Multi-modal robotic architecture for object referring tasks aimed at designing new rehabilitation strategies

Chiara Falagario[1,†], Shiva Hanifi[2,†], Maria Lombardi[1,*] and Lorenzo Natale[1,*]

[1] *Humanoid Sensing and Perception Group, Istituito Italiano di Tecnologia, Genoa, Italy*
[2] *Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg*

Abstract

The integration of robotics and Artificial Intelligence (AI) in healthcare applications holds significant potential for the development of innovative rehabilitation strategies. Great advantage of these new emerging technologies is the possibility to offer a rehabilitation plan that is personalised to each patient, especially in aiding individuals with neurodevelopmental disorders, such as Autism Spectrum Disorder (ASD). In this context, a significant challenge is to endow robots with abilities to understand and replicate human social skills during interactions, while concurrently adapting to environmental stimuli. This extended abstract proposes a preliminary robotic architecture capable of estimating the human partner's attention and recognizing the object to which the human is referring. Our work demonstrates how the robot's ability to interpret human social cues, such as gaze, enhances system usability during object referring tasks.

Keywords

attentive learning architecture, visual-language model, object referring, social assistive robotics, rehabilitation training

1. Introduction

The use of social assistive robots in Healthcare is rapidly expanding due to their potential to support individuals with special needs and enhance engagement during rehabilitation sessions, leading to improved therapy outcomes [1, 2, 3, 4]. In this context, the ability of robots to understand human mental status has a crucial and pivotal role in designing new rehabilitation strategies to assist frail people and patients. Designing and implementing a robust robotic visual system capable of perceiving and interpreting typical human social cues is essential for enabling natural and effective interactions between humans and robots. Visual perception enables the robot to understand the surrounding environment, anticipate human intentions, and help them appropriately even with a simple task (for example, reach and grasp an object). The availability of such technologies will open the possibility to offer rehabilitation plans that are personalised to each patient and that can best fit individual needs.

Among the multitude of social cues characterising human-human interactions that can be endowed in an assistive robot, attention and referring understanding are crucial abilities for any task-oriented interaction, raising great attention in the computer vision community [5, 6, 7]. Referring understanding tasks aim at localising objects (or regions of interest) in images or videos by using natural language description as input by humans. However, in a real-world scenario, the referring expression could be ambiguous or incomplete. For example, an ambiguous referring expression can be *"Could you pass me that cracker-box, please?"* if there are more than one cracker-box in the scene. In this case in order to improve the referring accuracy, gaze signal can be used together with the natural language as complementary cue (people often utilise gaze to confirm the referred target while interacting).

Workshop on Advanced AI Methods and Interfaces for Human-Centered Assistive and Rehabilitation Robotics (a Fit4MedRob event) - AIxIA 2024

[*] Corresponding authors.

[†] These authors contributed equally.

chiara.falagario@iit.it (C. Falagario); shiva.hanifi@uni.lu (S. Hanifi); maria.lombardi1@iit.it (M. Lombardi); lorenzo.natale@iit.it (L. Natale)
0009-0007-9381-8985 (C. Falagario); 0009-0001-9719-7342 (S. Hanifi); 0000-0001-5792-5889 (M. Lombardi); 0000-0002-8777-5233 (L. Natale)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3932/paper4.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Having a multi-modal attentive robotic system able to integrate natural language with the social cue of gaze can be a valuable tool, especially in rehabilitation from social disorders like Autism Spectrum Disorder (ASD). Studies suggest that children suffering from ASD prefer robots to interact with exhibiting increased engagement, specifically human-like verbal-featured robots, since they are more predictable and with more controlled visual stimuli [8, 9, 10, 11]. This suggests that robots can be effective tools for assessing and potentially improving social interaction and communication abilities in children with ASD. Children with ASD may experience challenges with both verbal and nonverbal skills. For example, some children may be very limited in communicating using speech or language, and some may have difficulties in establishing the correct visual focus of attention [12, 13].

The work presented in this extended abstract is part of a broader project aiming at developing new robot-assisted rehabilitation strategies for children with neuro-developmental disorders based on face-to-face human-robot interactions involving manipulation of physical objects. Within the scope of the project, the considered training protocol consists in the child and the robot collaborating to fulfil a shared task, like a pick and place objects or handle and pass to each other a series of different objects. In order to make the robot aware of the object of interest while interacting also with children with reduced communication skills, the proposed robotic perception system has been designed to address object referring tasks by integrating language description with the human attention estimation. Specifically, the system takes in input an image with a caption in natural language and gives in output the object the human is referring to. Combining verbal with non verbal cues in one multi-modal architecture, the robot can understand the object referred by the human even with incomplete or ambiguous description, increasing its usability and helping to perform the task in a more efficient way.

