21, 23–35. doi: 10.1016/S0921-8890(97)00004-3

Blackwell, P. (1997). Random diffusion models for animal movement. *Ecol. Model*. 100, 87–102. doi: 10.1016/S0304-3800(97)00153-1

Blythe, P. W., Todd, P. M., and Miller, G. F. (1999). "How motion reveals intention: categorizing social interactions," in *Simple Heuristics That Make Us Smart*, eds G. Gigerenzer and P. M. Todd (Oxford University Press). pp. 257–285.

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Honolulu, HI), 6299–6308. doi: 10.1109/CVPR.2017.502

Castelli, F., Frith, C., Happé, F., and Frith, U. (2002). Autism, asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 125, 1839–1849. doi: 10.1093/brain/awf189

Castelli, F., Happé, F., Frith, U., and Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12, 314–325. doi: 10.1006/nimg.2000.0612

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition* 107, 705–717. doi: 10.1016/j.cognition.2007.08.001

Dittrich, W. H., and Lea, S. E. (1994). Visual perception of intentional motion. *Perception* 23, 253–268. doi: 10.1068/p230253

Fajen, B. R., and Warren, W. H. (2003). Behavioral dynamics of steering, obstacle avoidance, and route selection. *J. Exp. Psychol.* 29:343. doi: 10.1037/0096-1523.29.2.343

Freiwald, W. A. (2020). The neural mechanisms of face processing: cells, areas, networks, and models. *Curr. Opin. Neurobiol.* 60, 184–191. doi: 10.1016/j.conb.2019.12.007

Gao, S., Zhou, M., Wang, Y., Cheng, J., Yachi, H., and Wang, J. (2018). Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction. *IEEE Trans. Neural Netw. Learn. Syst.* 30, 601–614. doi: 10.1109/TNNLS.2018.2846646

Gao, T., McCarthy, G., and Scholl, B. J. (2010). The wolfpack effect: perception of animacy irresistibly influences interactive behavior. *Psychol. Sci.* 21, 1845–1853. doi: 10.1177/0956797610388814

Gao, T., Newman, G. E., and Scholl, B. J. (2009). The psychophysics of chasing: a case study in the perception of animacy. *Cogn. Psychol.* 59, 154–179. doi: 10.1016/j.cogpsych.2009.03.001

Gelman, R., Durgin, F., and Kaufman, L. (1995). "Distinguishing between animates and inanimates: not by motion alone," in *Causal Cognition: A Multidisciplinary Debate*, eds, D, Sperber, D. Premack, and A. J. Premack (Oxford: Clarendon Press), 150–184. doi: 10.1093/acprof:oso/9780198524021.003.0006

Gonçalves, P. J., Lueckmann, J.-M., Deistler, M., Nonnenmacher, M., Öcal, K., Bassetto, G., et al. (2020). Training deep neural density estimators to identify mechanistic models of neural dynamics. *Elife* 9:e56261. doi: 10.7554/eLife.56261

Gordon, A. S., and Roemmele, M. (2014). "An authoring tool for movies in the style of Heider and Simmel," in *International Conference on Interactive Digital Storytelling* (Singapore: Springer), 49–60. doi: 10.1007/978-3-319-12337-0_5

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. (2018). "Social GAN: socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 2255–2264. doi: 10.1109/CVPR.2018.00240

Heider, F., and Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychol.* 57, 243–259. doi: 10.2307/1416950

Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proc. Natl. Acad. Sci. U.S.A.* 114, E9145–E9152. doi: 10.1073/pnas.1714471114

Kaduk, K., Elsner, B., and Reid, V. M. (2013). Discrimination of animate and inanimate motion in 9-month-old infants: an ERP study. *Dev. Cogn. Neurosci.* 6, 14–22. doi: 10.1016/j.dcn.2013.05.003

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (Columbus, OH), 1725–1732. doi: 10.1109/CVPR.2014.223

Kerr, W., and Cohen, P. (2010). "Recognizing behaviors and the internal state of the participants," in *2010 IEEE 9th International Conference on Development and Learning* (Ann Arbor, MI), 33–38. doi: 10.1109/DEVLRN.2010.5578868

Kumar, S. S., and John, M. (2016). "Human activity recognition using optical flow based feature set," in *2016 IEEE International Carnahan Conference on Security Technology (ICCST)* (Orlando, FL), 1–5. doi: 10.1109/CCST.2016.7815694

Luo, L., Zhou, S., Cai, W., Low, M. Y. H., Tian, F., Wang, Y., et al. (2008). Agent-based human behavior modeling for crowd simulation. *Comput. Anim. Virt. Worlds* 19, 271–281. doi: 10.1002/cav.238

McAleer, P., Kay, J. W., Pollick, F. E., and Rutherford, M. (2011). Intention perception in high functioning people with autism spectrum disorders using animacy displays derived from human actions. *J. Autism Dev. Disord.* 41, 1053–1063. doi: 10.1007/s10803-010-1130-8

McAleer, P., and Pollick, F. E. (2008). Understanding intention from minimal displays of human activity. *Behav. Res. Methods* 40, 830–839. doi: 10.3758/BRM.40.3.830

Michotte, A. (1946). *The Perception of Causality, Vol. 21*. New York, NY: Basic Books.

Mukovskiy, A., Slotine, J.-J. E., and Giese, M. A. (2013). Dynamically stable control of articulated crowds. *J. Comput. Sci.* 4, 304–310. doi: 10.1016/j.jocs.2012.08.019

Oatley, K., and Yuill, N. (1985). Perception of personal and interpersonal action in a cartoon film. *Br. J. Soc. Psychol.* 24, 115–124. doi: 10.1111/j.2044-8309.1985.tb00670.x

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C.-C., et al. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition* 130, 360–379. doi: 10.1016/j.cognition.2013.11.011

Paris, S., Pettré, J., and Donikian, S. (2007). "Pedestrian reactive navigation for crowd simulation: a predictive approach," in *Computer Graphics Forum, Vol. 26* (Prague: Wiley Online Library), 665–674. doi: 10.1111/j.1467-8659.2007.01090.x

Reichardt, W., and Poggio, T. (1976). Visual control of orientation behaviour in the fly: Part I. A quantitative analysis. *Q. Rev. Biophys.* 9, 311–375. doi: 10.1017/S0033583500002523

Reimann, H., Iossifidis, I., and Schöner, G. (2011). "Autonomous movement generation for manipulators with multiple simultaneous constraints using the attractor dynamics approach," in *2011 IEEE International Conference on Robotics and Automation* (Shanghai), 5470–5477. doi: 10.1109/ICRA.2011.5980184

Rimé, B., Boulanger, B., Laubin, P., Richir, M., and Stroobants, K. (1985). The perception of interpersonal emotions originated by patterns of movement. *Motiv. Emot.* 9, 241–260. doi: 10.1007/BF00991830

Rodríguez-Moreno, I., Martínez-Otzeta, J. M., Sierra, B., Rodriguez, I., and Jauregi, E. (2019). Video activity recognition: state-of-the-art. *Sensors* 19:3160. doi: 10.3390/s19143160

Roether, C. L., Omlor, L., Christensen, A., and Giese, M. A. (2009). Critical features for the perception of emotion from gait. *J. Vis.* 9:15. doi: 10.1167/9.6.15

Russell, J. C., Hanks, E. M., Modlmeier, A. P., and Hughes, D. P. (2017). Modeling collective animal movement through interactions in behavioral states. *J. Agric. Biol. Environ. Stat.* 22, 313–334. doi: 10.1007/s13253-017-0296-3

Saxe, R., and Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." *Neuroimage* 19, 1835–1842. doi: 10.1016/S1053-8119(03)00230-1

Scholl, B. J., and Gao, T. (2013). "Perceiving animacy and intentionality: visual processing or higher-level judgment," in *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention* eds, M. D. Rutherford and V. A. Kuhlmeier (Cambridge, MA: MIT Press), 197–230. doi: 10.7551/mitpress/9780262019279.003.0009

Scholl, B. J., and Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends Cogn. Sci.* 4, 299–309. doi: 10.1016/S1364-6613(00)01506-0

Schön, T. B., Wills, A., and Ninness, B. (2011). System identification of nonlinear state-space models. *Automatica* 47, 39–49. doi: 10.1016/j.automatica.2010.10.013

Schöner, G., and Dose, M. (1992). A dynamical systems approach to task-level system integration used to plan and control autonomous vehicle motion. *Robot. Auton. Syst.* 10, 253–267. doi: 10.1016/0921-8890(92)90004-I

Schöner, G., Dose, M., and Engels, C. (1995). Dynamics of behavior: theory and applications for autonomous robot architectures. *Robot. Auton. Syst.* 16, 213–245. doi: 10.1016/0921-8890(95)00049-6

Schultz, J., and Bülthoff, H. H. (2019). Perceiving animacy purely from visual motion cues involves intraparietal sulcus. *NeuroImage* 197, 120–132. doi: 10.1016/j.neuroimage.2019.04.058

Sehgal, S. (2018). "Human activity recognition using BPNN classifier on hog features," in *2018 International Conference on Intelligent Circuits and Systems (ICICS)* (Phagwara), 286–289. doi: 10.1109/ICICS.2018.00065

Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika* 27, 125–140. doi: 10.1007/BF02289630

Shepard, R. N. (1962b). The analysis of proximities: multidimensional scaling with an unknown distance function. II. *Psychometrika* 27, 219–246. doi: 10.1007/BF02289621

Shu, T., Kryven, M., Ullman, T. D., and Tenenbaum, J. B. (2020). "Adventures in flatland: perceiving social interactions under physical dynamics," in *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (Toronto).

Shu, T., Peng, Y., Fan, L., Lu, H., and Zhu, S.-C. (2018). Perception of human interaction based on motion trajectories: from aerial videos to decontextualized animations. *Top. Cogn. Sci.* 10, 225–241. doi: 10.1111/tops.12313

Shu, T., Peng, Y., Lu, H., and Zhu, S. (2019). "Partitioning the perception of physical and social events within a unified psychological space," in *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (Montreal).

Sliwa, J., and Freiwald, W. A. (2017). A dedicated network for social interaction processing in the primate brain. *Science* 356, 745–749. doi: 10.1126/science.aam6383

Springer, K., Meier, J. A., and Berry, D. S. (1996). Nonverbal bases of social perception: developmental change in sensitivity to patterns of motion that reveal interpersonal events. *J. Nonverb. Behav*. 20, 199–211. doi: 10.1007/BF02248673

Stergiou, A., and Poppe, R. (2019). Analyzing human-human interactions: a survey. *Comput. Vis. Image Understand*. 188:102799. doi: 10.1016/j.cviu.2019.102799

Szego, P. A., and Rutherford, M. D. (2008). Dissociating the perception of speed and the perception of animacy: a functional approach. *Evol. Hum. Behav*. 29, 335–342. doi: 10.1016/j.evolhumbehav.2008.04.002

Thurman, S. M., and Lu, H. (2014). Perception of social interactions for spatially scrambled biological motion. *PLoS ONE* 9:e112539. doi: 10.1371/journal.pone.0112539

Träuble, B., Pauen, S., and Poulin-Dubois, D. (2014). Speed and direction changes induce the perception of animacy in 7-month-old infants. *Front. Psychol*. 5:1141. doi: 10.3389/fpsyg.2014.01141

Tremoulet, P. D., and Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception* 29, 943–951. doi: 10.1068/p3101

Tremoulet, P. D., and Feldman, J. (2006). The influence of spatial context and the role of intentionality in the interpretation of animacy from motion. *Percept. Psychophys*. 68, 1047–1058. doi: 10.3758/BF03193364

Troje, N., Simion, F., Bardi, L., Mascalzoni, E., Regolin, L., Grossman, E., et al. (2013). *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention*. Cambridge, MA: MIT Press.

van Buren, B., Gao, T., and Scholl, B. J. (2017). What are the underlying units of perceived animacy? Chasing detection is intrinsically object-based. *Psychon. Bull. Rev*. 24, 1604–1610. doi: 10.3758/s13423-017-1229-4

van Buren, B., Uddenberg, S., and Scholl, B. J. (2016). The automaticity of perceiving animacy: goal-directed motion in simple shapes influences visuomotor behavior even when task-irrelevant. *Psychon. Bull. Rev*. 23, 797–802. doi: 10.3758/s13423-015-0966-5

Walbrin, J., Downing, P., and Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia* 112, 31–39. doi: 10.1016/j.neuropsychologia.2018.02.023

Ward, J. H. Jr. (1963). Hierarchical grouping to optimize an objective function. *J. Am. Stat. Assoc*. 58, 236–244. doi: 10.1080/01621459.1963.10500845

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci. U.S.A*. 111, 8619–8624. doi: 10.1073/pnas.1403112111

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 11 August 2021
doi: 10.3389/fnbot.2021.686010

Coordinating With a Robot Partner Affects Neural Processing Related to Action Monitoring

Artur Czeszumski[1†], Anna L. Gert[1†], Ashima Keshava[1†], Ali Ghadirzadeh[2], Tilman Kalthoff[1], Benedikt V. Ehinger[1,3,4], Max Tiessen[1], Mårten Björkman[2], Danica Kragic[2] and Peter König[1,5]*

[1] Institute of Cognitive Science, Universität Osnabrück, Osnabrück, Germany, [2] Robotics, Perception and Learning, School of Electrical Engineering and Computer Science, Kungliga Tekniska Högskolan Royal Institute of Technology, Stockholm, Sweden, [3] Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands, [4] Stuttgart Center for Simulation Science, University of Stuttgart, Stuttgart, Germany, [5] Institut für Neurophysiologie und Pathophysiologie, Universitätsklinikum Hamburg-Eppendorf, Hamburg, Germany

Robots start to play a role in our social landscape, and they are progressively becoming responsive, both physically and socially. It begs the question of how humans react to and interact with robots in a coordinated manner and what the neural underpinnings of such behavior are. This exploratory study aims to understand the differences in human-human and human-robot interactions at a behavioral level and from a neurophysiological perspective. For this purpose, we adapted a collaborative dynamical paradigm from the literature. We asked 12 participants to hold two corners of a tablet while collaboratively guiding a ball around a circular track either with another participant or a robot. In irregular intervals, the ball was perturbed outward creating an artificial error in the behavior, which required corrective measures to return to the circular track again. Concurrently, we recorded electroencephalography (EEG). In the behavioral data, we found an increased velocity and positional error of the ball from the track in the human-human condition vs. human-robot condition. For the EEG data, we computed event-related potentials. We found a significant difference between human and robot partners driven by significant clusters at fronto-central electrodes. The amplitudes were stronger with a robot partner, suggesting a different neural processing. All in all, our exploratory study suggests that coordinating with robots affects action monitoring related processing. In the investigated paradigm, human participants treat errors during human-robot interaction differently from those made during interactions with other humans. These results can improve communication between humans and robot with the use of neural activity in real-time.

Keywords: human-robot interaction, social neuroscience, joint action, ERP, EEG, embodied cognition, action monitoring

OPEN ACCESS

Edited by:
Tom Foulsham, University of Essex, United Kingdom

Reviewed by:
Stefano Tortora, University of Padua, Italy
Zhihao Zhu, Yancheng Institute of Technology, China
Stefan Ehrlich, Technical University of Munich, Germany

***Correspondence:**
Artur Czeszumski
aczeszumski@uni-osnabrueck.de

[†]These authors have contributed equally to this work and share first authorship

***Received:** 26 March 2021*
***Accepted:** 30 June 2021*
***Published:** 11 August 2021*

Citation:
Czeszumski A, Gert AL, Keshava A, Ghadirzadeh A, Kalthoff T, Ehinger BV, Tiessen M, Björkman M, Kragic D and König P (2021) Coordinating With a Robot Partner Affects Neural Processing Related to Action Monitoring. Front. Neurorobot. 15:686010. doi: 10.3389/fnbot.2021.686010

1. INTRODUCTION

We constantly interact with other humans, animals, and machines in our daily lives. Many everyday activities involve more than one actor at once, and groups of interacting co-actors have different size. Especially, interactions between two humans (so-called dyadic interactions) are the most prevalent in social settings (Peperkoorn et al., 2020). During such situations, we spend most of

our time trying to coordinate our behavior and actions with other humans. Until recently, human cognition was mostly studied in non-interactive and single participant conditions. However, due to novel conceptual and empirical developments, we are now able to bring dyads instead of single participants to our labs (Schilbach et al., 2013). This approach is called Second-person neuroscience (Schilbach et al., 2013; Redcay and Schilbach, 2019). It suggests that we need to study the social aspect of our cognition with paradigms that include real-time interactions between participants instead of the passive observation of socially relevant stimuli (Redcay and Schilbach, 2019). Such an approach can reveal a new perspective on human social cognition.

Coordination between members of a dyad is achieved by joint actions (Sebanz and Knoblich, 2021). There are different aspects of coordination that facilitate achieving common goals between co-actors. Firstly, Loehr et al. (2013) showed in pairs of pianists performing solo and duets that monitoring of our actions, our partner's actions, and our joint actions is required to coordinate successfully. Second, being familiar with each co-actors individual contributions in the dyad helps to form predictions about the partner's actions, which further improves coordination (Wolf et al., 2018). Third, recently proposed action-based communication serves as a fundamental block of coordination (Pezzulo et al., 2013). In comparison to verbal communication, this low-level sensorimotor communication is implicit and faster. Experiments by Vesper et al. (2017) serve as examples of sensorimotor communication in the temporal dimension. Their results have shown that participants adjusted their actions to communicate task-relevant information. Fourth, while both co-actors are engaged in a constant flow of perceptual information, they create coupled predictions about each other's actions that are necessary to achieve fruitful coordination (Sebanz and Knoblich, 2021). Curioni et al. (2019b) investigated coordination tasks with incongruent demands between partners, and their results suggested the benefits of reciprocal information flow between participants. In sum, there are different aspects of human cognition that allow for the maintenance of dyadic coordination: Action monitoring, predictions based on familiarity of partner's actions, action-based communication, and reciprocal information flow.

So far, most dyadic interaction studies investigated the coordination between human co-actors (Sebanz et al., 2006; Vesper et al., 2010). However, in recent years we are more and more surrounded by robotic co-actors (Ben-Ari and Mondada, 2018). Furthermore, there are many different predictions for the future of robotics, but all point into the same direction: there will be more robots among us (Stone et al., 2016; Diamond, 2020; Wiederhold, 2021). In line with this, humanoid robots are getting progressively better at socially relevant tasks (Campa, 2016). It is thought that these social robots will be used in many different fields of our everyday life in the upcoming years (Enz et al., 2011). One of the main challenges in robotics is creating robots that can dynamically interact with humans and read human emotions (Yang et al., 2018). Concerning these changes in our environment, a new research line has emerged and already substantially contributed to our understanding of human-robot interactions (Sheridan, 2016). As many different scientists are slowly approaching this topic, the field of human-robot interaction until now focused on human thoughts, feelings, and behavior toward the robots (Broadbent, 2017). Studying these specific aspects is essential and further, we believe that the scientific community has to investigate real-life interactions between humans and robots in order to fully understand the dynamics that underlie this field. Therefore, we propose to use both human and robot partners in experimental paradigms as this will help to close the gap in understanding dyadic interactions.

There are different tools and methods to study the social brain and behavior (Krakauer et al., 2017): EEG (Luck and Hillyard, 1994), fMRI (Eisenberger, 2003), MEG (Baillet, 2017), and fNIRS (Ferrari and Quaresima, 2012). From this list, Electroencephalography (EEG) stands out as particularly useful for studying dynamical interactions, as it not only aligns with the temporal resolution of social interactions, but also allows for free movement and thereby allows for dynamic interactions. This temporal resolution allows studying brain processes with milliseconds precision. One of the methods that are classically used within EEG research are event-related potentials (ERPs) (Luck and Hillyard, 1994). ERPs are suitable to study different components of brain processes while they evolve over time. The classic study by Miltner et al. (1997) showed different brain signatures for correctly and incorrectly performed trials at around 200-300 milliseconds after the feedback about an action was perceived. This brain component was named Feedback related negativity (FRN). In similar studies, van Schie et al. (2004) showed that the FRN is sensitive not only to our own actions but also those of others. Czeszumski et al. (2019) further extended this finding to different social contexts (cooperation and competition). Thus, EEG and specifically ERPs have been proven valuable tools to investigate the physiological basis of social interactions.

Therefore, we have a good understanding of EEG-based markers of action monitoring. Nonetheless, it is only in recent years that human behavior and its neural basis are studied together with robotic partners (Wykowska et al., 2016; Cheng et al., 2020). Based on more than 20 years of research on action monitoring in humans, similar ERP components (E/FRN) were expected to be elicited in human-robot paradigms. Namely, the difference between brain responses to correct and incorrect actions of a robotic arm was found (Iturrate et al., 2015; Kim et al., 2017). Furthermore, these differences in midfrontal ERP components were used to improve co-adaptation between human and robot behavior in turn-taking tasks (Salazar-Gomez et al., 2017; Ehrlich and Cheng, 2018, 2019a,b; Iwane et al., 2019), and real-world driving (Zhang et al., 2015; Chavarriaga et al., 2018). Such EEG based interfaces highlight the importance of studying the neural basis of human-robot interactions. The results confirm that similar brain mechanisms are involved when we observe actions of the robot. Yet, little is known about action monitoring in dynamic situations with non-human, robotic partners. The goal of this study was to test whether the same neural mechanisms are present when we interact with robots in a dynamic paradigm and if there are differences between human and robotic partners.

To answer these questions, we adapted a dynamic dyadic interaction paradigm for human-robot interactions. We chose the paradigm from Hwang et al. (2018) and Trendafilov et al. (2020), in which two human participants had to manipulate a virtual ball on a circular elliptic target displayed on a tablet and received audio feedback of the ball's movement. Participants used their fingers to move the tablet and manipulate the position of the ball. We changed the paradigm, by adapting the tablet to enable coordination with the robot and to fit the requirements for EEG measurements. On the one hand, this paradigm allows for coordination similar to a real-life situation; on the other hand, it allows for the analysis of neural underpinnings of cognitive functions required for coordination. In this study, we specifically focused on the aspect of action monitoring with human and robot partners. Thus, to extend our knowledge the present study investigates action monitoring in a dynamic interaction task between humans and robots. Additionally, based on the results from Hwang et al. (2018) we decided to test whether auditory feedback about actions (sonification) influences coordinated behavior and cognitive processes. Taken together, this study tries to approach a novel problem with interdisciplinary methods and sheds new light on the neural processes involved in dynamic human-robot interactions.

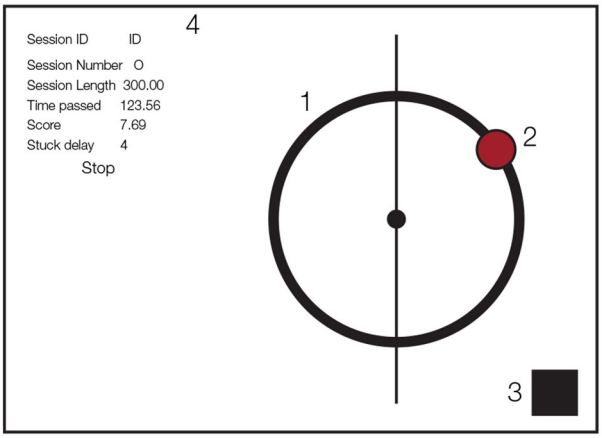


FIGURE 1 | Schematic of game design on the tablet. (1) Circular track, (2) ball, (3) flashing rectangle indicating experimental events (covered by luminance sensor), (4) text box for experiment monitoring (only used by experimenter).

2. METHODS

2.1. Participants

We recruited 16 participants (7 female, mean age = 25.31 ± 1.92 years) from KTH Stockholm Royal Institute of Technology. We had to exclude two dyads from further analysis, one due to measurement errors in the robot control and one due to excessive movements from participants which led to large artifacts in the EEG data, leaving data from 12 participants in 6 recording sessions. Participants had normal or corrected-to-normal vision and no history of neurological or psychological impairments. They received course credits for their participation in the study. Before each experimental session, subjects gave their informed consent in writing. Once we obtained their informed consent, we briefed them on the experimental setup and task. All instructions and questionnaires were administered to the participants in English. The Swedish Ethical Review Authority (Etikprövningsnämnden) approved the study.

2.2. Task and Apparatus

During each recording session, participants performed the task in four blocks of 10 min each, twice with a human partner and twice with the robot. Further, each dyad (partner human or robot) performed the task with or without auditory feedback (sonification on or sonification off). The task was based on a tablet game where the dyads cooperated with each other to balance a ball on a circular track as they simultaneously moved it in counter-clockwise direction (Hwang et al., 2018) (**Figure 1**). At random intervals, we added perturbations that radially dispersed the position of the ball away from the current position. In order to reduce the subjects' expectations of the occurrence of the perturbations, we sampled its rate of occurrence from a Poisson distribution with $\lambda = 4$ s.

The experimental task was implemented on an Apple IPad Air tablet (v2, 2048 × 1536 pixel resolution, refresh rate 60 Hz) using Objective-C for iOS. During the task, subjects saw a red ball of 76.8 pixel radius on a circular track with a radius of 256 pixels and a thickness of 42.67 pixels. The ball position was represented as the horizontal and vertical coordinates with respect to the center of the circular track (0,0). The tablet was mounted on a metal frame of size 540 × 900 mm. We further added a square of size 100 × 100 pixels that was used as a signal source for, and covered by, a luminance sensor.The luminance sensor is a light-sensitive diode that converts light into electrical current. We changed the color of a small patch on the tablet for the different events in the experiment (start of the experiment, start of a perturbation, end of the experiment) over which the luminance sensor was placed. **Figure 1** shows all the visual components displayed to the participants (the text box on the left side was used by the experimenter to monitor the experiment status).

During the periods with another human partner, we asked the participants to not verbally interact with each other. During the task, they sat face-to-face at 1m distance as they held handles connected to the short end of the frame. Similarly, while performing the task with the robot, subjects held the short end of frame while the other end of the frame was clamped to the grip effectors of the robot. **Figure 2** shows the physical setup of the subjects and the robot during the experiment.

For the periods involving sonification, the position and angular velocity of the ball were sonified. The auditory feedback was created by a Gaussian noise generator with a band-pass filter (cut-off frequency: ±25 Hz). The horizontal and vertical coordinates of the ball modulated the pitch of the auditory feedback, while its angular velocity modulated the loudness. The sonification procedure was implemented using the specifications provided in Hwang et al. (2018).

Lastly, we used a self-manufactured luminance sensor that synchronized the experimental events (experiment start and end,

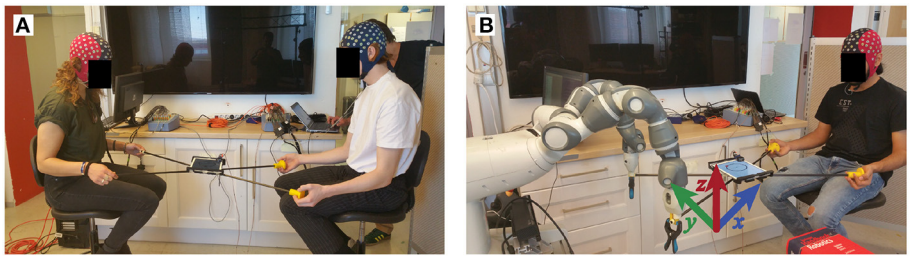


FIGURE 2 | Experimental setup. Participants performed the experiment with another participant **(A)** or a robot partner **(B)**. In each condition they played a tablet game by balancing a virtual ball on a circular track while moving it in the counter-clockwise direction.

and perturbation) between the tablet and the EEG amplifier. We changed the luminance source color from black to white to mark the start of the trials, white to black to mark the end of the trials. During a session the patch was white, except at the frame where the perturbation happened, which was marked with gray (RGB = 134, 134, 134).

2.3. Robot Control

We used the YuMi robot (ABB, Västerås, Sweden) shown in **Figure 2** for our experiments. We implemented a Cartesian space controller based on the original joint-level velocity controllers provided by the manufacturer. The robot had direct access to the tablet data and no active sensing was necessary. Starting the robot at the joint position depicted in the **Figure 2**, we send Cartesian space velocity commands to both arms at 10 Hz. The Cartesian controller was designed such that the *X, Y* positions of both end-effectors are kept constant during an execution, and only the *Z* position of the end-effectors are adjusted to move the ball. We denote the left and right end-effector velocity commands in the *z* axis by v_l^z and v_r^z and the current *X, Y* position of the ball on the game by (b_x, b_y), respectively. We first obtain the angle θ corresponding to the current position of the ball in the polar coordinate system by $\theta = arctan(b_y, b_x)$. Then, we obtain the next target angle $\hat{\theta} = \theta + \pi/12$ to let the ball move in the counterclockwise direction. The next target *X,Y* positions of the ball are found as $\hat{b}_x = G_p(R \times cos(\hat{\theta}) - b_x)$, $\hat{b}_y = G_p(R \times sin(\hat{\theta}) - b_y)$, where R is the radius of the circle on the IPad game and $G_p = 0.1$ is a constant gain. The velocity commands in the *z* axis are then found as $v_l^z = -G_v(\hat{b}_x - \alpha_x) - G_v(\hat{b}_y - \alpha_y)$, $v_r^z = G_v(\hat{b}_x - \alpha_x) - G_v(\hat{b}_y - \alpha_y)$, where, α_x, α_y are gravity acceleration in the *X, Y* directions measured by the IPad, and $G_v = 0.5$ is a constant gain. The command velocities are then clipped to have an absolute value less than 0.02 m/s, and the clipped values are sent to the Cartesian velocity controller.

2.4. Experimental Protocol

We prepared both participants for the EEG recording together, which took around 45 min to complete. Once the subjects were ready to start the experiment, we led them to a room that housed the robot. Depending on the dyad combination, we provided oral instructions about the task and clarified any remaining questions. For human-human dyads, we started the task on the tablet with either of the sonification conditions depending on the experiment session. To counterbalance the sonification and partner sequence for the combinations of dyads (human-human or human-robot), we permuted the combinations. Each experimental session was sequenced based on this permutation. We also counter-balanced the sonification during the task, so that every even numbered experiment session started with the sonification condition for all the dyad combinations. For the human-robot dyads, we first reset the limbs of the robot to its initial conditions and then started the task on the tablet. After each block, the participants were given a short break and then repeated the task with the alternate sonification condition. The whole experimental session lasted for about 4 h.

2.5. EEG Data Acquisition

We recorded the EEG using two 64-Ag/AgCl electrode systems (ANT Neuro, Enschede, Netherlands), and two REFA8 amplifiers (TMSi, Enschede, Netherlands) at a sampling rate of 1,024 Hz. The EEG cap consists of 64 electrodes placed according to the extended international 10/20 system (Waveguard, eemagine, Berlin, Germany). We placed the ground electrode on the collar-bone. We manually adjusted the impedance of each electrode to be below 10kΩ before each session. The recording reference was the average reference, which, only in the single-brain recordings, was later programatically re-referenced to Cz. During human-human interactions, two brains were recorded simultaneously with the separate amplifiers, synchronized through the ANT-link (Synfi, TMSi, Enschede, Netherlands). VEOGs were recorded with two additional electrodes, one placed below and one above the eye.

2.6. Pre-processing

The analysis of the EEG data was performed in MATLAB 2016b and the behavioral analyses in Python 3.7.

We preprocessed the data using the EEGLAB toolbox (v2019.0) (Delorme and Makeig, 2004). As a first step before preprocessing, we programmatically extracted the trigger events from the luminance sensor and added them to the recorded data. Then, the data from each condition was downsampled to 512Hz, followed by referencing all datasets to Cz electrode. We then high-pass filtered the dataset at 0.1Hz and then low-pass filtered it at 120 Hz in order to not unnecessarily discard gamma frequency

activity (6 dB cutoff at 0.5 Hz, 1 Hz transition bandwidth, FIRFILT, EEGLAB plugin, Widmann et al., 2015). Following this, we manually removed channels that showed strong drift behavior or excessive noise (mean: 7, SD: 2.7, range: 1–13). We manually inspected the continuous data stream and rejected the portions which exhibited strong muscle artifacts or jumps. To remove further noise from eye and muscle movements, we used independent component analysis (ICA) based on the AMICA algorithm (Palmer et al., 2008). Before performing ICA, we applied a high-pass filter to the data at 2Hz cut-off to improve the ICA decomposition (Dimigen, 2020). We visually inspected the resulting components in combination with using ICLabel (Pion-Tonachini et al., 2019) classifier. IClabel was run on epoched data, 200 ms before and 500 ms after the perturbation. Based on the categorization provided by ICLabel, and a visual inspection of the time course, spectra, and topography, we marked ICs corresponding to eye, heart and muscle movements for rejection (mean: 26.5, SD: 5.2, range: 18–44). We copied the ICA decomposition weights to the cleaned, continuous data and rejected the artifactual components. Finally, using spherical interpolation, we interpolated the missing channels based on activity recorded from the neighboring channels.

2.7. Behavioral Analysis

To understand the behavioral differences for the factors partner and sonification, we used measures of mean angular velocity and mean error produced. These behavioral differences indicate how well the partners coordinated with each other. Furthermore, as the velocity and position of the ball were sonified, these measures are indicative of the effect of sonification on the dyadic performance. We first calculated the instantaneous angular position θ (in degrees) of the ball using the horizontal and vertical (X, Y) positions of the ball on the tablet as follows:

$$\theta_t = \frac{180}{\pi} * \arctan \frac{y_t}{x_t} \tag{1}$$

We used the atan2 function to take into account the X, Y position in the negative coordinate axes. θ_t values were transformed from $[-\pi, \pi]$ to range $[0, 2\pi]$. Next, we computed the instantaneous angular velocity ω of the ball using the following formula where t is the sample time-point:

$$\omega = \frac{\Delta\theta}{\Delta t} \tag{2}$$

We, subsequently, calculated the mean ω for each participant for the four different conditions. Next, We calculated the error as the difference of the instantaneous radial distance between the radius of the track and the ball's current position measured as the distance from the track's center as follows:

$$error_t = \sqrt{x_t^2 + y_t^2} - Radius_{track} \tag{3}$$

2.8. Deconvolution and EEG Analysis

Even though the perturbations were sampled from a Poisson distribution with $\lambda = 4$, the corresponding neural responses might overlap in time and bias the evoked potentials (Ehinger and Dimigen, 2019; Dimigen and Ehinger, 2021). Further, experimental block onset and offset typically elicit very strong ERPs overlapping with the perturbations. Finally, we see clear, systematic differences in the behavior depending on the condition (e.g., higher velocity with a human partner), which could lead to spurious effects in the ERPs. We further added eccentricity (distance from the circles midpoint), in order to control for the ball's trajectory. In order to control both temporal overlap and covariate confounds, we used linear deconvolution based on time-regression as implemented in the unfold toolbox v1.0 (Ehinger and Dimigen, 2019). Consequently, we modeled the effects of the partner (human or robot), the sonification (off = 0, on = 1) and their interaction as binary, categorical variables, the eccentricity and the velocity were coded using B-spline basis functions and the angular position using a set of circular B-splines. The block on- and offsets were modeled as intercept only models. The complete model can be described by the Wilkinson notation below (Wilkinson and Rogers, 1973).

$$\begin{aligned} \textit{perturbation ERP} \sim\ & 1 + \textit{partner} + \textit{sonification} + \textit{partner} : \textit{sonification} \\ & + \textit{circularspline}(\textit{angular position}, 8) \\ & + \textit{spline}(\textit{eccentricity}, 5) \\ & + \textit{spline}(\textit{velocity}, 5) \\ \textit{block onset ERP} \sim\ & 1 \\ \textit{block offset ERP} \sim\ & 1 \end{aligned}$$

This model was applied on the average referenced continuous EEG data, and each event was modeled in the time range of −500 to 700 ms with respect to the event onset. We collected a mean value of 640 trials per subject.

Similar to the two-stage mass univariate approach, we calculated the t-value over subjects for each of the resulting regression coefficients (similar to difference waves between two conditions) for all electrodes and time points (time-range of −500 to 700 ms). That is, for the purpose of comparison of two conditions, they are preferable as they avoid confounds by other factors. The multiple comparison problem was corrected using a permutation based test with threshold-free cluster enhancement (TFCE) (Mensen and Khatami, 2013; Ehinger et al., 2015) with 10,000 permutations(default parameters $E = 0.5$ and $H = 2$). We used the eegvis toolbox (Ehinger, 2018) to visualize all evoked response potentials.

3. RESULTS

3.1. Behavioral

In this study, humans played a collaborative game either with other humans or with robots. We further added sonification of the ball's movement as a supplementary auditory feedback to the participants. **Figure 3** shows the raw positions of the ball overlaid for all subjects and the partner and sonification conditions. The behavior we analyse here, is the mean velocity of the ball during each session and the mean deviation of the ball from the circular track. These measures indicate how fast the participants

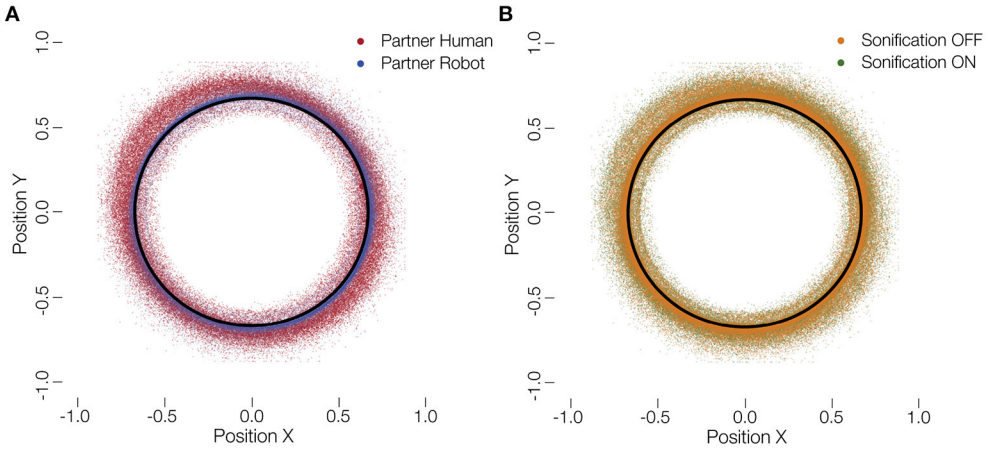


FIGURE 3 | Distribution of ball positions. **(A)** Ball positions on the tablet with a human (red) and robot (blue) partner. **(B)** Ball positions on the tablet for sonified (green) and not sonified movements (orange). The black circle represents optimal trajectory. It can be seen that participants deviated more with a human partner. No such difference is visible for a change in the sonification.

performed the task and how much error they produced, both a proxy of the success of the collaboration.

We calculated the mean angular velocity (ω) for each participant for the four different conditions (**Figure 4A**). To test the statistical significance of these findings, we computed a 2 × 2 factorial repeated measures ANOVA with the factors partner and sonification. The ANOVA showed a significant main effect of partner, $F_{(1,\ 11)} = 87.09, p < .0001$ where subjects exhibited a mean angular velocity of 265.20 degrees/second and SD ±0.28.29 with a human partner, conversely, with a robot partner subjects showed a mean angular velocity of 159.23 degrees/second ±29.40. The ANOVA did not reveal a significant main effect of sonification, $F_{(1,\ 11)} = 1.00, p = 0.33$, with mean angular velocity 210.06 degrees/second ±65.51 with sonification off and the mean angular velocity was 214.36 degrees/second ±62.53 with sonification on. There was no significant interaction of factors partner and sonification, $F_{(1,\ 11)} = 0.04, p = 0.83$. Hence, we can conclude that participants were faster at moving the ball on the circular track while performing the task with a human partner.

Next, we analyzed the mean error produced by participants during a session. **Figure 4B** shows the mean error across participants for the four different conditions. To statistically assess these differences, we performed a 2 × 2 factorial repeated measures ANOVA with factors partner and sonification. The ANOVA revealed significant main effect for partner $F_{(1,\ 11)} = 42.61, p < 0.0001$ where subjects had a mean error of $0.04 \pm SD = 0.012$ while performing with a human partner, conversely, they had a mean error of 0.01±0.012 while cooperating with the robot. We did not find a significant main effect of sonification $F_{(1,\ 11)} = 1.75, p = 0.21$ where subjects had a mean error of 0.032 ± 0.017 with the sonification off and mean error of 0.033 ± 0.018 with sonification on. There was no significant interaction of factors partner and sonification, $F_{(1,\ 11)} = 0.51, p = 0.48$. We can conclude that subjects made larger errors while performing the task with a human partner compared to the robot partner.

Lastly, we were interested in the correlation between the behavioral measures we analyzed. **Figure 4C** shows the correlation of mean error and mean velocity for the partner and sonification conditions. For human partner with sonification off the Pearson correlation showed a correlation coefficient $\rho = 0.98, p < 0.001$ and for sonification on $\rho = 0.89, p < 0.001$. For robot partner with sonification off $\rho = 0.97, p < 0.001$ and with sonification on $\rho = 0.97, p < 0.001$. These results show that the mean error and mean velocity were positively correlated during the task.

3.2. EEG

Next, we look at the overlap- and behavior-corrected brain activity during the task. Using a overlap-corrected time regression approach, we investigate the main effect and interaction ERPs from the 2 × 2 design, while adjusting for eccentricity, velocity and position of the ball (see section 2 for details). For the effect of the behavioral data on the ERP, please see the **Supplementary Material**. We only report ERPs time-locked to perturbation events. Descriptively, in electrode Cz (**Figure 5A**), we see the typical pattern of a positive deflection, followed by a negative and a second positive deflection after the perturbation onset. We did not have a specific hypothesis to a predefined component and analyzed all electrodes and time points simultaneously. The TFCE analysis reveals two clusters for the main effect of the factor of partner (**Figure 5B**). The first cluster is likely to represent the activity between 230 and 270 ms with its maximum amplitude being −2.8μV at electrode FC1 (median p: 0.025, minimal p: 0.018). The second cluster most likely represents the time range of 515–605 ms with a peak at −1.2μV at electrode FC2 (median p: 0.026, minimal p: 0.002). Both clusters are found in the central region. No significant clusters were found for neither the factor sonification nor the interaction term.

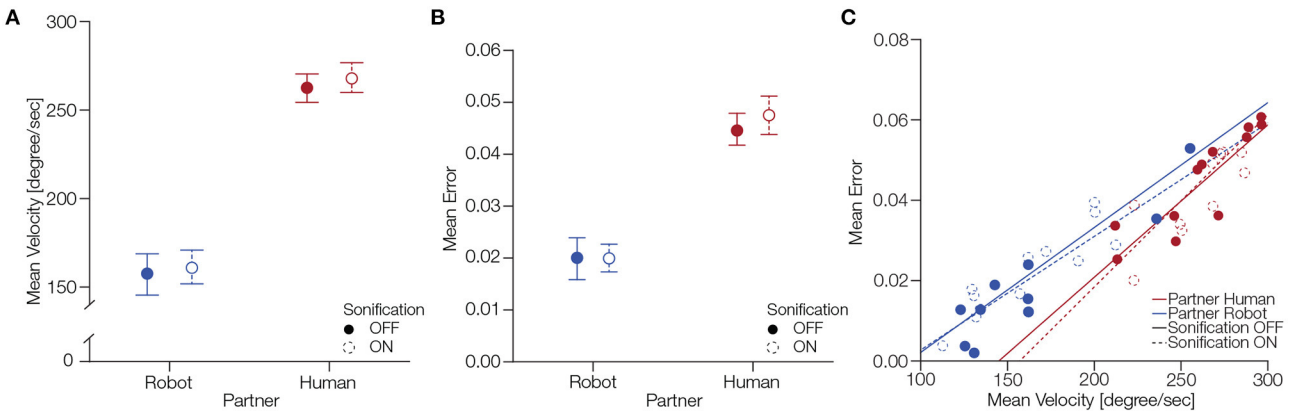


FIGURE 4 | Behavioral differences between conditions partner and sonification. **(A)** Differences in mean angular velocity across different participants. The error bars indicate standard error of mean. **(B)** Differences in mean error across different participants. The error bars indicate standard error of mean. Panel **(C)** shows the correlation of mean error and mean velocity for partner and sonification conditions.

These results show that we find differences in the participants' ERPs with respect to their current partner independently of their differences in behavior: When interacting with a robot partner the ERP will have a stronger amplitude indicating a systematically different processing.

4. DISCUSSION

Our experiment investigated neural correlates of action monitoring in a dynamic collaboration task that involves two co-actors. Participants performed the task with another human and robot partner while we measured EEG signals. Co-actors tried to keep a virtual ball on the circle displayed on a tablet; they used their hands (human arm or robotic arm) to manipulate independent orientation axes of the tablet. We perturbed the ball to investigate neural action monitoring processes of the participants. We found fronto-central ERP components at around 200–300 ms after the ball was perturbed. The components were stronger for human and robot partner compared to interactions with another human. These results suggest that the dynamic processing of our actions is influenced by whether we collaborate with a robot or a human.

The behavioral measures of our participants' actions were different between human and robot partners. We focused our analysis on two aspects of collaboration: The speed which is represented by the ball's velocity and the accuracy as indicated by the mean error. Our results suggest that participants perform slower when paired with the robot and achieve higher accuracy (ball closer to the circular track). There is a trade-off relation between these factors; this is why we discuss them together (**Figure 4C**). One simple explanation could be that the robot's control were themselves slow and prone to error. The human participants might have restrained themselves and thereby executed artificially slow movements. Another interpretation of why our participants slow down (and increased accuracy) while performing with the robot is that they had less trust in the robot than a human partner. This is in line with past research that suggests that level of trust changes during real-time interactions with robots (Desai et al., 2013) and that, in general, trust levels are different for human and robot partners (Lewis et al., 2018). Another interpretation for slower movements is that it is challenging to create a model of a partner's actions during a joint collaborative task with a robot. Based on work suggesting that we represent others' actions as our own (Sebanz et al., 2003), it is possible that in the case of interacting with a robot we need more time to create such representations. There is much space for interpretations why having a robot partner triggered slower movements; however, we would like to point that the main goal of our study was to investigate neural correlates of different partners, and behavioral responses were collected to exclude their influence on neural responses (see section 2.8 for details).

After adjusting for behavioral differences in the EEG analysis, we see that robot partners affect neural correlates of action monitoring differently in comparison to a human partner. We found that between 200-300 ms after the perturbation event disturbing the collaboration, the EEG amplitudes differ at the fronto-central sites. The literature on single participants at these electrodes and time window suggests that it is when and where monitoring our errors or feedback about our actions unravels (Miltner et al., 1997; Cavanagh et al., 2009). Similarly, when it comes to neural activity involved in action monitoring in dyadic situations, the same activations play a role (van Schie et al., 2004; Czeszumski et al., 2019). If the error is committed by the participant and can be inferred from his action (e.g., making a typo), the brain component involved is called Error-related negativity, with more negative activation for erroneous actions than correct ones (Yeung et al., 2004). In case of behavior that needs feedback to understand the consequences of the action (for example, gambling task), it is called Feedback related negativity (Hajcak et al., 2006). In comparison to these classic, static, and passive experiments, we had real-time collaboration between two participants, and we observed similar component peaking around 200-300 ms after the perturbation happened.

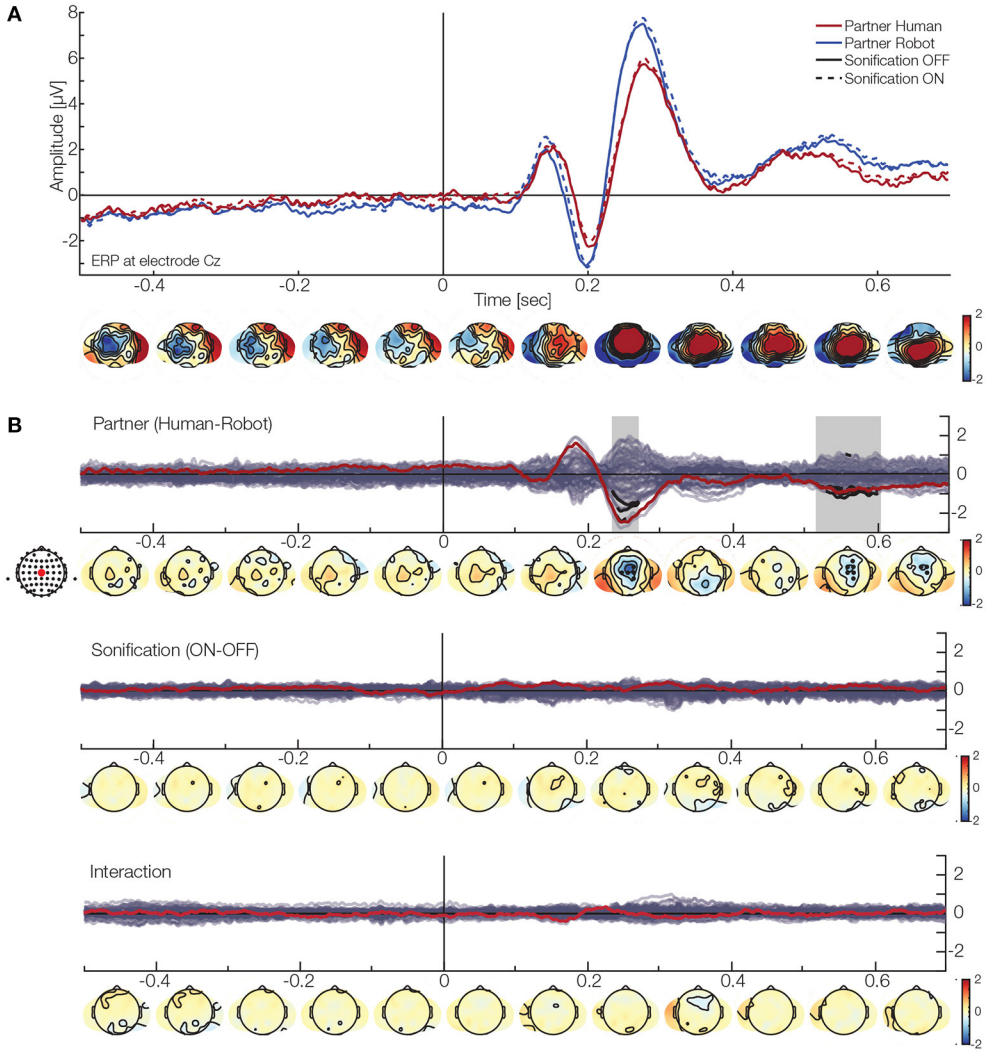


FIGURE 5 | EEG results. **(A)** ERP at electrode Cz. The red lines show the activation when interacting with a human partner, while blue lines indicate a robot partner. The solid line are the ERP when the sonification was off, while the dashed line represent sonification on. Below, are the topographies for the grand average (mean over all conditions). **(B)** Clustering results for the different factors (red line and dot represents electrode Cz). Top: Effect of partner. The analysis finds two clusters in the central area (black dots and segments). One is likely due to a difference at around 230–270 ms, while the second one is present later (around 510–600 ms). These results indicate that the ERP will have a smaller amplitude when interacting with a human partner. Middle: Effect of sonification. No cluster was found here. Bottom: Interaction. No cluster was found here.

Our participants were not informed about the perturbations, so they could have been treated as participants' own or the partner's error. Therefore, we suggest that the neural activation we observe in our study resembles classic components. Our finding that robotic partners modulate action monitoring corroborates recent study (Hinz et al., 2021). However, there is a crucial difference between both studies: Participants in Hinz et al. (2021) study performed a task sequentially (turn-taking), while in our study, participants interacted with each other in real-life. Both studies point in the same direction. Robot partners modulate neural activity. We speculate that differences in the amplitudes of the ERP for robotic and human partners may arise from differences in how we represent actions of artificial and human-like agents. Such differences might involve partly non-overlapping neuronal substrates with different visibility to EEG recordings. Furthermore, the perceived options to optimize performance in the joint interaction by adjusting to the behavior of the partner might differ. Such differences can elicit different neural patterns that we are able to measure with EEG.

Our results suggest that robot partners can modulate neural activity in a dyadic experiment. Concerning that there is not many studies that focused on neural underpinnings of human-robot interactions, the results we present here have a value for research topics in the field of join-action. They are a first exploratory step toward a theoretical and methodological foundation. We showed the feasibility of conducting a human-robot interaction study while measuring EEG from the human participant in a dynamical paradigm. With full experimental

control, we explored neural correlations of human-robot interactions in an ecologically valid setup (Matusz et al., 2019; Czeszumski et al., 2020; Nastase et al., 2020). There is vast literature on the topic of joint actions between humans and robot partners (Curioni et al., 2019a; Schellen et al., 2021; Wahn and Kingstone, 2021). Neural markers of action monitoring during human-robot interactions were studied in turn-taking tasks and utilized for brain-computer interfaces to improve communication between robots and humans (Ehrlich and Cheng, 2018, 2019a,b). Our study shows that it is possible to conduct studies with non-human agents collaborating with humans in real-time and measure brain activity and that the neural basis of action monitoring is affected by the robot partner.

Lastly, we observed small differences between human and robot partners at later time points (between 500-600 ms after the perturbation) around the midline electrodes. These differences are difficult to interpret. The topography suggest similar source as the component discussed above. However, based on time we speculate it could be P3b component. Huberth et al. (2019) reported similar component in study that investigated self and other (human vs. computer) generated actions in pianists. They found that P3b component was present only for self generated actions, suggesting greater monitoring of self generated actions. It is important to highlight that in our study, participant had to dynamically perform the task, while in the Huberth et al. (2019) study participants took turns to perform joint actions. What is similar is that they had to generate actions to achieve a common joint goal (Vesper et al., 2010). It is possible that the late effect we found in our experiment has the same function (greater monitoring of self generated actions). However, in comparison to the earlier effect (200–300 ms after the perturbation), the size of the effect in our study is small. Therefore, we have to be careful with interpretations. Future researcher with bigger sample size can help to understand the function of late ERP components in joint actions with robots.

4.1. Limitations

The exploratory aspect of investigating neural underpinnings of human-robot interactions pose many challenges and questions. In the present study, we tried our best to reconcile all of them. However, there are limitations that have to be addressed. First, our sample size was small in terms of number of dyads. However, it was not small in terms of recordings and total amount of gathered data. Thus, the effects reported are significant. Second we did not perform statistic al analyses with a predefined hypothesis. Instead, we performed an exploratory analysis that encompasses all electrodes and time points. It is important to understand that it is the first study of its kind. Therefore, it has to be replicated and evaluated by future research (Pavlov et al., 2021). Third our results could be dependent on the robot used in the study. We suggest that different types of robots (less/more humanoid) could modulate action monitoring differently. The robot used in the present study was clearly not-humanoid. Participants could clearly recognize it as a robot and devoid of typical human traits that are often used in communication/collaboration. Nonetheless, using this robot helped us to maximize the difference between conditions. Additionally, our claim is supported by research on a different level of trust depending on the appearance of humanoid robots (Haring et al., 2013; van Pinxteren et al., 2019). Therefore, it would interesting to perform a similar experiment and compare the results with a more human-like robot. Fourth, as discussed below, our robot did not have a model of the human actor. By this, the robot's behavior helped to boost the characteristic differences between the player's partners. Fifth, our statistical analysis does not take the dyadic dependency into account, possibly biasing the estimated model parameters of the human-human condition downward. In the future, study with a bigger sample size, could answer the question whether dyadic dependencies play a role in the effects reported in our study. Sixth, even though participants were asked to keep their eyes on the center of the circular track, we did not control for eye-movements in this study, which could result in biased viewing-behavior on the tablet. However, we adjusted for ball position while modeling the ERPs, which is likely to be a proxy for current eye position and also remove eye movement and blink related ICs. Furthermore, the game required constant attention and engagement, so it was assured that participants did not look away from the tablet and the ball. Additionally, we are interested in the EEG signal related to the behavior, rather than the visual stimulus. All in all, we addressed the limitations, and are convinced that they do not impede the interpretations of our results as presented in next paragraphs.

5. CONCLUSIONS

Taken together, this study explored and described event-related potentials related to action monitoring in humans collaborating with other humans or robots. We used a dynamic real-time collaborative task and found that around 200–300 ms after our actions are disturbed, our brain activity is modulated by the type of partner. Our results corroborate previous research on the neural basis of human-robot interactions. Furthermore, we show the feasibility of conducting research on collaboration between human and non-human partners with EEG. The results of our study suggest that non-human partners modulate how we perceive and evaluate joint actions. It is crucial that we found the differences between human and robotic partners during a dynamical coordination task, as it can have implications on the future of human-robot interactions and brain-computer interfaces. We speculate that our findings could improve already existing interfaces that use recognition of errors in real-time. It could be especially useful in situations when robots and humans have multiple interactions and it is important to distinguish between different partners. Further research into the origin of the observed differences might elucidate the neuronal substrate of understanding the behavior of a partner during joint action. Such research and application could further facilitate interactions between humans and robots in many environments.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://osf.io/s6zbm/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by The Swedish Ethical Review Authority (Etikprövningsnämnden). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

PK, DK, and MB: conceived the study. AC, ALG, AK, and PK: designed the study. AG and MB: programmed the tablet and the robot. AC, ALG, AK, AG, and MT data collection. ALG and AK: major data analysis. AC, TK, BE, and PK: minor data analysis. AC, ALG, and AK: initial draft of the manuscript. AC, ALG, AK, AG, BE, MB, and PK: revision and finalizing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

We gratefully acknowledge the support by the European Commission Horizon H2020-FETPROACT-2014 641321-socSMCs, Deutsche Forschungsgemeinschaft (DFG) funded Research Training Group Situated Cognition (GRK 2185/1), Niedersächsischen Innovationsförderprogramms für Forschung und Entwicklung in Unternehmen (NBank)—EyeTrax, the German Federal Ministry of Education and Research funded project ErgoVR-16SV8052 and the DFG Open Access Publishing Fund of Osnabrück University. We acknowledge the support of Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2075—390740016 for BE.

ACKNOWLEDGMENTS

We would like to thank all partners in the socSMC consortium. Especially, we thank Alfred O. Effenberg and Tong-Hun Hwang for providing the sonification algorithm, and the help with implementing it.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnbot.2021.686010/full#supplementary-material

REFERENCES

Baillet, S. (2017). Magnetoencephalography for brain electrophysiology and imaging. *Nat. Neurosci.* 20, 327–339. doi: 10.1038/nn.4504

Ben-Ari, M., and Mondada, F. (2018). *Robots and Their Applications*. Cham: Springer International Publishing. doi: 10.1007/978-3-319-62533-1_1

Broadbent, E. (2017). Interactions with robots: the truths we reveal about ourselves. *Annu. Rev. Psychol.* 68, 627–652. doi: 10.1146/annurev-psych-010416-043958

Campa, R. (2016). The rise of social robots: a review of the recent literature. *J. Evol. Technol.* 26, 106–113.

Cavanagh, J. F., Cohen, M. X., and Allen, J. J. B. (2009). Prelude to and resolution of an error: EEG phase synchrony reveals cognitive control dynamics during action monitoring. *J. Neurosci.* 29, 98–105. doi: 10.1523/JNEUROSCI.4137-08.2009

Chavarriaga, R., Uščumlić, M., Zhang, H., Khaliliardali, Z., Aydarkhanov, R., Saeedi, S., et al. (2018). Decoding neural correlates of cognitive states to enhance driving experience. *IEEE Trans. Emerg. Top. Comput. Intell.* 2, 288–297. doi: 10.1109/TETCI.2018.2848289

Cheng, G., Ehrlich, S. K., Lebedev, M., and Nicolelis, M. A. (2020). Neuroengineering challenges of fusing robotics and neuroscience. *Sci. Robot.* 5, 7–10. doi: 10.1126/scirobotics.abd1911

Curioni, A., Knoblich, G., Sebanz, N., Goswami, A., and Vadakkepat, P. (2019a). "Joint action in humans: a model for human-robot interactions," in *Humanoid Robotics: A Reference*, eds A. Goswami and P. Vadakkepat (Dordrecht: Springer), 2149–2167. doi: 10.1007/978-94-007-6046-2_126

Curioni, A., Vesper, C., Knoblich, G., and Sebanz, N. (2019b). Reciprocal information flow and role distribution support joint action coordination. *Cognition* 187, 21–31. doi: 10.1016/j.cognition.2019.02.006

Czeszumski, A., Ehinger, B. V., Wahn, B., and König, P. (2019). The social situation affects how we process feedback about our actions. *Front. Psychol.* 10:361. doi: 10.3389/fpsyg.2019.00361

Czeszumski, A., Eustergerling, S., Lang, A., Menrath, D., Gerstenberger, M., Schuberth, S., et al. (2020). Hyperscanning: a valid method to study neural inter-brain underpinnings of social interaction. *Front. Hum. Neurosci.* 14:39. doi: 10.3389/fnhum.2020.00039

Delorme, A., and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* 134, 9–21. doi: 10.1016/j.jneumeth.2003.10.009

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). "Impact of robot failures and feedback on real-time trust," in *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (Tokyo: IEEE), 251–258. doi: 10.1109/HRI.2013.6483596

Diamond, A. M. Jr. (2020). Robots and computers enhance us more than they replace us. *Am. Econ.* 65, 4–10. doi: 10.1177/0569434518792674

Dimigen, O. (2020). Optimizing the ICA-based removal of ocular EEG artifacts from free viewing experiments. *Neuroimage* 207:116117. doi: 10.1016/j.neuroimage.2019.116117

Dimigen, O., and Ehinger, B. V. (2021). Regression-based analysis of combined EEG and eye-tracking data: theory and applications. *J. Vis.* 21, 3–3. doi: 10.1167/jov.21.1.3

Ehinger, B. V. (2018). *EEGVIS Toolbox*. Osnabrück. doi: 10.5281/zenodo.1312813

Ehinger, B. V., and Dimigen, O. (2019). Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. *PeerJ* 7:e7838. doi: 10.7717/peerj.7838

Ehinger, B. V., Konig, P., and Ossandon, J. P. (2015). Predictions of visual content across eye movements and their modulation by inferred information. *J. Neurosci.* 35, 7403–7413. doi: 10.1523/JNEUROSCI.5114-14.2015

Ehrlich, S. K., and Cheng, G. (2018). Human-agent co-adaptation using error-related potentials. *J. Neural Eng.* 15:066014. doi: 10.1088/1741-2552/aae069

Ehrlich, S. K., and Cheng, G. (2019a). "A computational model of human decision making and learning for assessment of co-adaptation in neuro-adaptive human-robot interaction," in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)* (Bari: IEEE), 264–271. doi: 10.1109/SMC.2019.8913872

Ehrlich, S. K., and Cheng, G. (2019b). A feasibility study for validating robot actions using EEG-based error-related potentials. *Int. J. Soc. Robot.* 11, 271–283. doi: 10.1007/s12369-018-0501-8

Eisenberger, N. I. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science* 302, 290–292. doi: 10.1126/science.1089134

Enz, S., Diruf, M., Spielhagen, C., Zoll, C., and Vargas, P. A. (2011). The social role of robots in the future-explorative measurement of hopes and fears. *Int. J. Soc. Robot.* 3, 263–271. doi: 10.1007/s12369-011-0094-y

Ferrari, M., and Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application. *Neuroimage* 63, 921–935. doi: 10.1016/j.neuroimage.2012.03.049

Hajcak, G., Moser, J. S., Holroyd, C. B., and Simons, R. F. (2006). The feedback-related negativity reflects the binary evaluation of good versus bad outcomes. *Biol. Psychol.* 71, 148–154. doi: 10.1016/j.biopsycho.2005.04.001

Haring, K. S., Matsumoto, Y., and Watanabe, K. (2013). "How do people perceive and trust a lifelike robot," in *Proceedings of the World Congress on Engineering and Computer Science*, 6, San Francisco.

Hinz, N.-A., Ciardo, F., and Wykowska, A. (2021). ERP markers of action planning and outcome monitoring in human-robot interaction. *Acta Psychol.* 212:103216. doi: 10.1016/j.actpsy.2020.103216

Huberth, M., Dauer, T., Nanou, C., Román, I., Gang, N., Reid, W., et al. (2019). Performance monitoring of self and other in a turn-taking piano duet: a dual-EEG study. *Soc. Neurosci.* 14, 449–461. doi: 10.1080/17470919.2018.1492968

Hwang, T.-H., Schmitz, G., Klemmt, K., Brinkop, L., Ghai, S., Stoica, M., et al. (2018). Effect- and performance-based auditory feedback on interpersonal coordination. *Front. Psychol.* 9:404. doi: 10.3389/fpsyg.2018.00404

Iturrate, I., Chavarriaga, R., Montesano, L., Minguez, J., and Millán, J. D. R. (2015). Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control. *Sci. Rep.* 5, 1–10. doi: 10.1038/srep13893

Iwane, F., Halvagal, M. S., Iturrate, I., Batzianoulis, I., Chavarriaga, R., Billard, A., et al. (2019). "Inferring subjective preferences on robot trajectories using EEG signals," in *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)* (San Francisco, CA: IEEE), 255–258. doi: 10.1109/NER.2019.8717025

Kim, S. K., Kirchner, E. A., Stefes, A., and Kirchner, F. (2017). Intrinsic interactive reinforcement learning-using error-related potentials for real world human-robot interaction. *Sci. Rep.* 7, 1–16. doi: 10.1038/s41598-017-17682-7

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., and Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron* 93, 480–490. doi: 10.1016/j.neuron.2016.12.041

Lewis, M., Sycara, K., and Walker, P. (2018). "The role of trust in human-robot interaction," in *Foundations of Trusted Autonomy*, eds H. A. Abbass, J. Scholz, and D. J. Reid (Cham: Springer International Publishing), 135–159. doi: 10.1007/978-3-319-64816-3_8

Loehr, J. D., Kourtis, D., Vesper, C., Sebanz, N., and Knoblich, G. (2013). Monitoring individual and joint action outcomes in duet music performance. *J. Cogn. Neurosci.* 25, 1049–1061. doi: 10.1162/jocn_a_00388

Luck, S. J., and Hillyard, S. A. (1994). Spatial filtering during visual search: Evidence from human electrophysiology. *J. Exp. Psychol. Hum. Percept. Perform.* 20:1000. doi: 10.1037/0096-1523.20.5.1000

Matusz, P. J., Dikker, S., Huth, A. G., and Perrodin, C. (2019). Are we ready for real-world neuroscience? *J. Cogn. Neurosci.* 31, 327–338. doi: 10.1162/jocn_e_01276

Mensen, A., and Khatami, R. (2013). Advanced EEG analysis using threshold-free cluster-enhancement and non-parametric statistics. *Neuroimage* 67, 111–118. doi: 10.1016/j.neuroimage.2012.10.027

Miltner, W. H. R., Braun, C. H., and Coles, M. G. H. (1997). Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a "generic" neural system for error detection. *J. Cogn. Neurosci.* 9, 788–798. doi: 10.1162/jocn.1997.9.6.788

Nastase, S. A., Goldstein, A., and Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *Neuroimage* 222:117254. doi: 10.1016/j.neuroimage.2020.117254

Palmer, J. A., Makeig, S., Kreutz-Delgado, K., and Rao, B. D. (2008). "Newton method for the ICA mixture model," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing* (Las Vegas, NV: IEEE), 1805–1808. doi: 10.1109/ICASSP.2008.4517982

Pavlov, Y. G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C. S., Beste, C., et al. (2021). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex*. doi: 10.1016/j.cortex.2021.03.013. [Epub ahead of print].

Peperkoorn, L. S., Becker, D. V., Balliet, D., Columbus, S., and Molho, C. (2020). The prevalence of dyads in social life. *PLoS ONE* 15:e0244188. doi: 10.1371/journal.pone.0244188

Pezzulo, G., Donnarumma, F., and Dindo, H. (2013). Human sensorimotor communication: a theory of signaling in online social interactions. *PLoS ONE* 8:e79876. doi: 10.1371/journal.pone.0079876

Pion-Tonachini, L., Kreutz-Delgado, K., and Makeig, S. (2019). ICLabel: An automated electroencephalographic independent component classifier, dataset, and website. *Neuroimage* 198, 181–197. doi: 10.1016/j.neuroimage.2019.05.026

Redcay, E., and Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat. Rev. Neurosci.* 20, 495–505. doi: 10.1038/s41583-019-0179-4

Salazar-Gomez, A. F., DelPreto, J., Gil, S., Guenther, F. H., and Rus, D. (2017). "Correcting robot mistakes in real time using EEG signals," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Singapore: IEEE), 6570–6577. doi: 10.1109/ICRA.2017.7989777

Schellen, E., Bossi, F., and Wykowska, A. (2021). Robot gaze behavior affects honesty in human-robot interaction. *Front. Artif. Intell.* 4:51. doi: 10.3389/frai.2021.663190

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., et al. (2013). Toward a second-person neuroscience. *Behav. Brain Sci.* 36, 393–414. doi: 10.1017/S0140525X12000660

Sebanz, N., Bekkering, H., and Knoblich, G. (2006). Joint action: Bodies and minds moving together. *Trends Cogn. Sci.* 10, 70–76. doi: 10.1016/j.tics.2005.12.009

Sebanz, N., and Knoblich, G. (2021). Progress in joint-action research. *Curr. Direct. Psychol. Sci.* 30, 138–143. doi: 10.1177/0963721420984425

Sebanz, N., Knoblich, G., and Prinz, W. (2003). Representing others' actions: just like one's own? *Cognition* 88, B11-B21. doi: 10.1016/S0010-0277(03)00043-X

Sheridan, T. B. (2016). Human-robot interaction: status and challenges. *Hum. Fact.* 58, 525–532. doi: 10.1177/0018720816644364

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., et al. (2016). *Artificial Intelligence and Life in 2030. One hundred year study on artificial intelligence: Report of the 2015-2016 Study Panel.* Stanford University, Stanford, CA. Available online at: http://ai100.stanford.edu/2016-report (accessed September 6, 2016).

Trendafilov, D., Schmitz, G., Hwang, T.-H., Effenberg, A. O., and Polani, D. (2020). Tilting together: an information-theoretic characterization of behavioral roles in rhythmic dyadic interaction. *Front. Hum. Neurosci.* 14:185. doi: 10.3389/fnhum.2020.00185

van Pinxteren, M. M., Wetzels, R. W., Rüger, J., Pluymaekers, M., and Wetzels, M. (2019). Trust in humanoid robots: implications for services marketing. *J. Serv. Market.* 33, 507–518. doi: 10.1108/JSM-01-2018-0045

van Schie, H. T., Mars, R. B., Coles, M. G. H., and Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nat. Neurosci.* 7, 549–554. doi: 10.1038/nn1239

Vesper, C., Butterfill, S., Knoblich, G., and Sebanz, N. (2010). A minimal architecture for joint action. *Neural Netw.* 23, 998–1003. doi: 10.1016/j.neunet.2010.06.002

Vesper, C., Schmitz, L., and Knoblich, G. (2017). Modulating action duration to establish nonconventional communication. *J. Exp. Psychol. Gen.* 146, 1722–1737. doi: 10.1037/xge0000379

Wahn, B., and Kingstone, A. (2021). Humans share task load with a computer partner if (they believe that) it acts human-like. *Acta Psychol.* 212:103205. doi: 10.1016/j.actpsy.2020.103205

Widmann, A., Schröger, E., and Maess, B. (2015). Digital filter design for electrophysiological data-a practical approach. *J. Neurosci. Methods* 250, 34–46. doi: 10.1016/j.jneumeth.2014.08.002

Wiederhold, B. K. *Cyberpsychology, Behavior, and Social Networking*. (2021) 24, 289-290. doi: 10.1089/cyber.2021.29213.editorial

Wilkinson, G., and Rogers, C. (1973). Symbolic description of factorial models for analysis of variance. *J. R. Stat. Soc.* 22, 392–399. doi: 10.2307/2346786

Wolf, T., Sebanz, N., and Knoblich, G. (2018). Joint action coordination in expert-novice pairs: can experts predict novices' suboptimal timing? *Cognition* 178, 103–108. doi: 10.1016/j.cognition.2018.05.012

Wykowska, A., Chaminade, T., and Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philos. Trans. R. Soc. B Biol. Sci.* 371:20150375. doi: 10.1098/rstb.2015.0375

Yang, G.-Z., Bellingham, J., Dupont, P. E., Fischer, P., Floridi, L., Full, R., et al. (2018). The grand challenges of *Science Robotics*. *Sci. Robot*. 3:eaar7650. doi: 10.1126/scirobotics.aar7650

Yeung, N., Botvinick, M. M., and Cohen, J. D. (2004). The neural basis of error detection: conflict monitoring and the error-related negativity. *Psychol. Rev*. 111:931. doi: 10.1037/0033-295X.111.4.931

Zhang, H., Chavarriaga, R., Khaliliardali, Z., Gheorghe, L., Iturrate, I., d, R., et al. (2015). EEG-based decoding of error-related brain activity in a real-world driving task. *J. Neural Eng*. 12:066028. doi: 10.1088/1741-2560/12/6/066028

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 09 September 2021
doi: 10.3389/fnbot.2021.630386

Client-Server Approach for Managing Visual Attention, Integrated in a Cognitive Architecture for a Social Robot

*Francisco Martín[1], Jonatan Ginés[1], Francisco J. Rodríguez-Lera[2], Angel M. Guerrero-Higueras[2] and Vicente Matellán Olivera[2]**

[1] Intelligent Robotics Lab, Universidad Rey Juan Carlos, Fuenlabrada, Spain, [2] Grupo de Robótica, Universidad de León, León, Spain

This paper proposes a novel system for managing visual attention in social robots. This system is based on a client/server approach that allows integration with a cognitive architecture controlling the robot. The core of this architecture is a distributed knowledge graph, in which the perceptual needs are expressed by the presence of arcs to stimuli that need to be perceived. The attention server sends motion commands to the actuators of the robot, while the attention clients send requests through the common knowledge representation. The common knowledge graph is shared by all levels of the architecture. This system has been implemented on ROS and tested on a social robot to verify the validity of the approach and was used to solve the tests proposed in RoboCup @ Home and SciROc robotic competitions. The tests have been used to quantitatively compare the proposal to traditional visual attention mechanisms.

Keywords: visual attention, cognitive architectures, social robots, object-based visual attention, robotic cognition, robot vision

OPEN ACCESS

Edited by:
Letizia Marchegiani, Aalborg University, Denmark

Reviewed by:
Csar Renn–Costa, Federal University of Rio Grande do Norte, Brazil
Roseli Aparecida Francelin Romero, University of São Paulo, Brazil

***Correspondence:**
Vicente Matellán Olivera
vicente.matellan@unileon.es

Received: *17 November 2020*
Accepted: *12 May 2021*
Published: *09 September 2021*

Citation:
Martín F, Ginés J, Rodríguez-Lera FJ, Guerrero-Higueras AM and Matellán Olivera V (2021) Client-Server Approach for Managing Visual Attention, Integrated in a Cognitive Architecture for a Social Robot. Front. Neurorobot. 15:630386. doi: 10.3389/fnbot.2021.630386

1. INTRODUCTION

Mobile social robots incorporate a myriad of sensors and actuators (Kanda and Ishiguro, 2017), for example sonar and LIDAR sensors for obstacle detection, autonomous location and navigation, microphones and speakers for human-robot interaction, and more and more commonly, different types of cameras. Unlike other sensors, the portion of the space that cameras can perceive is limited by their field of vision, which is usually quite narrow compared to the entire space surrounding the robot. Besides, the design of most mobile social robots resembles human morphology. Even those non-humanoid robots place cameras on the robot's head, which is attached to the body by an articulated neck. These actuated cameras overcome the limitations of the narrow field of vision but need to implement an attention management system because it is not possible to simultaneously cover the entire space around the robot. Even if many cameras could be placed in the robot, along with enough computer power to analyze the images, the cognitive system would need to focus on the more relevant elements detected using an attention-managing system. For the sake of this paper, the problem faced is to define where the attention systems have to direct the camera a mobile service robot according to the perceptual needs thereof.

Visual attention systems based on fixed patterns for scanning a scene were the first approaches made, but they have proven to not be very efficient, and more sophisticated approaches, such as the one proposed in this paper, are required (Nguyen et al., 2018). Another aspect to take into

account in the evolution of visual attention systems is the integration into complex robotic software architectures which are in charge of selecting the most adequate behavior to fulfill the robot's task. This integration requires the attention system to be modular, parametrizable, and able to share a common way of representing information.

In a previous work, Agüero et al. (2012), a method for managing visual attention, integrated in the cognitive architecture, was initially proposed. In that seminal work, cognitive behaviors indicate their perceptual needs, and the attention system organizes these needs according to their salience. The new approach presented in this paper differs from that work in that the attention system does not arbitrate among behavioral needs, but among elements to be perceived that are indicated by the planning system at the highest level of the cognitive architecture. In order to do so, the system relies on a centralized repository of information that has been implemented as a "knowledge graph." The robot stores all relevant internal and external knowledge in this repository. The graph contains nodes that represent the elements of the environment, and arcs that indicate the relationships, symbolic or geometric, among them.

The software design of the proposed attention system is modular, allowing specialization in the way different types of stimuli are dealt with. Modularity has been achieved using a client-server systems. This approach is also scalable, meaning that it can be easily expanded with more types of stimuli by adding separate clients. Monolithic approaches to the visual attention systems make them much more difficult to extend.

This implementation has been validated in the RoboCup@Home[1] and European Robotics League[2] competitions, which consist of a set of tests which take place in a simulated domestic environment. The performance of each robot is evaluated for tasks focused on assistance or collaboration with humans, which is an excellent way to contrast different research approaches. For all the tasks in these competitions, robots need to visually perceive the scene. Some tasks require the robot to look at a person's face while talking with them or to follow them around a house. For other tasks, the robot must search for objects on a table, or check if it is carrying the correct objects on its tray. All of the task in the competitions require a challenging management of visual attention. In particular, guaranteeing that no interest point remains unattended for a long period of time is one of the most relevant requirements; the time for answering questions about the environment is limited, and the total time for accomplishing the task is also limited. We consider that the maximum time that an interest point remains unattended is the most relevant parameter when comparing different solutions.

In summary, the main contribution made in this paper is the design of the visual attention system. This system is integrated into the cognitive architecture through the knowledge graph, where visual perception requirements are expressed through the creation of arcs between nodes that indicate these requirements.

[1]http://www.robocupathome.org/

[2]https://www.eu-robotics.net/robotics_league/

The remainder of the paper is structured as follows: In section 2 we review the state of the art of visual attention systems, and how different cognitive architectures address their integration. In section 3 we describe the proposed cognitive architecture. Next, we describe the visual attention system. In section 5 real examples of its operation are presented. Finally, in section 6, conclusions are drawn and future work is discussed.

2. STATE OF THE ART

Visual attention management systems have been a recurrent research topic in mobile robotics. Historically, there have been different approaches to this problem, from basic ones, where segmentation was directly used to focus the attention of the robot, as in Scheier (1997), to ones using methods borrowed from other scientific fields, such as psychology and ethology. For instance, Butko and Movellan (2009) proposed a method of driving a robot that scans scenes based on the model of visual searches in humans. This method predicts scanpaths to maximize the long-term information about the location of the target of interest. In Meger et al. (2008) a combination of a peripheral-foveal vision system, and the attention system that combines bottom-up visual saliency with structure from vision allowed the "Curious George" robot to build a semantic map of the region explored, thereby labeling objects.

The use of visual attention in social robots is widespread. For instance, Kismet (Breazeal and Scassellati, 1999) the famous robotic head which popularized the "affective computing" paradigm, included an attention system capable of directing the robot's eyes toward the areas of interest of an image. These areas of interest, or high salience, were calculated by combining several filters (face detection, color, and movement), which allowed the robot to pay attention to different scene elements. These are the basic questions (Treue, 2003),: *where*, *what* and *how* such as how to recognize a point of interest, and why it is needed it for scene understanding.

The WABIAN humanoid robot (Hashimoto et al., 2002) was also equipped with an active vision system that directed its gaze toward people, based on images and sound. The work of Wolfe (1994), studying how to determine relevant areas in images, inspired these approaches. This concept of salience is addressed in a multitude of works (Itti et al., 1998; Harel et al., 2006; Hou and Zhang, 2007; Goferman et al., 2012; Grotz et al., 2017), although most focus on which parts of the image are relevant, without spatial information beyond the image. The salience-based approach was previously explored by the authors of this paper, as described in Garcıa et al. (2010), and is still present in the current proposal.

In Bachiller et al. (2008), "Regions of Interest" are used in an image to determine where to direct the camera of a robot. In this case, the robot's active tasks determine the attention zones. Recent works (Stefanov et al., 2019) combine bio-inspired models with Neural Networks to obtain saliency maps, as opposed to spatial areas of the environment. Our approach is not based on the detection of exciting areas in images, but rather areas of

space where to direct the robot's camera. We enrich the image-processing with 3-D information. Our attentive system never works on image coordinates but orients itself on the real world. Another relevant difference is that our system is intended to avoid that any interest point identified by the cognitive level could remain unattended for long periods of time.

Integrating the attention system into the cognitive architecture is one of the major problems when using it as a social robot. Some of the systems already mentioned are effective in managing visual attention but are hardly integratable into cognitive architectures.

In Agüero et al. (2012), the authors had proposed a method of visual attention management applied to humanoid robots integrated within the cognitive architecture base on salience. The term salience ceases to be used only for areas of an image and applies to points in space. The salience indicates the need to look at them and increases proportionally to the need to see them. Current behaviors determine this need. In this work, a subsumption architecture (Brooks, 1986), developed for soccer robots (Martin et al., 2010), integrates the attention system. The different execution units that form the behaviors indicate their perceptual needs, and it is the attention system that merges these needs through salience. The current proposal differs from this work in that the attention system does not arbitrate between behavioral needs, but between elements to be perceived by a single behavior.

Cortex is another cognitive architecture, closely related to our own. Its attention system described in Manso et al. (2018) has some similarities to our proposal. The main difference is that Cortex indicates where to find items, instead of determining search points. The system thereby determines which areas of the environment can contain it, thus directing the robot's gaze there. Although similar in many aspects, the system presented in this paper solves the problem of scanning an area and seeing what can be found in it.

iCub robot (Ruesch et al., 2008) applies the concept of EgoSphere, originally by Kawamura et al. (2005). This sphere stores the orientation of the perceived elements to the robot. Saliency and spatial information (angles only) determine the orientation of the robot head. In this case, salience applies to the areas of the self-sphere relevant to the robot. A highly-valued contribution of this approach is sensory fusion. The attention system also adds auditory information to modify the salience of specific areas. Our proposed system is also able to perform advanced spatial reasoning, not limiting the angle of visual stimulation.

Visual attention can be influenced by other sensors. For instance, in Aliasghari et al. (2020), visual attention is used in a social robot to control where the gaze is directed within the context of a group conversation. In order to make this decision, the system uses other stimuli, such as where the people are, where the sound comes from, hand movements, and pointing gestures. It also uses some concepts of proxemic. For instance, it is more natural to look at people who are closer than those further away from the beholder. These stimuli are incorporated into a logical control unit with long-and short-term memory. This control unit decides the neck's movement.