In our study, we chose to use the humanoid robot iCub [14]. Its design strikes a balance between being sufficiently human-like and avoiding the uncanny valley effect (see [15]), which can occur with too human-like android robots [16]. Studies presented in [17] have shown that children with ASD respond well to the iCub robot, making it an ideal choice for our research.

2. Related works

Very few learning architectures exists in the current literature addressing the problem of object referring by combining natural language with additional inputs. Among them, Vasudevan et al. [18] proposed a multi-modal architecture combining the text description with different input sources such as gaze estimation, optical flow for motion feature and depth map. However, not all the aforementioned input sources are always available if considering different application scenarios. For example in the considered rehabilitation scenario, the iCub humanoid robot is equipped with low-resolution only RGB camera making the depth estimation from the image a challenging task. The work proposed in [6] overcame the problem in [18] combining the text description only with the gaze signal reaching even higher object referring accuracy. However, the proposed pipeline was designed to detect human attention targets while using looking images on screens-based devices, like tablets and smartphones. This scenario does not align with the conditions of a rehabilitation session, where the child and the humanoid robot are required to interact online on a collaborative task. To overcome the aforementioned limitations and meet the needs of a rehabilitation setting, the framework proposed in this extended abstract is specifically designed to run online on a robotic platform like iCub while using only RGB information coming from the cameras.

3. Attentive robotic architecture for object referring tasks

The proposed system is composed of two main blocks, each one based upon a different computer vision architecture: a human attention model, designed to estimate the human target of attention (Gaze), and an object detection model (MDETR - Modulated Detection Transformer [19]), responsible for detecting and recognizing objects in the scene. For that, we refer to our system as GazeMDETR.

Human attention estimation. The human attention model is responsible for estimating the focus of a human's gaze in a given scene. In this study, we use the fine-tuned VTD (Visual Target Detection) as proposed in [20], which provides a more comprehensive gaze target distribution within the scene. This refinement is particularly suited to tabletop scenarios, a common setting in healthcare applications. The work was based on the Visual Target Detection (VTD) system [21], which uses a spatio-temporal architecture to predict gaze targets in real-time video streams. VTD combines both head orientation and scene features by leveraging a EfficientNetB5 convolutional network as a feature extractor, enhanced with an attention mechanism. Specifically, the module takes in input the image and the human face bounding box (extracted by using [22, 23]) and provides as output an attention heatmap representing the image area that more likely contains the target of human attention. The returned heatmap is an image-sized matrix, where each cell corresponds to an image pixel. The value of each cell ranges from 0 to 1 (respectively, the lowest and the highest probability to be –or to be close to– the target of human attention).

Object detection. The object detection model is based on the MDETR [19], end-to-end framework detecting objects within images conditioned on natural language text given in input, such as captions or questions. Briefly, MDETR uses a combination of convolutional neural networks (CNNs) and transformer-based encoders to fuse visual and textual data, allowing the model to align objects with free-form text descriptions. MDETR is able to detect nuanced concepts from free-form text, and generalizes to unseen combinations of categories and attributes. MDETR has been pre-trained on large multi-modal datasets, and then fine-tuned in order to solve different downstream tasks, such as phrase grounding, visual question answering, referring expression detection and segmentation. In this work, MDETR is used with the reference to *referring expression detection* task (i.e., given an image and a referring expression in plain text, the system returns the bounding box around the referred object).

Combining attention with object detection. GazeMDETR integrates the *Human attention estimation* module and the *Object detection* module in one multi-modal architecture, as shown in Figure 1. Specifically, in order to merge the gaze information in the object detection, the attention heatmap produced by the Human attention module was first downsampled to match the dimensions of the feature map produced by the MDETR backbone, and then normalised in the range of $(0.5, 1)$. The resulting heatmap was finally multiplied with the convolutional features map (Figure 1). By integrating the gaze information from the VTD module with the object detection capabilities of MDETR, GazeMDETR provides a more context-aware detection framework. The fusion of these two systems enables GazeMDETR to detect objects within complex scenes while also inferring the primary focus of human attention. This means that the model is able to prioritize relevant objects based on the social cue of gaze (also in cluttered scenarios), offering enhanced accuracy in object detection tasks.

4. Methods and Preliminary results

In order to evaluate the performance of the proposed system, a testset was collected having different human participants looking at several objects in different cluttered scenarios. The same testset was then used also to make a comparison between our system and MDETR (used as baseline).

Data collection. A total of 4 participants were involved for the data collection (2 females, 2 males, age: mean 27, sd 3.54). All participants had normal or corrected normal vision and provided written informed consent. The data collection was conducted using the camera of the iCub robot [14] positioned on one side of a table, while the human participant stands on the other side. On the table were placed up to 11 objects chosen from the YCB dataset [24] together with regular office objects, thus to increase the difficulty of the task. The participants were instructed to look at the requested object in a natural and spontaneous manner. For each session and for each trial, each object was gazed at for 5 seconds by the participant. Each participant completed three recording sessions, each one characterised by a

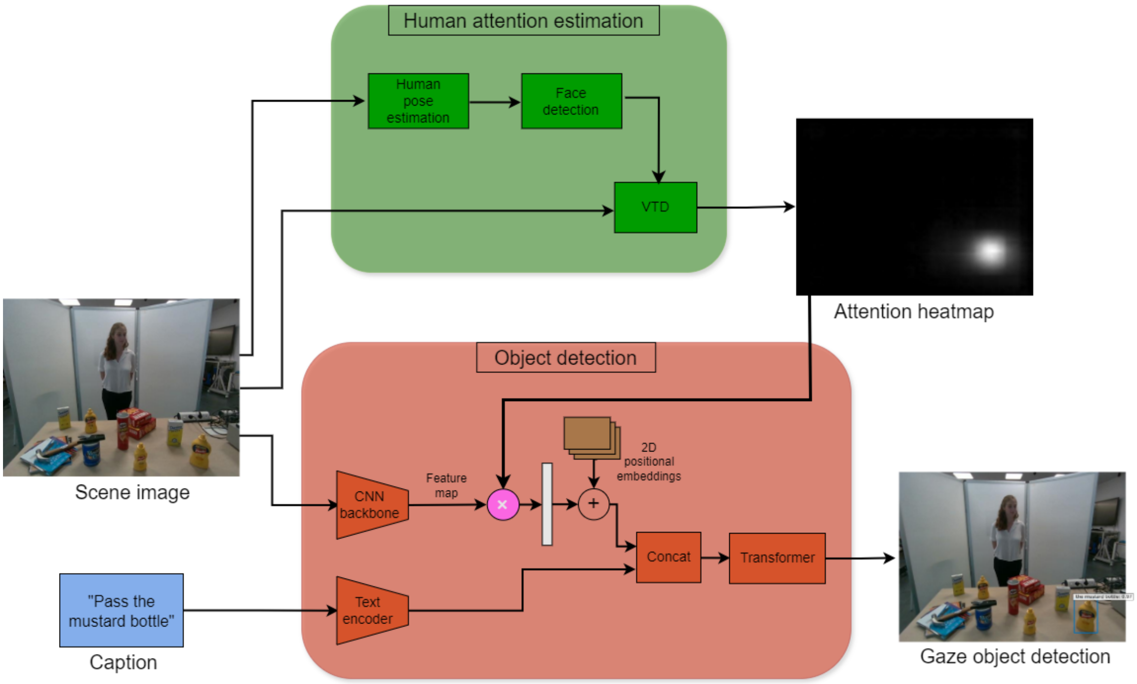

Figure 1: GazeMDETR architecture. The *Human attention estimation* module is composed by the pose estimation model [22], the face detection model [23] and the VTD architecture [21]. The attention heatmap in output is then used as input to the *Object detection* module to weight the feature map extracted by the MDETR's convolutional backbone [19]. Final output is the bounding box of the object the human is referring to.

specific arrangement of objects (Figure 2) - note that in a single session, the same object can be present multiple times:

1. *Heterogeneous cluttered scenario*: coffee can, stapler, journals, mustard bottles, chips can, sugar boxes, crackers boxes;
2. *Scenario with only boxes*: baby food boxes, pudding boxes, crackers boxes, sugar boxes;
3. *Scenario with only repeated objects*: crackers boxes, mustard bottles.