3. THE COGNITIVE ARCHITECTURE FOR SOCIAL ROBOTS

The cognitive architecture, in which the proposed attention system is integrated, is organized in concentric layers, named tiers, as shown in **Figure 1**. A more detailed description of the architecture can be found in Martin et al. (2020). Describing from the outermost layer to the innermost layer, which can also be considered as a bottom-up description:

- `Tier 5` represents the bare metal, the robot hardware and the programming interfaces of the basic controllers of sensors and actuators.
- `Tier 4` interacts directly with these controllers to offer a higher level of abstraction in defining the robot's basic *skills*, such as navigating to a place, picking up an object, talking with a person, wandering, detecting objects, etc. The innermost layers use the skills in this layer as primitives.
- `Tier 3` is the operational level of the robot. These operations are defined as *actions*. An action uses different skills from tier 4 to accomplish a unit task, e.g., getting the robot to move from one room to another using the skill of navigation. In addition, the robot should take into account whether the door is open, using its perceptual and probably attention skills as well. If the door is closed, the robot will use its manipulation skills to open it and enter the destination room. The actions' implementation defines how the skills are named and which specific parameters (metric destination point, position of the element to manipulate, phrase to speak, etc.) are used. Actions, loops, branches, and sequences can be used to define the control logic for achieving the task.
- `Tier 2` is the task manager level where *plans* (ordered executions of actions) are generated. It is based on a symbolic planner which uses PDDL to define what types, symbolic predicates, and actions are used to solve a problem. This knowledge base is accessible by other tiers.
- `Tier 1` manages the high-level *mission* of the robot. This level is built using hierarchical state machines which define the different stages of the robot mission at a high level of abstraction. Transitions between states are implemented by consulting predicates in the knowledge base, and the goals to be solved by the planner in Tier 2 are defined in the states.

Tiers 1 and 2 mainly use symbolic information for facing the process of information abstraction, while Tiers 3 and 4 use sub-symbolic information, mainly sensor readings. When a state machine at Tier 1 establishes a goal, the planner at Tier 2 creates a plan using the content of its knowledge base. This plan is built as a sequence of domain actions. The planner delivers the actions at Tier 3 one at a time. Each time an action indicates that it has been successfully completed, the next one is delivered until the plan finishes. If an action ends with an error, it forces a replanning.

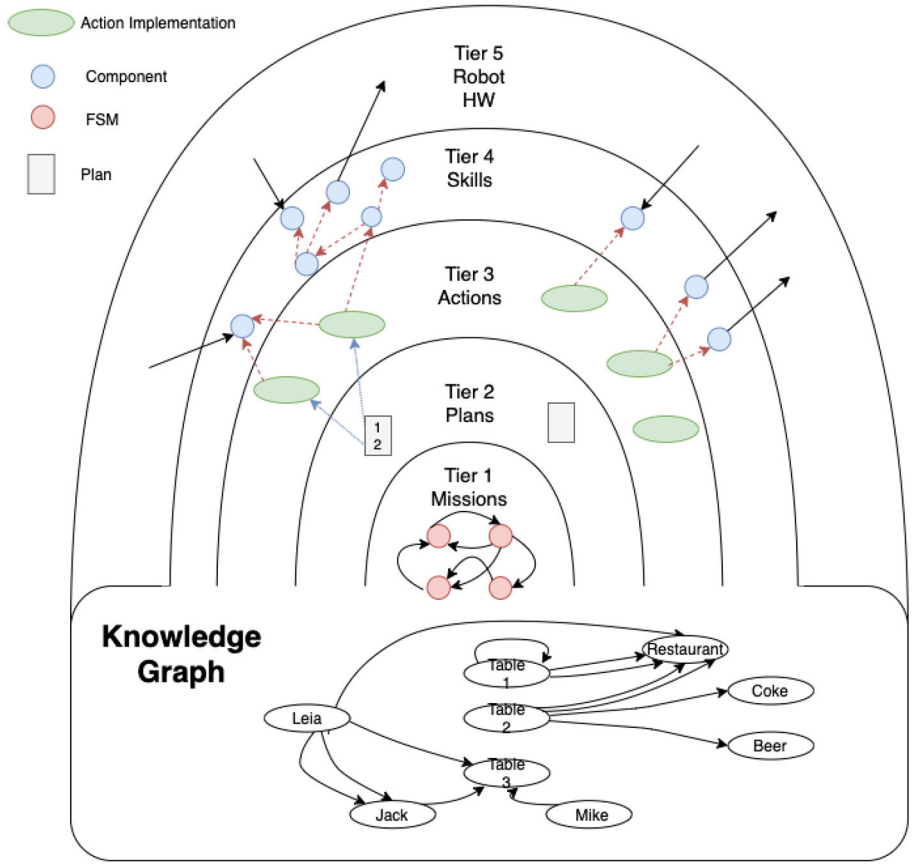


FIGURE 1 | Layered cognitive architecture shown as concentric tiers. The innermost layers (Tier 1) control the mission using a state machine, setting the goals to be achieved by the planning layer (Tier 2). This layer activates the actions that are part of the plan (in Tier 3). Actions can use skills (at Tier 4) to carry out their task. These skills can be perceptual or acting, directly sending information to the actuators, or receiving information from the sensors.

As mentioned previously, Tier 3 contains the implementation of the actions defined in the PDDL domain at Tier 2. This level is the bridge between both paradigms. The planner activates actions according to the generated plan. When activated, the planner passes the parameters to the actions (instances of a type). Usually, the action must translate symbols into specific data. For example, a *move* action could receive *kitchen* as a parameter. The action must then obtain the metric coordinate corresponding to the *kitchen* symbol and send it to the navigation module.

In order to manage the information between layers the Knowledge Graph is defined. It stores all the information relevant to the operation of the robot, and is accessible from all the Tiers. This shared representation of data disengages some components from others, especially among different layers. For instance, an action in Tier 3 uses the result of computing a skill in Tier 4 by reading it from the knowledge graph. Tier 1 can also use the symbolic information contained in the graph.

The elements of the knowledge graph are nodes and labeled arcs. The nodes represent instances of a specific type. The arches can contain a text, or they can provide a geometric transformation. The visual attention requirements are expressed as arcs of a special type "want_see," as explained in the next section, where the knowledge graph of **Figure 2** is depicted in more detail.

The relationship between the symbolic information at Tier 2 and the global knowledge graph is based on a synchronization process. This process adds nodes to the knowledge graph when the symbolic knowledge base creates instances of a relevant type. It also creates arches when the symbolic knowledge base inserts a relevant predicate. If the predicate has two arguments of related types, the arc connects two nodes with a text corresponding to the predicate. If the predicate has only one argument it is represented as a self arc (*need_check* arc in **Figure 2**). Updates only go one way, from the symbolic knowledge base to the graph. Updates from the graph to the symbolic knowledge base are not permitted.

ROSPlan (Cashmore et al., 2015) is the planner used in Tier2 and a BICA (Martín et al., 2016) framework was used for the implementation of the actions and skills, as BICA components. A BICA component is an independent process which can declare that it depends on other BICA components. When a BICA component is activated, it automatically activates all its dependencies. When all components which enable a dependency are deactivated, the dependency is deactivated. This mechanism is a simple way to save computation time when the results of

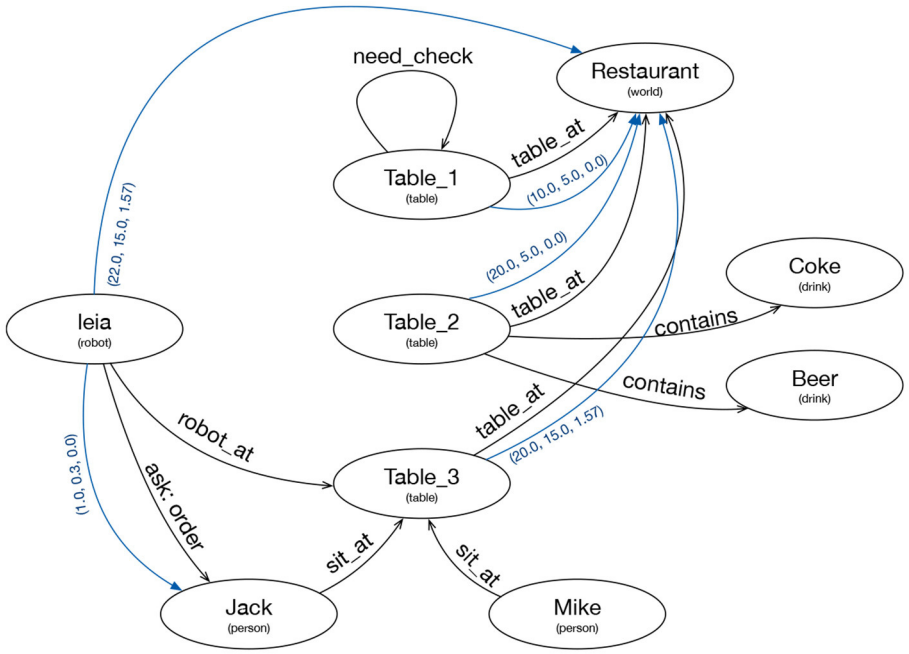


FIGURE 2 | Knowledge Graph representing the internal and external knowledge of the robot. Ellipses represent nodes with an ID and a type. Black lines are text arcs and blue lines are geometric arcs.

certain computations are not being used, and permits different compositions of skills as shown in **Figure 2**.

4. THE ATTENTION SYSTEM

The attention system is integrated into the cognitive architecture as a skill in Tier 4. It can be activated by actions in Tier 3 which require attention to different stimuli. Actions set perceptual needs in the knowledge graph by creating `wants_see` arcs from one robot node to another, as shown by the red arrow in the left part of **Figure 3**. Other skills in Tier 4 can also add arches requesting attention, as well as in the innermost tiers if it is considered necessary.

The attention mechanism is built as a client-server system, as shown by the orange boxes in **Figure 3**. The three clients in the figure send *attention points* to the attention server. The *attention server* sends motion commands to the robot's actuators in order to direct the camera to a position in the space. Its main task is to select the next pose to look at among all the requests received and the length of time to maintain that position in the robot's field of view. The attention clients make requests to the server by sending attention points.

Each attention point sent to the server is labeled with the stimulus type and the id to perceive. A client can communicate to the server that it no longer requires attending to a specific type and/or instance. The server then removes these points from its list.

There are as many clients as types of stimulus to deal with. If the robot wants to perceive a table, it can mean that it is interested in either scanning the entire surface, or in determining the existence of the table itself. Different clients should be built for each one. For example, if interested in the objects on the table, the attention points will cover the table's surface, apart from assuming that the table is a static element of the environment. If the robot wants to perceive a small object, there will be a single attention point in the center of the object, if already detected. If the object has not been detected yet, the points will be placed where it is more likely that the object could be. Each type of stimulus requires a custom client specialized in perceiving it adequately. For example, while perceiving a person, it can be enough to look at his face, but these are dynamic points.

Attention clients scrutinize the Knowledge graph in case their participation is required, that is, they look for the existence of `wants_see` arcs from one node to an element of the type that this client manages. If so, it generates a set of attention points with geometric information that indicates where to direct the robot's camera to perceive the stimulus.

In the case of **Figure 3**, there are nodes of types *robot*, *table*, and *person*. There are also geometric and symbolic arches. Several processes, named `[stimulus]_attention_client` one for each type of node which the robot can pay attention to, are shown in the center. Each one is aware of the changes in the graph. In this case, `table_attention_client` should be active because there is an arc from a node of type robot to a node of type table. When active, `table_attention_client` sends a set of `attention_points` in the frame of `Table_1` to check.

Furthermore, the `attention_server` receives the attention points of all the clients and iterates among each one of them. The `attention_server` maintains a list with all the

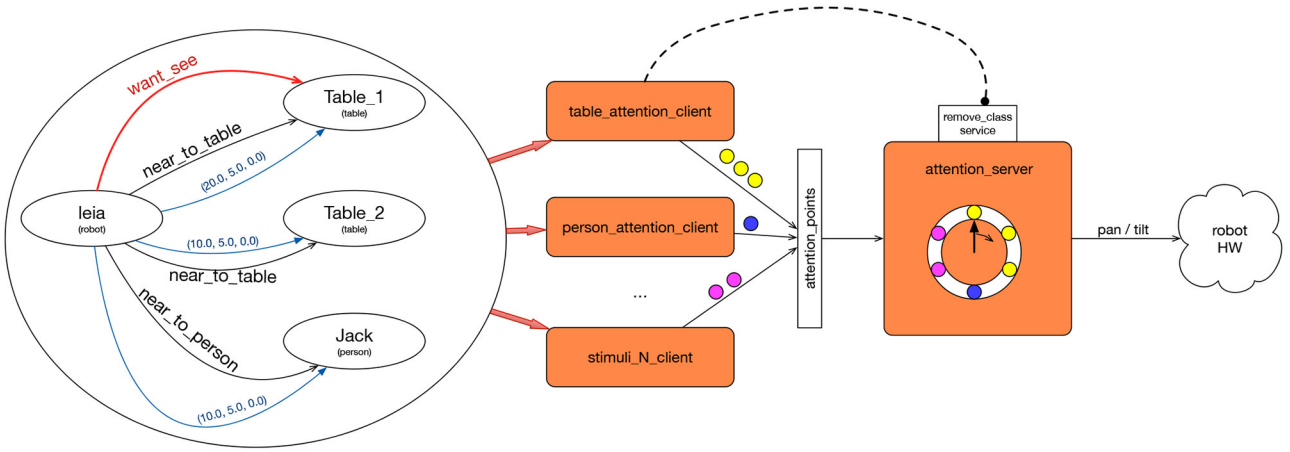


FIGURE 3 | Attention system architecture. The attention clients (orange rectangles in the center) observe the graph **(left)** looking for "want_see" edges. Each client is specialized in one type of goal node for the "want_see" edge. When any of them activates, it sends the attention points to the attention_server **(right)**. The attention_server sends motion commands directly to the neck motors.

attention points received. For each point, it transforms it into a coordinate related to the axes of the robot, and generates a pan and a tilt that it sends to the motors of the robot's neck, visiting each point for a few seconds.

The attention module is not responsible for image detection, only for looking there. If an action requires detecting objects on a table, this action must activate both the attention and the module that perceives the objects in the image. When an object is detected in the image, this is written down in the graph, thereby allowing eliminating the corresponding attention arcs from the action.

Attention clients send the points of attention of each one of the elements to attend to (small circles of **Figure 3**) to the server. Messages sent to the server contain the class and identifier of the element. In addition, they contain a vector of 3D points. For each point, we specify the reference axis (frame_id) of its coordinates.

The attention server receives the set of new points, $\mathcal{NP}$, sent by the clients. The server stores the points received in a list. Each point, p, on the list $\mathcal{P}$ that the server stores contains the following fields:

- `point_id`: The attention server must be able to attend to requests to eliminate points of attention, either by specifying an entire class or just one instance of a class. This field contains an identifier `class.instance_id.n`, where n is the i-th point of attention of one of the elements. In this way, it is easy to determine the points that belong to each class and instance.
- `point`: The point coordinates, stamped with its frame_id and time.
- `tilt` and `pan`: The axes of reference of the points can move with regard to the robot. That means that points could be coordinates of a global map, and the robot could be moving, or points could be coordinates of the robot arm, and the robot could be moving its arm.
- `epoch`. Attention cannot be paid to one point again until the rest of the points have been attended to. Epoch represents the current iteration of the attention system. Attentions server does not attend to a point if there is another point with a lower epoch value. Each point attended increases its epoch by one.

Algorithm 1 Attention Server algorithm.

while *robot_operation* **do**
  for all $p_i \in \mathcal{NP}$ **do**
    $p_i^{epoch} = p_j^{epoch}$, where $p_j = last(\mathcal{P})$
    $\mathcal{P} \leftarrow p_i$
  end for
  for all $p_i \in \mathcal{P}$ **do**
    $p_{aux1} = RT^{4\times4}(p_i^{frame} \rightarrow pan\_frame) * p_i$
    $p_{aux2} = RT^{4\times4}(p_i^{frame} \rightarrow tilt\_frame) * p_i$
    $p_i^{pan} = \arctan(p_{aux1}^{y}, p_{aux1}^{x}))$
    $p_i^{tilt} = \arctan(p_{aux2}^{z}, p_{aux2}^{x}))$
  end for
  sort($\mathcal{P}$)
  repeat
    $p = first(\mathcal{P})$
    $p^{epoch} = p^{epoch+1}$
  until is_in_fovea(p)
  pan_t = p^{pan}
  tilt_t = p^{tilt}
  send_command_to_neck($pan_t, tilt_t$)
  t_{flight} = flight_time($pan_{t-1}, pan_t, tilt_{t-1}, tilt_t$)
  t_{in_point} = 1.0 s
  wait($t_{flight} + t_{in\_point}$)
end while

The attention server calculates the *pan* and *tilt* values to send to the robot's neck actuators, calculated from P. pan_{t-1} is the

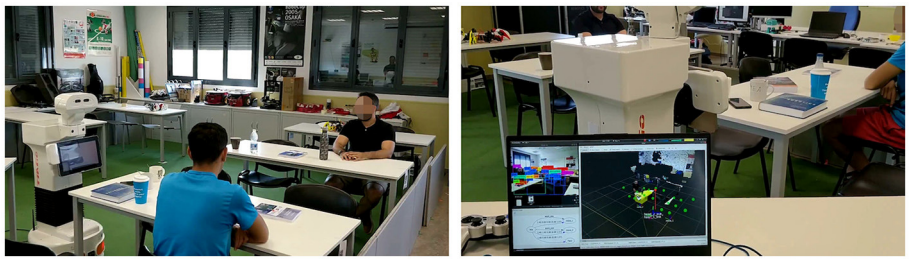


FIGURE 4 | Left image: Experimental setup. Right image: Laptop screen showing the objects being perceived (left part of the screen and their spatial location).

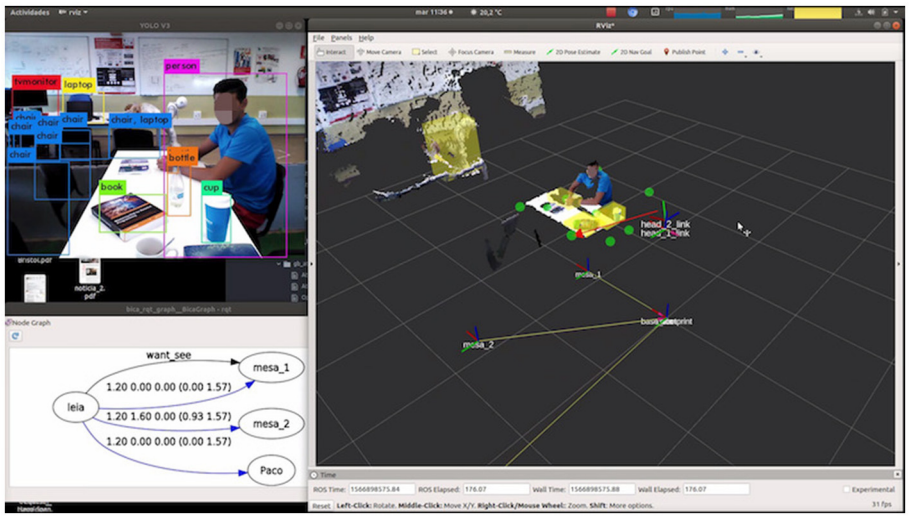


FIGURE 5 | Visual debugging of detections (left upper image), knowledge graph (left bottom figure), attention points (green circles), and attending point (red arrow) in right part of the figure.

current pan value, and pan_t is the new pan value to send to the actuators.

This algorithm is summarized in Algorithm 1 and is based on these three rules:

1. The robot cannot handle an attention point again until after handling the other attention points.
2. The next attention point is the point that implies the least head movement.
3. If the next attention point is already in the fovea, it is considered handled.

In more detail, the algorithm works as follows:

- Lines 2-5 show how the server incorporates the new points $\mathcal{NP}$ received from clients to the list $\mathcal{P}$ of attention points.
- Lines 6-11 recalculate the pan and tilt of each point before sorting. Many points could be defined in frames that have changed with regard to the robot's neck. If we define points in `map` frame, their new pan/tilt values depend on the robot displacement and the localization.
- Line 12 sorts $\mathcal{P}$ using the operator "$<$" defined as:

$$p_i < p_j = \begin{cases} \text{if } p_i^{epoch} < p_j^{epoch}, \\ \text{or} \\ \text{if } p_i^{epoch} = p_j^{epoch} \text{and } j(p_i) < j(p_j) \end{cases}$$

where,

$$j(p) = |p^{pan} - pan_{t-1}| + |p^{tilt} - tilt_{t-1}|$$

From now on, the most appropriate points to pay attention to are at the top of the list.
- Lines 13–16 select the point p to attend on the 13–16. Starting at the beginning of the list, we take the first one that is not in the fovea.
- p point contains the new pan and tilt values. After sending them to the actuators, the waiting time before sending other values to the actuators must be calculated. This waiting time depends on two time periods. The first one is the duration of positioning in the new pan/tilt values (line 20). The second one is the time in which the robot maintains attention to a point. It is convenient for the robot to stop at a point for a

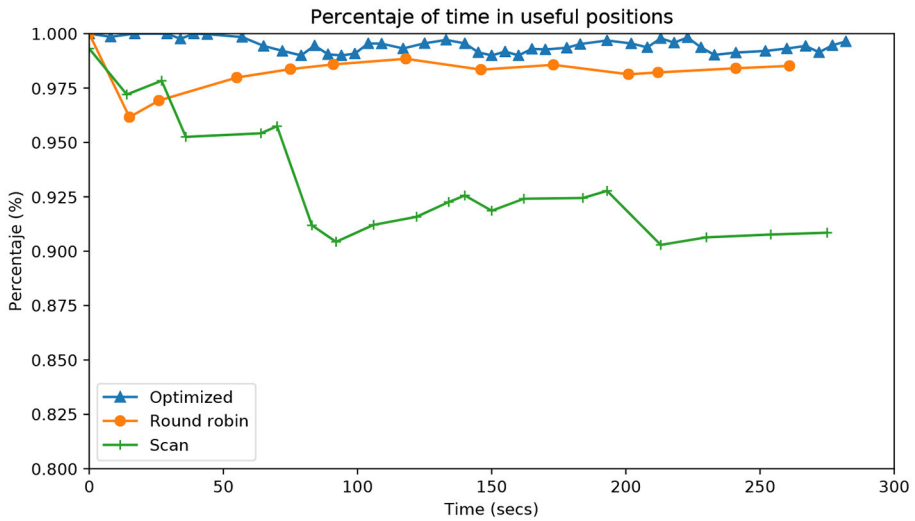


FIGURE 6 | Evolution during the entire time of the experiment (*X*-axis) as a percentage of time (*Y*-axis) that there is any attention point in the fovea (central part of the image). Each line represents a different algorithm, comparing our contribution (Optimized) with respect to the two usual algorithms of attention (Simple scanning and Round robin).

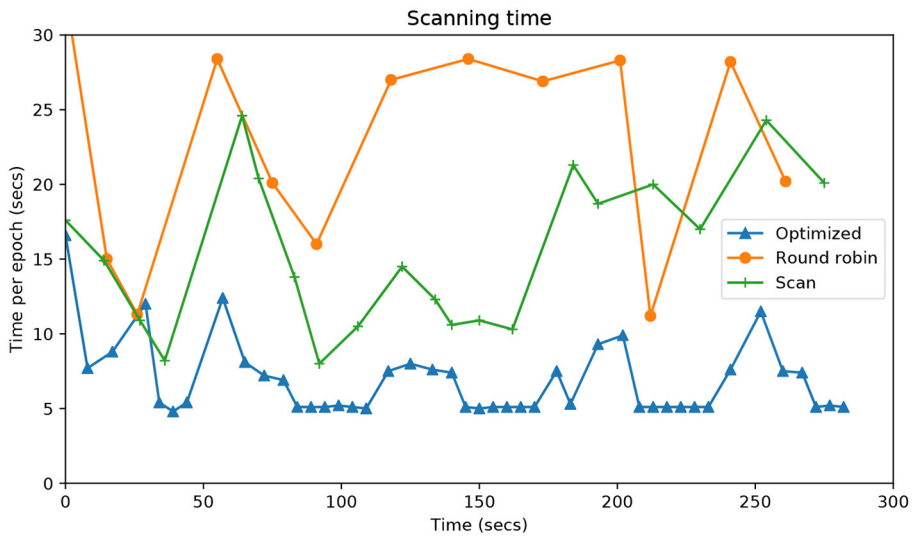


FIGURE 7 | Evolution during the entire time of the experiment (*X*- axis) vs. the time in seconds (*Y*-axis) that it takes each algorithm to visit all the attention points.

short moment. The image could be moving or degenerating its processing. We consider a second to be an appropriate value.

Our attention system's design allows us to efficiently attend to visual stimuli (objects, people, areas, etc.) since it personalizes the attention for each stimulus type. The system is also scalable: a new kind of stimulus to attend to requires only creating an attention client that defines the points of attention in the stimulus's reference axis. In the next section we will show the experiments carried out to validate our proposal, a simple way to save computation time when the results of certain computations are not being used and which allows different compositions of skills as shown in **Figure 2**.

5. EXPERIMENTAL VALIDATION

This section describes the experiments carried out with a real robot to evaluate the validity of our proposal. The main feature of an attention system is attending to the relevant areas of robot operation. In order to determine what the advantages of the proposed system are, two other classical approaches have also been implemented:

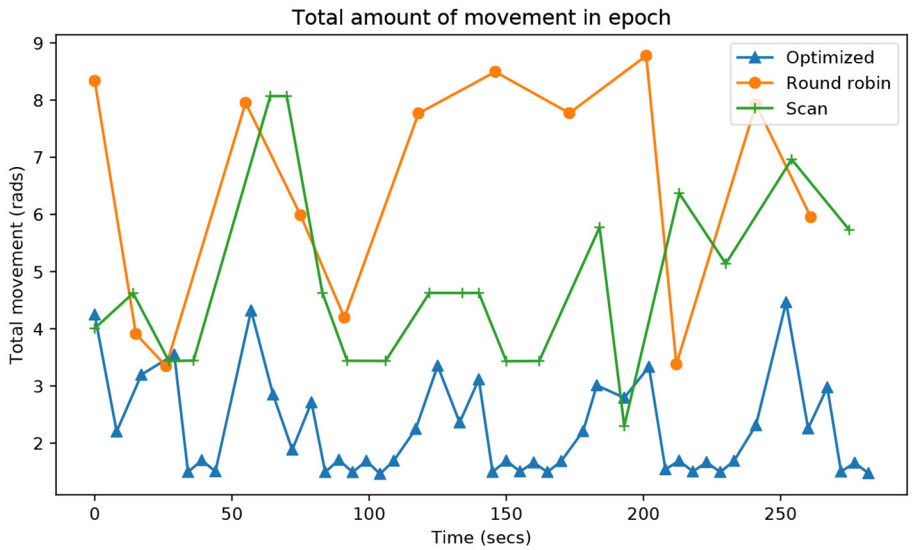


FIGURE 8 | Evolution during the entire time of the experiment (*X*-axis) of the accumulated neck rotation, in radians (*Y*-axis), that it takes each algorithm to visit all the attention points.

1. Slow scan across the environment around the robot. This mechanism would be activated whenever the robot wants to perceive something, moving the robot's neck in a fixed pattern.
2. Selection of the point of attention, p_i, from the list $\mathcal{P}$ using round-robin without ordering the points to optimize the movement of the robot's neck.

The metrics used to compare this proposal vs. these other two approaches are:

- The percentage of time that the robot is attending relevant areas.
- The time to cover all relevant areas.
- The amount of energy used to cover all relevant areas.

The correctness of the detections has not been included as criterion for the experimentation because it depends on other modules in charge of perception. Neither has a quantitative analysis of the integration of the attention system in the cognitive architecture been included because this analysis can only be done qualitatively. In our system was validated by integrating it into the software of the SciRoc Robotics competition.

The SciRoc competition environment **Figure 4** was reproduced for the evaluation. It simulates a restaurant in which the robot should check how many persons are sitting at the tables and which objects are on the tables. The robot is in front of one of the tables (mesa_1), and to its left there is another table (mesa_2). There are several objects to be detected on the tables, and two people sitting, one at each table. The setup of this experiment can be seen in this video [3].

The robot knows *a priori* its relative position to the table. This information is introduced in the knowledge graph (**Figure 5**). A table_attention_client was implemented which establishes 10 points of attention per table: 6 on the surface and 4 in positions where there could be people. Attention points can be seen in **Figure 5** depicted as green circles.

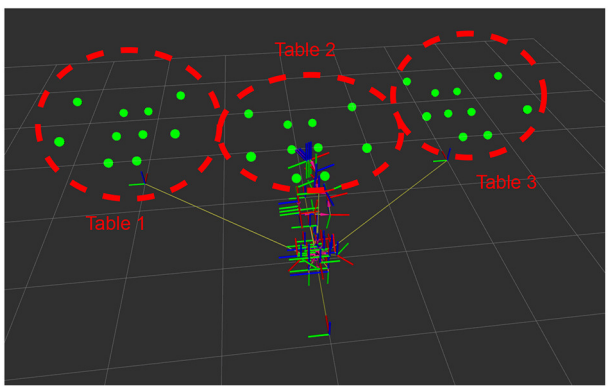


FIGURE 9 | Attention points (green circles) corresponding to the setup for the second experiment, with the position of the robot (denoted by the axes of its actuators). In this case, the attention points were located on three tables.

A skill that adds "want_see" arcs in the graph from the robot to the tables was specifically designed for this setup. Every 30 s, the "want_see" arc is added to or deleted from table_2 (*mesa_2* in the figure). When both arches are active, there are two points of attention on each table. Initially, we activate attention for two tables, so there are 20 attention points. At time 30, we remove an arc, so the robot is attending to only one table, and there are 10 points' the cycle loop restarts at time 60.

The attention mechanisms were compared with two alternatives previously mentioned:

[3]https://youtu.be/IyCyx_HfdrE

TABLE 1 | Results of the second experiment, showing the time (in seconds) to deal with an attention point, and the number of points attended.

	1 Table			2 Tables			3 Tables		
	Opt	**RR**	**Scan**	**Opt**	**RR**	**Scan**	**Opt**	**RR**	**Scan**
Mean	6.38 s	5.95 s	9.44 s	9.11 s	6.70 s	12.01 s	9.42 s	7.69 s	14.50 s
Stdev	4.40 s	6.90 s	7.76 s	5.71 s	6.84 s	0.83 s	5.47 s	7.69 s	13.20 s
Median	3.99 s	3.09 s	6.29 s	7.94 s	4.29 s	6.19 s	8.99 s	4.69 s	6.20 s
Max	**18.89 s**	28.49 s	31.09 s	**28,39 s**	37.39 s	37.89 s	**25.69 s**	50.09 s	41.39 s
Points	131	134	90	192	265	135	285	345	191

- *Round Robin*: the robot evaluates attention points in the order in which they are stored on the server, which can be expressed as a new operator $<'$ defined as:

$$p_i <' p_j \text{ if } p_i^{epoch} < p_j^{epoch}$$

- *Scan*: The robot continuously moves its neck to cover the robot's environment. This approach was optimized so that it only scans the areas where there are points of attention. Before scanning, it calculates the range of pan/tilt angles based on the current points of attention.

Figure 6 shows the accumulated percentage of time that the robot has any attention point in the fovea. The fovea is the central area of the image, half the size of the total image. The marks on the lines of each approach indicate when each epoch is completed. As expected, the Scan approach is significantly worse than the others.

Figure 7 shows the time it takes for each approach to complete an epoch, that is, to visit each of the points of attention. The system proposed shows times of around 5 s when there is only one table active. In the case of two tables, the time only exceeds 15 s once. The Round Robin method takes more than double the time in virtually all epochs. In any case, these results are much better than those of the Scan method.

The last indicator is the energy required in each epoch. As it is difficult to obtain energy measurements, the difference between the current angle and the commanded angle was measured. The smaller the displacement of the head, the lower the energy required to visit a point of attention. **Figure 8** shows that the proposed system is also, by far, the one that preserves the most energy to complete each epoch.

We carried out a second experiment to measure the time it takes for each algorithm to return to an attention point. **Figure 9** shows the distribution of the points of attentions (green dots) and the coordinates transformation from the three tables to the robot using the ROS `tf` visualization tool. The robot is in front of three tables, each one with the same attention points. As in the previous experiment, we consider that the robot deals with a point of attention when it is in the fovea. An algorithm is considered best if it does not allow points to be unattended for a long time.

We have carried out multiple runs of the three algorithms attending one, two and three tables. In each case, we measured the time that a point of interest is unattended. Each trial lasts 2 min, and the results of the experiment are shown in **Table 1**.

The table's analysis reveals that the Round Robin algorithm yields the best times in the mean and the median. Also, more points are served in the 2 min that each trial lasts. The numbers are very similar to the algorithm that optimizes attention, although it deals with fewer points during these 2 min. Still, the maximum time a point has waited to be observed is much longer with the Round Robin algorithm, which is the critical factor that we tried to minimize with our proposal. The scan algorithm, which is the baseline in this work, has the worst statistics, showing that our proposal significantly improves a robot's attention.

In order to illustrate the criticality of the maximum time parameter, it has to be noted that in the competitions, the time that the robot is inactive is very limited. For instance, the rulebook[4] of the RoboCup competition states that 30 s of inactivity disqualifies a robot. In the same way, the maximum time for each trial is also limited.

6. CONCLUSIONS

This paper has presented a visual attention system integrated into a cognitive architecture. This attention system calculates the head movements necessary to perceive the elements of the robot's environment. The cognitive architecture integrates the attention system as a robot skill. Perceptual needs are expressed in a degree of knowledge, adding arcs that indicate this. The attention system is aware of these attention arcs. The way of attending to an element of the environment depends on the objective of this arc.

The experiments carried out show that the system proposed improves conventional systems based on scanning the environment. The robot's gaze always goes to relevant areas without wasting time in areas where the searched items cannot be found. Attention to these areas is always given using the lowest possible energy and stabilizing the image in a position to perform the detection with sharp images.

We have tested this approach on a real Tiago robot. We have proven its validity in one of the tests of the SciRoc competition. To validate our approach, we have implemented two representative attention methods. In this experimentation, we have shown that our approach improves the other methods according to the maximum time, which is the main factor in this problem and has been highlighted in **Table 1**.

Finally, as a further work, we think that the energy consumed by each method should be analyzed, as well as the relevance of the order of the points in the RR method, fixed by the designer in this method.

[4]http://www.robocupathome.org/rules

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

FM and FR-L were the main designers of the software architecture. JG and AG-H were responsible for the experimentation, and VM was the coordinator of the team and contributed in the design of the cognitive architecture. All authors contributed to the article and approved the submitted version.

FUNDING

This work has been partially funded by Ministerio de Ciencia, Innovación y Universidades through grant RTI2018-100683-B-I00.

ACKNOWLEDGMENTS

We want to thank the rest of the members of the Gentlebots team for their work in the competition.

REFERENCES

Agüero, C. E., Martn, F., Rubio, L., and Caas, J. M. (2012). Comparison of smart visual attention mechanisms for humanoid robots. *Int. J. Adv. Rob. Syst.* 9:233. doi: 10.5772/53571

Aliasghari, P., Taheri, A., Meghdari, A., and Maghsoodi, E. (2020). Implementing a gaze control system on a social robot in multi-person interactions. *SN Appl. Sci.* 2, 1–13. doi: 10.1007/s42452-020-2911-0

Bachiller, P., Bustos, P., and Manso, L. J. (2008). "Attentional selection for action in mobile robots," in *Advances in Robotics, Automation and Control*, Chapter 7, eds J. Aramburo and A. R. Trevino (Rijeka: IntechOpen), 111–136.

Breazeal, C., and Scassellati, B. (1999). "A context-dependent attention system for a social robot," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, (San Francisco, CA: Morgan Kaufmann Publishers Inc.), 1146–1153.

Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE J. Rob. Autom.* 2, 14–23. doi: 10.1109/JRA.1986.1087032

Butko, N., and Movellan, J. (2009). "Optimal scanning for faster object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Miami, FL), 2751–2758.

Cashmore, M., Fox, M., Long, D., Magazzeni, D., Ridder, B., Carreraa, A., et al. (2015). "Rosplan: Planning in the robot operating system," in *Proceedings of the Twenty-Fifth International Conference on International Conference on Automated Planning and Scheduling*, ICAPS'15, (Jerusalem: AAAI Press), 333–341.

Garcıa, J. F., Rodrıguez, F. J., Martın, F., and Matellán, V. (2010). "Using visual attention in a nao humanoid to face the robocup any-ball challenge," in *5th Workshop on Humanoids Soccer Robots*, (Nashville, TN), 1–6.

Goferman, S., Zelnik-Manor, L., and Tal, A. (2012). Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1915–1926. doi: 10.1109/TPAMI.2011.272

Grotz, M., Habra, T., Ronsse, R., and Asfour, T. (2017). "Autonomous view selection and gaze stabilization for humanoid robots," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC), 1427–1434.

Harel, J., Koch, C., and Perona, P. (2006). "Graph-based visual saliency," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, NIPS'06 (Cambridge, MA: MIT Press), 545–552.

Hashimoto, S., Narita, S., Kasahara, H., Shirai, K., Kobayashi, T., Takanishi, A., et al. (2002). Humanoid robots in waseda university—hadaly-2 and WABIAN. *Auton. Rob.* 12, 25–38. doi: 10.1023/A:1013202723953

Hou, X., and Zhang, L. (2007). "Saliency detection: a spectral residual approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN), 1–8.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 1254–1259. doi: 10.1109/34.730558

Kanda, T., and Ishiguro, H. (2017). *Human-Robot Interaction in Social Robotics*. CRC Press.

Kawamura, K., Dodd, W., Ratanaswasd, P., and Gutierrez, R. A. (2005). "Development of a robot with a sense of self," in *2005 International Symposium on Computational Intelligence in Robotics and Automation* (San Diego, CA), 211–217.

Manso, L. J., Gutierrez, M. A., Bustos, P., and Bachiller, P. (2018). Integrating planning perception and action for informed object search. *Cogn. Proc.* 19, 285–296. doi: 10.1007/s10339-017-0828-3

Martin, F., Aguero, C., Canas, J. M., and Perdices, E. (2010). "Humanoid soccer player design," in *Robot Soccer*, Chapter 4 (Rijeka: IntechOpen), 1–15.

Martín, F., Agüero, C. E., and Cañas, J. M. (2016). "A simple, efficient, and scalable behavior-based architecture for robotic applications," in *Robot 2015: Second Iberian Robotics Conference*, eds L. P. Reis, A. P. Moreira, P. U. Lima, L. Montano and V. Muñoz-Martinez (Cham: Springer International Publishing), 611–622.

Martin, F., Rodrguez Lera, F., Gins, J., and Matellin, V. (2020). Evolution of a cognitive architecture for social robots: integrating behaviors and symbolic knowledge. *Appl. Sci.* 10:6067. doi: 10.3390/app10176067

Meger, D., Forssén, P.-E., Lai, K., Helmer, S., McCann, S., Southey, T., et al. (2008). Curious george: an attentive semantic robot. *Rob. Auton. Syst.* 56, 503–511. doi: 10.1016/j.robot.2008.03.008

Nguyen, T., Zhao, Q., and Yan, S. (2018). Attentive systems: a survey. *Int. J. Comput. Vis.* 126, 86–110. doi: 10.1007/s11263-017-1042-6

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *2008 IEEE International Conference on Robotics and Automation* (Pasadena, CA), 962–967.

S tefanov, K., Salvi, G., Kontogiorgos, D., Kjellström, H., and Beskow, J. (2019). Modeling of human visual attention in multiparty open-world dialogues. *ACM Trans. Hum. Robot Interact.* 8, 8:1–8:21. doi: 10.1145/3323231

Scheier, C.. S. E. (1997). "Visual attention in a mobile robot," in *ISIE '97 Proceeding of the IEEE International Symposium on Industrial Electronics*, vol.1, SS48–SS52.

Treue, S. (2003). Visual attention: the where, what, how and why of saliency. *Curr. Opin. Neurobiol.* 13, 428–432. doi: 10.1016/S0959-4388(03)00105-3

Wolfe, J. M. (1994). Guided search 2.0 a revised model of visual search. *Psychon. Bull. Rev.* 1, 202–238. doi: 10.3758/BF03200774

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 23 November 2021
doi: 10.3389/fnbot.2021.703545

Egocentric Gesture Recognition Using 3D Convolutional Neural Networks for the Spatiotemporal Adaptation of Collaborative Robots

*Dimitris Papanagiotou, Gavriela Senteri and Sotiris Manitsaris**

Centre for Robotics, MINES ParisTech, PSL Université, Paris, France

Collaborative robots are currently deployed in professional environments, in collaboration with professional human operators, helping to strike the right balance between mechanization and manual intervention in manufacturing processes required by Industry 4.0. In this paper, the contribution of gesture recognition and pose estimation to the smooth introduction of cobots into an industrial assembly line is described, with a view to performing actions in parallel with the human operators and enabling interaction between them. The proposed active vision system uses two RGB-D cameras that record different points of view of gestures and poses of the operator, to build an external perception layer for the robot that facilitates spatiotemporal adaptation, in accordance with the human's behavior. The use-case of this work is concerned with LCD TV assembly of an appliance manufacturer, comprising of two parts. The first part of the above-mentioned operation is assigned to a robot, strengthening the assembly line. The second part is assigned to a human operator. Gesture recognition, pose estimation, physical interaction, and sonic notification, create a multimodal human-robot interaction system. Five experiments are performed, to test if gesture recognition and pose estimation can reduce the cycle time and range of motion of the operator, respectively. Physical interaction is achieved using the force sensor of the cobot. Pose estimation through a skeleton-tracking algorithm provides the cobot with human pose information and makes it spatially adjustable. Sonic notification is added for the case of unexpected incidents. A real-time gesture recognition module is implemented through a Deep Learning architecture consisting of Convolutional layers, trained in an egocentric view and reducing the cycle time of the routine by almost 20%. This constitutes an added value in this work, as it affords the potential of recognizing gestures independently of the anthropometric characteristics and the background. Common metrics derived from the literature are used for the evaluation of the proposed system. The percentage of spatial adaptation of the cobot is proposed as a new KPI for a collaborative system and the opinion of the human operator is measured through a questionnaire that concerns the various affective states of the operator during the collaboration.

Keywords: human-robot collaboration, gestures, actions, recognition, CNN, egocentric vision, collaborative robot, pose estimation

OPEN ACCESS

Edited by:
Dimitri Ognibene,
University of Milano-Bicocca, Italy

Reviewed by:
Guido Schillaci,
Joint Research Centre (JRC), Italy
Francesco Rea,
Italian Institute of Technology (IIT), Italy
Rukiye Kiziltepe,
University of Essex, United Kingdom

****Correspondence:***
Sotiris Manitsaris
sotiris.manitsaris@mines-paristech.fr

Received: *30 April 2021*
Accepted: *19 October 2021*
Published: *23 November 2021*

Citation:
Papanagiotou D, Senteri G and Manitsaris S (2021) Egocentric Gesture Recognition Using 3D Convolutional Neural Networks for the Spatiotemporal Adaptation of Collaborative Robots. Front. Neurorobot. 15:703545. doi: 10.3389/fnbot.2021.703545

1. INTRODUCTION

Robots were first introduced to industrial environments in the mid-1950s and consequent advancements in the areas of perception of humans and of the environment, during the last few decades, have led to the evolution of a new area of research, named Human-Robot Interaction (HRI). The International Federation of Robotics (IFR)[1] reports a record of 2.7 million industrial robots operating in factories around the world, which indicates an overall increase of 12% for the year 2020 alone.

Until quite recently, conventional automation of Industry 3.0 has been trying to insert more and more robots into production processes to perform repetitive and hazardous tasks which have traditionally been performed by humans. The translation to Industry 4.0, using means such as cyber-physical systems (CPS), cloud computing and Industrial Internet of Things (IIoT)s, aims to insert human-robot collaboration (HRC) frameworks into the manufacturing process. There are different categories of HRI, depending on the workspace, the aims, the working times of the robot and the operator.

The current work aims at the development of a Human-centered Artificial Intelligence perception layer of a robot, which is inserted in an industrial HRC scenario. Active vision through gesture recognition and pose estimation enables the spatiotemporal adaptation of the robot to each user. We focus on the insertion of a smaller, lightweight robot which facilitates HRC, without the need for physically separated workspaces. Different types of interaction are implemented and ultimately the goal of this paper is to evaluate their impact on both human-robot collaboration and user experience. On the way to safer and more effective HRC scenarios, touchless interaction is implemented.

Egocentric computer vision for action/gesture recognition unleashes great potential for touchless HRI. The proposed human egocentric system constitutes an initial step in active vision. It is not affected by some critical issues for active vision as the camera is unique, on the top of the operator, and moves according to operator's head motion. There is no change or motion of the camera for better field of vision. In addition, occlusion or limited resolution are improbable as the operator's actions are executed in front of her/his body. These actions are communicated to the robot so as to dynamically adapt its behavior. Therefore, both the temporal and spatial profile of the motion of the robot depend on the rhythm and the pose of the human operator, respectively. From an industrial point of view, the production cycle time becomes adaptable.

This paper presents a gesture recognition module based on 3D Convolutional Neural Networks (3DCNNs), trained on an egocentric view, for a natural collaboration between the human and the robot. Deep Learning (DL) is a field of Machine Learning (ML) with impressive results in pattern detection, speech recognition and many more applications and can provide the necessary robustness that an HRI scenario requires. The two hypotheses that are tested, are the reduction of cycle time of the assembly routine through the insertion of gesture recognition, as well as the improvement of the handover position via the implementation of pose estimation.

This paper consists of nine sections in total. Following the Introduction in section 1, section 2 presents human-robot interaction that can be achieved either physically or by touchless means. In section 3, the routine which was implemented, together with the experiments which are used to evaluate the contribution of the proposed modules (gesture recognition, pose estimation, sound notification) are described. In sections 4 and 5, the modules and their respective implementation methodologies are presented. section 6 describes the way that the robot performs and presents a variety of metrics that are commonly used for the evaluation of an HRI system. In section 7, each type of collaboration is described and evaluated, while, in section 8, future work perspectives emerging from the various types of collaboration are examined, with areas for future research suggested in Conclusion, in section 9.

2. STATE OF THE ART

Since the initial establishment of robots in industry, the aim has been to assist humans in their heavy-duty tasks, and to keep everyone safe at the same time. The limitations of robots, in this early period, in conjunction with the ever-increasing levels of safety which have had to be observed in industry, have served to create a somewhat primitive workplace for industrial robots. Traditionally, they have been installed in assembly-lines and have been assigned to undertaking the tasks which are repetitive, heavy-duty and dangerous for human operators, as described by Hentout et al. (2019). Regardless of their efficiency and velocity, the assembly-lines that use this type of robot have been lacking in flexibility, especially when the presence of a human operator is required.

Humans, on the other hand, have the flexibility and the intelligence to consider different approaches to solve a problem and can choose the best option from among a range of possibilities. They can then command robots to perform assigned tasks, since robots can be more precise and more consistent in performing repetitive and dangerous work. This collaboration between industrial robots and humans demonstrates that robots have the capabilities to ensure maximum efficiency in manufacturing and assembly; however, the evolution of technology, together with the ongoing automation of traditional manufacturing and industrial practices, has shown that there are many tasks which are too complex to be fully executed by robots, or are too financially burdensome to be fully automated.

This is the reason why the research agenda in the past few years has focused on creating appropriate industrial working environments, where robots and human operators can interact effectively. Nowadays, mixed environments are being created and industries aim to explore and create the ideal working environment through combining the cognitive skills of the human operators (intelligence, flexibility, and ability to act when confronted with unexpected events) with the ergonomic advantages of the robots (high precision, repeatability, and strength) (Prati et al., 2021).

[1] https://ifr.org/

The creation of mixed industrial environments, where humans and robots co-exist and work for a common goal reinforces the necessity of the insertion of cobots in manufacturing process. IFR in accordance with ISO 8373 describes two different types of robots (industrial and service). Cobots could be considered to be service robots, since they are intended to work alongside humans; however, there are different definitions of cobots, depending on the applications they are used for. In the beginning, a cobot was defined as "an apparatus and method for direct physical interaction between a person and a general purpose manipulator controlled by a computer" (Bicchi et al., 2008); however, due to the development of the sensors that cobots use and because of the way they interact with humans, this definition has evolved.

Active vision is mentioned as the capability of a robot to actively perceive the environment and obtain useful information for various tasks (Chen et al., 2011). It is used in plenty of use-cases such as collaborative robotics (Queralta et al., 2020) and industrial applications (Muhammad et al., 2017). The workflow of a typical active vision or perception system, includes view planning, motion planning, sensor scanning and map updating (Zeng et al., 2020). After each stage information is collected and update the status of the robot and its task goal.

In recent research, a cobot is referred to as a robot that has been designed and built to collaborate with humans (Schmidtler et al., 2015), or as a robot intended to physically interact with humans in a shared workspace (Colgate and Peshkin, 2010). For this reason, the discussion has shifted to Human-Robot Interaction and the way this interaction is achieved in each application.

2.1. Categories of Human Robot Interaction

HRI research has attracted the attention of numerous research domains. For this reason, HRI can be classified into many categories depending on the criteria that are used. Kopp et al. (2021) and El Zaatari et al. (2019) distinguish HRI as functioning on different levels, according to the workspace (separated or common), the working time/steps (sequential or simultaneous) and the aims (different or common) of the robot and the human operator respectively. At the lowest level, human and robot work alongside each other without a shared workspace (Long et al., 2018). They have neither common tasks, nor actions, nor intentions. Traditional industrial robots are used extensively in such cases. At the second level, however, the human and the robot share all or part of a workspace, they do not work on a part or on a machine at the same time. Unhelkar et al. (2020) name this type of collaboration as sequential, which implies that the human operator adapts to the rhythm and the orientation of the robot, since its velocity and its trajectories are pre-defined.

In a few industries, in recent years, humans and robots have been working on the same part or machine at the same time, and both are in motion (Cherubini et al., 2016). This level of interaction is called human-robot co-operation and requires advanced sensors and technology, like force/torque sensors or computer vision. Despite the sharing of workspace and aim, the human operator must adapt to the pre-defined temporal and spatial profile of the robot. That makes this type of interaction less natural than interaction between humans and, because of this, different types of communication and collaboration are established within the framework of Industry 4.0. Finally, at the upper level, the robot responds in real-time to the worker's motion which is called responsive HRC. The combination of artificial intelligence and high-tech sensors make robots able to adapt their rhythm and motion to unpredictable incidents and the anthropometric characteristics of the operator. The purpose of this category is the transformation of the robot, from being more than just a useful machine, to being a real collaborator.

Responsive Human-Robot Collaboration can be classified into physical (pHRC, Ajoudani et al., 2018) and touchless (tHRC, Khatib et al., 2017). pHRC can be divided into two different categories, depending on the intended purpose of the touching. On the one hand, there are operations which were intended to be without contact, but where instinctively the operator touches the robot. On the other hand, there are operations where the operator presses or touches the robot on purpose and the robot reacts in a particular way, depending on the amount and the direction of the operator's force. In the first case, the robot should perceive the presence and the velocity of the human operator inside its workspace and react correspondingly, either by reducing its velocity or protectively stopping its motion in order to avoid a collision, as noted by Michalos et al. (2015). In the second case, Bo et al. (2016) note that the robot can either be used as a tool which extends the capabilities of the human operator (strength, precision etc.), or can be taught by demonstration in order to be able to repeat a certain task precisely.

Long before the outbreak of Covid-19, which has necessitated social distancing, industries were using technologies that minimize the need for physical interaction among industrial workers, enabling device operation at a safe distance[2]. Contactless technology is a branch of control technology, which has as its aim the establishing of communication between computers/machines and human operators, without the need for any contact whatsoever. It relies on the interpretation of human movement and behavior, using ML algorithms and sensors, namely RGB-D cameras, thermal cameras or Inertia Measurement Units (IMUs, Zhang et al., 2020). The sensors and algorithms provide the machines/cobots with commands or instructions derived from the detection of facial patterns (Melinte and Vladareanu, 2020), voice translation (Gustavsson et al., 2017) and gesture recognition (El Makrini et al., 2017).

Contactless technology allows users to control digital or industrial systems, using their anthropometric characteristics or motion. It has gained a lot of attention in the gaming and medical worlds, as well as in other fields, such as the automotive and cultural industries. Human action recognition is one of the tools used to achieve contactless communication between a computer/machine and a human operator, and can be defined as the conversion of a human/humanoid movement or signal to a form understandable to a machine. Action recognition

[2]https://www.intel.com/content/dam/www/public/us/en/documents/pdf/the-need-for-enabling-touchless-technologies-whitepaper.pdf

enables the human operator to interrelate with a machine in an environment characterized by the absence of means for physical interaction.

2.2. Movement-Based Implicit and Explicit Interaction

With a view to more natural HRC, the adaptation of robots, in accordance with the temporal and spatial profile of the human operator, has evolved into a very meaningful research topic. Humans can be involved, beyond their traditional offline role, as they can now interact with a cobot either explicitly or implicitly (El Zaatari et al., 2019). Explicit interaction, on the one hand, is what is referred to as direct communication between the robot and the human. Implicit interaction, on the other, involves an action (or practical behavior), which represents a message in itself, rather than a message being conveyed through language, codified gestures (such as a thumbs-up or nod of the head) (Gildert et al., 2018) or other non-verbal, sensorimotor forms of communication to send coordination signals (Pezzulo et al., 2013; Vesper and Sevdalis, 2020).

Temporal adaptation can be achieved either explicitly or implicitly. There is research where explicit interaction for temporal adaptation is achieved through the use of a button[3] or a smartwatch (Michalos et al., 2018), thanks to which the operator can inform the robot that he/she has executed a task. However, if judged according to the previously-given definitions of HRI, this case matches more with human-robot co-operation, as the insertion of a button makes the interaction less natural. In the research of Cherubini et al. (2016), force feedback and pointing gestures are introduced as a means of HRC, in order to adapt the temporal profile of the robot and create hybrid interaction. In the present case, a totally implicit interaction is presented from our previous research (Coupet et al., 2019), which uses gesture recognition as a means to inform the robot about the percentage rate of completion of the human gesture, in order for it to react correspondingly. Such implicit interaction scenarios are also implemented outside of industrial workspaces, as described by Gabler et al. (2017) and Vogt et al. (2017).

The spatial adaptation of a robot to an industrial environment is commonly presented as collision avoidance between the robot and the human operator who share the same workspace (Mohammed et al., 2017; Safeea et al., 2019). Apart from their applications in industry, such adaptations are reported in other research, such as that of Canal et al. (2018), where the creation of a daily living assistant is presented. This research describes a cobot that is able to readjust its trajectories according to user movements and can thus handle incidents which are unpredictable. In the context of the present article, a spatiotemporal adaptation of a cobot, working according to the desired handover positions and rhythm of a human operator, is described. The goal of this research is to improve the perception of robots, using professional gesture recognition in cooperation with ergonomic parameters, with a view to creating a better and more natural HRC.

[3]http://roboticsandautomationnews.com/2017/03/04/bmw-shows-off-its-smart-factory-technologies-at-its-plants-worldwide/11696/

2.3. Machine Learning for Professional Gesture Recognition

A significant amount of scientific work aims at making machines smarter, improving their perception, enabling them to interpret human behavior, and to learn and react in a way similar to the human brain. In order to achieve these goals, solid results in the field of activity and, more specifically, in the field of gesture recognition are necessary, since this will permit more natural Human-Robot Collaboration (HRC). Indeed, an essential goal of the research community is the development of algorithms that can accurately recognize and understand human actions. Research on human action recognition focuses mainly on two strategies; namely, Pose- (Skeleton-) based recognition and Appearance-based recognition methods.

2.3.1. Pose-Based Methods for Gesture Recognition

The main goal of pose-based methods is gesture recognition through the extraction of feature vectors which provide input to the corresponding ML algorithm. Essentially, those features are a set of coordinates able to describe the pose of a person and give explicit details about their position within a space. Pose estimation is usually performed using RGB-D cameras, such as the Kinect camera[4], or optical hand tracking sensors such as the LeapMotion sensor[5], algorithms, and modules such as Openpose (Cao et al., 2021), Alphapose (Fang et al., 2017), and Densepose (Güler et al., 2018), that use Deep Learning architectures themselves, performing either 2D or 3D pose estimation for both offline and online purposes, for the extraction of body joints. In general, recovering 3D pose from RGB images is considered more difficult than 2D pose estimation, due to the larger 3D pose space and other ambiguities. A number of factors can cause these ambiguities, such as body occlusions (Cheng et al., 2019), skin color, clothing, an overloaded background or quality of lighting (Rahmat et al., 2019).

Stochastic methods, such as Hidden Markov models (HMMs) (Borghi et al., 2016; Bui et al., 2018) and Random regression forests (Canavan et al., 2017), as well as DL methods, such as Recurrent Neural Networks (RNNs) (Shahroudy et al., 2016; Chalasani et al., 2018) have been used in various implementations for gesture classification. In the works cited, the aforementioned ML methods were used for the temporal correlation of the body features, leading to satisfactory classification results. Yan et al. (2018), in an attempt to create an algorithm that automatically learns both the spatial and temporal patterns from data, leading to a stronger generalization capability of the algorithm, propose a novel model of dynamic skeletons called Spatial-Temporal Graph Convolution Networks. Even though satisfactory results can be achieved, extracting features from data can lead to the loss of important information. The estimation of the human joints, and thus the skeletization of the whole body, must not only be absolutely accurate, but must be able to anticipate estimation problems caused by any of the factors mentioned previously (i.e., lighting, occlusions etc.). Thus, what constitute the challenges in these methods is not only the way that classification is

[4]https://en.wikipedia.org/wiki/Kinect
[5]https://developer.leapmotion.com/

performed, but also the way in which accurate pose-estimation is to be accomplished.

2.3.2. Appearance-Based Methods for Gesture Recognition

In contrast with Pose-based recognition methods, Appearance-based ones consider visual cues (i.e., color and edges), to reach a gesture recognition result. Action recognition with these kinds of methods, can achieve results end-to-end, through mostly using sensors that extract visual information, such as RGB-D or thermal cameras. The end-to-end results are obtained by the hierarchical analysis of the characteristics of the visual input (edges, lines etc.) and algorithms, such as 3D CNNs (Tran et al., 2015), two stream fusion networks (Feichtenhofer et al., 2016) and inflated 3D convolution (I3D) (Carreira and Zisserman, 2017). One could say that the two-stream (RGB and optical flow) I3D models, based on 2D ConvNet Inflation, were a breakthrough in this field, as such models made it possible to learn seamless spatiotemporal feature extractors from videos, while leveraging successful ImageNet architecture designs and even their parameters.

There are many cases where the two categories of Pose-based recognition and Appearance-based recognition methods have been combined. Song et al. (2016) propose a multi-modal, multi-stream DL framework for egocentric activity recognition, using video and sensor data. They extend a multi-stream CNN to learn spatial and temporal features from egocentric videos, by using a multi-stream LSTM architecture to learn the features from multiple sensor streams (accelerometer, gyroscope etc.). Cao et al. (2017) perform egocentric gesture recognition, combining traditional CNN architectures with spatiotemporal transformer modules in order to address problems that arise from the global camera motion, caused by the spontaneous head movement of the device wearer. More specifically, a spatiotemporal transformer module (STTM) is proposed, that is able to transform 3D feature maps to a canonical view in both spatial and temporal dimensions. The challenge of capturing and recognizing images, from an egocentric view, lies in the fact that we can identify two parallel movements, that of the background and of the person themselves, and that of the camera that follows the motion of the head, with the motion of the head not always aligned to the motion of the rest of the body.

2.3.3. Human-Robot Collaboration With Artificial Intelligence

Cobots are becoming ever more present in industrial environments, as an automated solution, enabling industrial workspaces to become more cost-effective, flexible, and ergonomic. For this to be accomplished successfully, cobots need to be equipped with tools that will make them adjust to the workspace and help the industrial operator, without creating an extra burden during the work process. These tools include ML algorithms, such as Markov chains or HMMs, and DL architectures, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) and deep Reinforcement Learning (RL) for gesture recognition, voice detection, working environment surveillance, to mention only a few.

Machine learning architectures such as Markov chains or HMMs are known for their applications in signal processing and pattern detection. They are used to estimate the probability of going from one state of a system to another and, therefore, lead to data classification. The limitation of these methods is connected to the fact that inserting images as input to be classified according to their state probabilities, demands preprocessing. This preprocessing concerns the extraction of features for the creation of vectors that will constitute the required input.

Many research projects (Liu and Hao, 2019; Sharkawy et al., 2019; Sharkawy et al., 2020) have used such approaches to detect a collision based on robot sensor stream data, or perform continuous gesture recognition (Tao and Liu, 2013). In order to enable a smooth Human-Robot collaboration, where the robot is able to synchronize, adapt its speed and detect any unexpected incident, Coupeté (2016) implements gesture recognition of professional gestures in an automotive assembly-line using Discrete HMMs and inertia sensors to finetune the results. Dröder et al. (2018) use an ML-enhanced robot control strategy, combining also a nearest neighbor approach, for obstacle detection in an HRC scenario. All of the cases mentioned above, require either the use of specific sensors that provide with-time-series, or involve time-consuming pre-processing, as previously discussed.

On the other hand, DL architectures, such as CNNs and RNNs, are widely used nowadays in finance, robotics and medicine. Such methods require a large amount of data in order to be trained properly and, in most cases, require a great deal of computational power and time. However, in some DL methods, such as CNNs, preprocessing is not necessary, ensuring there is no loss of information.

El-Shamouty et al. (2020), in trying to minimize the risk of accidents in HRC scenarios, propose a deep RL framework that encodes all the task and safety requirements of the scenario into RL settings, and also takes into account components such as the behavior of the human operator. Liu and Hao (2019) work on a scenario of multimodal CNNs and use a Leap Motion sensor for hand motion detection, as well as voice and body posture recognition. Amin et al. (2020) aim to upgrade safety and security in an HRC scenario, by using a combination of human action recognition and tactile perception in order to distinguish between intentional and incidental interactions if physical contact between human operators and cobots takes place. A Panda robot, along with a 3D-CNN for human action recognition and a 1D-CNN for contact-type detection, was deployed.

Most of the methods presented above are focused on specific factors (safety, accident prevention, fast response from a cobot in an HRC laboratory-implemented scenario), without considering all the limitations, as well as the spatiotemporal variations that might occur in a real-life scenario. Different users of the same set-up have different anthropometric characteristics and different behaviors when asked to perform the same action. The aim in the present work, however, is also to examine the contribution of an egocentric gesture recognition module with a Deep Learning architecture in an HRC industrial scenario.

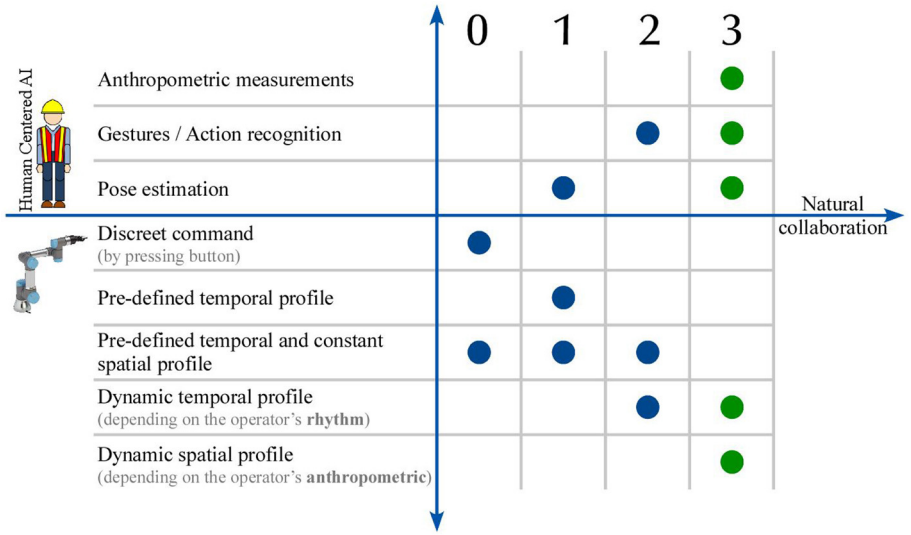


FIGURE 1 | Evolution of Human-Robot Interaction (HRI) toward natural collaboration through human-centered Artificial Intelligence.

Figure 1 illustrates the potential of Human-centered AI in contributing toward a more natural HRC. The more anthropocentric the information that is extracted, the richer the perception of the robot is. The more its perception is enriched, the more it can predict human actions. In order for the robot to collaborate with the human, it has to understand not only its tasks but also human actions and intentions. At the beginning (level 0), the introduction of traditional industrial robots is the baseline, and the most common case in industry currently. There is no interaction between the robot and the operator and the robot completes a task very quickly and precisely. The first step toward interaction (level 1) was achieved by giving the robot information about the human's presence inside its workspace. Both spatial and temporal profiles remain constant and predefined, but when it perceives that an operator is inside its workspace, it reacts either by protectively stopping its motion or by reducing its velocity. Moreover, the human action and gesture recognition (level 2), converts the temporal profile to dynamic, adapting to the operator's rhythm. In the present research, the development of a dynamic spatiotemporal HRC framework is presented, receiving the human's actions and poses as input parameters (level 3).

3. PILOT SCENARIO

The use-case that was used for this research was derived from industry and, in particular, from Arçelik's TV assembly factory. In the actual assembly line, the task is executed manually by two different operators. The first operator has the task of picking up the electronic cards and placing them on the TV panel. The electronic cards are divided into two different types: the power supply (PSU) and the main board (chassis), that are located in two different boxes next to the conveyor belt. The second operator is responsible for the screwing of the cards onto the TV. The insertion of a temporal and spatial adjustable cobot, which can perform the first part of the operation, is proposed.

Factories in the Industry 4.0 era need the high efficiency and repeatability of the robots, together with the flexibility and variety of products that a human operator can provide. The parallel operation of a cobot and an operator on an assembly line was examined. The experiments were as follows:

1. Physical Interaction
2. Physical Interaction and Spatial adaptation (Operator's pose estimation)
3. Physical Interaction, Spatial adaptation (Operator's pose estimation), and Sound notification
4. Physical Interaction and Gesture recognition
5. Combination of spatiotemporal adaptation and sound feedback

Initially, the operator interacts with the robot only physically (pHRC). This is accomplished through a Force Sensor (FS) which is placed on the robotic arm, just above the end-effector (Gripper). Every time the operator finishes with a task, s/he presses the FS in order to inform the robot and make it advance to the next position. The operator presses the FS to start the routine. When the robot grasps the card, it brings the card to a particular position. Then, the operator presses the FS again to release the card and the robot advances to a waiting position. The operator decides if the card is functional or not and presses the FS accordingly. If the FS is pressed on the horizontal axis, the card is not functional and the robotic arm returns to take the next card from the same box. If otherwise, and the card is functional, the operator presses the FS at the vertical axis, as always, and the robotic arm continues and grasps a card from the second box. When the operator takes the first card, s/he places it on the TV board and s/he screws it in place. The same procedure is followed also for the second card and when it is well-positioned on the TV,

the operator presses the FS to inform the robot that the routine is finished.

The physical interaction that has just been described is complemented with a pose estimation module, during the second experiment. The robotic arm does not place the cards in a particular position, as previously, but instead is spatially adapted to the anthropometric characteristics of each operator. This procedure improves the pose of the operator ergonomically. The skeleton of the operator is extracted, and the position and velocity of each wrist is recorded. When the human's hand is motionless and in a position which is reachable for the robot, the robotic arm records it and approaches it holding the card. A natural HRC also demands the exchange of information. Thus, for the third experiment, a sonic notification is inserted. This notification is activated when the operator asks for the card at a position that is not reachable for the robot.

Gesture recognition is implemented in the fourth experiment. Physical interaction is used only for the release of the card. Each card is delivered to a particular position, with the aim of evaluating the added value of the gesture recognition module in the HRC scenario. The camera that records the operator's gestures is placed on the operator's helmet, thus offering an egocentric view. The final experiment brings all of the 4 modules together. Physical interaction is used for the release of the card, pose estimation for the spatial adaptation of the robotic arm, with sound notification and gesture recognition used in the ways previously referred to. The aim of the final experiment is the evaluation of all the modalities together, in order to see what the positive contribution is for the human operator.

Through the execution of the aforementioned five experiments, this research aims to evaluate the dynamic temporal profile that is achieved through the implementation of gesture recognition and the dynamic spatial profile that is achieved through the implementation of pose estimation. In addition, every experiment is executed twice, in order to indicate the compliance of the robot to unpredicted incidents (actions not corresponding with the work sequence). In **Figure 2**, the architecture of the system is presented. Physical interaction demands that the operator stop his/her task in order to inform the robotic arm about the work sequence. The cycle time is therefore expected to increase. The insertion of pose estimation is expected to improve the handover position of the card for each user; however, it will necessarily increase the cycle time because of the path calculation for the position of the operator's hand. Sonic notification is supposed to decrease the average cycle time, as each operator knows where to place the hand in order to ask for the card. Adding gesture recognition will reduce the cycle time as the operator interacts more implicitly with the robotic arm in comparison with other types of interaction. Thus, it is expected that experiments that contain gesture recognition will improve the naturalness of the HRC scenario, with users' responses gathered via questionnaires. The hypotheses that are extracted from the above expectations are the following:

H1: Can gesture recognition facilitate the temporal adaptation of the robot for reducing the cycle time in assembly lines?

H2: Can human pose estimation facilitate the spatial adaptation of the robot for reducing the range of motion of the operator and improving the handover position?

4. POSE ESTIMATION OF HANDOVER POSITION FOR ROBOTIC ARM

During the execution of the experiments "*Pose Estimation*", "*Sonic Notification*", and "*Combination*", pose estimation is used as a mean of interaction between human and robot. The OpenPose[6] framework is used for the skeleton extraction of the operator. This framework detects the body key points on RGB images and concludes with the extraction of 2D positions for each body joint, using DL architectures. Pose estimation, in the context of these experiments, was used both to estimate the position of the operator's right hand and to calculate its velocity. The coordinates of the right wrist, as extracted from the framework, are used. The camera that is used for the pose estimation is placed parallel to the operator, next to the conveyor belt. The framework extracts the position of the wrist in the image frame counted in pixels (X,Y) and an estimation of the distance on the Z axis is counted in meters. The procedure of providing the robot with the coordinates of the operator's wrist consists of two steps:

1. The first step is the conversion of the camera pixels to meters. Initially, the Intel-RealSense RGB-D camera is positioned so that the X and Y axes of the camera are parallel to the X and Z axes of the robot, accordingly. Using the parameters of the RGB-D camera that was used (focal length, principal point and distortion coefficients) it was possible to convert pixels to meters for each different depth value. The equations that were used are the following:

$$x = \frac{(X - c_x) * z}{f_x} \qquad y = \frac{(Y - c_y) * z}{f_y} \tag{1}$$

   Where c_x, c_y is the central - principal point of the camera (956, 538) and f_x, f_y is the focal point of the camera (973, 973). The camera that was used has no distortion coefficient.
2. As the position of the operator's wrist was defined in meters for the coordinate system (CS) of the camera (X_C, Y_C, Z_C), the second step was the transformation of this CS to the robot CS. For this transformation, the homogeneous transformation matrix was used. For the X axis there is only transfer for d_1, for the Y axis there is rotation of 90° and transfer for d_2, and for the Z axis there is rotation of 90° and transfer for d_3. Using the direction cosines of the initial point of camera CS to robot CS, the homogeneous transformation matrix is calculated and

[6]https://github.com/CMU-Perceptual-Computing-Lab/openpose

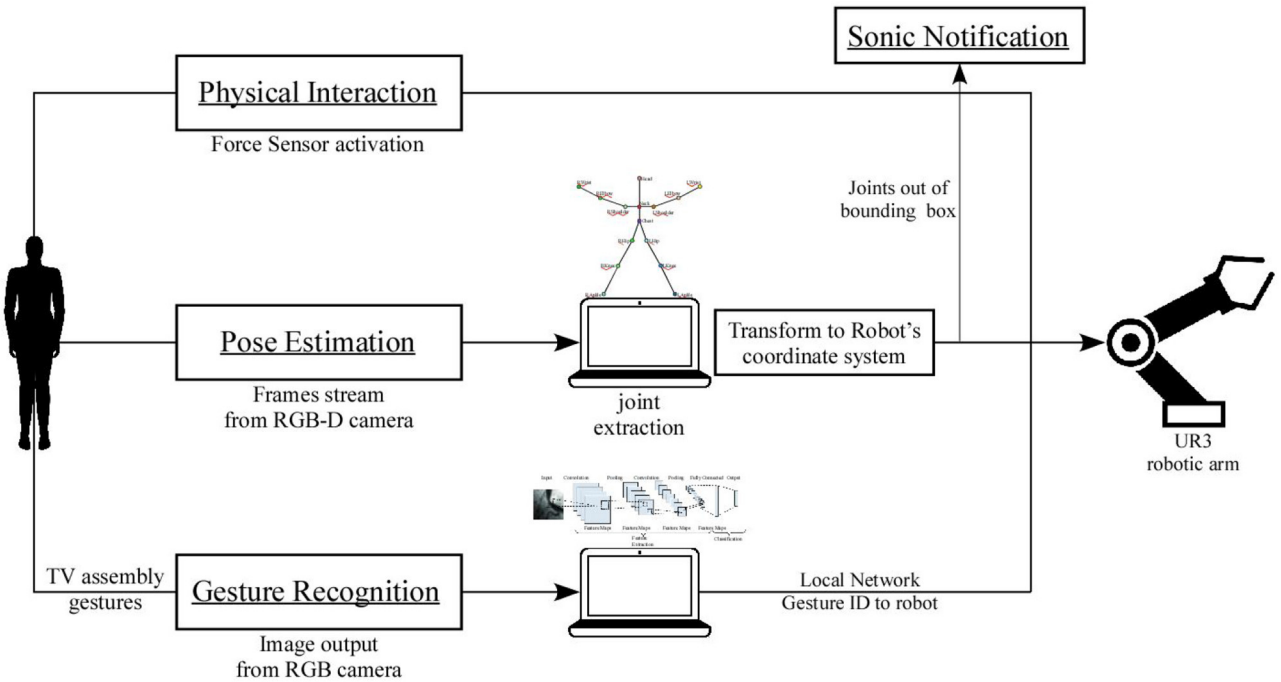


FIGURE 2 | Architecture of the system.

presented in the following equation:

$$\begin{bmatrix} X_R \\ Y_R \\ Z_R \\ 0 \end{bmatrix} = \begin{bmatrix} \cos(X_C, X_R) & cos(Y_C, X_R) & cos(Z_C, X_R) & d_1 \\ \cos(X_C, Y_R) & cos(Y_C, Y_R) & cos(Z_C, Y_R) & d_2 \\ \cos(X_C, Z_R) & cos(Y_C, Z_R) & cos(Z_C, Z_R) & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix}$$
$$= \begin{bmatrix} 1 & 0 & 0 & d_1 \\ 0 & 0 & -1 & d_2 \\ 0 & -1 & 0 & d_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} X_C \\ Y_C \\ Z_C \\ 1 \end{bmatrix} (2)$$

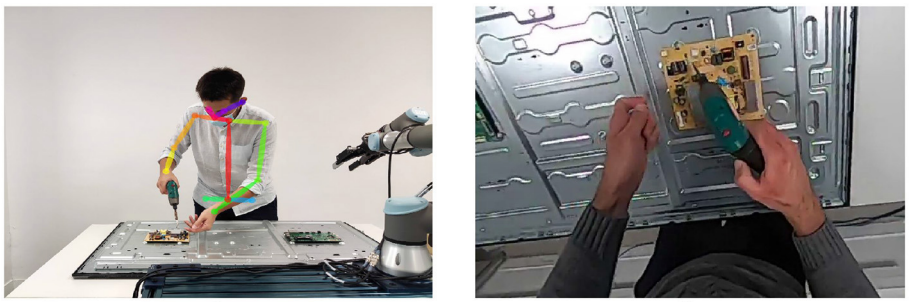
Figure 3 shows the experimental setup during the execution of the experiments. **Figure 3i** presents the view of the camera that is used for the pose estimation. The skeleton that is executed through OpenPose during the 2nd, 3rd, and 5th experiment is demonstrated. In the meantime in **Figure 3ii** the egocentric view that is used for gesture recognition during the 4th and 5th experiment is presented.

5. EGOCENTRIC GESTURE RECOGNITION USING 3DCNNS

For the temporal adaptation of the cobot to the behavior of the human operator, a gesture recognition module was used in the experiments "*Gesture Recognition*" and "*Combination*", which are described in detail below. Briefly, the gestures and postures of different human operators, during the TV assembly routine in an assembly line, were captured with a GoPro RGB camera, segmented and used for the training of a Deep Neural network with Convolutional Layers. The goal of this module was the exploration of the contribution of gesture recognition to an HRC professional scenario. The initial step for this module was the creation of a collaboration protocol between the human operator and the cobot. The parts of the use-case described that included decision-making, were assigned to the human operator, and those that did not, were assigned to the cobot. The gesture recognition results were sent as IDs to the cobot, which interpreted them and acted according to the defined protocol.

5.1. Network Architecture

The DL method used for egocentric gesture recognition in this work was 3D Convolutional Neural Networks (3DCNNs). 3DCNNs are the 3D equivalent of 2DCNNs, taking as input a 3D volume or a sequence of 2D frames. Image sequences with a size of c×l×h×w were used, where c was the number of channels, l was length in number of frames, h and w were the height and width of the frame, respectively. We also refer to 3D convolution and pooling kernel size by d × k × k, where d was kernel temporal depth and k was kernel spatial size. All image frames were resized to 84 × 48, so the input dimensions were finally 5 × 84 × 48 × 3. The network used had 6 convolution layers and 3 pooling layers, 1 fully-connected layer and a softmax loss layer to predict action labels. The number of filters for 4 convolution layers from 1 to 6 were 32, 32, 64, 64, 64, and 64, respectively. All convolution kernels had a size of 3 × 3 × 3, where $d = 3$, $k = 3$. All of these convolution layers were applied with appropriate padding (both spatial and



i) ii)

FIGURE 3 | (i) View of the experimental setup from the camera that is used for pose estimation. **(ii)** Egocentric view of the experimental setup from the camera that is placed on the head of the user.

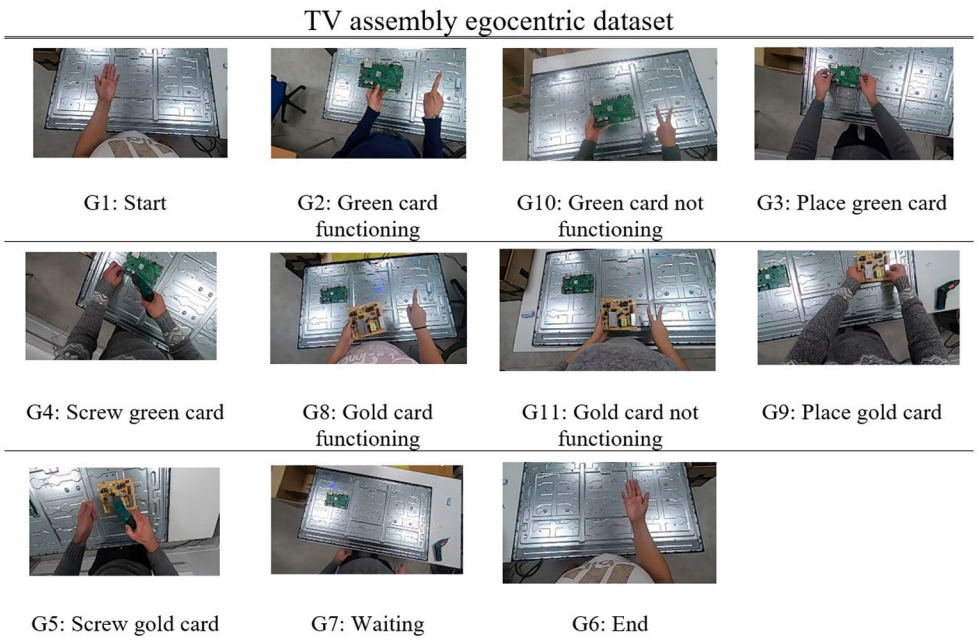


FIGURE 4 | Presentation of the TV assembly dataset, consisting of 11 classes in total, 6 gestures, and 5 postures.

temporal) and stride 1; thus, there was no change in terms of the size from the input to the output of these convolution layers. The standard ReLu activation function was used. All pooling layers were max pooling, with kernel size 3 × 3 × 3. The fully-connected layers had 512 outputs and dropout was not used. The output of the network was a softmax with 11 nodes, like the number of the gesture and pose classes. This network proved to be the most effective, in terms of recognition accuracy, after many experiments with network parameters and layers were performed.

5.2. Industrial Dataset and Gestural Vocabulary

The performance of 3DCNNs was evaluated by recording an egocentric dataset, inspired by an industrial TV assembly scenario. The main routine in assembling a TV is separated into sub-tasks, performed by either a human operator or a robotic arm. The objects that are involved in the assembly routine are a TV frame and two TV cards, one green and one gold. The human operator only performs gestures to interact with the robotic arm, while physical interaction (activation of the force torque sensor) is used only for the TV cards to be released by the gripper of the cobot.

The dataset includes RGB sequences of images recorded at a resolution of 848 × 480 and 20 frames per second, presenting 13 users performing six different gestures that correspond to six different commands. These commands, given to the robotic agent by the human operator, along with five postures that were captured during the TV assembly routine, consist of a total of 11 classes, which are used as input to the classification algorithm. The gestural vocabulary is given in **Figure 4**.

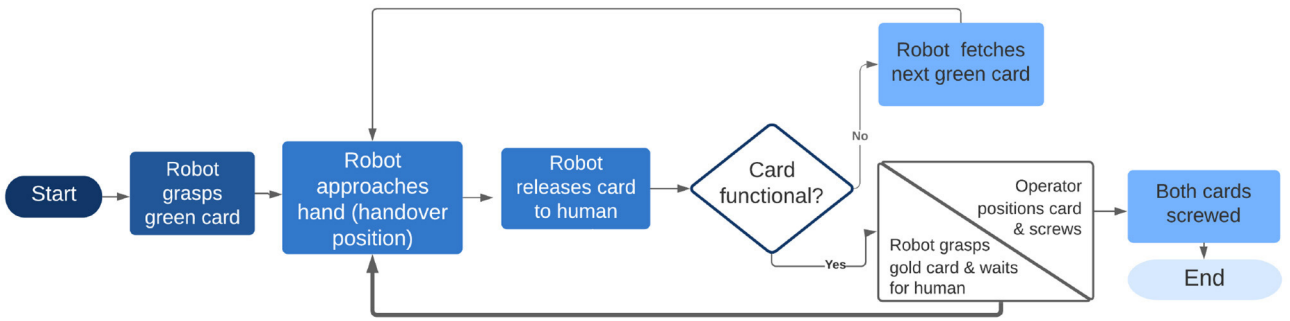


FIGURE 5 | Work-flow of the HRC system.