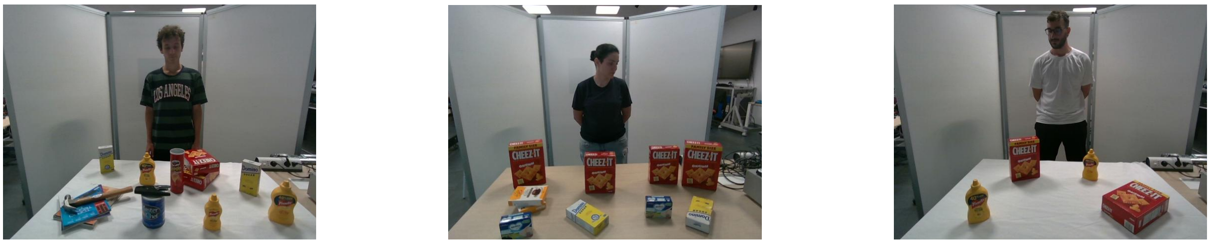

Figure 2: Sample frame for each scenario. From the left, the participant was asked to gaze: at the *coffee can* object (session 1); at the big *cracker box* on the right (session 2); at the *mustard bottle* on the left (session 3).

Evaluation on the cluttered testset. We evaluate and compare the performance between MDETR and GazeMDETR using Accuracy@1 (Acc@1). For each image the bounding box of the predicted referred object is compared with the ground truth: if thee bounding box overlaps the gazed object over a certain threshold, the prediction is counted as true positive, otherwise, it is predicted as a false positive. Note that if more than one bounding box is returned as output, only the bounding box with the highest confidence value is selected. The overlapping between the bounding boxes was evaluated in terms of Intersection over Union (IoU) and the threshold was set at 0.5.

The accuracy is reported as average value evaluated across all participants and all objects within a session, using captions at different level of detail. Specifically, we considered 4 different captions having a different number of attributes related to the referred object: *1*) pose + color + name + placement, *2*) pose + name + placement, *3*) color + name, *4*) name. "Pose" refers to the object orientation (e.g., vertical/horizontal), while "placement" refers to the object position (e.g., on the left/on the right). This degree of detail is useful to study the performance of the models with ambiguous or incomplete sentences. Table 1 reports the accuracy of MDETR and GazeMDETR for each caption and each session.

Session	Caption	**GazeMDETR** [Acc@1]	**MDETR** [Acc@1]
1	A1	**0.69**	0.45
	A2	0.32	0.32
	A3	**0.78**	0.61
	A4	**0.54**	0.43
2	A1	0.47	**0.53**
	A2	0.43	**0.46**
	A3	**0.51**	0.29
	A4	**0.34**	0.14
3	A1	**0.94**	0.89
	A2	0.79	**0.99**
	A3	**0.92**	0.46
	A4	**0.86**	0.46

Table 1
Comparison between MDETR and GazeMDETR. Mean of accuracy@1 is reported for the most significant captions for all the sessions. Specifically: A1 = *The + pose + color + name + placement*, A2 = *The + pose + name + placement*, A3 = *The + color + name* , A4 = *The + name*. Highest performance for each caption is in bold.

5. Discussion and future directions

The results in Table 1 compare GazeMDETR and MDETR in terms of accuracy across the three sessions, with captions varying in complexity from detailed descriptions to simpler ones. While for the captions *A1* and *A2* (more detailed captions) GazeMDETR and MDETR can be considered comparable alternatives, GazeMDETR reports a major improvement for the captions *A3* and *A4* (less detailed captions) for all the sessions. For example, in session 3 GazeMDETR scores respectively an accuracy value of 0.92 and 0.86 for A3 and A4, while MDETR performance drastically drops to 0.46 for both cases.

Given the promising results and given the effect that the caption has on the object detection accuracy, ongoing work is focused on further analysing the capabilities of GazeMDETR with a more natural input text trying to simulate a human request in an interaction. Examples of input description that can be considered with different level of detail are: "Please, could you pass me the + object", "Look at the + object", "Point at the + object". Having a perception system robust to the level of detail in object referring is crucial to enhance the usability and the user experience, especially for people suffering from ASD with reduced verbal skills, resulting in a smoother communication and greater engagement during the rehabilitation sessions.

Next step will be the implementation of the GazeMDETR model on a robotic platform like iCub humanoid robot (in this work the robot's camera has been used only for data collection). Having such an architecture endowed in iCub will allow the robot to be aware of the surrounding environment and of the patient while performing the training trials. In order to have a socially assistive humanoid robot, the proposed perception system will be combined also with other learning algorithms implementing further social cues, such as action recognition, mutual gaze estimation and so on.

Funding

This work received funding under the project Fit for Medical Robotics (Fit4MedRob) - PNRR MUR Cod. PNC0000007 - CUP: B53C22006960001.