Gestures are performed in a predefined working space, with a conveyor between the robotic agent and the human operator. A GoPro camera[7] is mounted with a headband on the head of the operator, providing an egocentric view of the TV assembly process. There are two main challenges connected to capturing a dataset from an egocentric view. The first challenge concerns the "double" movement of the hands and the head. The hands of the operator move during the execution of the gesture, while the camera moves along with the head, and is therefore not always in accordance with the hands. The second challenge concerns the fact that due to the short distance, from the camera to the hands, and the field-of-view the camera has, the hands are usually prominent in the frame, but can also be partly, or even totally, out of the field-of-view.

More specifically, during the performance of the TV assembly scenario, the operator performs Gesture 1 (G1) to indicate the start of the assembly routine to the cobot. The cobot goes above the box with the TV cards, then toward the green card, takes it and hands it to the operator, who checks the card for functionality problems. In cases where this particular card is not functional (e.g., is broken, or has a missing part) the human operator performs Gesture 10 (G10) to notify the cobot, which in turn fetches the next green card. The operator verifies that the new green card functions and performs Gesture 2 (G2) to confirm the functionality of the card to the cobot. The operator places the green card (Posture G3) on the TV frame and starts screwing it in place (Posture G4). At the same time, the cobot approaches the gold card and gives it to the operator, as soon as the screwing procedure with the green card has finished. The operator then performs Gesture 8 or 11 (G8 or G11), depending on whether the gold card is functional or not. The above steps are repeated until the two cards are placed appropriately on the TV frame, and the TV is assembled. Finally, the human operator performs Gesture 6 (G6) to confirm the end of the assembly routine, until a new one starts again with Gesture 1 (G1). The captured gestures have the same duration, on average, apart from G4 and G5, during which the operator screws the green and the gold cards respectively.

[7]https://gopro.com/en/fr/shop/cameras/hero9-black/CHDHX-901-master.html?gclsrc=aw.ds&&gclid=Cj0KCQiAjKqABhDLARIsABbJrGnysxveH64ikG8aUbTJACoVucx259TEujqz_3cDlwFAZuE5Yhgi5zKoaAmbHEALw_wcB

To ensure the safety of the human operator, errors must be avoided; thus, two control layers were employed in decision-making. The recognized gesture ID was taken into consideration only if the same recognition accuracy result, with a probability of 100%, was extracted for twenty consecutive frames. The time between the capture of the frame, up to the correct classification of a gesture, was between 0 and 800 ms, thus leading to the conclusion that no important latency was observed during the performance of the HRC scenario. The extracted recognition result was transformed to an ID from 1 to 11 and the result was then sent to the cobot, through the use of a UDP communication protocol. At this point, the second layer of security was added. The thought behind this specific layer was based on the idea of a specific sequence performed during assembling a TV, without any important variations to be taken into consideration. Thus, the received accuracy result was checked by the cobot and was accepted only in cases where it corresponded with the expected gesture ID that was defined according to the work-flow and the scenario presented in **Figure 5**.

5.3. Gesture Recognition Results

For the evaluation of the performance of the gesture recognition algorithm and the proposed methodology, the metrics of *accuracy* and $f - score$ were calculated. The $f - score$ metric is derived by a combination of the metrics *recall* and *precision*. Those metrics are defined as shown below:

$$precision = \frac{\#(true\_positives)}{\#(true\_positives) + \#(false\_positives)} \tag{3}$$

$$recall = \frac{\#(true\_positives)}{\#(true\_positives) + \#(false\_negatives)} \tag{4}$$

$$f - score = 2\frac{precision * recall}{precision + recall} \tag{5}$$

Concerning the *accuracy*, if $\hat{y}_i$ is the predicted value of the i-th sample and y_i is the corresponding true value, then the fraction

of correct predictions over n_{samples} is defined as:

$$\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i) \qquad (6)$$

The network presented was initially trained on the TV assembly custom dataset that was created as part of this work. The dataset was split into training and validation sets with a ratio of 80:20. The network was trained from scratch with a batch size of 32 frames and an Adam optimizer for 40 epochs. The accuracy results for offline gesture recognition, with a sliding window of 5 frames can be found in **Table 1**.

After performing the same experiment, using only new users to whose gestures the recognition module was not trained (**Table 1**), it was observed that the network possibly needed to be trained to a larger amount of data in order to be able to distinguish the differences between the hands of the operator for each gesture. For this to be achieved, an egocentric gestural dataset, that was created during the work of Chalasani et al. (2018), was used for transfer learning. The dataset consisted of 10 classes of gestures captured in an egocentric view in front of a green background. The specific dataset included three iterations per user, for 22 users in total. Even though the size of this dataset cannot be considered appropriate for transfer learning, it had the advantage of being easily customized and turned into a larger dataset. In order for this to be achieved, around 100 images that provided a view of the TV assembly background (TV frame and TV cards), from different angles, were recorded. The green background of the original dataset was removed and replaced by a custom background, leading to a new, larger dataset, to be used for transfer learning. The process involved in the preparation of this dataset is shown in **Figure 6**.

To reach the final number of layers to be frozen, several experiments were performed. It was noticed that freezing network layers did not improve the recognition accuracy results, so after the initial training of the network with the improved dataset from Chalasani et al. (2018), the network was retrained, using the egocentric TV assembly dataset. The 80:20 approach was used again, and the stratification parameter was deployed to split it in such a way that the proportion of values, in the training set, would be the same as the proportion of values in the test set, leading to a balanced proportion in the classes within each. The recognition accuracy results, along with the f-score, with both the 80:20 approach and the testing of the network with completely new users, are shown in **Table 1**. Two diagrams of the accuracy and loss for an experiment using transfer learning to perform gesture recognition, with 40 epochs in total, with the 80:20 method is shown in **Figure 7** for the visualization of the convergence of the training and testing phases.

It was thus observed that transfer learning led to an improvement of 11% in the accuracy results, in cases where new users were introduced to the dataset, which is rather significant. After running the same experiment, using early stopping, the accuracy increased to 98.5%. Also, in **Figure 8**, the confusion matrices are presented with only new users in the testing set without the use of transfer learning (above) and with transfer learning (below).

TABLE 1 | Recognition accuracy and f-score with and without transfer learning.

		Accuracy (%)	F-score (%)
Test with no new users	No transfer learning	99.8	99.8
	Transfer learning	99.9	99.8
Test with new users	No transfer learning	84.68	60
	Transfer learning, 40 epochs	95.7	97.2
	Transfer learning, early stopping	98.5	98.6

In the two confusion matrices presented, a significant total improvement of 11% is observed, as already mentioned. More specifically, for each gesture, in the case where transfer learning was not used, G1 and G2 were not recognized correctly at all, while when transfer learning is used, the recognition level rises to 100%. Even if these gestures are considered as simple and rather static, transfer learning was required for the 3DCNN network to be able to perform accurate recognition. Concerning G5 (Screw gold card) and G9 (Place gold card), satisfactory results can be observed even without transfer learning, which can be explained by the fact that these two classes have the characteristic of the introduction of the gold card, which makes them much more discrete for the network than G1, G2, G3, G4, and G10.

We can indeed foresee that when the learning base contains examples of an operator's gestures, his/her future gestures will be better recognized by the system. However, since the implementation of 3DCNNs is a method with high computational time demands, one of the goals of this work is to examine if the proposed gesture recognition module can be used in the assembly-plant directly, without any further training. At the same time, we had to ask ourselves how many iterations of the same operator were necessary in order to have an improvement in the recognition rate. In **Table 2** and **Figure 9**, the improvement rates in recognition accuracy are presented in the cases where 1, 3, 6, or 9 iterations of the test user were added in the training phase. The baseline for these experiments is the result extracted when there are no iterations for this operator in the training set. At that point, a recognition accuracy of 95.76% was achieved, leading to the conclusion that, indeed, the proposed gesture recognition module could be used in an assembly line, without the need for it to be trained with samples from each new human operator. The rest of the results extracted provide an idea of what can be deemed a sufficient amount of data to be used in the training phase for the desired recognition results. In this particular case, it was the number of 3 sets that gave the best results and reached an accuracy level of up to 99.8%.

Other experiments performed using the same network architecture, but with a TV assembly dataset recorded not from an egocentric, but from a top view, provided results that reached up to 96% with an 80–20 approach. Thus, the results with only new users in the test set were much lower than the ones provided

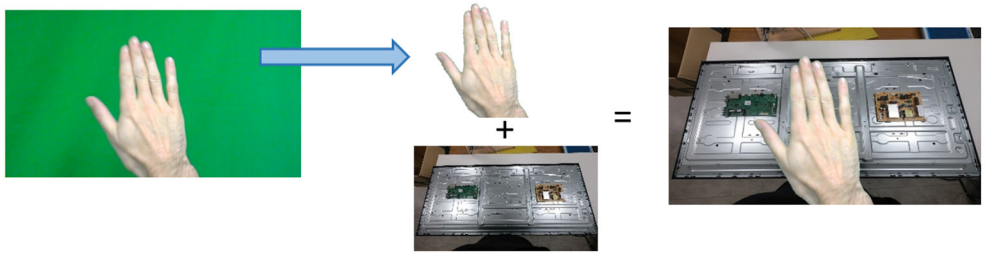


FIGURE 6 | Preparation process of the dataset used for transfer learning.

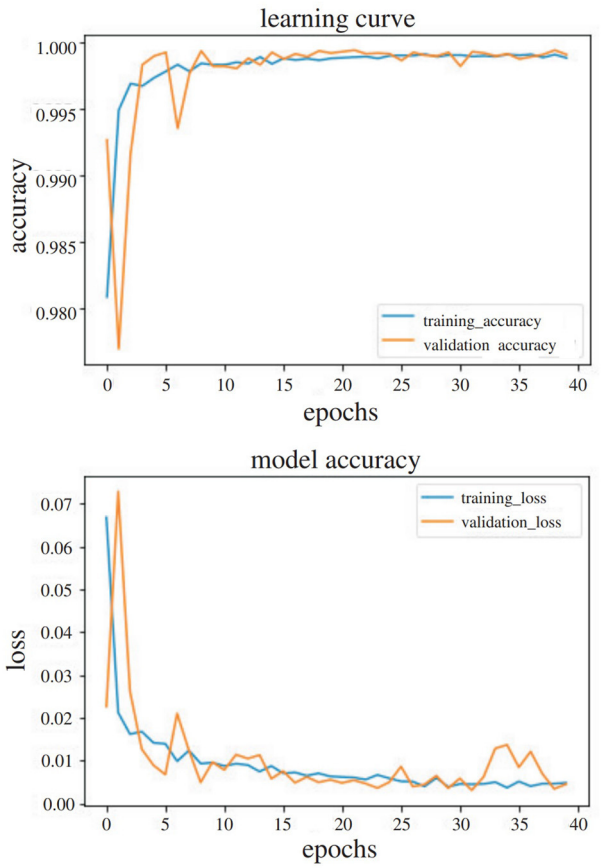


FIGURE 7 | Accuracy and loss diagrams of the experiment with transfer learning, trained for 40 epochs in total.

in this work. This result enforces the added value of an egocentric dataset, as in the top-view it was observed that in many frames, important information for the gesture recognition module was either occluded or out of the frame.

The results presented are compared to the work of Coupeté et al. (2016) that used a Hidden Markov Models gesture recognition engine, in a HRC assembly scenario. The authors in this work, deployed Nearest Neighbors (k-NNs), geodesic distances, as well as Hidden Markov Models, to perform gesture recognition, reaching recognition accuracy results of 85%, with a split of the training and testing data using the 80:20 method, while when testing with unknown operators, the accuracy results concluded to an accuracy of 80% in total. The method presented in this specific paper, outperforms the recognition results of Coupeté et al. (2016), showing very satisfying results.

6. CONTROL OF THE ROBOT AND EVALUATION OF HRC SCENARIO

The cobot used in this scenario was the UR3[8] robotic arm from Universal Robots. The external parts that were used for grasping the cards and for the introduction of physical interaction were from ROBOTIQ (gripper: 2F-140[9] & force torque sensor: FT-300-S[10]). For the control of the robotic arm, Robot Operating System (ROS) was used. Official ROS packages were used, in this instance, both for the control of the robotic arm (UR3[11]) and for the control of external parts (gripper & force sensor[12]).

As mentioned previously, during the execution of every experiment, there were two different types of robot goal points. First of all, there were the predefined points, like the waiting position or the handover position in the experiments "*Physical Interaction*" and "*Gesture Recognition*". On the other hand, when pose estimation was inserted, goal points that were estimated on-the-fly were sent to the robot. ROS provides plenty of libraries for the control of the robot. One of them, named ActionLib was used to allow the motion of the robot through a series of predefined poses. To be more specific, it takes a series of robot poses to form a ROS action. To achieve tasks using actions, the notion of a goal that can be sent to an ActionServer by an ActionClient is introduced. The goal is a PoseStamed message that contains information about where the robot should move within its environment. For each position, it computes the inverse kinematics solution to find the joint angles corresponding to the end effector position. Through this procedure, it creates a smooth trajectory and passes it to the drivers of the robot for execution. For experiments with

[8]https://www.universal-robots.com/products/ur3-robot/
[9]https://robotiq.com/products/ft-300-force-torque-sensor
[10]https://robotiq.com/products/2f85-140-adaptive-robot-gripper
[11]https://github.com/UniversalRobots/Universal_Robots_ROS_Driver
[12]https://github.com/ros-industrial/robotiq

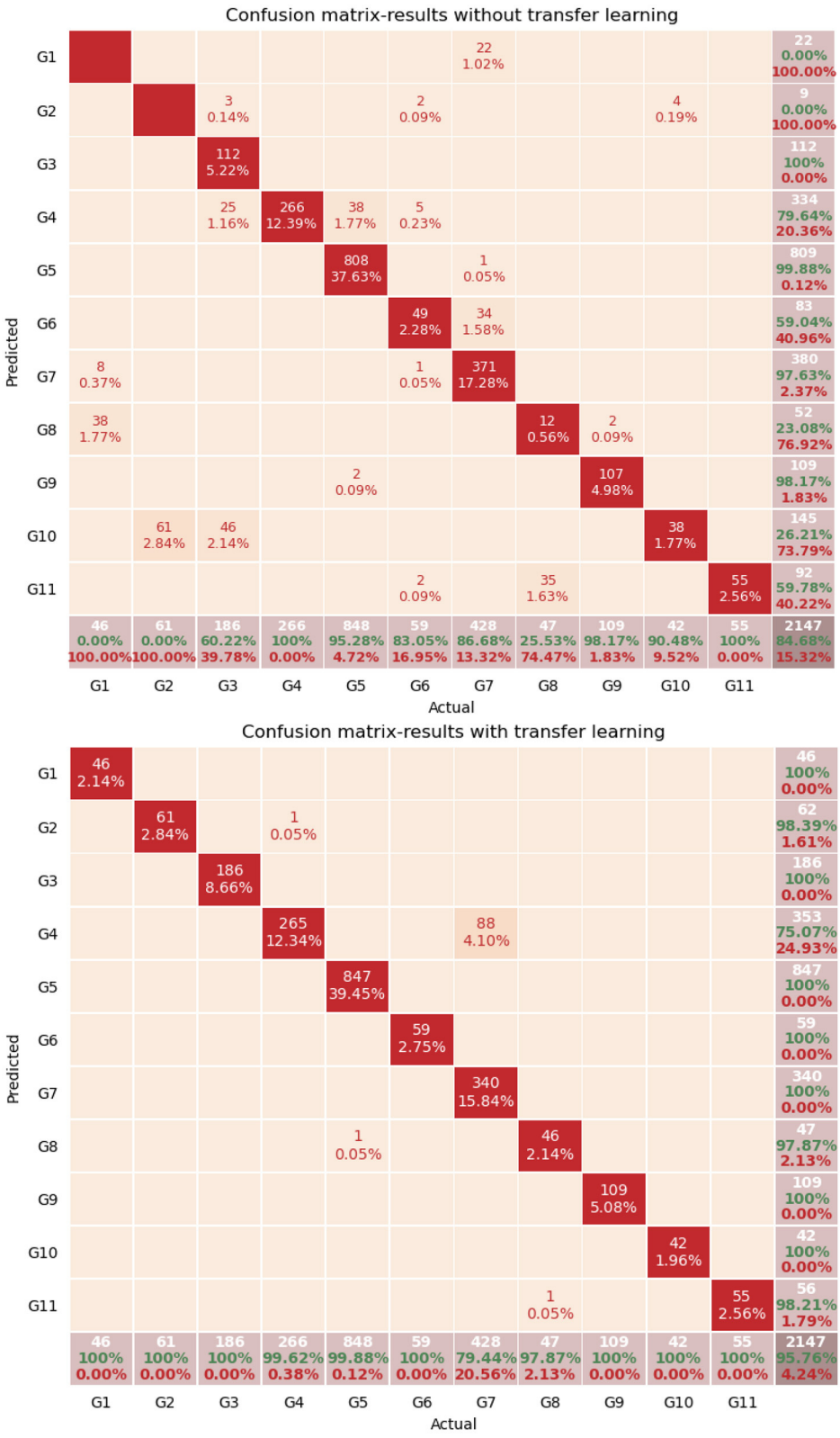


FIGURE 8 | Confusion matrices with only new users in the test set. Without transfer learning **(top)** and with transfer learning **(bottom)**.

TABLE 2 | Contribution of the number of sets from the new user in improving recognition accuracy.

	Number of added sets			
Improvement rates of recognition	1	3	6	9
	+2.2%	+1.9%	+0%	+0.2%

pose estimation MoveIt[13], a motion planning framework named Open Motion Planning Library (OMPL) was used. From the aforementioned pose estimation procedure, the position of the operator's hand was perceivable and was sent to the robot in the Cartesian space. Therefore, for the specific motion of the robot, the Cartesian path was computed using the MoveIt framework, with specific constraints (same orientation of end effector and safety restriction of velocity). Cartesian planning supports a type of constraint that keeps the robot end effector upright, in order to reduce the possibility of injuring the operator. Cartesian path planning, through the MoveIt framework, satisfied the use-case constraints, as the end effector moved along a straight line, using waypoints interpolation.

Common metrics derived from the literature were used for the evaluation of the proposed Human-Robot Interaction (HRI) system as a whole, the effectiveness of the cobot, and the opinion of the human operators about the Human-Robot (HR) interaction. The evaluation of the system as a whole was able to be measured by specific metrics, such as the efficiency of the robotic arm. This included the time it took for the cobot to move, in relation to the time that the whole routine needed. Had this been extremely small, then this would have revealed that the specific cobot could be used in two assembly lines in parallel, thus speeding up the production process. The evaluation of the effectiveness of the robot was able to be measured by metrics such as neglect tolerance (NT), which is concerned with the amount of time that a human can ignore a cobot, and also robot attention demand (RAD), which measures the attention that the cobot demands from the operator, depending on the degree of Interaction Effort (IE) that is expected from the user. The smaller this number is, the more realistic the interaction between the human and the cobot is.

The NASA Task Load Index (TLX)[14], is widely used as a subjective workload assessment tool, which rates perceived workload (both mental and physical) in order to assess a task. A version adapted to the specific use case was implemented, in order to evaluate the workload of the task of screwing of electronic cards on a frame. In addition, for every experiment, users were questioned about the relationship that was developed between the robotic arm and them. Finally, users responded concerning which experiment provided the most natural and realistic collaboration.

[13]https://moveit.ros.org/

[14]https://humansystems.arc.nasa.gov/groups/tlx/

7. RESULTS

Every experiment was executed twice by 14 operators (the group consists of 4 women and 10 men, aged from 23 to 44 with little and medium experience of the execution of TV assembly). During each execution, the operator followed a particular sequence of actions. Initially, s/he asked for the first green card. The robotic arm brought it and the operator checked to see whether the card was functional or not. S/he informed the robotic arm, concerning the functionality of the card, and it reacted accordingly. When the operator had a functional card, s/he started screwing it in place. When the operator was finished, s/he asked for the second card and the robotic arm brought it. The same sequence of actions was executed until both cards were screwed onto the TV panel.

In the first execution of each experiment, for each operator, the first card of both types (green and gold) was deliberately not functional. As mentioned before, the operator had to inform the robotic arm about the functionality of the electronic cards, depending on the type of interaction that was used in each experiment. In the second execution, every card (of both types) delivered was functional. The purpose of these two types of experiments was to present the adaptation of the robotic arm with a predicted interruption in the procedure. The cycle time for each experiment is presented in **Figure 10** by a whisker plot showing and comparing distributions. Experiments with non-functional cards of both kinds are referred to as Form A and the ones with both cards functional are referred to as Form B. A one-way ANOVA for experiments of Form B revealed that there is not a statistically significant difference in cycle time between different types of interaction [F(between groups df, within groups df) = [0,29], p = [0,88]]. This can be justified as the time of interaction is small and the main parameter that affects the cycle time is the duration of card-screwing operation. However, the questionnaires, which are presented later, proved that the insertion of gesture recognition and pose estimation improves the sense of collaboration and reduce the motion of every user. A second one-way ANOVA test for experiments of Form A is executed and confirms that cycle time changes significantly for different types of interaction (F = [10,71], p = [9,91e-06]). During the execution of the experiments of Form A, the completion of the routine, when gesture recognition was implemented, lasted 20% less than the experiment "*Physical Interaction*" and about 13% less than the experiments where pose estimation was used.

Figure 10 presents the adaptation of the average cycle time of the routine, depending on the sequence followed. The cycle time of the routine is dynamic and from **Figure 10** one more interesting result appears. The fastest execution of the routine takes place when gesture recognition is used as the means of interaction between the operator and the robotic arm. Furthermore, it is important to mention that in **Figure 10**, the gap between the cycle time of the experiments of Form A and B appears to be about 20% less in the experiment "*Gesture Recognition*". This metric is an indication that the implementation of gesture recognition in this HRC scenario can reduce the cycle time of the routine, even though predicted or unpredicted incidents occur.

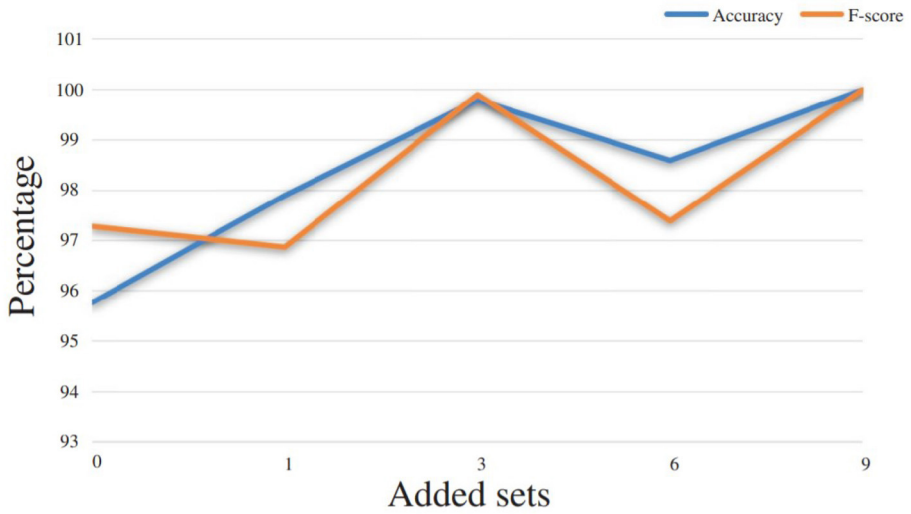


FIGURE 9 | Performance improvement according to the number of gesture examples that are added in the training set and provided by a given user.

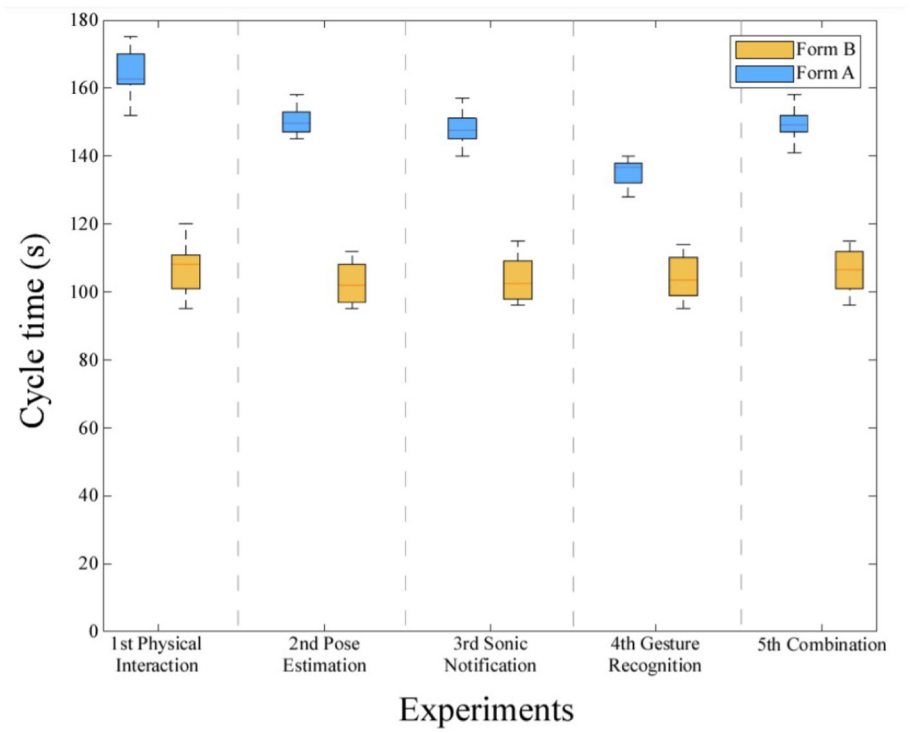


FIGURE 10 | Dynamic cycle time depending on the sequence. Form A: Experiments where the first card of each kind is non-functional. Form B: Experiments where every card is functional.

Figure 11 presents the average timeline of execution of each experiment of Form A. The purpose of this figure is the presentation of the task of each participant in this collaboration and the way that they interact. By way of comparison of each of the interaction types used, the response time for each interaction is given. Response time, in order to facilitate this comparison, is defined as the time from the beginning of the motion of the operator up to the moment that the robot starts moving. The average response time of Physical Interaction (PI) in all experiments is 1.63 s, as compared with 2.87 s, which is the response time for Pose Estimation. Gesture recognition is located between these two and amounts

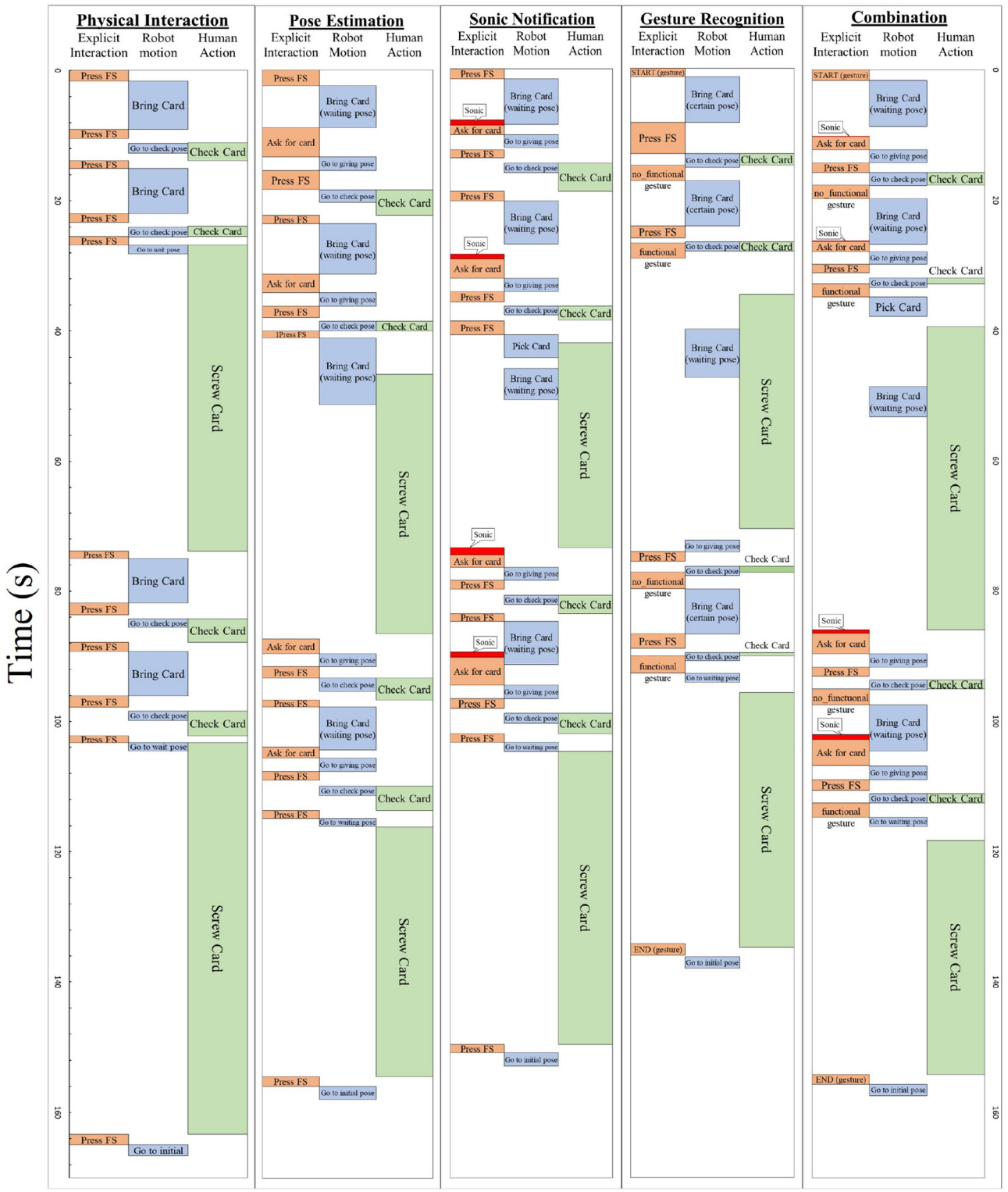


FIGURE 11 | Average timeline of each experiment (Form A).

to 2.01 s. Despite the fact that PI seems to display the best response time, in **Figure 11** the average cycle time of the experiment "*Physical Interaction*" seems to be greater. This can be explained by the fact that the robot interacts more implicitly and less explicitly (i.e., through direct commands from the operator).

In the experiment "*Physical Interaction*", the operator has to stop his routine and inform the robot about his current action (by pressing the button) which is replaced by recognition of professional gestures and pose estimation in later experiments. These two modules make the execution of the experiment faster. In the case of the experiments "*Pose Estimation*" and "*Sonic Notification*", the information about the completion of screwing each card is provided to the robotic arm by the position of the operator's hand (the operator lets go of the screwdriver) and in the case of gesture recognition, the robotic arm is constantly provided with information about the actions that are being executed by the operator. In every experiment, there is a particular sequence of actions that is followed by the operator, which is a safeguard for the smooth execution of the routine.

In **Table 3**, the metrics that are used for the evaluation of the HRC scenario are presented. Initially, the efficiency of the robot (i.e., the percentage of time that the robot moves while running a program-routine) is measured. During the execution of experiment 1 and 4, the handover position is predefined and the time that the robot moves during the execution of the experiment doesn't change (it is represented by an * symbol in **Table 3**). The control box of the robotic arm certifies that while the robotic arm is in motion, the power demanded is approximately 100 W, whereas during the time that the robotic arm remains motionless, it is about 75 W. Thus, this metric informs the operator concerning the time during which the robot is moving and, therefore, concerning the power demands of the robot. As the cycle time of the routine of the experiment "*Gesture Recognition*" is lower and the efficiency of the robot remains at the same level, the motion time of the robot is less, compared to the other experiments. However, by adding the gesture recognition module to within the scenario, and thus a new computer that exploits its GPU to almost its maximum capacity, as well as a camera that provides streaming in real-time, this increases the total power demand by 60 W. As the routine for the whole TV assembly scenario using gesture recognition lasts 136 s, this means an increase of 2.28 Wh for every TV assembled.

Neglect Tolerance (NT) and Interaction Effort (IE), that were mentioned previously, are also presented in **Table 3** with their standard deviation in parenthesis. Robot Attention Demand (RAD) is calculated using the following equation:

$$RAD = \frac{IE}{NT + IE} \tag{7}$$

This metric provides us with information about how many times the operator has paid attention to the robot and has provided it with commands concerning the next step of the routine. As NT contains the time when the screwing of cards is executed, RAD is a metric that depends on the rhythm of each operator. The greater the value of NT is, the less rich information the robot receives about the human's actions and intentions. Moreover, the larger the RAD, the more the robot is able to understand and adapt to its partner. The average, as presented in **Table 3** and its standard deviation in parenthesis, shows that RAD is stable among the last four experiments, despite the fact that NT is significantly smaller during the experiments in which gesture recognition is implemented. The reason that RAD is smaller during the execution of the experiment "*Physical Interaction*" is that the NT is greater, as the operator interacts only explicitly with the robotic arm.

During the execution of the experiments without spatial adaptation (*SA*), the operator receives the cards from a particular handover position (*PHP*). The KPI that is proposed in equation 8 indicates the percentage of robot spatial adaptation in the case of every operator. For the calculation of the KPI, the distance from the waiting point (*WP*), to the particular handover position for the experiments that is stable, is compared to the adaptable handover position (*AHP*) for the experiments in which pose estimation is implemented. Distances are measured in centimeters. The higher the adaptation rate, the greater the effort that was demanded of the operator during the experiments, without spatial adjustment. As **Table 4** shows, operators 3 and 12 asked for the card to be brought closer to the particular handover position and as a result the robotic arm had to adapt less than for the other operators. This KPI could also be useful for discovering the position that each individual user prefers as the handover position.

$$SA(\%) = \frac{\|AHP - WP\| - \|PHP - WP\|}{\|PHP - WP\|} \tag{8}$$

Where *SA*: spatial adaptation, *AHP*: adapted handover position, *WP*: waiting point and *PHP*: particular handover position.

Human factors (or ergonomics) are defined by ISO 26800 as the "scientific discipline concerned with the understanding of interactions among human and other elements of a system, and the profession that applies theory, principles, data, and methods to design in order to optimize human well-being and overall system performance". In order to achieve an optimal level of collaboration, it is essential to take into account the opinion of the human involved in operations with the robot. To evaluate the execution of the experiments users responded to two different questionnaires. Initially the workload of the TV assembling task was estimated through the NASA-TLX. This tool consists of a questionnaire with six items for evaluation: mental demand, physical demand, temporal demand, effort, frustration and performance.

11 out of 14 participants replied that the task they undertook was neither physically nor mentally demanding. In addition, none of the them felt that the pace of the task was hurried. Thus, the reason that a cobot is used to substitute a human operator for this task is the need for repeatability and the fact that a cobot can not only repeat the same task many times, but can perform the task precisely and fast. Every participant was able to accomplish all the experiments and responded that they did not find it difficult to interact with the robot and understand its reactions. Due to the inexperience of some users, some errors occurred during the execution of the experiments; however, this did not affect the accomplishment of the task, as the robotic arm was following a particular sequence of actions.

In addition, the participants were asked to categorize the type of HRI of each experiment and to characterize the relationship between the robot and the operator during the execution of each

TABLE 3 | Average of HRC metrics for each experiment. Neglect Tolerance (s), Interaction effort (s), Robot Attention Demand (RAD), and Efficiency of the robot(%) (*no spatial adaptation).

	Neglect tolerance	Interaction effort	RAD	Efficiency of the robot
Physical interaction	118.9 (σ: 5.4)	15.5 (σ: 3.2)	0.12 (σ: 0.001)	23*
Pose estimation	90.6 (σ: 3.8)	27.6 (σ: 4.1)	0.23 (σ: 0.001)	31 (σ: 0.3)
Sound notification	88.8 (σ: 4.1)	30.9 (σ: 4.7)	0.26 (σ: 0.002)	30 (σ: 0.8)
Gesture recognition	61.4 (σ: 3.2)	23.1 (σ: 2.6)	0.27 (σ: 0.001)	28*
Combination	88.1 (σ: 3.6)	30.0 (σ: 4.0)	0.25 (σ: 0.001)	29 (σ: 0.4)

TABLE 4 | Spatial adaptation (%) of each operator.

Operators	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Spatial adaptation	40.2	33.9	19.6	45.6	36.5	40.2	41.9	39.5	27.5	28.9	33.1	21.5	39.2	23.9

experiment. In section 2.1, the categories of HRI are analyzed (Coexistance, Synchronized Cooperation and Collaboration). In the first part of **Figure 12**, the types, from among which the participants chose the category of HRI, are presented. The majority of the respondents thought that the implementation of gesture recognition in the experiment "*Gesture Recognition*" and "*Combination*" strengthened the sense of collaboration, while they felt that the first three experiments belonged to the category of synchronized cooperation. All the participants considered that with only "*Physical Interaction*", the robotic arm was simply following the human operator, which led to slower execution of the task. Finally, as was mentioned before, the aim of this research is to convert the robot from a useful machine to a real collaborator. When only physical interaction was used (1st experiment) most of the users felt that the robot had a supporting role. However, 9 out of 14 participants declared that the insertion of "*Pose estimation*" or "*Gesture Recognition*" made them feel that their contribution to the task was equal to that of the robot.

8. DISCUSSION

The proposed methodology and the experiments concerning the contribution of different modalities to an HRC scenario, concludes by showing great potential for the future. Both hypotheses that were defined at the beginning of this work, are evaluated. Through the experiments performed, it was validated that on what concerns the temporal adaptation of the robotic arm, the insertion of gesture recognition reduces the cycle time of the routine of every operator (by 20% on average), adding a relatively small increase in energy consumption by the system. The second hypothesis was concerned with the implementation of pose estimation in order to achieve the spatial adaptation of the robotic arm. According to the results collected, the hypotheses presented are both valid. For 9 of the 14 operators, the percentage of spatial adaptation is more than 30%, which shows the importance of this modality regarding reducing the operator's effort.

Concerning the different modalities, gesture recognition is proved to be capable of accelerating an assembly line and of providing the human operator with a sense of true cooperation with the cobot and not just coexistence. Meanwhile, pose estimation offers the prospect of converting the cobot to a partner who adapts to every operator. A significant observation for pose estimation is that robot attention demand is increasing while the average motion time of the robot decreases in contrast to physical interaction. The argument given above proves that the cobot possesses more information about the human operator and as a result it moves less during the routine, as it can predict human's motions.

In both gesture recognition and pose estimation modules, the response time is satisfactory, within a challenging task, that facilitates the spatiotemporal adaptation. A great improvement in the accuracy of gesture recognition was noted after the implementation of transfer learning, proving that the initial amount of acquired data was not sufficient, even after a few sessions of recording. 3DCNNs have to be robust and extract confident results, even in real-time, with operators that the network has not been trained with. Egocentric gesture recognition might be a challenging task, but it can lead to impressive results, independent of anthropometric characteristics and clothing. The most important observation that the gestural module provided, was the fact that it can be used in a real-life assembly line with great results, without retraining the network each time that a new human operator was introduced to it. Even though handling data from an egocentric point of view was a challenging task in order for an accurate classification to be performed, and for the safety of the human operator to be ensured, it provided great potential for the future. Apart from this, the TV assembly dataset created can be enriched with more classes in an egocentric view from different professional environments, in order for the proposed approach to respond to different professional setups.