References

[1] H. I. Krebs, J. J. Palazzolo, L. Dipietro, M. Ferraro, J. Krol, K. Rannekleiv, B. T. Volpe, N. Hogan, Rehabilitation robotics: Performance-based progressive robot-assisted therapy, Autonomous robots 15 (2003) 7–20.

[2] S. Boucenna, A. Narzisi, E. Tilmont, F. Muratori, G. Pioggia, D. Cohen, M. Chetouani, Interactive technologies for autistic children: A review, Cognitive Computation 6 (2014) 722–740.

[3] J. Fan, L. C. Mion, L. Beuscher, A. Ullal, P. A. Newhouse, N. Sarkar, Sar-connect: a socially assistive robotic system to support activity and social engagement of older adults, IEEE Transactions on Robotics 38 (2021) 1250–1269.

[4] X. Yang, X. Shi, X. Xue, Z. Deng, Efficacy of robot-assisted training on rehabilitation of upper limb function in patients with stroke: a systematic review and meta-analysis, Archives of Physical Medicine and Rehabilitation 104 (2023) 1498–1513.

[5] A. Khoreva, A. Rohrbach, B. Schiele, Video object segmentation with language referring expressions, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part IV 14, Springer, 2019, pp. 123–141.

[6] J. Chen, X. Zhang, Y. Wu, S. Ghosh, P. Natarajan, S.-F. Chang, J. Allebach, One-stage object referring with gaze estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5021–5030.

[7] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, J. Shen, Referring multi-object tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14633–14642.

[8] S. Baron-Cohen, Empathizing, systemizing, and the extreme male brain theory of autism, Progress in brain research 186 (2010) 167–175.

[9] M. Hart, Autism/excel study, in: Proceedings of the 7th International ACM SIGACCESS Conference on Computers and Accessibility, 2005, pp. 136–141.

[10] J. Lee, H. Takehashi, C. Nagai, G. Obinata, D. Stefanov, Which robot features can stimulate better responses from children with autism in robot-assisted therapy?, International Journal of Advanced Robotic Systems 9 (2012) 72.

[11] L. V. Calderita, L. J. Manso, P. Bustos, C. Suárez-Mejías, F. Fernández, A. Bandera, Therapist: towards an autonomous socially interactive robot for motor and neurorehabilitation therapies for children, JMIR rehabilitation and assistive technologies 1 (2014) e3151.

[12] A. Di Nuovo, D. Conti, G. Trubia, S. Buono, S. Di Nuovo, Deep learning systems for estimating visual attention in robot-assisted therapy of children with autism and intellectual disability, Robotics 7 (2018) 25.

[13] A. Alabdulkareem, N. Alhakbani, A. Al-Nafjan, A systematic review of research on robot-assisted therapy for children with autism, Sensors 22 (2022) 944.

[14] G. Metta, L. Natale, F. Nori, G. Sandini, The icub project: An open source platform for research in embodied cognition, in: Advanced Robotics and its Social Impacts, 2011, pp. 24–26.

[15] M. Mori, K. F. MacDorman, N. Kageki, The uncanny valley [from the field], IEEE Robotics & automation magazine 19 (2012) 98–100.

[16] M. Mara, M. Appel, T. Gnambs, Human-like robots and the uncanny valley, Zeitschrift für Psychologie (2022).

[17] D. Ghiglino, F. Floris, D. De Tommaso, K. Kompatsiari, P. Chevalier, T. Priolo, A. Wykowska, Artificial scaffolding: Augmenting social cognition by means of robot technology, Autism Research 16 (2023) 997–1008.

[18] A. B. Vasudevan, D. Dai, L. Van Gool, Object referring in videos with language and human gaze, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[19] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, N. Carion, Mdetr-modulated detection for end-to-end multi-modal understanding, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1780–1790.

[20] S. Hanifi, E. Maiettini, M. Lombardi, L. Natale, icub detecting gazed objects: A pipeline estimating human attention, 2024. URL: https://arxiv.org/abs/2308.13318. arXiv:2308.13318.

[21] E. Chong, Y. Wang, N. Ruiz, J. M. Rehg, Detecting attended visual targets in video, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 5396–5406.

[22] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019. URL: https://arxiv.org/abs/1812.08008. arXiv:1812.08008.

[23] M. Lombardi, E. Maiettini, V. Tikhanoff, L. Natale, icub knows where you look: Exploiting social cues for interactive object detection learning, in: 2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids), 2022, pp. 480–487. doi:10.1109/Humanoids53995.2022.10000163.

[24] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, A. M. Dollar, Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set, IEEE Robotics & Automation Magazine 22 (2015) 36–52.