The operator's sense of collaboration with the cobot improved significantly because of the sonic notification. It could be enriched with many different kinds of messages; however, due to the fact that this use-case is intended for

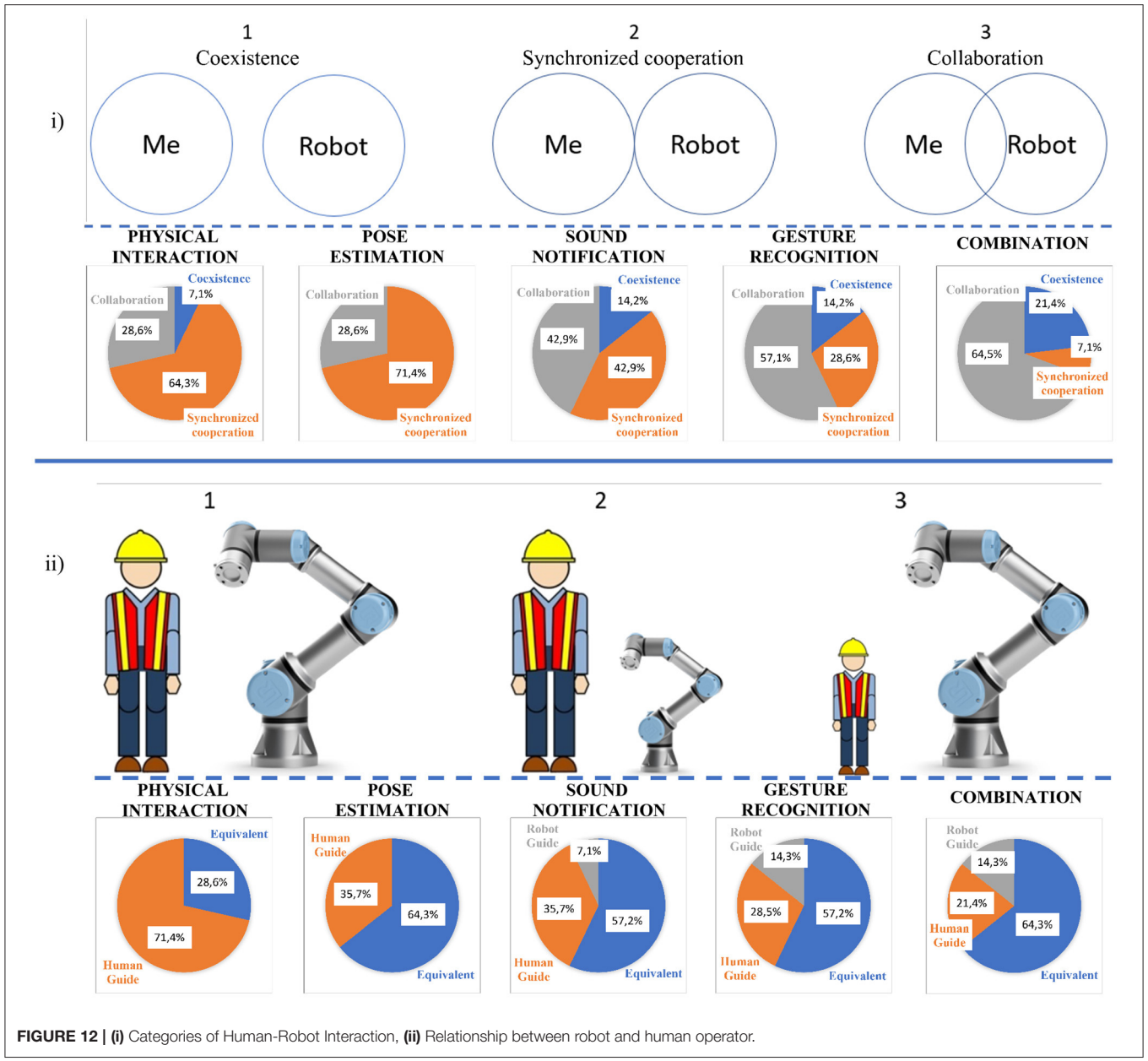


FIGURE 12 | (i) Categories of Human-Robot Interaction, **(ii)** Relationship between robot and human operator.

an industrial environment, simplicity has to be preserved. Furthermore, the existence of many different sonic notifications could create comprehension problems in the case of many parallel assembly lines. The questionnaire validated the fact that the task was neither mentally nor physically demanding. The reason that it was used for this research was its repetitiveness, because robots tend to take over the dull, dirty, dangerous and dear (i.e., costly) tasks from humans, otherwise known as the 4 Ds of robotization. Finally, according to the answers of the participants, the implementation of pose estimation made them feel that they participated equally with the robot in the routine of TV assembly, while gesture recognition enhanced the sense of collaboration in contrast to synchronized cooperation.

9. CONCLUSION AND FUTURE WORK

In this paper, an HRC scenario is defined and different modalities are evaluated concerning the cycle time of the execution of a TV assembly routine and the naturalness of this collaboration, according to the human operators. The insertion of gesture recognition accelerates the execution of the proposed routine by about 20%, reducing, in parallel, the effort required of the operator, in order to perform.

In this research, a new KPI regarding spatial adaptation is proposed and shows that the insertion of a cobot with a dynamic spatial profile that adjusted to the operators, changes the handover position of the experiment by up to 40%.

The ergonomic parameters of a task can be analyzed and the robot adjusts its motion not only to avoid collisions with the operator, but also in order to ergonomically improve the pose of the operator during the execution of their task.

Moreover, this paper opens up potential for investigating industrial HRC scenarios and proposing intelligent and efficient solutions on the road to Industry 4.0. This research could have been enriched with experiments executed by professional users from the industry; however, due to the conditions imposed by Covid-19 restrictions, this was impossible. Our future work will be focused on the upgrading of the robot's perception of the user and their environment, with an aim to improving their collaboration. To this end, the way that the robot can make best use of pose estimation is investigated. Finally, the fact that the robot is able to perceive through pose estimation, and to follow the position and every action of the operator in real time, undoubtedly improves their collaboration and further facilitates the insertion of robots in common industrial work-spaces with human operators.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article can be made available by the authors upon request.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

DP: leader of cobot control, the pose estimation module, and its integration with the cobot, and co-leader of the experimental protocol, architecture, and realization. GS: leader of the recordings of the used dataset, segmentation, and labeling, defined the gestural vocabulary, and interpreted the recognition results, developer of the gesture recognition module and its integration with the cobot, and co-leader of the experimental protocol, architecture, and realization. SM: conceptualization of the methodology and definition of the scientific and industrial needs to be addressed with respect to human-robot collaboration, machine learning, and pose estimation. All authors contributed to the article and approved the submitted version.

FUNDING

The research leading to these results has received funding from the EU Horizon 2020 Research and Innovation Programme under Grant Agreement No. 820767, CoLLaboratE project.

REFERENCES

Ajoudani, A., Zanchettin, A. M., Ivaldi, S., Albu-Schoffer, A., Kosuge, K., and Khatib, O. (2018). Progress and prospects of the human-robot collaboration. *Auton. Robots* 42, 957–975. doi: 10.1007/s10514-017-9677-2

Amin, F. M., Rezayati, M., Wernher van de Venn, H., and Karimpour, H. (2020). A mixed-perception approach for safe human-robot collaboration in industrial automation. *Sensors* 20:6347. doi: 10.3390/s20216347

Bicchi, A., Peshkin, M. A., and Colgate, J. E. (2008). "Safety for physical human-robot interaction," in *Springer Handbook of Robotics*, eds B. Siciliano and O. Khatib (Berlin; Heidelberg: Springer), 1335–1348. doi: 10.1007/978-3-540-30301-5_58

Bo, H., Mohan, D. M., Azhar, M., Sreekanth, K., and Campolo, D. (2016). "Human robot collaboration for tooling path guidance," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)* (Singapore), 1340–1345. doi: 10.1109/BIOROB.2016.7523818

Borghi, G., Vezzani, R., and Cucchiara, R. (2016). "Fast gesture recognition with multiple stream discrete HMMs on 3D skeletons," in *2016 23rd International Conference on Pattern Recognition (ICPR)* (Cancun), 997–1002. doi: 10.1109/ICPR.2016.7899766

Bui, G., An, T., Anh, N., and Ho Nhut, M. (2018). Hidden Markov model for recognition of skeletal data-based hand movement gestures. *EAI Endorsed Trans. Contex Aware Syst. Appl.* 4:154819. doi: 10.4108/eai.18-6-2018.154819

Canal, G., Pignat, E., Alenya, G., Calinon, S., and Torras, C. (2018). "Joining high-level symbolic planning with low-level motion primitives in adaptive HRI: application to dressing assistance," in *2018 IEEE International Conference on Robotics and Automation (ICRA)* (Brisbane, QLD), 3273–3278. doi: 10.1109/ICRA.2018.8460606

Canavan, S., Keyes, W., Mccormick, R., Kunnumpurath, J., Hoelzel, T., and Yin, L. (2017). "Hand gesture recognition using a skeleton based feature representation with a random regression forest," in *2017 IEEE International Conference on Image Processing (ICIP)* (Beijing), 2364–2368. doi: 10.1109/ICIP.2017.8296705

Cao, C., Zhang, Y., Wu, Y., Lu, H., and Cheng, J. (2017). "Egocentric gesture recognition using recurrent 3D convolutional neural networks with spatiotemporal transformer modules," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice), 3783–3791. doi: 10.1109/ICCV.2017.406

Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2021). OpenPose: realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 172–186. doi: 10.1109/TPAMI.2019.2929257

Carreira, J., and Zisserman, A. (2017). "Quo vadis, action recognition? A new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI), 4724–4733. doi: 10.1109/CVPR.2017.502

Chalasani, T., Ondrej, J., and Smolic, A. (2018). "Egocentric gesture recognition for head-mounted ar devices," in *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)* (Munich), 109–114. doi: 10.1109/ISMAR-Adjunct.2018.00045

Chen, S., Li, Y., and Kwok, N. M. (2011). Active vision in robotic systems: a survey of recent developments. *Int. J. Robot. Res.* 30, 1343–1377. doi: 10.1177/0278364911410755

Cheng, Y., Yang, B., Wang, B., Wending, Y., and Tan, R. (2019). "Occlusion aware networks for 3D human pose estimation in video," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul), 723–732. doi: 10.1109/ICCV.2019.00081

Cherubini, A., Passama, R., Crosnier, A., Lasnier, A., and Fraisse, P. (2016). Collaborative manufacturing with physical human-robot interaction. *Front. Neurosci.* 40:7. doi: 10.1016/j.rcim.2015.12.007

Colgate, J. E., and Peshkin, M. A. (2010). *Cobots*. US5952796A. Available online at: https://patents.google.com/patent/US5952796A/en

Coupeté, E. (2016). *Reconnaissance de gestes et actions pour la collaboration homme-robot sur chaîne de montage* (theses). Université Paris sciences et lettres, Paris, France.

Coupeté, E., Moutarde, F., and Manitsaris, S. (2016). "A user-adaptive gesture recognition system applied to human-robot collaboration in factories," in *MOCO '16* (New York, NY: Association for Computing Machinery). doi: 10.1145/2948910.2948933

Coupete, E., Moutarde, F., and Manitsaris, S. (2019). Multi-users online recognition of technical gestures for natural human-robot collaboration in manufacturing. *Auton. Robots* 43, 1309–1325. doi: 10.1007/s10514-018-9704-y

Dröder, K., Bobka, P., Germann, T., Gabriel, F., and Dietrich, F. (2018). A machine learning-enhanced digital twin approach for human-robot-collaboration. *Proc. CIRP* 76, 187–192. doi: 10.1016/j.procir.2018.02.010

El Makrini, I., Merckaert, K., Lefeber, D., and Vanderborght, B. (2017). "Design of a collaborative architecture for human-robot assembly tasks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC), 1624–1629. doi: 10.1109/IROS.2017.8205971

El Zaatari, S., Marei, M., Li, W., and Usman, Z. (2019). Cobot programming for collaborative industrial tasks: an overview. *Robot. Auton. Syst*. 116, 162–180. doi: 10.1016/j.robot.2019.03.003

El-Shamouty, M., Wu, X., Yang, S., Albus, M., and Huber, M. F. (2020). "Towards safe human-robot collaboration using deep reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)* (Paris), 4899–4905. doi: 10.1109/ICRA40945.2020.9196924

Fang, H.-S., Xie, S., Tai, Y.-W., and Lu, C. (2017). "RMPE: regional multi-person pose estimation," in *2017 IEEE International Conference on Computer Vision (ICCV)* (Venice), 2353–2362. doi: 10.1109/ICCV.2017.256

Feichtenhofer, C., Pinz, A., and Zisserman, A. (2016). "Convolutional two stream network fusion for video action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 1933–1941. doi: 10.1109/CVPR.2016.213

Gabler, V., Stahl, T., Huber, G., Oguz, O., and Wollherr, D. (2017). "A game theoretic approach for adaptive action selection in close proximity human robot-collaboration," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Marina Bay Sands), 2897–2903. doi: 10.1109/ICRA.2017.7989336

Gildert, N., Millard, A. G., Pomfret, A., and Timmis, J. (2018). The need for combining implicit and explicit communication in cooperative robotic systems. *Front. Robot. AI* 5:65. doi: 10.3389/frobt.2018.00065

Güler, R. A., Neverova, N., and Kokkinos, I. (2018). "Densepose: dense human pose estimation in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Salt Lake City, UT), 7297–7306. doi: 10.1109/CVPR.2018.00762

Gustavsson, P., Syberfeldt, A., Brewster, R., and Wang, L. (2017). Human-robot collaboration demonstrator combining speech recognition and haptic control. *Proc. CIRP* 63, 396–401. doi: 10.1016/j.procir.2017.03.126

Hentout, A., Aouache, M., Maoudj, A., and Akli, I. (2019). Human-robot interaction in industrial collaborative robotics: a literature review of the decade 2008–2017. *Adv. Robot.* 33, 764–799. doi: 10.1080/01691864.2019.1636714

Khatib, M., Al Khudir, K., and De Luca, A. (2017). "Visual coordination task for human-robot collaboration," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (Vancouver, BC), 3762–3768. doi: 10.1109/IROS.2017.8206225

Kopp, T., Baumgartner, M., and Kinkel, S. (2021). Success factors for introducing industrial human-robot interaction in practice: an empirically driven framework. *Int. J. Adv. Manufact. Technol.* 112, 685–704. doi: 10.1007/s00170-020-06398-0

Liu, Z., and Hao, J. (2019). Intention recognition in physical human-robot interaction based on radial basis function neural network. *J. Robot.* 2019, 1–8. doi: 10.1155/2019/4141269

Long, P., Chevallereau, C., Chablat, D., and Girin, A. (2018). An industrial security system for human-robot coexistence. *Indus. Robot Int. J.* 45, 220–226. doi: 10.1108/IR-09-2017-0165

Melinte, D. O., and Vladareanu, L. (2020). Facial expressions recognition for human-robot interaction using deep convolutional neural networks with rectified adam optimizer. *Sensors* 20:2393. doi: 10.3390/s20082393

Michalos, G., Kousi, N., Karagiannis, P., Gkournelos, C., Dimoulas, K., Koukas, S., et al. (2018). Seamless human robot collaborative assembly–an automotive case study. *Mechatronics* 55, 194–211. doi: 10.1016/j.mechatronics.2018.08.006

Michalos, G., Makris, S., Tsarouchi, P., Guasch, T., Kontovrakis, D., and Chryssolouris, G. (2015). Design considerations for safe human-robot collaborative workplaces. *Proc. CIRP* 37, 248–253. doi: 10.1016/j.procir.2015.08.014

Mohammed, A., Schmidt, B., and Wang, L. (2017). Active collision avoidance for human-robot collaboration driven by vision sensors. *Int. J. Comput. Integr. Manufact.* 30, 970–980. doi: 10.1080/0951192X.2016.1268269

Muhammad, J., Altun, H., and Abo-Serie, E. (2017). Welding seam profiling techniques based on active vision sensing for intelligent robotic welding. *Int. J. Adv. Manufact. Technol.* 88, 127–145. doi: 10.1007/s00170-016-8707-0

Pezzulo, G., Donnarumma, F., and Dindo, H. (2013). Human sensorimotor communication: a theory of signaling in online social interactions. *PLoS ONE* 8:e79876. doi: 10.1371/journal.pone.0079876

Prati, E., Peruzzini, M., Pellicciari, M., and Raffaeli, R. (2021). How to include user experience in the design of human-robot interaction. *Robot. Comput. Integr. Manufact.* 68:102072. doi: 10.1016/j.rcim.2020.102072

Queralta, J. P., Taipalmaa, J., Can Pullinen, B., Sarker, V. K., Nguyen Gia, T., Tenhunen, H., et al. (2020). Collaborative multi-robot search and rescue: planning, coordination, perception, and active vision. *IEEE Access* 8, 191617–191643. doi: 10.1109/ACCESS.2020.3030190

Rahmat, R., Chairunnisa, T., Gunawan, D., Pasha, M. F., and Budiarto, R. (2019). Hand gestures recognition with improved skin color segmentation in human-computer interaction applications. *J. Theoret. Appl. Inform. Technol.* 97, 727–739.

Safeea, M., Neto, P., and Bearee, R. (2019). On-line collision avoidance for collaborative robot manipulators by adjusting off-line generated paths: an industrial use case. *Robot. Auton. Syst*. 119, 278–288. doi: 10.1016/j.robot.2019.07.013

Schmidtler, J., Knott, V., Holzel, C., and Bengler, K. (2015). Human centered assistance applications for the working environment of the future. *Occup. Ergon*. 12, 83–95. doi: 10.3233/OER-150226

Shahroudy, A., Liu, J., Ng, T., and Wang, G. (2016). "NTU RGB+D: a large scale dataset for 3D human activity analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV), 1010–1019. doi: 10.1109/CVPR.2016.115

Sharkawy, A., Koustoumpardis, P., and Aspragathos, N. (2020). Human-robot collisions detection for safe human-robot interaction using one multi-input-output neural network. *Soft Comput*. 24, 6687–6719. doi: 10.1007/s00500-019-04306-7

Sharkawy, A.-N., Koustoumpardis, P., and Aspragathos, N. (2019). Neural network design for manipulator collision detection based only on the joint position sensors. *Robotica* 38, 1–19. doi: 10.1017/S0263574719000985

Song, S., Chandrasekhar, V., Mandal, B., Li, L., Lim, J., Babu, G. S., et al. (2016). "Multimodal multi-stream deep learning for egocentric activity recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (Las Vegas, NV), 378–385. doi: 10.1109/CVPRW.2016.54

Tao, C., and Liu, G. (2013). A multilayer hidden Markov models-based method for human-robot interaction. *Math. Problems Eng.* 2013:384865. doi: 10.1155/2013/384865

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). "Learning spatiotemporal features with 3D convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 4489-4497. doi: 10.1109/ICCV.2015.510

Unhelkar, V. V., Li, S., and Shah, J. A. (2020). "Decision-making for bidirectional communication in sequential human-robot collaborative tasks," in *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20* (Cambridge: Association for Computing Machinery), 329–341. doi: 10.1145/3319502.3374779

Vesper, C., and Sevdalis, V. (2020). Informing, coordinating, and performing: a perspective on functions of sensorimotor communication. *Front. Hum. Neurosci*. 14:168. doi: 10.3389/fnhum.2020.00168

Vogt, D., Stepputtis, S., Grehl, S., Jung, B., and Amor, H. B. (2017). "A system for learning continuous human-robot interactions from human-human demonstrations," in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (Marina Bay Sands), 2882–2889. doi: 10.1109/ICRA.2017.7989334

Yan, S., Xiong, Y., and Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Thirty-Second AAAI Conference on Artificial Intelligence* (New Orleans, LO).

Zeng, R., Wen, Y., Zhao, W., and Liu, Y.-J. (2020). View planning in robot active vision: a survey of systems, algorithms, and applications. *Comput. Visual Media* 6, 225–245. doi: 10.1007/s41095-020-0179-3

Zhang, J., Li, P., Zhu, T., Zhang, W.-A., and Liu, S. (2020). "Human motion capture based on kinect and imus and its application to human-robot collaboration," in *2020 5th International Conference on Advanced Robotics and Mechatronics (ICARM)* (Shenzhen), 392–397. doi: 10.1109/ICARM49381.2020.9195342

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.

frontiers in Neurorobotics

ORIGINAL RESEARCH
published: 23 November 2021
doi: 10.3389/fnbot.2021.648595

A Biological Inspired Cognitive Framework for Memory-Based Multi-Sensory Joint Attention in Human-Robot Interactive Tasks

Omar Eldardeer[1,2]*, *Jonas Gonzalez-Billandon*[1,3], *Lukas Grasse*[4], *Matthew Tata*[4] and *Francesco Rea*[2]

[1] Dipartimento di Informatica, Bioingegneria, Robotica e Ingegneria dei Sistemi, Università di Genova, Genova, Italy, [2] Robotics, Brain, and Cognitive Science Department, Istituto Italiano di Tecnologia, Genova, Italy, [3] COgNiTive Architecture for Collaborative Technologies, Istituto Italiano di Tecnologia, Genova, Italy, [4] Neuroscience/CCBN Department, The University of Lethbridge, Lethbridge, AB, Canada

One of the fundamental prerequisites for effective collaborations between interactive partners is the mutual sharing of the attentional focus on the same perceptual events. This is referred to as joint attention. In psychological, cognitive, and social sciences, its defining elements have been widely pinpointed. Also the field of human-robot interaction has extensively exploited joint attention which has been identified as a fundamental prerequisite for proficient human-robot collaborations. However, joint attention between robots and human partners is often encoded in prefixed robot behaviours that do not fully address the dynamics of interactive scenarios. We provide autonomous attentional behaviour for robotics based on a multi-sensory perception that robustly relocates the focus of attention on the same targets the human partner attends. Further, we investigated how such joint attention between a human and a robot partner improved with a new biologically-inspired memory-based attention component. We assessed the model with the humanoid robot iCub involved in performing a joint task with a human partner in a real-world unstructured scenario. The model showed a robust performance on capturing the stimulation, making a localisation decision in the right time frame, and then executing the right action. We then compared the attention performance of the robot against the human performance when stimulated from the same source across different modalities (audio-visual and audio only). The comparison showed that the model is behaving with temporal dynamics compatible with those of humans. This provides an effective solution for memory-based joint attention in real-world unstructured environments. Further, we analyzed the localisation performances (reaction time and accuracy), the results showed that the robot performed better in an audio-visual condition than an audio only condition. The performance of the robot in the audio-visual condition was relatively comparable with the behaviour of the human participants whereas it was less efficient in audio-only localisation. After a detailed analysis of the internal components of the architecture, we conclude that the differences in performance are due to egonoise which significantly affects the audio-only localisation performance.

Keywords: joint attention, multisensory integration, memory, decision-making, computational neuroscience, human robot interaction, active perception, biological motion control

OPEN ACCESS

Edited by:
Dimitri Ognibene,
University of Milano-Bicocca, Italy

Reviewed by:
Elio Tuci,
University of Namur, Belgium
Davide Marocco,
University of Naples Federico II, Italy

***Correspondence:**
Omar Eldardeer
omar.eldardeer@iit.it

Received: *30 December 2020*
Accepted: *10 September 2021*
Published: *23 November 2021*

Citation:
Eldardeer O, Gonzalez-Billandon J, Grasse L, Tata M and Rea F (2021) A Biological Inspired Cognitive Framework for Memory-Based Multi-Sensory Joint Attention in Human-Robot Interactive Tasks. Front. Neurorobot. 15:648595. doi: 10.3389/fnbot.2021.648595

1. INTRODUCTION

Robots approach a stage of technological advancement at which they will become a frequent partner in our daily lives. At this stage they regularly interact and engage in collaborative tasks with us. Humans and robots have to coordinate their actions in a shared environment in order to efficiently collaborate in these diverse scenarios. While humans are good at coordinating perception and action planning with their movements to achieve a common goal, such complex coordination is still an open challenge in robotics. When we collaborate with another human partner we recruit typical perceptual and action coordination skills. One of the most important coordination skills we use is joint attention as a fundamental mechanism to coordinate our actions (Schnier et al., 2011).

Joint attention can be defined as a shared attentional focus on the same perceptual events between multiple individuals (Reddy, 2005). It is used to coordinate between each of the agents toward a common object or event. Thus joint attention occurs as an emergent condition when a salient event captures the attention of both partners without a priori negotiation of the attentive target. For example, when two people are discussing a painting they are jointly seeing, the shared perception of the same painting allows them to exchange information about the same object. Joint attention is a natural phenomenon that we experience every day and can be triggered by different means: environmental-based (e.g., the appearance of a visual-auditory salient object in the environment) and social-based (e.g., eye-gazing, pointing, or other verbal or non-verbal indications) events (Mundy and Acra, 2006). Mastering correct joint attention with a partner is an important skill that facilitates collaborative interactions. It allows us to share our focus with another partner, enabling us to reason on a common basis. However joint attention not only must be correctly shared between interactants, but the timing of the focus shift also has to be comparable between the human and robot. Jointly shifting attention to the correct location is not necessarily useful if the timing fails to match human timing, as the interaction will fall out-of-sync. Joint attention has been studied extensively in humans, for its role in the development of children (Moore et al., 2014), in language acquisition (Tomasello and Farrar, 1986) and also as a way to identify autism (Bruinsma et al., 2004). Most of the studies on joint attention have been carried out in controlled environments, due to its complex nature and the diversity of scenarios under which it can occur. Current studies in joint attention between a human and an artificial system have mostly focused either on the human or the artificial agent performance. The assessment of combined performance (including mutual influence) across all the agents involved in the task is not common. A thorough assessment of both human attention and the attention of artificial agents would be relevant to the research community. In fact, research evidence shows frequently that both agents influence each other in joint collaborative tasks (Vannucci et al., 2017).

In cognitive architectures that take into account joint attention processes in order to create rich collaborative behaviours, other functionalities such as working memory might participate in attentional refocusing. Such components provide correct and accurate attention-timing and more importantly promote the intelligent behaviour of an attentive capable robotic agent. The influence that working memory has on the attentional mechanism is relevant (Mayer et al., 2007; Shipstead et al., 2014; Oberauer, 2019) but is rarely addressed in cognitive architectures for collaborative robots. Working memory has been defined as short-term memory used in order to proactively reinterpret the information in order to better operate in the environment (Miyake and Shah, 1999; Oberauer, 2019). Different computational models of attention for artificial agents have been proposed (Nagai et al., 2003; Triesch et al., 2006; Ognibene and Demiris, 2013) to respond to visual (Itti and Koch, 2001) and auditory stimuli (Treisman, 1996). However, these models do not fully consider the potential role of working memory related to the process of attentional focus redeployment. Some attention systems have been designed and evaluated to specifically address the context of collaboration between the human and the physically present robot partner (Admoni and Scassellati, 2017) but the potential role of memory remains only partially explored. In this work, we intend to endow the robot with the ability to rely on working memory, to reinterpret the information acquired in previous instances and states in order to better attend to the environment. Different possible computational models of working memory have been provided in different cognitive studies (Repovš and Baddeley, 2006) and in robotics applications (Phillips and Noelle, 2005). Inspired by these previous works,we provided the robot with a simple implementation of working memory that improves the attentive performance of the cognitive architecture for the humanoid robot iCub (Metta et al., 2008). The implementation engage the working memory component in a bio inspired decision making process.

Thus, we propose and evaluate the performance of a computational cognitive architecture for memory-based multi-sensory joint attention. Our goal with this study is to validate emergent joint attention guided by our cognitive framework. The architecture includes a multi-sensory attentional model, a working memory, a decision-making element, and an action executor (motor controller) to solve audio-visual stimuli localisation with human-like performance. We implemented a bio-inspired decision-making strategy (Murphy et al., 2016) for multi-sensory integration that will take into consideration both cognitive models of attention and the processing of working memory. We aimed at studying how the cognitive architecture responds in collaborative tasks between the iCub robot (Metta et al., 2008) and a human partner. We address the concept of joint attention emerging from a biologically-inspired multi-sensory selective attentional process defined as the selection of the relevant stimulus while ignoring irrelevant stimuli in the current environmental state (Nothdurft, 1991). With the goal of endowing an artificial agent with the ability to attend salient objects as humans do (accurate in location estimation and with optimal timing), we can promote emergent memory-based joint attention in collaborative scenarios. To evaluate the joint attention performance during unconstrained interaction and to exploit mutual influence between the parts, we compared human performance with the robot performance in a task in which both agents are exposed to the same salient audio or audio-visual

stimuli. In particular, we focused on decision making as our main contribution, and we then addressed perceptual performance (localisation accuracy and reaction time) during the task. Our main testing and performance analysis is structured around three main hypotheses: H1-Memory-based Decision Making Process: The memory-based cognitive architecture is able to attend to multi-sensory stimulation and correctly take a decision based on the localisation process; H2-audio-visual vs. Audio only: The stimulus localisation accuracy and reaction time of the robot in audio visual task is better than in audio only tasks; H3-Robot Performance: The performance (accuracy and reaction time) of the robot will be as good as the performance of the human participants in localising the stimulus.

In section 2, we give a high-level description of the cognitive architecture and we describe the details of all the different components developed for each cognitive architecture layer (section 2.1). Then, section 2.2 describes the experimental design that tests the performance of the cognitive architecture. In section 3, we describe the results of the experimental session, and in section 4 we discuss the main results, drawing at the same time, some conclusions on the performance of the proposed cognitive architecture.

2. MATERIALS AND METHODS

2.1. The Cognitive Architecture

We designed the cognitive architecture (see **Figure 1**) with three main goals in mind. The first goal was to build a multi-modal (audio-visual) attention computational system to facilitate joint attention between a robot and a human during an interactive task. The second goal was to address the accuracy-time trade-off in decision making inspired by human behaviour. The third goal was to improve the attention, decision-making, and action execution cycle by including a working memory component. The first goal relates to the audio-visual perception component while, the second goal concerns the decision making process. Finally, the third one addresses the role of working memory in the decision making process The cognitive architecture is composed of four main building blocks. In this section, we will explain in details the four blocks (Audio-Visual Perception, Decision Making, Working Memory, and Action Execution). The details will include the biological inspiration, the overall process, and the connections between the different blocks.

The perception block uses early features from both of the sensory inputs (the audio and the vision) to trigger the start of the decision making process. The decision making process modulates perception to meet the task requirements and further sends commands to the motor control for action execution. Finally, the memory governs the entire process and is shared between all of the units. We will also explain the technical implementation for each component of the cognitive architecture after mentioned the overall functionalities of the component. **Figure 2** outlines the structure and connections of our model's modules. Starting with the middleware, a software infrastructure that supports the integration of different cognitive modules, we used YARP Metta et al. (2010) (Yet Another Robot Platform) as our base. It is a multi-language middleware designed for robotic platforms. It is based on building multiple programs that run together in parallel and connect with peer-to-peer communication. We implemented our YARP modules using the C++ and python programming languages.

2.1.1. Audio-Visual Perception

To facilitate attending to auditory stimulus, we built the audio attention component based on an existing bio-inspired Bayesian audio localisation model (Kothig et al., 2019). The auditory attention component redirects the attention of the robot toward salient auditory signals. The system is based on the biological basis of how humans perform sound localisation. Humans use different cues to localise sound sources: the interaural time difference (ITD) and interaural level difference (ILD). Both are differently recruited by the auditory system to derive the direction of sound arrival. In our implementation we focused on the ITD cue as the principal computational method since there is a robust literature that uses ITD for sound localisation in artificial systems (Argentieri et al., 2015). The general idea behind ITD is to infer the direction of a sound from the difference in time of arrival (TOA) between the two ears. Different approaches have been proposed in robotics to compute the TOA, the most common one is based on correlation metrics (Hosangadi, 2019). This approach performs well but is sensitive to noise and reverberation, which is problematic, especially in presence of ego noise produced by robots. Other biological systems in nature use ITD cues to localise sound by employing either banks of coincidence detectors connected by delay lines, as in the avian brainstem (Jeffress, 1948), or more complex phase-tuned mechanisms as in the mammalian brainstem (Grothe et al., 2010). The audio localisation model used in this research modelled the spectral decomposition of the human basilar membrane with a Gammatone filterbank and model delay-tuned units in the auditory pathway as banks of narrow-band delay-and-sum beamformers. To further deal with the spatial ambiguities associated with interaural cues (Blauert, 1997), the model uses a Bayesian regression model that infers the location of the sound source using the previous results of the spatial localisation values. As a result the location is reliably estimated in robot's allocentric coordinate frame as a probability distribution of sound source locations across azimuthal angles. This probability distribution is used to create an allocentric saliency map of the sound locations.

Another important aspect of attention is selective visual attention which allows an agent to focus on salient points in a visual scene. It acts as a filter, discarding non-essential information and retaining only important information for further higher cognitive processing. Itti and Koch (Itti and Koch, 2001) proposed a computational model of selective visual attention based on Treisman's (Treisman and Gelade, 1980) Feature Integration model of human visual attention. This model uses bottom-up flows of information, which are combined into a unified saliency map (Itti and Koch, 2001). In the Feature Integration model (Treisman and Gelade, 1980; Ruesch et al., 2008), the bottom-up information is processed to extract visual features such as edges, intensity, motion, and chrominance. High saliency within one of these low-visual feature maps allows the

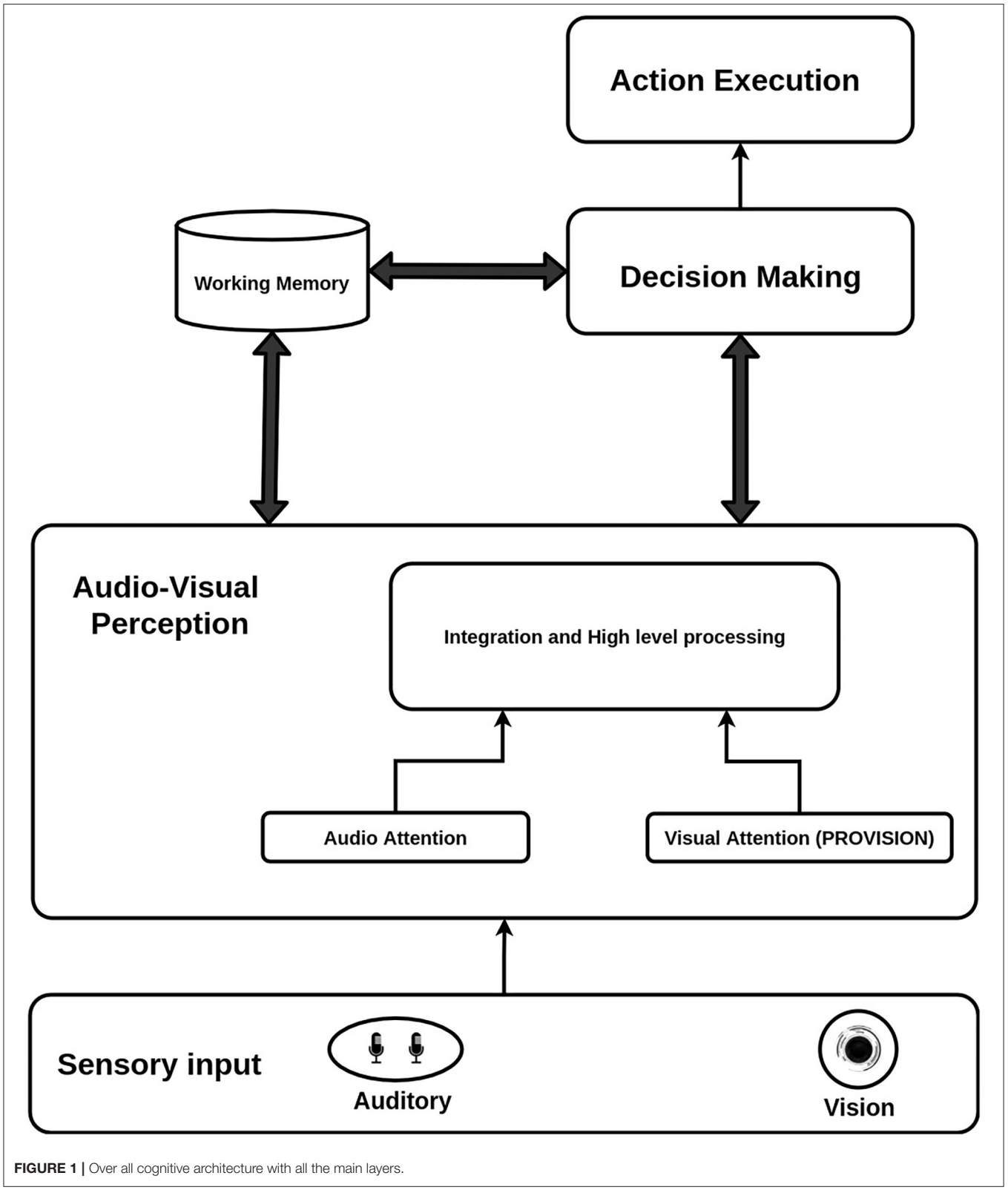


FIGURE 1 | Over all cognitive architecture with all the main layers.

model to orient the focus of an agent toward salient points such as colourful objects, geometric forms, or moving objects. Following this idea, in this work, we used the PROVISION attention model developed for the iCub robotic platform (Rea et al., 2014). PROVISION is an implementation of the attention model of Itti and Koch for the robot iCub (Itti and Koch, 2001). It provides

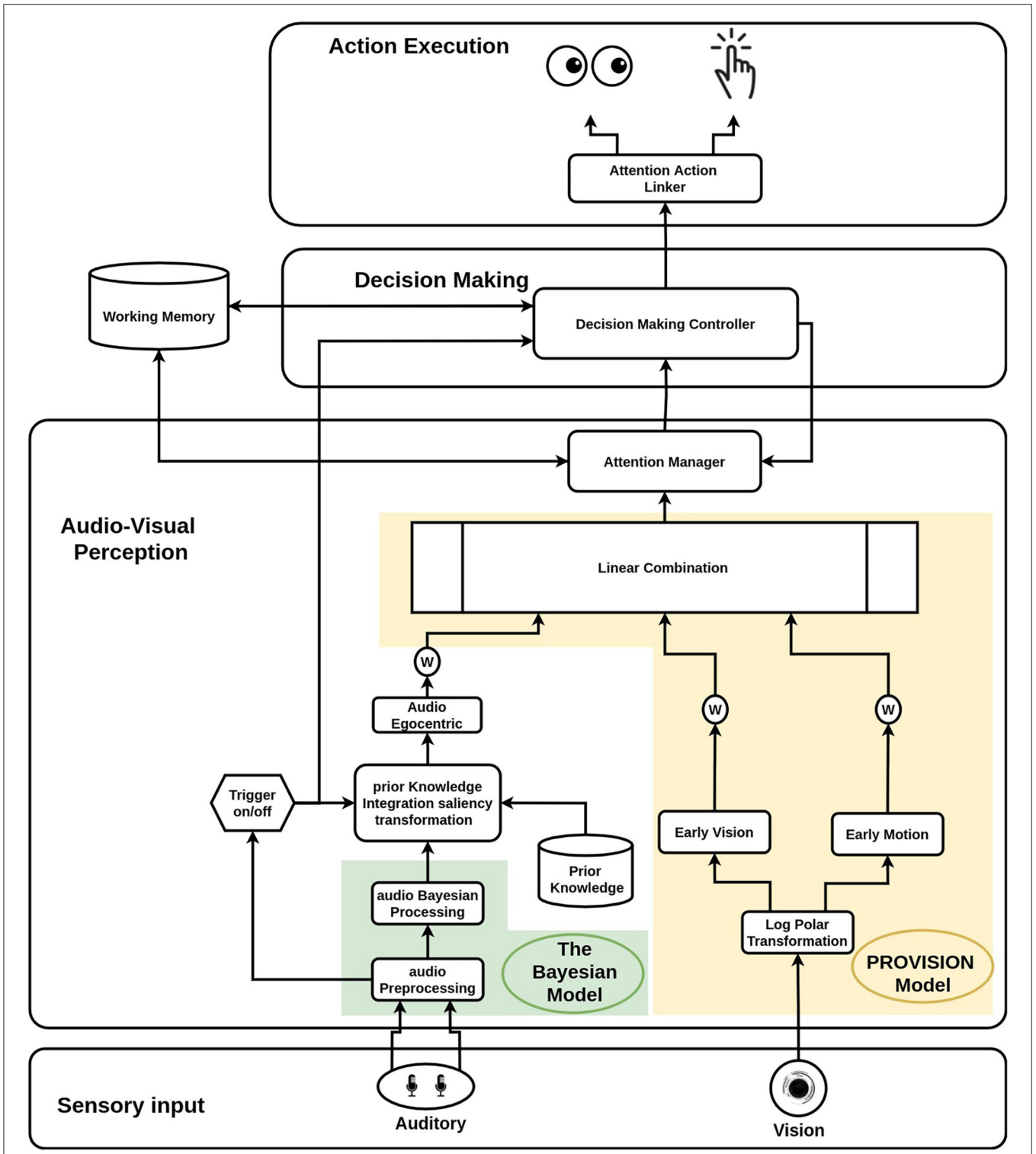


FIGURE 2 | Detailed representation of the system implementation architecture.

a modular tool for bottom-up attention, PROVISION integrates the different visual features with a weighted linear combination, enabling the ability to tune the importance of a particular visual stimuli, for example, forcing the attention toward a bright object by putting more weight on the intensity value. For the audio visual model, we had to implement an integration algorithm

where both visual attention and audio attention are aligned and have the same representation. This integration is designed to be processed in the integration and high level processing component of the audio-visual perception block. In this component of the architecture the auditory attention is integrated together with the visual attention system. We remapped the allocentric auditory map into a visual egocentric saliency map. The map is then added as a feature to the linear combination of the attention system (already developed in the visual attention PROVISION model Rea et al., 2014). The sound then reinforces the visual saliency map at the corresponding azimuthal location only if the source of sound is located within the field of view. The aim of this process is to provide a unified multi-sensory saliency map which enables identification of salient points from both auditory and visual signals. After sensory integration, the output of the integration process is a saliency integrated map. Next, the saliency selection process happens, in which the system selects the point the model needs to attend to. As found in other attention models (Ognibene and Baldassare, 2015; Baldassarre et al., 2019), we moved from the cyclical selective attention systems which are typically used in robots to a temporally asynchronous method for selective attention. We therefore implemented the temporal asynchronous attention at salient changes in the landscape of the perceptual sensors. This allows the system to resemble the asynchronous attentional redeployment of humans. This selection is performed on the integrated scene and based on a time variant threshold which is defined based on a confidence-urgency trade off from the decision making block. When the selection process is finished, the selected point is then processed by the decision making block. This is where the Audio-Visual Perception block is connected to the decision making block. It is also connected to the working memory, in order to update the perceptual states in the memory for a better memory based decision making process. In this process a confidence-urgency trade off is performed based on the state time and the stimulation states. More details concerning the decision making block will be discussed in the following part (Decision Making).

Another added component in the audio-visual perception is the integration of prior knowledge for audio perception. The prior knowledge is the spatial locations of possible stimulation sources. This knowledge influences the perceptual abilities of the robot. This process is inspired by biological evidence about the importance of the prior knowledge in decreasing cognitive load, improving learning abilities, and improving perception (Cook, 2006; De Lange et al., 2018).

In **Figure 2**, the PROVISION model is highlighted with a yellow background colour and the audio Bayesian model is highlighted in a green background colour. The following part of this section is explaining in details the implementation of the added components to the audio-visual perception block which was mentioned above in brief.

2.1.1.1. Trigger, and Prior Knowledge Integration

In order to overcome false positives coming from ambient sound in the environment, we integrated a power detection algorithm along with our sound localisation system as a relevant attentive mechanism in human audition (Rohl and Uppenkamp, 2012). We aimed to test the reliability of the sound power as an early informative feature. We added the calculations of the sound power in an early stage (audio prepossessing module) of the audio input. Using a fixed threshold on the total power for both audio channels, the system can determine whether the audio signal is high enough to be considered a valid sound or is just ambient noise. The threshold is autonomously extracted from the environment. The instantaneous sound power is used as an input to the trigger block. The trigger module receives the audio power processed by the audio prepossessing module. Based on a defined threshold for the instantaneous power, the trigger outputs signal to a higher level audio perception module (Prior Knowledge integration & saliency transformation) and also to the decision making block. Additionally, it updates the working memory which will be explained in a separate section.

Moving to the prior knowledge integration and saliency transformation module, we define two aspects of prior information for the audio stimulation. The first aspect is the possible locations of the stimulation. As the current audio system only considers the azimuth angle, this information is in a form of two lists. The first list is of angles describing where in azimuthal space the audio stimulation might be occurring and the second list is the spatial resolution of the angles, which reflects the size of the stimulation source. Thus for each stimulation source in the scene, we express the location in azimuthal allocentric angles from the robot's head axis as (X degrees $\pm$ resolution). These angles and their resolutions are the only locations that are considered from the allocentric probability map and the rest are ignored. The allocentric probability map is the output of the audio localisation model, which is a set of 360 values that represent the probability of the sound source's location at any arrival angle around the robot. These probabilistic values correspond to the 360 degrees centralised around the head axis. After considering the prior defined locations only, the resulting map is normalised to keep the Bayesian representation in the form of a probability distribution. By integrating this prior knowledge, we force the model to only focus on pre-biased defined locations. The second prior for the audio stimulation is the stimulation audio power. It is used to identify the threshold level of the sensitivity of the trigger. The trigger gives a high output if the audio power exceeded the threshold, which is the defined stimulation power level. Conversely, the trigger gives a low output if the audio power is less than this threshold. This signal is used to activate the transmission of the Bayesian map after adding the priors to the next stages. Otherwise, the transmitted map is a zero map. The trigger supports the prior knowledge module with the trigger signal to activate and deactivate the map transmission.

The next process is saliency transformation. The input of this process is the resultant Bayesian map after adding both priors (the stimulation activation level and the sources angles). The whole map is then multiplied by a total audio power and a scale factor. The audio power multiplication gives more importance to high stimulation than low stimulation (both are above the threshold level) and the scale factor transforms from Bayesian values (0–1) to the values of the monocular image (0–255).

2.1.1.2. Audio Egocentric

The input of this module is an allocentric audio map, created after biasing the possible locations of audio sources. The allocentric map is 360 values for the 360 degrees of the azimuth plane. On the other hand, the visual attention system is egocentric with a retinotopic reference. The camera moves as well as the head of the robot, and based on these movements the robot sees different parts in the space. The aim of this module is to align the allocentric output of the spatial auditory system with the egocentric spatial vision. To integrate the audio to the visual attention system we had to perform this alignment. To achieve this task, the module needs to know the current state of the locations of both the head and camera in the azimuth direction. The process of extracting the egocentric map is based on the current locations on the camera and head in azimuthal direction and the camera parameters. The camera parameters specify the width of the area of vision, while the location states of the camera and the head specify the middle value in the area of vision range. Knowing the middle angular value and the angular width of the sound source, the module computes the starting and ending degree angles which then are extracted from the allocentric map. This is the first stage of the audio egocentric module which has an output of a subset from the allocentric saliency map of the audio. The second stage involves scaling these values vertically and horizontally to be in equal size with the frame size of the visual image. The horizontal scaling assumes that the audio source is from the horizontal level in the scene as we only consider the azimuth plane in the audio localisation module. The output of the scaling stage is now ready to be integrated as a feature in the PROVISION attention system with a defined weight in the linear combination part.

2.1.1.3. Attention Manager

The attention manager is a central control module. It is responsible for analysing the combined scene from the output of the linear combination block of attention. The analysis is basically computing a confidence level. We propose a novel approach of recognising the unique target point of the scene to avoid continuous movement between different points. It is a measurement of the confidence level of uniqueness for the most salient point of the scene. We called this measure gamma value (Γ). The gamma value (Γ) represents how much the most salient point differs from the average salience across the scene. If the (Γ) value exceeds a threshold, then this point is identified as unique point of attentional interest. We call it a "hot point." Γ is computed by calculating how far is the saliency of the maximum point from the triple of the standard deviation:

$$\Gamma = max\_value - mean\_value - 3\sigma \quad (1)$$

Where σ is the standard deviation of the combined saliency image. The Γ value gives information about the confidence level of uniqueness. Higher values are more likely to be a unique target whereas low values mean that in the scene there are multiple salient points with similar level of saliency. When a unique target is recognised [(Γ) value is greater than the current confidence threshold], it sends the selected point to the next connected elements in the architecture which is the decision making controller in the decision making block.

Additionally, the attention manager block receives manipulation commands for the threshold value from the decision-making layer. The threshold here represents the level of the confidence in which action is required. Therefore, the attention manager here can be presented as a trigger that acquires an action execution process for that current scene from the decision making block. Also, the module is able to fully control the process of suspending and resuming the attention process as well as the linear combination parameters. To summarise this part, the attention manager presents the main control unit of attention. It has the ability to change the attention parameter. It receives commands from other parts in the system, and finally it communicates with the other parts of the system and sends combined information about the current scene.

2.1.2. Decision Making

From research theories elaborated on in the previous decade, visual processing in humans and animals triggers a decision-making mechanism in the form of a higher-level process, relying on the extraction of low-level features and properties from visual input (Vanrullen and Thorpe, 2001). This process is meant to evaluate the perceptual output properties and their relevance to the current goal and expectations.

Decision-making processes inspired by time-invariant models have been adopted for decades by the computational neuroscience community (Ratcliff and Smith, 2004). These models are based on a decision-making signal, which is triggered by a fixed threshold. The process integrates confidence over time and once the confidence reaches the fixed threshold, the decision is made and the signal is executed. Recent studies, Murphy et al. (2016), Ditterich (2006), Churchland et al. (2008), and Saaty (2007) have shown that the time dependency of the decision making process and the urgency of signals are invoked by humans. These findings show that humans may make decisions with different levels of confidence based on urgency. The more urgent the decision, the less confidence may be accepted. This urgency-based process allows humans to adopt time-variant pressure to execute actions (execution pressure) as a time-variant variable. The first study also showed the existence of neural gain modulation for urgency generation in humans, which implies the existence of a modulation signals. These signals are initiated to express urgency and modulate the confidence level.

Inspired by the biological evidence of the time-variant decision making processes, we propose a model for the multi-sensory decision-making process that recruits a time-variant decision-making signal. The model performs four main tasks. The first one is tracking the changes in the working memory to detect the state change of the stimulation. The second task is threshold manipulation based on the urgency. This second process is the main element which addresses the time-variant feature of the decision making block. The third task is analysing the relevance of the received spatial location within a predefined task by the experimenter. The experimenter should define the relevant working area and the required information to perform the projection. The last task is sending the action execution

signal to the action execution block based on the required actions which are also defined by the experimenter. These tasks are defined within three parallel processes. The first process aims to respond to the signal coming from the audio visual perception block that shows the presence of the stimulation and that the urgency of taking a decision should start. The second process is to predict the spatial location of the source of the stimulation from the 2D response of the audio visual perception. Finally, the third one is evaluating the relevance of this stimulation based on its 3D location. The first process works as the urgency trigger which starts a modulation signal for the threshold value of the confidence level for the localisation task. Once the confidence exceeds the threshold, and based on the defined task, the evaluation of the signal starts. If it is relevant to the task then the action is executed.

2.1.2.1. Decision Making Controller

The decision-making controller block is the module responsible to control the flow of decisions, manipulate the threshold of the confidence level in the attention manager, analyse the salient perception output based on the context and finally send the request to the action execution system. The control flow consists of two parallel processes. Each process has events that trigger behaviours. The aims of the first process is receiving the salient hot point from the attention manager, analysing the relevance of this point based on the task information, and finally sending action execution commands if it fulfills the action requirements. The second process is responsible for the control flow and manipulation of the threshold. The following part will explain the events and the behaviours for both of the processes.

In the first process, we have three events that set and reset behaviours. The first event is a trigger event from the audio stimulation. This event sets the decreasing threshold behaviour which sends commands to the attention manager to subtract a defined decreasing rate from the current threshold value. This signal is an urgency signal to the perception block. The second event is the action execution. If the action is executed, the action state in the working memory is set and the decreasing threshold behaviour is reset. The final event is the off trigger of the stimulation. This event sends a resetting signal to the attention manager to reset the threshold and to the action execution block to return the robot to the home position. The resetting signals have two different delays. The threshold reset signal is sent after 0.5 s after the off trigger of the stimulation. The home reset signal is sent after 4 s from the off trigger of the stimulation. These delays are chosen to maintain the stability of the system. The setting and resetting flags for the action, thresholds, and the stimulation are saved and recalled in the working memory which will be explained in the next section.

In the second process, there is only one event, which is receiving a salient hot point under the condition of the idle state. This event starts the evaluation of this point in the task context. The evaluation is the relevance of the 3D projection of this point to the predefined working area in the environment. Knowing the 2D coordinates of the hot point received from the attention manager and the equation of the plane of the working area, we calculate the 3D location in the environment. Based on the defined task, the decision is made whether to do the action or not, and which action to do based on the projected 3D location of the hot point. If the action is done then the action execution event is triggered.

As explained here the processes are parallel. However, they are interconnected, and both are dependent on each other. So the second process is only running when the robot is in the idle mode and the mode of the robot is controlled by the second process. And in the first process, there is a behaviour that is triggered by the second process which is action execution when the mode is changed by the second process. Following the assumption of ignoring the vertical component in the audio stimulation, we implement a function to force the vertical component of the 2D hot point to meet the location of the stimulation sources. This is done by estimating the vertical component given the current head altitude angle and the vertical field of vision. The robot identifies the stimulation source by calculating the distances between the projected 3D location and all the stimulation source. The source corresponding to the minimum distance is the winning location. Finally, the decision-making block sends an action execution command with information about the localised stimulation source to execute an action.

There is stored information related to the task and environment. This means that in this block, the task is defined with its requirement. The task is a defined action under a certain stimulation condition. The task related information is information about the stimulation conditions, the starting level of confidence of the stimulation, the modulation rate which defines the urgency-accuracy trade off, and finally the required action when the conditions are applied. On the other hand, the environment related information in the action execution layer is a higher level information. It includes the locations of the relevant stimulation sources, the working plane, and the action execution parameters. This information helps the robot to project the action from the 2D egocentric frame of the vision to the 3D world and execute it in a proper way. More information related to this section will be explained in the experimental setup section of the paper.

2.1.3. Working Memory

The concept of working memory has emerged in psychology literature as a broad set of mechanisms that explain this accumulation of perceptual information over time. Psychology researchers have shown the relationship between attention and working memory (Schweizer and Moosbrugger, 2004; Phillips and Noelle, 2005). They have shown the irreplaceable role of working memory in solving cognitive problems by maintaining some essential information for certain tasks that involve monitoring the environment. Based on this information, we added a working memory element in our model to endow the robot with this ability. The working memory in our model maintains essential environmental and internal states for understanding the current scenario and for executing the correct action in the defined task. In our working memory model there are two main memory components. The first one is the stimulation states and the second one is the actions

states. The stimulation states define whether the stimulation is currently on or off and track it, whereas the actions states define whether the robot is executing the action or has finished the execution or still hasn't executed it for the current active stimulation.

As shown in the **Figure 2** the working memory block is bidirectionally connected to both the decision making and perception components. In our implementation, we developed a state working memory. It stores the states of the stimulation, action, and confidence level to enable better interaction with the environment. The stimulation states define whether the stimulation is currently on or off and track it (for both vision and audio). The audio stimulation state is set based on the audio trigger, while the visual stimulation state is defined by the gamma value of the scene. If the gamma value exceeds the threshold, there is a visual stimulation. The attention manager block is responsible for maintaining the stimulation state. Whereas the actions states define whether the robot is executing the action or has finished the execution or still hasn't executed it for the current active stimulation. The decision making controller maintains the state of the action execution as well as the confidence threshold. The attention manager and the decision making blocks are recalling these states in their processes. The working memory block ensures a stable robotic behaviour for attention, decision making, and action execution cycle.

Another aspect of the working memory system is the habituation process. It is a perceptual stage necessary for the humanoid robot iCub to memorise the specific conditions of the environment, as well as details about the human partner. Habituation is a well-studied process in psychology and neuroscience. It is the simplest form of learning (Rankin et al., 2009). It is defined as the process of learning how to filter out irrelevant stimulation and focus only on the important stimulation. (Groves and Thompson, 1970; Wagner, 1979). It is an important biological process for an effective learning. In this work we implement a simple form of habituation which allows the robot to learn the baseline sensorial characteristics of the environment and of the human partner in order to properly compensate during the task.

From the implementation point of view, the cognitive architecture comprises of a habituation signal that is sent to the decision-making block. This signal changes the current task to calculate some parameters from the scene in a defined time period. This signal also informs the process that the stimulation will be presented, and it is required to see the effect of this stimulation and memorise it. When this signal is received, the decision-making block starts to analyse the scene and records the changes. More specifically the Γ value changes. After the defined time period for the habituation process, the initial threshold of the confidence is set by the maximum Γ value during the habituation process, minus a fixed value as a sensitivity zone. The initial threshold value is one of the relevant details in the human robot collaboration with the human partner. In particular, this threshold changes based on the visual environment, which includes the presence of the human subject.

2.1.4. Action Execution

The action execution block receives commands from the decision making block and then executes these commands by performing whole-body motor execution of a required action. The action is previously learned by the robot. The motor action execution is expecting an allocentric location in the working environment. By providing a reasonable assumption about the task, its context, and working area, we were able to define the attentive plane in a geometrical representation. Applying projection on this plane we estimate the allocentric representation of the required point. Based on the task, we assess the spatial relevance of this point and check if this point relies on the predefined working area of the current task. The implemented module for the action execution is called attention action linker.

2.1.4.1. Attention Action Linker

The attention action linker interprets the decision and executes the motor commands. The decision-making layer gives the command to the action execution layer with the result of the decision task. The linker also controls the motor action by enabling or by disabling it. The actions are predefined in the current task. In corresponding to the stimulation source there are two actions, the gaze action, and the point action. This part of the architecture is more task oriented. In this module, the response actions of the robot are defined based on the stimulus location. The main goal of putting this module in the architecture is to enable taking actions after finishing the perception process and making an attentional decision. In the Experimental part we will talk about the Implemented actions for the defined task in the experiment.

2.1.5. Incremental Approach

To sum up, our main contribution is the integration of: perceptual processes, working memory and its rule in attention, time-variant decision making, and finally the action execution into a complete cognitive architecture. Delving deeper into the details, the audio-visual perception has four main contributions. The first one is adding new modules on the top of the audio Bayesian localisation model to create an audio salient based allocentric attention representation. Secondly, the multi-sensory integration, by embedding the audio saliency map as another feature map in the linear combination of the PROVISION model. The third contribution, is the implementation of the asynchronous selection of the saliency. The last contribution in the perception block is the integration of the prior knowledge into the audio attention component to improve the localisation abilities of the robot.

For the decision making block, our contribution relies on the computational implementation of the time-variant threshold manipulation which addresses the confidence-urgency trade off in perceptual decisions. Our final contribution is the integration of working memory in the cognitive architecture, which is inspired by human cognition.

2.2. The Experimental Setup

We test our hypothesis by performing a joint human-robot attentional task in an unstructured environment. The rationale

behind the design of this experiment is the facilitation of the decision making process evaluation, the performance of the system in different stimulation modes (audio-visual vs. audio only), and finally, the comparison between human and robot performances. **Figure 3** shows the experimental setup. The robot is facing the human participant. In between, there is a table that has the stimulation board and a keyboard in front of the human participant. The stimulation board is approximately centralised between the robot and the human with 57 centimeters distance to both. The height of the chair where the participant sits is configured so that the human is on the same level as the robot. This height places the stimulation board within an optimal location for the field of vision for both the robot and the participant.

2.2.1. Participants

We conducted the experiment with 21 healthy participants (female: 14, male: 9) aged between 26 and 43 years old, with an average age equal to 30.5 ± 4. All participants voluntarily participated and signed an ethical and information consent approved by an ethical committee at San Martino Hospital in Genoa, Italy. All the participants work within the institution with no direct involvement to the research.

2.2.2. Stimulation

We built a stimulation setup which consists of four identical boxes. The boxes are placed horizontally on the same line. We noted the names of the boxes with respect to the robot's frame of reference: (FL) for the far left box, (ML) for the middle left box, (MR) for the middle right box, and (FR) for the far right box. Each box can produce both audio stimuli and visual stimuli. The visual stimuli are produced by a smart bulb. The smart bulb emits up to 800 luminous flux. We use red colour with the maximum luminous. The audio stimuli are produced by a three watt Bluetooth speaker. Both the bulb and speakers are embedded inside the box. The top layer of each box has holes where the light and sound waves can prppagate through, but that hide the smart bulb. The width of the box is 9 cm. The boxes are placed with 15 centimeter separation distance (center to center). Therefore, the distance that separates the boxes is 6 cm. We placed the stimulation boxes in this configuration with the given spacing to make sure that all boxes are within the direct field of view (the view with a zero yaw angle for the face) of both the robot and the human participant. Additionally, we made the task more challenging by minimising the distance between the boxes. As it is proven that human perception matches sound sources and visual sources for angles as large as 30 degrees apart (Jack and Thurlow, 1973). we selected a long distance as half of 30 degrees and a short distance as one fourth of these 30 degrees. This drove our choice for the configuration setup. We use a complex tone with a 1 KHz fundamental frequency and 3 harmonics for audio stimulation. The visual stimuli is a red light emitted from a smart bulb. The choice of the complex frequency and the red colour is because of their high saliency compared to other colours for the vision, and simple tone for the audio. This was chosen to ease detection for both the human and robot.

2.2.3. Task Description

The task for both the human and the robot is to identify the active stimulation box and react as quickly as possible. There are two types of activation for the stimulation boxes. The first type is audio only stimuli and the second type is audio-visual stimuli. Only one box can be activated at a time. The stimuli are activated for a fixed time (10 s). The time between rounds is also fixed at 10 s. The stimulation trials were distributed equally over the four boxes. So, each box was turned on 25% of all trials. Also, the stimulation types were distributed equally. 50% of the trials were auditory-only and the other 50% are audio-visual. Each box was activated for 8 trials, 4 of them were audio-only and the other 4 were audio-visual. The sequence of trials and the type of stimulation were randomised, but fixed across participants.

In the implementation section, we mentioned that the user defines the task for the robot and gives to the system the required information for the task and its environment. Therefore, we defined the task on the top of the attention system. The task is to localise the stimulation from a set of defined sources located horizontally in front of the robot. After localising the location, the robot should execute gaze action (to look to the stimulation source) and point action (to point with the arms to the stimulation source. We provided the robot with the environment related info which are the working plane where the stimulation sources are located, and the working area on this plane. Additionally, we informed the robot that the stimulation sources are in that defined area in space. Consequently any localised stimulation within this area is considered as relevant to the task. If the localised stimulation is outside this area, then the robot ignores it as it is irrelevant stimulation. Extra environment information was added to the robot here, including the stimulation sources count and location. After localising the 3D location of stimulation, the robot should identify the source of this stimulation from the defined set of sources. To sum up, the task is stimulation localisation which is estimated in the decision making layer. This task divided into 2 stages, the first stage is localising the stimulation within the 2D frame and the second stage is to check the relevance of this stimulation when the 2D location is projected into the 3D world. If it is relevant, then the robot will execute the action. The next section is describing the defined actions for the robot and also for the human participant.

2.2.4. Human/Robot Reaction

We placed a keyboard in front of the human participant. On this keyboard, eight buttons were highlighted in four groups. Each group consisted of two side by side buttons. The human participants were requested to react as fast as possible by pressing any of the two buttons within the buttons group, which correlated to the activated stimulation box. We decided to use two buttons in the keyboard to increase the pressing area in order to simplify the action and minimise the execution time. On the other hand, we defined two actions associated with each localised stimulation box. The first action is a pointing action using the arm, the hand, and the fingers while the second action is a gaze action using the head and the cameras (eyes) of the robot. For the right side boxes the robot will point to the selected box (FR,MR) using its right hand. Similarly, the left hand is used for the left side

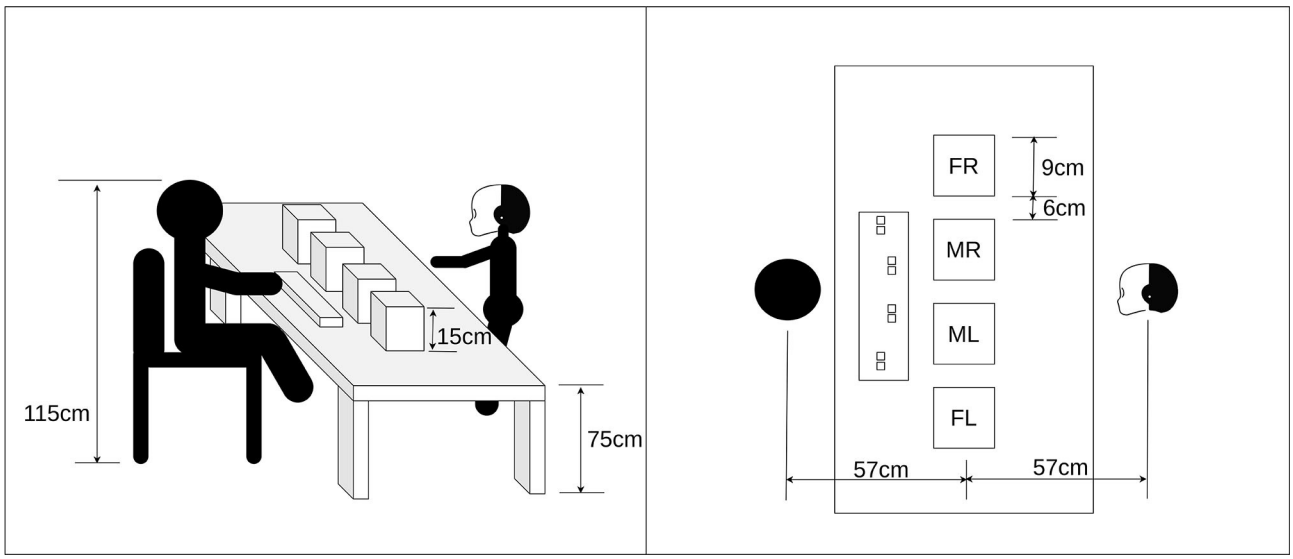


FIGURE 3 | Experiment setup showing the positioning of the robot and the participant. Also, the four stimulation boxes and their locations. Far left "FL," Middle left "ML," Middle right "MR," and Far right "FR."

boxes (FL,ML). For the gaze action, movements in head and the cameras are involved. The reaching action is biological and human-like movement that recruits not only the entire upper body of the humanoid robot iCub, but also the control of head and gaze of the robot. The gaze action brings the fixation point (line of sight) on the target with optimal coordination of the 6 degrees of freedom of head and eyes. The pointing with the index finger of the most opportune hand brings the robot to assume a new posture in less than 2 s. The coordination between head movement and upper body movement is designed in detail and makes the whole body movement look natural and human-like. It is possible that the human participant's attention is biased by this movement, but this is useful information in order to estimate the human-robot mutual influence in joint tasks.

2.2.5. Measurements and Rounds

The robot and the human do the task together at the same time. Before the first trial for each subject, we introduced the visual stimulation for the robot and the human. The robot performed the habituation process with the starting signal during this stimulation introduction period. Our first aim was comparing the performance of the robot vs. the performance of the human participants in terms of both accuracy and reaction time. In general, we were also interested in measuring how much one participant influences the other in human-robot collaboration. In order to measure accuracy and reaction time for the human participant we recorded the pressed keys and their correspondence to the target as well as the reaction time. For the robot accuracy and reaction time, we recorded the action execution commands of the robot and the internal triggering commands of these actions as relevant information about the timing and selected location. Additionally, we aimed to analyse all components of the decision making processes. Thus, we recorded the threshold profile (indicating the urgency to act) as well as the integrated scene analysis which includes the Γ value (indicating the confidence on target localisation process) during the whole trial. The second aim was to understand the behaviour of the human participants considering the presence of the robot. Specifically, in this experiment we focused on gaze behaviour. We recorded the gaze data during the whole experiment using Tobii pro glasses. This data includes the 2D gaze location within the field of view of the camera and the gaze event (Fixation/Saccade). This is the main data from the eye tracker that we focused on. For better analysis we developed a program to ensure synchronisation between the eye tracker time stamp and the time stamp from our system. The idea of the program is to send a timestamp instance from our system to the Tobii pro glasses, and in the analysis stage we map the timestamp of the eye tracker to our system's timestamp. The synchronisation process ensures the transfer of the trials' information to the gaze data. The trials' information mainly include the current state of the stimulation, the active box, the starting time of the trial, and the type of the stimulation.

3. RESULTS

We primarily focused on assessment of the performance of the memory-based cognitive architecture for joint attention. To perform an extensive evaluation of the system, we subdivided the analysis into two main sections. The first section is analysis related exclusively to the performance of the cognitive architecture. This includes the evaluation of the whole system dynamics which is mainly the decision making process and the overall performance (localisation accuracy and reaction time) by comparing it with human performance in a similar attentive challenge. The second parts of the results is a detailed analysis

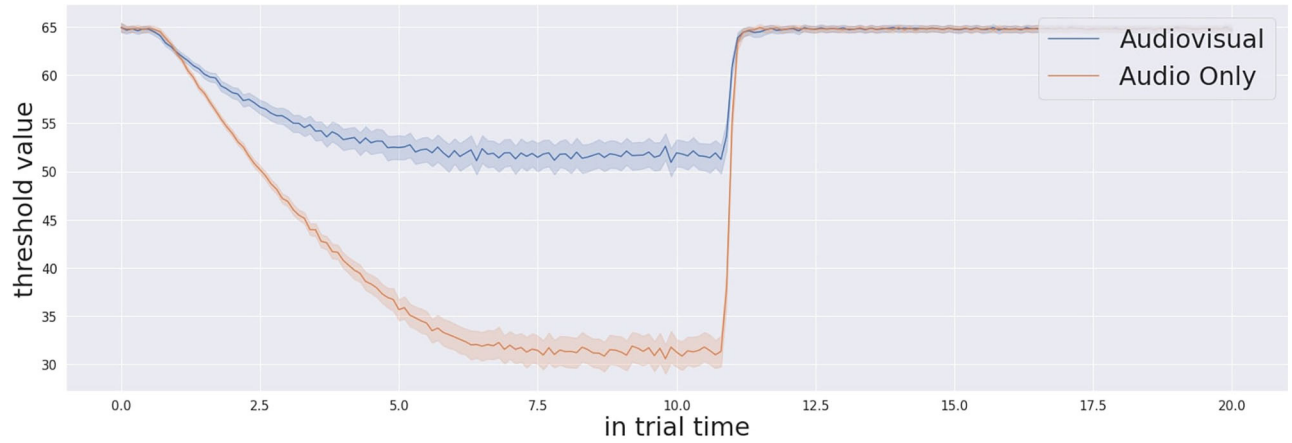


FIGURE 4 | Confidence threshold profile across the whole trial time (20 s) for both audio-visual and audio only trials.

of the gaze patterns. Given a thorough description of how the focus of attention was jointly redeployed, we focused our secondary analysis on the gaze patterns of both the robot and human participants. Such gaze behaviours are a direct result of attentional processing but more importantly tend to cause mutual influence between the robot and human. Humans tend to look where their partner directs their gaze (Frischen et al., 2007). Also, it is an important component in joint attention (Yu and Smith, 2013). So, the actions of the robot which are the gaze movement and pointing might influence the attention of the human toward a specific location. On the other hand, the gaze action of the human changes the visual features of the scene while the head moves. Consequently, this creates changes in the saliency map of the robot which might change its behaviour, and this what we want to analyse.

3.1. The Performance Analysis

3.1.1. The Memory Based Decision Making Process

We evaluated the memory-based decision-making process to report how the cognitive architecture makes the decision to act, averaged across all trials. The process is based on working memory, the confidence measure and the decision threshold (the threshold in which if the confidence reached, the agent will make a decision) as core factors of the decision-making process. The cognitive system makes the decision to act in presence of the event of crossing between the confidence measure and the threshold curve. Therefore, we analysed the decision-making behaviour to assess the effect of working memory as well as the performance of the confidence measure and the decision threshold, which are core factors of the decision-making process.

Adding working memory allowed the robot to track the stimulation state of the trial (presence of a stimulation), and the state of his own action (whether the action is done, or in progress, or not yet executed). This has a clear advantage with respect other work done in the recent past (Gonzalez-Billandon et al., 2019). Once the robot executed an action for a certain stimulus, it could realise that the task is done and there is no need to execute the action again until the current stimulus stops. This represents its internal working memory of the active motor actions. When the stimulus stops the working memory is updated, allowing the robot to reset and wait for another stimulus. Thanks to this mechanism the robot was successfully able to execute the action on the right time frame (after the stimulus turned on and before it turned off) in 95.8% of the trials.The working memory stabilises the action cycle and also allows the execution of the action based on meaningful environmental and internal states. This leads us to accept the first hypothesis, "The memory-based cognitive architecture is able to attend to multi-sensory stimulation and correctly make a decision based on the localisation process."

Moving to the analysis of the confidence measure and the threshold manipulation, **Figure 4** shows the average threshold profile with audio-only trials in blue and audio-visual trials in orange. The initial threshold is different for each participant. This is due to the habituation process, as the system memorises a different initial threshold for each participant. The process runs at the beginning of the experiment for each participant, because this initial threshold is dependent on the visual features of the environment including the human participant in the field of view. Thanks to the working memory, the robot retains important information of its task and this contextualisation is not only related to the environment. The starting time of the threshold modulation process is based on detecting the existence of the stimulation. Thus, the exact starting time of the modulation signal is different from one trial to another. Similarly, the confidence incremental process defines the action execution time together with the threshold decision. Therefore, the linear decreasing rate creates a curved, averaged response. After execution of the action, the decision making process slows down the threshold decreasing rate and this creates the flat part of the curve observed in **Figure 5**. In the audio-visual condition, the threshold decreasing rate slows down earlier. This is because the action is typically executed earlier due to the greater level of confidence in target localisation. After the multi-sensory stimulation stops (experimentally fixed in time after 10 s from

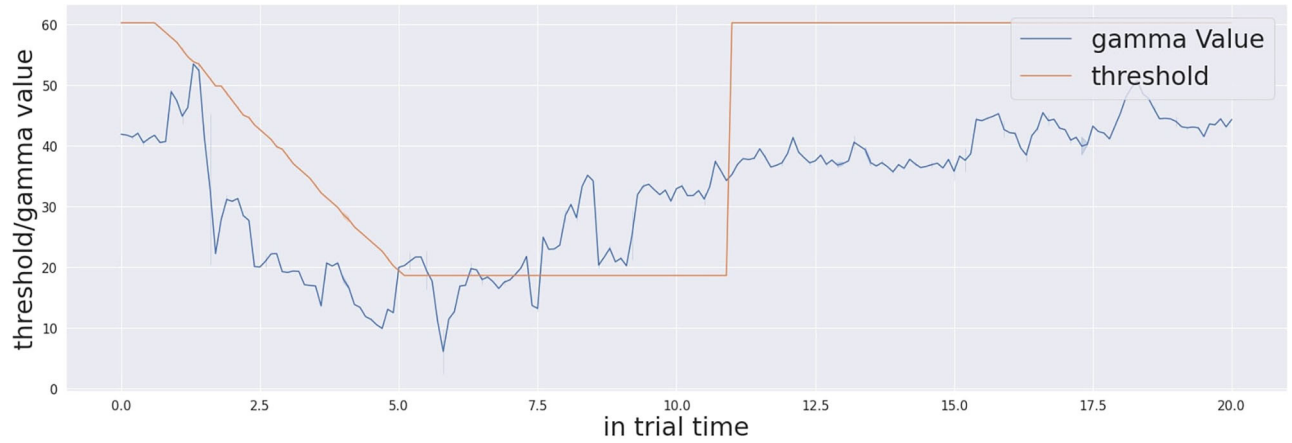


FIGURE 5 | Gamma measure ($\hat{W}$) and the threshold profiles in one of the trials across the whole trial time (20 s). The crossing occurs around 5 s.

the beginning of the stimulation), the threshold resets again to the initial value. In this exact moment, in the audio-only trials, the threshold starts from a lower value. This reflects the lower confidence and consequently the longer response time to take a decision to act. On the other hand, by looking at the Γ measure in **Figure 6**, we observe that the Γ function in audio-visual trials (orange curve) produces a spike almost instantaneously after the beginning of the stimulation. This is due to the visual saliency of the stimulation, which provides a strong, unique visual stimulation in the field of view. In the audio-only trials the Γ function shows that the confidence decreases at the beginning as causal effect of proactive sensing (the robot tries to eliminate the effect of the environment noise) and it starts to increase (after approximately 6 s in average) till the stimulation ends. When the threshold profile and the Γ measure cross one each other, the cognitive system makes a decision that triggers the action of pointing to the target stimulus.

It is also important to describe the decision-making process in detail by presenting an example trial. **Figure 5** shows a single trial taken from one participant. Once the simulation starts, the threshold of confidence starts to decrease in time with a decreasing factor from the initial value (the parameter is specific to the participant, computed during calibration, and kept in memory by the system). The level of confidence indicated by the Γ function and the threshold profile progresses in time under their proper temporal dynamics until the Γ value and the threshold cross each other. At this point, the cognitive architecture makes a decision and acts, by consequently pointing to the estimated source of stimulation. Once the stimulation ends (after 10 s from its beginning) the system waits 0.5 s and then resets the threshold to the initial value. The starting and stopping of the trial stimulation are autonomously detected by system based on the audio power in the audio signals received by both the microphones as presented here in **Figure 7**. The reset of the threshold profile to the original value occurs exactly 0.5 seconds after the end of stimulation is detected.

3.1.2. The Overall Performance (Accuracy and Reaction Time)

To assess the performance of the robot, we compared the attention system of the robot with human performance in response to the same multi-sensory stimulation and mutual sensorial influence. We analysed the overall performance based on (a) the reaction time and (b) accuracy as the primary source of evaluation. In particular, we characterised the performance based on the two stimulus typologies: audio-only stimulus or audio-visual stimulus. **Figure 8** shows the measure of the reaction time and accuracy for both the robot and the human participants, averaged across all the trials/participants. The bars in orange indicate the performance of the robot and the blue bars indicate the performance of the human participant. The participant and robot's choice is considered wrong if the identified box wasn't the active box or if the action didn't execute. Looking into the accuracy for each of the stimulus types separately, the robot records similar performance to the human in audio-visual attention tasks. The robot autonomously identified the source of the stimulation with 89% average accuracy. On the other hand, the robot performed with 43% average accuracy in the audio-only trials. The audio-only trials were more challenging for humans as well. To assess performance, we performed multiple t-tests to compare the behaviour of the human in the audio-visual task vs. audio only task, and similarly for the robot. The results of all the tests that demonstrated significant differences are the following:

- Human audio-visual reaction time vs. human audio only reaction time : $t_{(40)} = -3.7527, p < 0.01$.
- Human audio-visual accuracy vs. audio only accuracy: $t_{(40)} = 2.1436, p = 0.0382$.
- Robot audio-visual reaction time vs. robot audio only reaction time: $t_{(40)} = -9.6, p < 0.01$.
- Robot audio-visual accuracy vs. robot audio only accuracy: $t_{(40)} = 12.2, p < 0.01$.

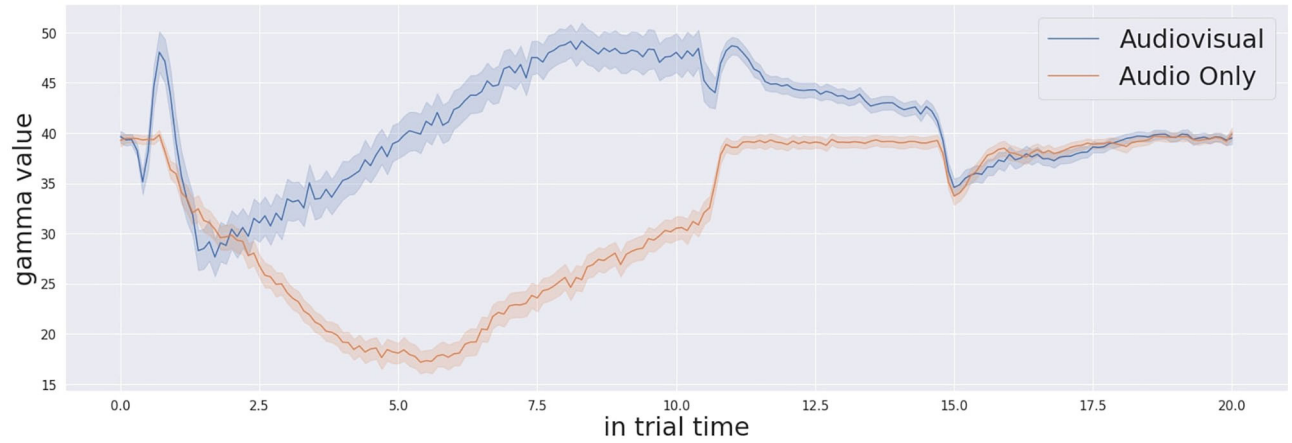


FIGURE 6 | Confidence profile (gamma measure Γ) across the whole trial time (20 s) for both audio-visual and audio only trials.

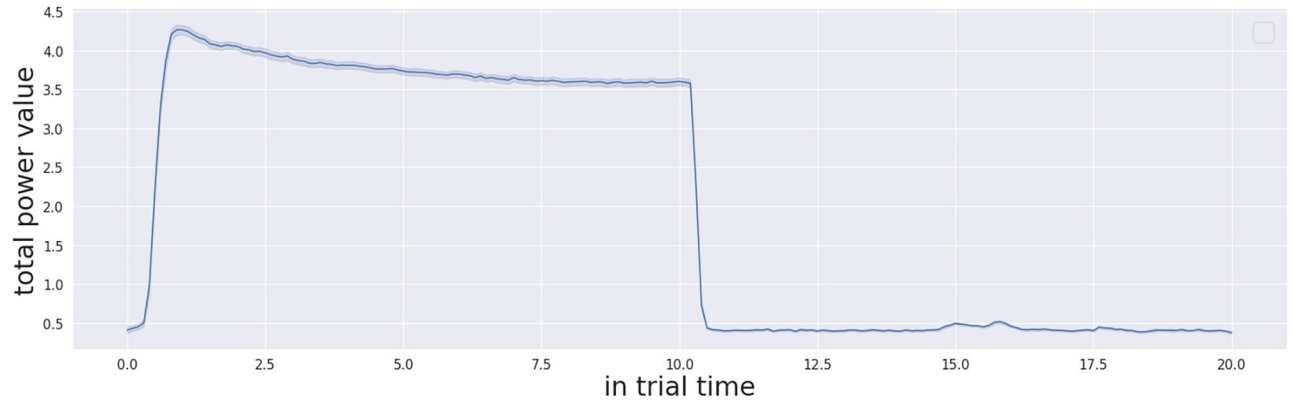


FIGURE 7 | Audio power profile across the whole trial time (20 s) during all the trials.

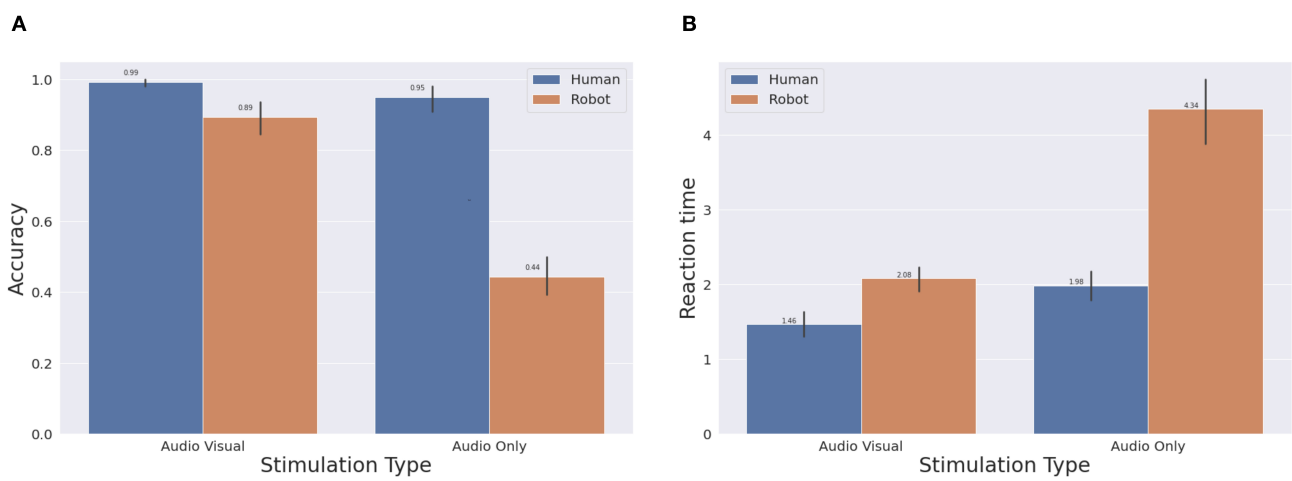


FIGURE 8 | Overall performance across the different types of stimulation for both the robot and the human participants. **(A)** Accuracy. **(B)** Reaction time.

So there are significant differences of both reaction time and accuracy between the audio-visual condition and audio only condition for the human participants and also for the robot. The differences in the case of the robot were all significant ($p < 0.01$). As the average accuracy value for audio-visual is higher, and the reaction time is lower compared to the audio only task (shown in **Figure 8**), we accept our second hypothesis that "The stimulus localisation accuracy and reaction time of the robot in the audio-visual task is better than in the audio only tasks."

We also performed t tests to compare the performance (reaction time and accuracy) of the robot vs. the performance of the human in the localisation task. The statistical tests showed significant differences between both performances as follows:

- Human audio-visual reaction time vs. Robot audio-visual reaction: $t_{(40)} = -4.99, p < 0.01$.
- Human audio-visual accuracy vs. Robot audio-visual accuracy : $t_{(40)} = 4.06, p < 0.01$.
- Human audio only reaction time vs. robot audio only reaction time: $t_{(40)} = -9.7, p < 0.01$.
- Human audio only accuracy vs. robot audio only accuracy: $t_{(40)} = 14.9, p < 0.01$.

Thus, we reject our third hypotheses The performance (accuracy and reaction time) of the robot will be as good as the performance of the human participants in localising the stimulus.

We did further statistical investigations using Wilcoxon signed ranked test (Rey and Neuhäuser, 2011) to test how different the performance of the robot was compared to the human. We found that the accuracy drop in the audio-visual condition is statistically less than 20% of the human accuracy. Also, the difference in reaction time of the robot in the audio-visual condition compared to the reaction time of the human is less than 1 s which is 70% of the increase in human reaction time. For the audio only condition, the difference was much bigger than the audio-visual condition. The differences in the audio-visual condition are comparable considering the complexity of the system and processing speed of the machine. The audio only condition is more complex compared to the audio-visual condition for both the human and the robot. However, the complexity of the audio-only localisation task does not entirely explain the considerable gap. To understand the reasons of this performance drop, we more thorougly investigated the conditions of wrong actions. The results are shown in **Table 1**. There are two conditions in which we consider the behaviour of the robot to be worse. The first condition is when the action is executed but the identification of the active box was wrong and is annotated with (wrong identification). The second condition occurs if the action never executed during the on time of the trial and we annotate this behaviour as (no action). For the human, all the wrong action trials were due to wrong identification. For the robot, the first condition occurred most of the time (89% of the total failures). On the other hand, there were two causes for no action failures. The first cause is when the robot executes an action in the off time of the stimulation due to some confusion from visual features in the scene. More specifically, it was observed that for some participants the robot got confused from the hand of the participant, indicating once again how mutual influence impacts attentive tasks. The hand worked as visual stimulation and the robot identified the closest box to the hand as a source of stimulation during the off time. If the robot executed an action during the off time, the robot does not reset the exception event before the end of stimulation of the next trial. The consequence of this is a (no action) failure for the trial next to the off time when the robot executed the action. This actually happened very few times (15 times) across all trials, which consists of 6% of the total failures. This is 60% of the second type of robot failures (No action failure). The remaining 40% of the no action failures are due to low confidence level. The robot did not execute an action few times because the confidence value (Γ value) never reached the threshold during the on time of the trial. This type of failure only forms 4% of the total failures.

Based on these analyses, the major cause of failure is wrong identification. Therefore, it is also important to analyse in detail the attentive process in time. More specifically, the audio components need to be analyzed, because the difference in performance lies in the temporal response of the attention system. So, in the next section we analyse the temporal responses of the audio probabilities, which are the base of the localisation process during the audio only condition.

3.1.3. Detailed Analysis of the Audio-Only Trials (Probability Profile)

Since the behaviour of the decision making process does not show erroneous behaviour, but instead the decision is made in the right time frame with a reasonable level of confidence, we believe that the reason for the worse performance in audio-only trials is to be found in the localisation process. As shown by a more detailed analysis for audio-only trials, the localisation process is based on the level of confidence that each box is the target, in other words the probability that each one of the four locations is the target. Such probability changes over time for each potential location of a stimulation source. In the audio-only condition, the probability profile is extracted from the Bayesian map, which is the output of the audio localisation system. The temporally detailed analysis of the probability profile is carried out during the 20 s time frame of the trials. During the first 10 s, the auditory stimulation is generated by the target box only.

Figures 9, **10** show the probability profiles for the 4 locations of the stimulation sources when the active box is the far left one and the middle left one respectively. The response is averaged across all trials. The first relevant point of these figure is that the shape of the curves are similar for the boxes located on the same side, independently from the location of the source of stimulation. In other words the probability profile over time of the far-right is similar to the one middle right and similarly the probability profile of the far left is similar to the middle left. Such results indicate that there are differences for time progression of the probability profile between the left and right boxes from the location where the robot is standing. Such difference has an impact on the localisation of the sound target since the certainty of sound location changes over time differently between the left and right boxes. Similar difficulty from one side over the other was actually reported by most of the participants. Another aspect

TABLE 1 | Robot's failure types percentages.

	Failure		Percentage from total failures (%)
Type 1	Wrong identification		89.4
Type 2	No action (Wrong action type in previous trial)	60%	6.4
	No action (low confidence)	40%	4.2

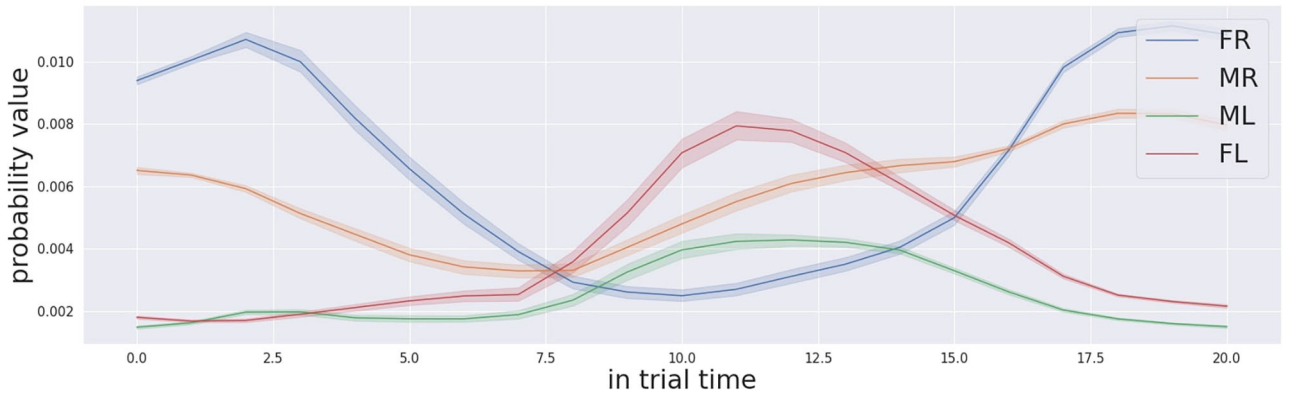


FIGURE 9 | Audio probabilities profile for the trials that the far left box (FL in red colour) was activated.

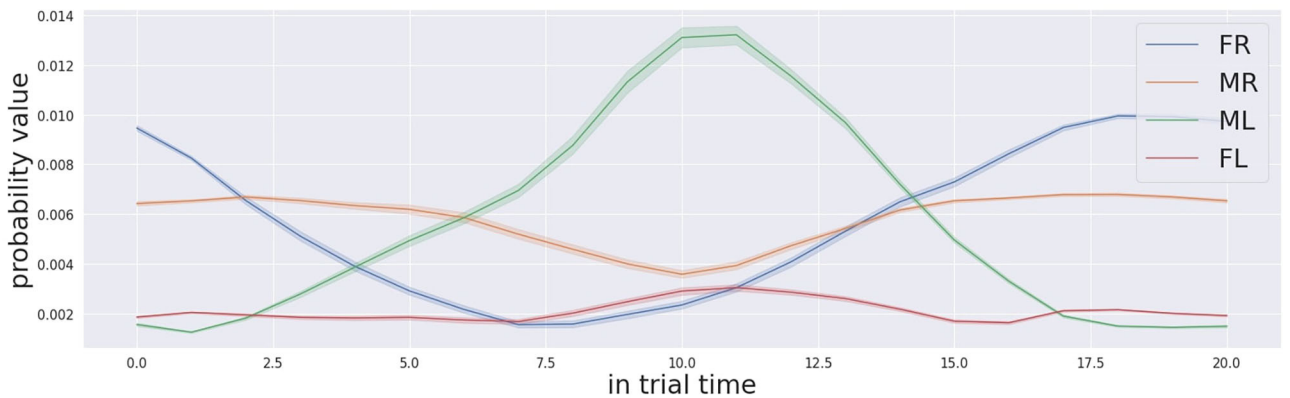


FIGURE 10 | Audio probabilities profile for the trials that the middle left box (ML in green colour) was activated.

that might have an impact on the localisation of the source of sound is that the probability profile of the sound sources from the same side evolve similarly. This makes the discrimination task complex for the robot, but also for the human participant. It was challenging for them to identify which box between the 2 boxes in the same side is the stimulation source in audio-only trials. The similarity between human robot participant in same-side during sound discrimination suggests that the Bayesian modelling implemented in the cognitive architecture shares some similarities with human behaviour.

Another relevant point relates to the temporal profile of the probabilities for the different salient locations. The probability corresponding to the right location increases with time as long as the stimulus is active, (in the first 10 s) which is the right and required behaviour. However, the probabilities of corresponding matches between the source of sound and different locations do not always start from zero and equal values. This indicates that before the activation of the stimulation, the localisation system believes that one location is more likely to produce sound than another location. Each probability goes to an initial value that is not equal to zero and also not equal to other location's probabilities. Our speculation explains the presence of these two phenomena as the result of acoustic noise in the environment. The acoustic noise equally affects the performance of the robot and of the human participant. It would be wise to remove the constant acoustic noise in the environment to eliminate its effect on the Bayesian map probabilities first, and then integrate evidence from the actual stimulation over time.

The final consideration regards the time the system requires to make the right decision. From both graphs, we observe that it takes in approximately 7.5 s for the far left box to be the box with the highest probability and 6.5 s for the middle left box. For the boxes located on the right side of the robot, the value for the middle right is similar and is approximately 7 s. For the far left box, the system struggles due to the noise, the probability for the far left never reaches the maximum when the box was activated within the on time frame (10 s). The decision-making process is tuned with some parameters to react faster than the required time. So the average reaction time of the robot for audio-only stimulus was measured to be around 4.34 s (STD: 1 s), definitely faster than the time necessary for the temporal probability profile to converge on the correct stimulation. Thus, we note that the attentive system can localise the target with a higher accuracy if the decision making process is allowed a longer reaction time. However given enough time, the auditory localisation process is always correct and the probability of the correct target always exceeds the probability of the others. For example, the audio probability profile for the far left box is the highest after 7.5 s. For the middle left box the audio probability profile for the middle left target is the highest after 6.5 s. Such fine refinement is actually doable in the cognitive architecture proposed, since by adjusting the tuning parameter we can refine the decision-making process and adjust the decreasing rate of the threshold.

In conclusion, we assessed that the task results are also difficult for the human participants according to an interview in the debriefing phase of the experiment. Another relevant observation in regards to the numerous comments of many participants indicates the change in the auditory landscape as the most meaningful cue to localise the target. The suggestion convinced us to looked in the change rate of the confidence level for the different possible targets. In the **Figures 11**, **12**, we show the change rate of the confidence probabilities for the four locations for the trial respectively when the target is FL and ML. We noticed that the attentive system can localise the target correctly in a shorter time if decision making process analyzes the change rate of the confidence probability instead of the confidence probability. For example, for the target in FL (see **Figure 11**) the correct detection of the target can occur as early as approximately 3.0 s, and for the target in ML (see **Figure 12**) the correction detection the target can occur as early as at approximately 2.5s.

3.2. The Behavioural Gaze Analysis of the Human and the Robot

The behavioural analysis of the human participants gives us a relevant insight on the mutual influence between the two partners. The behavioural analysis relies on data from the eye tracker. We were able to record the gaze data of the human participants. The gaze data is the 2D location of the gaze and the gaze event. The gaze events can be one of two types: fixation and saccade. We aimed to count the fixation events on the stimulation boxes and also on the robot's face during each trial. So, we had to define where the 2D location is projected in the 3D world. We are interested in 5 regions (the 4 stimulation boxes, the robot's head, and other areas). The eye tracker gives the 2D location of the gaze in the camera frame, which changes when the participant moves their head with respect to the world. In order to cluster the fixation events based on the 2D location into 6 clusters, we had to transfer the 2D location from the camera moving frame to a global fixed frame. We achieved this by extracting a reference point in the scene which always exists and then we track this point. This point works as a reference point and all interested regions are defined with respect to this point.

From the 21 subjects of the experiment, we could extract the gaze data perfectly from all 12 of them. Three subjects were moving their head very rapidly, and due to this the process of extracting the reference was not accurate enough. The gaze data of 5 participants weren't accurate enough to be considered because the eye tracker failed to calibrate their eyes. So in this section we only consider the data of the 12 subjects for which the calibration was accurate and the reference extraction process was sufficient. The robot behaviour in this experiment consists of its actions, which are the gaze movement toward the selected box and the pointing action with the arm. **Figure 13** is showing the fixation distribution in trials. It is divided into 4 panels based on the location of the simulation. (FL, ML, MR, and FR for top left, top right, bottom left, and bottom right panel respectively). The y axis shows the fixation counts. The x axis here is the five defined regions of interest (4 stimulation boxes and the robot's head). We also categorise it based on the stimulation type: audio only in blue and audio-visual in orange. Similarly, **Figure 14** shows the gaze of the robot. The robot only does one fixation event during each trial, which is the action of the task. So, the graph also represents the action distribution of the robot. The fixation counts on the active stimulation box is marked with a red rectangle surrounding the bars of this location in each of the panels for both the robot and the human participants. We divide our findings into two parts. The first part is for audio only trials and the second part is for audio-visual trials.

The first observed information is that in audio-visual trials the participants do fixation events on the active stimulation box more than other boxes in FL, ML, and MR trials. But in trials during which FR box was active, the participants do more fixation events on MR box on average. This drive us toward the second observation. Looking into the robot's gaze behaviour, we found that in the FR trials, the robot was confused toward the MR box and sometimes performs gaze actions toward the MR box instead of FR. The next three observations are in the audio-only trials. In the FL trials the robot mostly was driven toward the ML box. This records the highest average in comparison with the other boxes. Similarly, the participants also do more fixation events on the ML box, even more the correct active box which is FL. The second observation in audio only trials are in the ML trials. In these trials both the robot and the participants do fixation events on the correct box more than other boxes. Thirdly, in MR trials the confusion of the robot was between the right box (MR) and the ML box. But it is less than the confusion in FL trials. On the other hand the participants' gaze record the highest count on the right box (MR) and the second highest is the ML box. Finally, it is clearly shown that the participants also spend time looking to robot's head in all trials for all conditions.

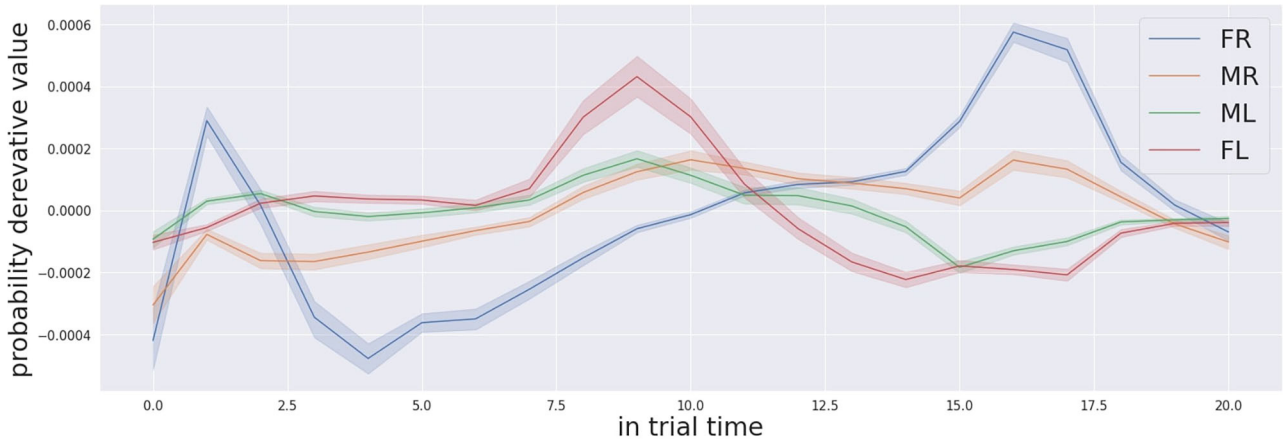


FIGURE 11 | Derivative of the audio probabilities profile for the trials that the far left box (FL in red colour) was activated.

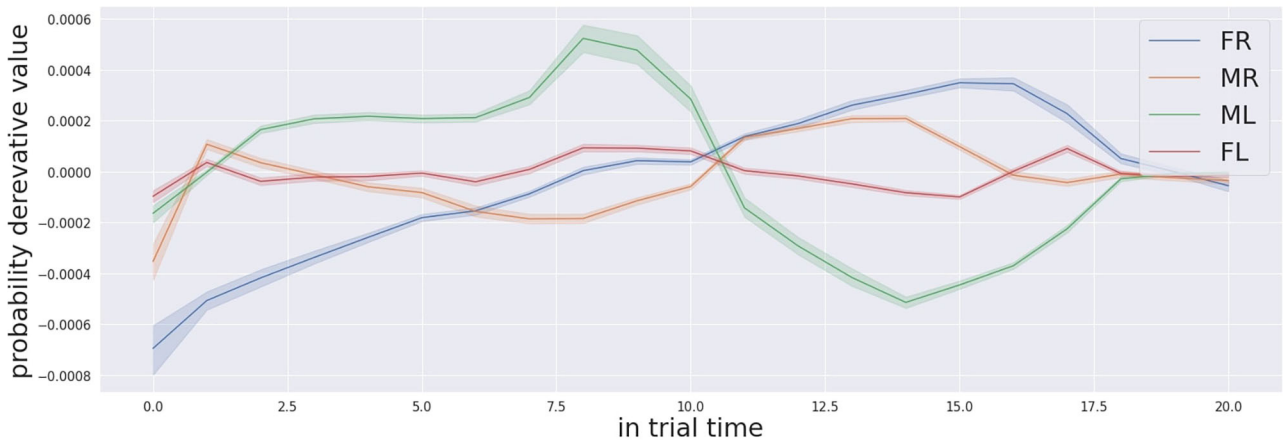


FIGURE 12 | The derivative of the audio probabilities profile for the trials that the middle left box (ML in green colour) was activated.

4. DISCUSSION

Joint attention is a fundamental component for better collaboration in real-world scenarios, such as in industrial environments where the robot and the human worker have to be aware of the products being manufactured (indicated by machinery through visual and audio features). They will be able to coordinate their actions and activities when initiated through their joint attention directed to the same target. The proposed biologically inspired cognitive framework, based on a multi-sensory attention system and supported by memory, constitutes the computational model used to evaluate emergent joint attention between the human participant and the artificial agent. The study had three main hypotheses. H1-Memory-based Decision Making process: The memory-based cognitive architecture is able to attend to multi-sensory stimulation and correctly make a decision based on the localisation process, H2-Audio-visual vs. Audio-only: The stimulus localisation accuracy and reaction time of the robot in the audio visual condition will perform better than in the audio only condition. H3-Robot performance: The performance of the robot will be as good as the performance of human participants. To answer the hypothesis we designed a multi-sensory task, and presented the task to the human participant and the robot. The setup includes stimulation boxes, which are a general model for real-world applications. Thus, we were able to compare the performance of the robot with the performance of a human participant in the same task which is an important aspect defining the quality of the interaction. The comparison focuses on the assessment of both agents in terms of the execution of the same localisation task with the same response time. The rationale behind the co-assessment of both the participants is that we intend to assess the performance of the robot and the human to measure how much they can coordinate in the joint task and to also measure the mutual influence between the robot and the participant.

The statistical analyses resulted on accepting the first two hypotheses (H1-Memory-based decision making process: The memory-based cognitive architecture is able to attend to multi-sensory stimulation and correctly make a decision based on the localisation process and H2-Audio-visual vs. Audio-only: The

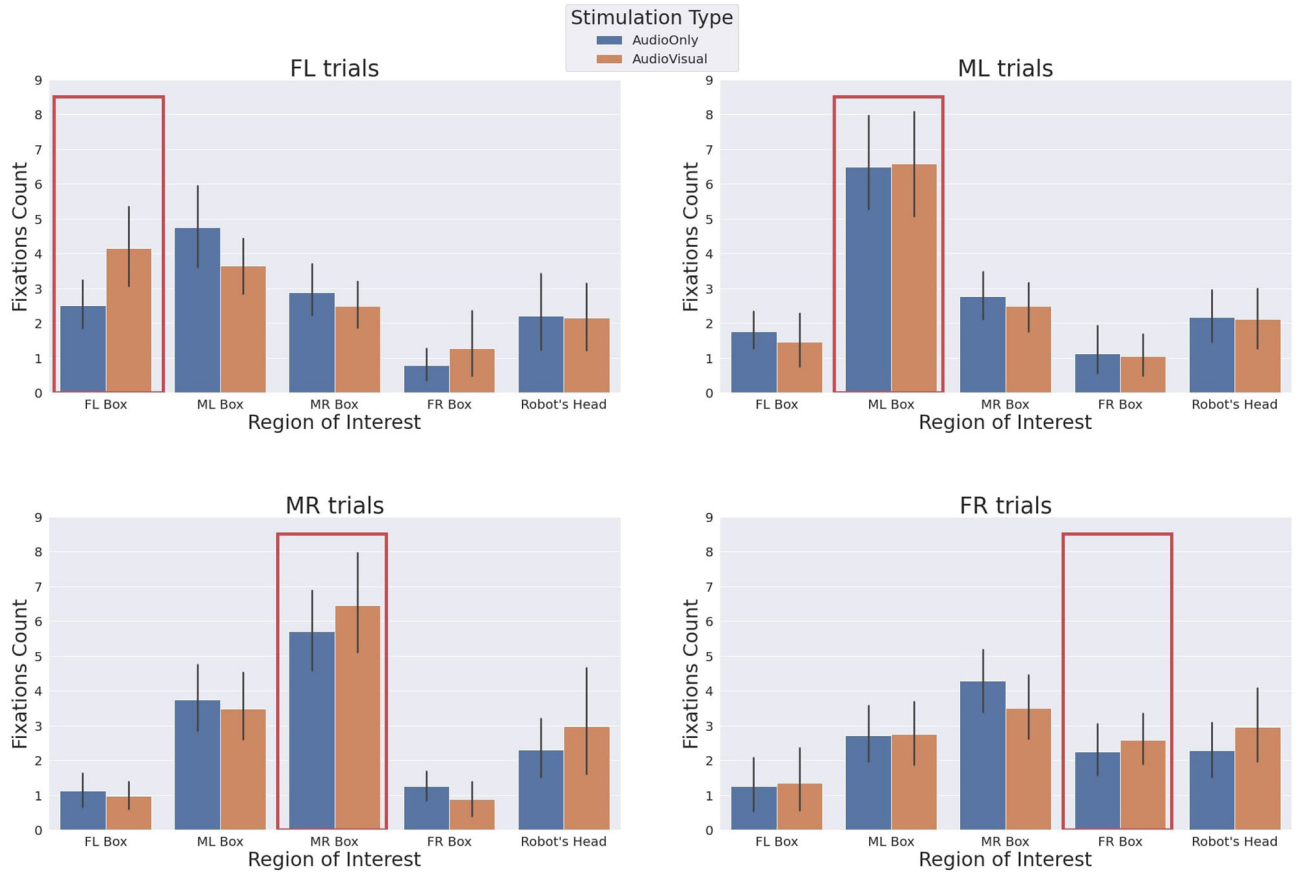


FIGURE 13 | The gaze behaviour of the human.

stimulus localisation accuracy and reaction time of the robot in audio visual task in better than in audio only tasks). and rejecting the last one H3-Robot performance: The performance (accuracy and reaction time) of the robot will be as good as the performance of the human participants in localising the stimulus.

However, further statistical analyses showed that the performance of the robot in the audio-visual condition is comparable, as the accuracy drop was less than 20% of the human accuracy and the reaction time differences were less than 1 s which is less than 170% of the human reaction time. These values are acceptable considering the machine processing speed of such complex computational processes. Indeed the cognitive system is less reactive in audio-only stimulation and only partially influenced in the different internal processes by the presence of the human partner. Although the audio only condition is in general a challenge for both the human and the robot participant, the analysis showed that the main cause of the performance drop in the audio only condition is the false audio localisation, which is caused by the acoustic egocentric noise.

Furthermore, we performed a more detailed analysis of the cognitive processes, and we realised that the decision-making process is robustly designed to swiftly guide the system to make a decision with excessively fast temporal dynamics. On the contrary, the auditory attention system requires longer time periods to make the Bayesian network converge, and thus localise the auditory target. Whereas the auditory localisation process is correct in inferring the location, also in presence of environmental noise (typical in robotic applications), the temporal dynamics of the system require longer periods for the processing of the auditory stimulation. However, the specific inefficiency is of simple resolution for two reasons that we intend to verify in future work. First, the specific modular structure of the developed cognitive architecture and its parametric configuration is designed to allow for fast re-adaptation of the decision-making process. As one possibility, by reducing the urgency to act parameter in the decision-making process, we can allow more time for the Bayesian network to converge, and consequently, we can guarantee improved accuracy. However, although the specific solution improves the accuracy it does not guarantee a faster reaction time. Secondly, thanks to the margin for faster response during auditory localisation, the process allows us to provide more auditory evidence for Bayesian integration in the same time interval. Faster processing of auditory stimulation is expected to improve the reaction time of the auditory localisation system and make it more similar to the reaction time of human participants.

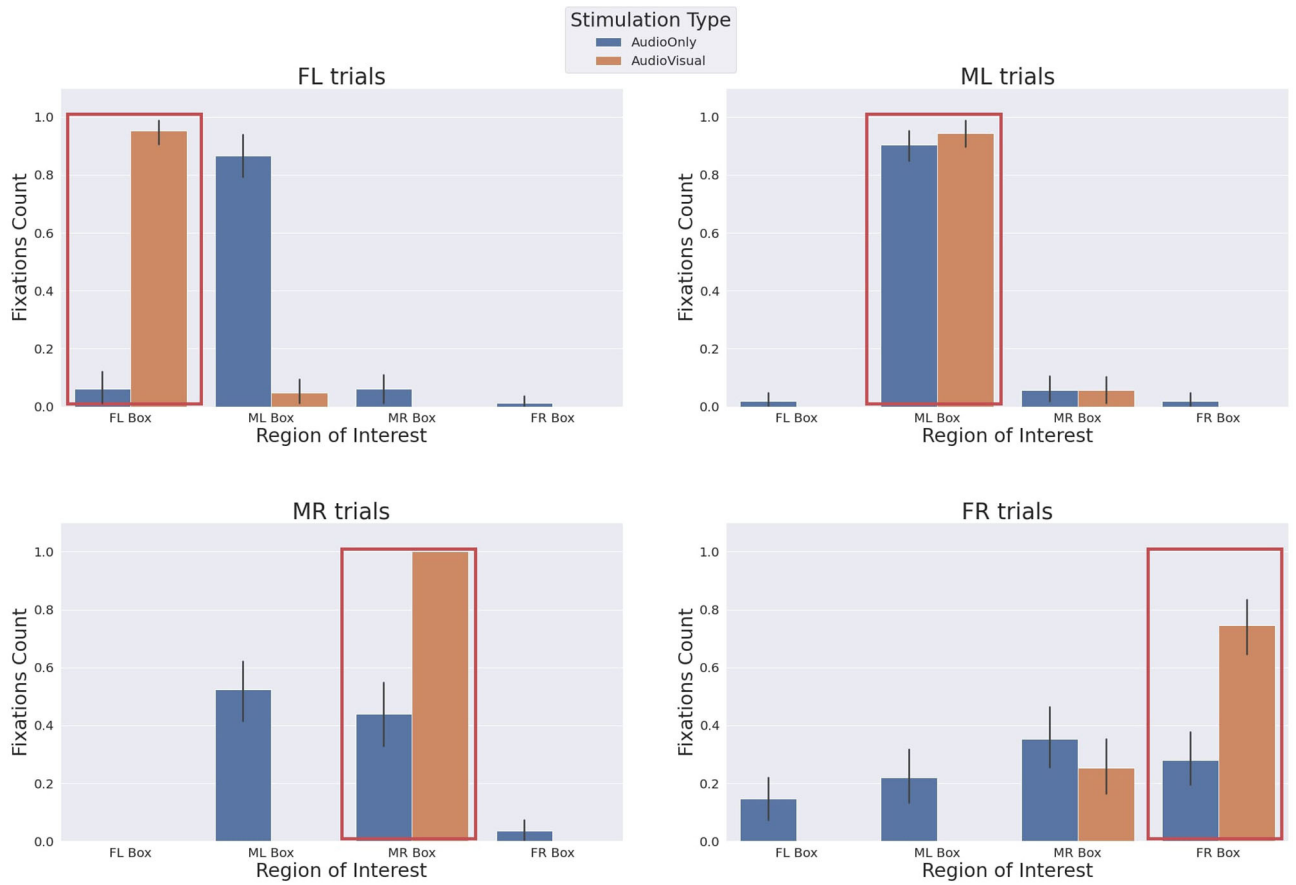


FIGURE 14 | The gaze behaviour of the robot.

Undoubtedly, the temporal dynamics of how auditory evidence is integrated is a very important aspect. We noticed in human participants that changes in the auditory landscape are more meaningful for target localisation than a static auditory landscape. The same process based on changes in the Bayesian network facilitates the process of inference over the stimuli localisation. The importance of relative changes in the auditory landscape, together with the importance of proactively creating such changes in the auditory landscape (self-programmed head movements) is a promising area of study, and we are planning to exploit it further in future work. Nevertheless, even without these improvements, the cognitive architecture has been demonstrated to be effective, and it shows a natural and robust joint attentive behaviour for Human-Robot interactive tasks. Furthermore, for a thorough understanding of behaviour related to mutual presence and its mutual influence, we also analysed the gaze behaviour of the human participants. The results showed that in the conditions in which the robot confused the location of the active box, the human participants tended to do more fixation events on the wrong box, suggested by the wrong behaviour of the robot. Also, the participants spend time looking at the head of the robot during the experiment, which shows how the human participant and the robot mutually influence each other in similar interactive tasks. This brings us to conclude that the behaviour of the robot may reinforce the gaze of the human toward the robot's chosen box. This is reinforced by the robot's behaviour which is both built on the directed gaze and the pointing actions. In the future, we intend to investigate this aspect further with more statistical evidence, and we intend to know whether this hypothesis of the mutual reinforcement is confirmed and what exactly drives it: whether the gaze or the pointing or a combination of both have a stronger effect on the human partner. Finally, we believe that the proposed system paves the way to human-robot collaboration, since coordinated joint attention is proven to facilitate coordination between the interacting parts. Such an optimal mechanism of coordination is considered one of the main facilitation mechanisms in multi-partner interaction tasks. We also showed that the robot affects the gaze behaviour of the participants. Furthermore, with this cognitive architecture, we demonstrate the importance of implementing a complete cognitive architecture (including working memory) in order to attend to salient targets in the environments as humans do. By sharing the same attentional focus redeployment mechanism with the human partner we provide effective joint attention that essentially emerges from environmental stimulation and reinforces natural human-robot collaboration.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by San Martino Hospital, Genova. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

OE designed and implemented the modules in the cognitive architecture whereas all authors equally contributed to the design of the cognitive architecture and the writing of the manuscript.

REFERENCES

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

Argentieri, S., Danes, P., and Soueres, P. (2015). A survey on sound source localization in robotics: from binaural to array processing methods. *Comput. Speech Lang*. 34, 87–112. doi: 10.1016/j.csl.2015.03.003

Baldassarre, G., Lord, W., Granato, G., and Santucci, V. G. (2019). An embodied agent learning affordances with intrinsic motivations and solving extrinsic tasks with attention and one-step planning. *Front. Neurorobot*. 13:45. doi: 10.3389/fnbot.2019.00045

Blauert, E. (1997). *Spatial Hearing: the Psychophysics of Human Sound Localization*. Cambridge, MA: The MIT Press. doi: 10.7551/mitpress/6391.001.0001

Bruinsma, Y., Koegel, R. L., and Koegel, L. K. (2004). Joint attention and children with autism: a review of the literature. *Mental Retardat. Dev. Disabil. Res. Rev*. 10, 169–175. doi: 10.1002/mrdd.20036

Churchland, A. K., Kiani, R., and Shadlen, M. N. (2008). Decision-making with multiple alternatives. *Nat. Neurosci*. 11, 693–702. doi: 10.1038/nn.2123

Cook, M. P. (2006). Visual representations in science education: the influence of prior knowledge and cognitive load theory on instructional design principles. *Sci. Educ*. 90, 1073–1091. doi: 10.1002/sce.20164

De Lange, F. P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception? *Trends Cogn. Sci*. 22, 764–779. doi: 10.1016/j.tics.2018.06.002

Ditterich, J. (2006). Evidence for time-variant decision making. *Eur. J. Neurosci*. 24, 3628–3641. doi: 10.1111/j.1460-9568.2006.05221.x

Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychol. Bull*. 133:694. doi: 10.1037/0033-2909.133.4.694

Gonzalez-Billandon, J., Grasse, L., Sciutti, A., Tata, M., and Rea, F. (2019). "Cognitive architecture for joint attentional learning of word-object mapping with a humanoid robot," in *IEEE/RSJ International Conference on Intelligent Robots and Systems, Workshop on Deep Probabilistic Generative Models for Cognitive Architecture in Robotics* (Macao).

Grothe, B., Pecka, M., and McAlpine, D. (2010). Mechanisms of sound localization in mammals. *Physiol. Rev*. 90, 983–1012. doi: 10.1152/physrev.00026.2009

Groves, P. M., and Thompson, R. F. (1970). Habituation: a dual-process theory. *Psychol. Rev*. 77:419. doi: 10.1037/h0029810

Hosangadi, R. (2019). "A proposed method for acoustic source localization in search and rescue robot," in *Proceedings of the 5th International Conference on Mechatronics and Robotics Engineering*, 134–140. doi: 10.1145/3314493.3314510

Itti, L., and Koch, C. (2001). Computational modelling of visual attention. *Nat. Rev. Neurosci*. 2, 194–203. doi: 10.1038/35058500

Jack, C. E., and Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Percept. Motor Skills* 37, 967–979. doi: 10.2466/pms.1973.37.3.967

Jeffress, L. A. (1948). A place theory of sound localization. *J. Comp. Physiol. Psychol*. 41, 35–39. doi: 10.1037/h0061495

Kothig, A., Ilievski, M., Grasse, L., Rea, F., and Tata, M. (2019). "A Bayesian system for noise-robust binaural sound localisation for humanoid robots," in *2019 IEEE International Symposium on Robotic and Sensors Environments (ROSE) (IEEE)* (Ottawa, ON), 1–7. doi: 10.1109/ROSE.2019.8790411

Mayer, J. S., Bittner, R. A., Nikolić, D., Bledowski, C., Goebel, R., and Linden, D. E. (2007). Common neural substrates for visual working memory and attention. *Neuroimage* 36, 441–453. doi: 10.1016/j.neuroimage.2007.03.007

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). The icub humanoid robot: an open-systems platform for research in cognitive development. *Neural Netw*. 23, 1125–1134. doi: 10.1016/j.neunet.2010.08.010

Metta, G., Sandini, G., Vernon, D., Natale, L., and Nori, F. (2008). "The icub humanoid robot: an open platform for research in embodied cognition," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, 50–56. doi: 10.1145/1774674.1774683

Miyake, A., and Shah, P. (1999). *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9781139174909

Moore, C., Dunham, P. J., and Dunham, P. (2014). *Joint Attention: Its Origins and Role in Development*. Hove: Psychology Press. doi: 10.4324/9781315806617

Mundy, P., and Acra, C. F. (2006). "Joint attention, social engagement, and the development of social competence," in *The Development of Social Engagement: Neurobiological Perspectives*, eds P. J. Marshall and N. A. Fox (New York, NY: Oxford University Press), 81–117. doi: 10.1093/acprof:oso/9780195168716.003.0004

Murphy, P. R., Boonstra, E., and Nieuwenhuis, S. (2016). Global gain modulation generates time-dependent urgency during perceptual choice in humans. *Nat. Commun*. 7, 1–15. doi: 10.1038/ncomms13526

Nagai, Y., Hosoda, K., Morita, A., and Asada, M. (2003). A constructive model for the development of joint attention. *Connect. Sci*. 15, 211–229. doi: 10.1080/0954009031000165510

Nothdurft, H. (1991). Texture segmentation and pop-out from orientation contrast. *Vis. Res*. 31, 1073–1078. doi: 10.1016/0042-6989(91)90211-M

Oberauer, K. (2019). Working memory and attention-a conceptual analysis and review. *J. Cogn*. 2:36. doi: 10.5334/joc.58

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. Mental Dev*. 7, 3–25. doi: 10.1109/TAMD.2014.2341351

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

Phillips, J., and Noelle, D. (2005). "A biologically inspired working memory framework for robots," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication* (Nashville, TN), 599–604.

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., et al. (2009). Habituation revisited: an updated and revised description of the behavioral characteristics of habituation. *Neurobiol. Learn. Mem*. 92, 135–138. doi: 10.1016/j.nlm.2008.09.012

Ratcliff, R., and Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychol. Rev*. 111:333. doi: 10.1037/0033-295X.111.2.333

Rea, F., Sandini, G., and Metta, G. (2014). "Motor biases in visual attention for a humanoid robot," in *2014 IEEE-RAS International Conference on Humanoid Robots (IEEE)* (Madrid), 779–786. doi: 10.1109/HUMANOIDS.2014.7041452

Reddy, V. (2005). Before the 'third element': understanding attention to self," in *Joint Attention: Communication and Other Minds: Issues in Philosophy and Psychology*, eds N. Eilan, C. Hoerl, T. McCormack, and J. Roessler (New York, NY: Oxford University Press), 85–109. doi: 10.1093/acprof:oso/9780199245635.003.0005

Repovš, G., and Baddeley, A. (2006). The multi-component model of working memory: explorations in experimental cognitive psychology. *Neuroscience* 139, 5–21. doi: 10.1016/j.neuroscience.2005.12.061

Rey, D., and Neuhäuser, M. (2011). *Wilcoxon-Signed-Rank Test*. Berlin; Heidelberg: Springer. doi: 10.1007/978-3-642-04898-2_616

Rohl, M., and Uppenkamp, S. (2012). Neural coding of sound intensity and loudness in the human auditory system. *J. Assoc. Res. Otolaryngol.* 13, 369–379. doi: 10.1007/s10162-012-0315-6

Ruesch, J., Lopes, M., Bernardino, A., Hornstein, J., Santos-Victor, J., and Pfeifer, R. (2008). *Multimodal Saliency-Based Bottom-Up Attention a Framework for the Humanoid Robot Icub*. Technical report. IEEE. doi: 10.1109/ROBOT.2008.4543329

Saaty, T. L. (2007). Time dependent decision-making; dynamic priorities in the AHP/ANP: generalizing from points to functions and from real to complex variables. *Math. Comput. Modell.* 46, 860–891. doi: 10.1016/j.mcm.2007.03.028

Schnier, C., Pitsch, K., Dierker, A., and Hermann, T. (2011). "Collaboration in augmented reality: How to establish coordination and joint attention?" in *ECSCW 2011: Proceedings of the 12th European Conference on Computer Supported Cooperative Work*, eds S. Bødker, N. O. Bouvin, V. Wulf, L. Ciolfi, and W. Lutters (London: Springer), 405–416. doi: 10.1007/978-0-85729-913-0_22

Schweizer, K., and Moosbrugger, H. (2004). Attention and working memory as predictors of intelligence. *Intelligence* 32, 329–347. doi: 10.1016/j.intell.2004.06.006

Shipstead, Z., Lindsey, D. R., Marshall, R. L., and Engle, R. W. (2014). The mechanisms of working memory capacity: primary memory, secondary memory, and attention control. *J. Mem. Lang.* 72, 116–141. doi: 10.1016/j.jml.2014.01.004

Tomasello, M., and Farrar, M. J. (1986). Joint attention and early language. *Child Dev.* 57, 1454–1463. doi: 10.2307/1130423

Treisman, A. (1996). The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178. doi: 10.1016/S0959-4388(96)80070-5

Treisman, A. M., and Gelade, G. (1980). A feature-integration theory of attention. *Cogn. Psychol.* 12, 97–136. doi: 10.1016/0010-0285(80)90005-5

Triesch, J., Teuscher, C., Deák, G. O., and Carlson, E. (2006). Gaze following: why (not) learn it? *Dev. Sci.* 9, 125–147. doi: 10.1111/j.1467-7687.2006.00470.x

Vannucci, F., Sciutti, A., Jacono,M., Sandini, G., and Rea, F. (2017). "Adaptation to a humanoid robot in a collaborative joint task," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* (Lisbon), 360–365. doi: 10.1109/ROMAN.2017.8172327

Vanrullen, R., and Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *J. Cogn. Neurosci.* 13, 454–461. doi: 10.1162/08989290152001880

Wagner, A. R. (1979). "Habituation and memory," in *Mechanisms of Learning and Motivation: A Memorial Volume for Jerzy Konorski*, eds A. Dickinson and R. A. Boakes (Jersey City, NJ: Erlbaum Hillsdale), 53–82.

Yu, C., and Smith, L. B. (2013). Joint attention without gaze following: human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PLoS ONE* 8:e79659. doi: 10.1371/journal.pone.0079659

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.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership