Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Robust and Scalable Task Selection for Humans and Robots with the Use of Pictograms as Fiducial Markers

ANKE FISCHER-JANZEN[1], Thomas M. Wendt[2], and Kristof Van Laerhoven[3]

[1]Work-Life Robotics Institute, Offenburg University of Applied Sciences, 77652 Offenburg, Germany (e-mail: anke.fischer-janzen@hs-offenburg.de)
[2]Work-Life Robotics Institute, Offenburg University of Applied Sciences, 77652 Offenburg, Germany (e-mail: thomas.wendt@hs-offenburg.de)
[3]Ubiquitous Computing, University of Siegen, 57076 Siegen, Germany (e-mail: kvl@eti.uni-siegen.de)

Corresponding author: Anke Fischer-Janzen (e-mail: anke.fischer-janzen@hs-offenburg.de).

This work did not receive any financial support.

ABSTRACT Controlling robots through eye-tracking enables people with severe physical disabilities to perform everyday tasks independently. Interpreting the intent of users from their gaze alone, considering the presence of multiple objects and diverse tasks, poses a significant challenge that influences the design of this control modality. The objective of error-free object or task selection, which enhances user safety, stands in contrast to the improved usability resulting from increased robot autonomy. This paper presents a task selection method that employs easily comprehensible pictograms as fiducial markers for robot control. The pictograms represent eight daily tasks that can be performed with a robotic arm. We present a pilot dataset that contains 1,008 in-the-wild images. We evaluated the performance of a YOLOv12n model trained on this dataset through a stress test to identify shortcomings in its real-world performance. The comprehensibility of the pictograms was evaluated in an online survey, resulting in an average correct response rate of 83%. Based on these results, we discuss the model's usability, accessibility, scalability, and robustness in the context of the anticipated application. Our method provides a robust and scalable way to select daily tasks that can be performed in domestic and public environments.

INDEX TERMS Assistive robots, Assistive technologies, Gaze Tracking, Object detection, Robust control, Scalability

I. INTRODUCTION

The level of autonomy exhibited by shared controls varies among different robotic applications, including robotic arms, social robots, and humanoid robots [1]. It empowers people with physical disabilities to achieve independence by assisting them with daily task performance. In these shared controls, task execution is delegated to the robot to reduce the cognitive workload of the user [2]. One of the ongoing challenges is achieving seamless and intuitive interaction between humans and robots [3]. Gaze has proven to be a promising input modality for realizing Human-Robot Interaction (HRI) in shared controls. In this research domain, a prevalent methodology is to use gaze as an explicit input modality by fixating gaze to directly point to objects, implying the intention to execute a designated task. Engaging with a tangible environment or using a display can achieve this objective [4]. In contrast to gaze gestures, which are predefined eye movements that instruct the robot, this method requires less cognitive effort and reduces user fatigue [5]. In real-world applications, a single object may be utilized for multiple tasks, creating ambiguity in the selection of the intended task. For example, users could intend to place, fill, or drink from the same glass, showing that the gaze-pointing method would not be applicable.

Observing gaze behavior in multi-object scenes has yielded encouraging outcomes for detecting user intent in the context of task prediction [6]–[8]. However, the efficacy of this approach is limited when attempting to extend it to a substantial number of tasks [9]. Moreover, the misinterpretation of gaze events due to unintentional fixations or incorrect interpretation of the context caused by natural gaze behavior interferes with the correct interpretation of user intent [10]. The utilization of fiducial markers, such as AprilTags or QR codes, can improve performance by providing the robotic system with additional information about tasks and objects, as well as assisting with robot positioning [11], [12]. The integration of a fiducial marker for each task and its subsequent linkage to an object has the potential to address the aforementioned challenges, such as robustness in selection and scalability. However, users may encounter challenges in interpreting and selecting the appropriate fiducial marker due to its visual characteristics.

In this paper, we present an approach based on pictograms to improve the robustness of daily task selection for communication between humans and robots. The pictograms' easy-to-understand appearance benefits legibility for the user. An object detection model creates interactive bounding boxes by detecting the associated task, from which the robot can obtain more information, and therefore repurposes the pictograms as fiducial markers.

The main contributions of this work are as follows:

- An object detection-based eye-tracking control strategy adapted for controlling robotic arms in daily tasks.
- A pilot dataset and object detection model trained to robustly distinguish eight common daily tasks that is systematically evaluated under real-world variability.

II. RELATED WORK

Pictograms are used to make environments more accessible for people with intellectual disabilities by providing alternative means of communication, as in augmentative and alternative communication. They provide rich information and enhance usability by replacing text and therefore overcoming language barriers [13], [14]. In robotics, they are used to localize the position of mobile robots [15] or to extend fiducial markers to improve comprehensibility in user-centered designs [16]. In contrast, while fiducial markers benefit robotics by providing precise object location and information about objects, the stored data is not interpretable by humans without an adequate device. Fiducial markers are used in many eye-tracking-related works revolving around task selection. These approaches use them to define the workspace and calculate the location of the desired object [7], [17], [18]. Other approaches enhance the accuracy of 3D gaze estimation via fiducial markers [12], [19]. Huang and Mutlu integrated fiducial markers to indicate desired ingredients for a task and move the robot towards the object [6]. Similar to this approach, Yang et al. and Wang et al. incorporated fiducial markers into their setup to improve detection rates and feature extraction [11], [20]. Lastly, Nunez Sardinha et al. created a system based on diegetic buttons, which realize interactive 3D-printed fiducial markers in real-world settings. Users found the system easy to learn and use [21].

Combining fiducial markers with pictograms would provide benefits from both domains: satisfactory interpretability for the user, higher information density, and enhanced accuracy in location estimation for the robot. To the best of the authors' knowledge, no approach exists that uses pictograms as fiducial markers to communicate task selection with a robot via gaze in a real-world setting.

Gaze, as an input modality, provides rich information and context for controlling robotic arms in assistive tasks, social robotics, and robotic surgery [4], [22]–[24]. One convenient solution for selecting tasks based on objects in a real-world scene is to interpret user intent based on natural gaze, as presented by Admoni et al., Zhang et al., and Baptista et al. [7], [8], [12]. While effective in small laboratory settings, this approach has struggled to scale to a large number of available tasks while maintaining robust selection capabilities [7]–[9]. Another emerging solution involves integrating vision and language, such as through the use of vision-language models (VLMs). VLMs translate images into labels and extract context from detected objects. Examples were presented by Chi et al. and Wu et al. [25], [26]. Depending on the scene and objects included, it can be challenging to conclude the correct task by identifying objects that relate to multiple tasks.

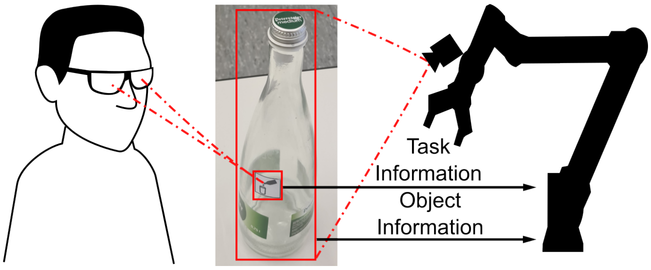

FIGURE 1. Control process: The user selects the desired task with their gaze. After a certain dwell time, the task is selected. Information is required by a camera mounted on the robot, which performs the task.

While these approaches yield practical solutions for individual tasks, it is impractical to scale them to a variety of daily tasks. Daily tasks that can be performed with a robotic arm include hobbies, household chores, and work-related tasks [27], [28], personal hygiene [29], [30], and eating and drinking independently [30], [31]. The goal of this work was to develop a reliable method for selecting between these tasks and others by using gaze information to guide the robot, while ensuring that the method can be scaled to incorporate new tasks.

III. METHODS

A. CONTROL STRATEGY

The control input uses the gaze location within the eye-tracker's world camera scene. This scene includes pictograms representing the desired task, which are attached to suitable objects. Since the intended user group includes individuals with severe physical disabilities, the system design assumes that pictograms may be applied with assistance from caregivers or family members when such support is available. The user selects the desired object for the task by fixating their gaze on the corresponding pictogram. Pictograms are detected in relation to the camera scene and are selected when the user's gaze overlaps with the pictogram's bounding box. This process is outlined in Figure 1. The selection information sent to the robot contains the task class, object class, and object location, as well as additional information defining environmental conditions, such as the need for additional objects.

B. PICTOGRAM DESIGN PROCESS

A subset of eight tasks was chosen based on their importance in everyday life [10]. An initial design was created for each task. Eighteen fellow researchers with mixed backgrounds in robotics and ubiquitous computing were asked to identify each pictogram's corresponding task using open-ended (free-text) responses. Based on their feedback, the designs were adjusted. Designs that appeared multiple times, such as the person, varied between pictograms to provide more distinguishing markers in the object detection training process.

A subsequent online survey evaluated the comprehensibility of the new set of pictograms. The goal was to determine if participants could deduce the task depicted in each pictogram. The survey was approved by the Ethics Council of the author's institution. Participants were shown each pictogram and could select one out of four task descriptions. To minimize bias, the order of the answers was randomized for each participant. The field for optional answers was removed. Participants could provide feedback after answering all questions. Appendix IX lists all selectable answers available for each pictogram in Table 4. The correct response rate was calculated to evaluate whether the pictograms are easy-to-understand. Additionally, the feedback was reviewed and summarized to identify shortcomings in the pictogram design.

C. MODEL SELECTION

The model was selected based on performance evaluated in the Computer Vision Model Leaderboard [32], which was benchmarked on COCO 2017. The model is expected to have a low inference time (less than 500 ms) to minimize user frustration during the control sequence by avoiding long response times of the system. The model has to have a high mean average precision (mAP50:90) at 50% and 90% for small object detection. mAP summarizes the precision and recall performance across different confidence thresholds and is used as a measure to verify if pictograms are detected robustly. Additionally, the model was chosen to achieve a high F1 score to ensure effective performance in a real-world application, where the camera position causes the size of the pictograms in the scene to be represented by only a few pixels. The YOLOv12x model [33] performed within the four best models (mAP50:95 (small) = 0.267, F1 50 (small) = 0.149, F1 75 (small) = 0.098, and general mAP50:95 = 0.527, ranking sixth).

Training performance tests revealed that YOLOv12n produced sufficiently satisfactory results, despite its lower ratings compared to the YOLOv12x model [32]. The authors decided to implement the YOLOv12n model to maintain a lightweight algorithm with low inference speed.

D. DATA COLLECTION, ANNOTATION AND AUGMENTATION

The revised pictograms were printed in two sizes: 1.5 by 1.5 cm and 2.5 by 2.5 cm. A total of 1,008 images were taken in domestic home and workplace settings. The pictograms were randomly placed in the scenes and partly attached to related objects. All images were captured using the same camera, resulting in a median resolution of 4032 px x 3024 px. Although using cameras with lower resolutions was considered, the idea was rejected since newer eye-tracking and robot cameras show rising performance and higher resolutions. Intentionally taking images with lower resolution to fit the current standard in data processing and object detection would result in a less suitable dataset and model for future applications. Although the number and placement of pictograms in each image were randomized, the dataset was balanced so that each pictogram class and the total number of pictograms within each image were distributed equally across samples. The number of pictograms ranged from one to five and, in some cases, represented the same task multiple times to more closely resemble real-world applications. The distance between the camera and the pictograms was randomized to reduce bias in the detected size and orientation of the pictograms.

Annotation was performed in Roboflow [34]. Due to the possibility of different levels of occlusion in both cameras, as indicated in Figure 1), pictograms overlapping with other pictograms or objects were annotated. Reflections were not annotated. The robot could pose a danger to the user when grasping an object due to reflections detected in glasses or cutlery, including knives, or when moving into a mirror. Pictograms that were out of focus were included to mimic real-world cases of unfocused camera images.

The augmentation was performed for the following reasons.

- Crop: 0 to 10% - Pictograms were rarely located at the edges of images. The boundaries were cropped to distribute the pictograms evenly throughout the images.
- Rotation: -15 to 15% - Simulates a higher or lower position of the camera in relation to the object and unevenly applied pictogram stickers.
- Shear: 10° - Simulates pictograms that are placed on the sides of cylindrical objects, such as cups or bottles.
- Brightness: -15 to 15% - Simulates illumination changes.
- Blur: 2% - Simulates pictograms out of focus.
- Noise 0.1% - Reducing overfit.

The total number of images in the dataset was increased from 1,008 to 2,420 with these augmentations. Three augmented images were generated for each of the 706 training images. Together with the 202 validation images and the 100 test images, the total number of images was 2,420. The images were split randomly into train (70%), valid (20%) and test (10%). The dataset is available in Roboflow[1].

E. MODEL TRAINING AND EVALUATION

The model was trained with the code provided in the GitHub repository[2]. The code was partially reused from Ponomarjova

[1] Dataset available at: https://app.roboflow.com/wlrisemanticlables/wlri-semantic_labels/8

[2] Model training code: https://github.com/AnkeLinus/PictogramFidMarkerSet/blob/main/README.md

et al. [35]. The dataset images were downloaded at a size of 1280x1280 px to minimize data loss due to compression. The model was trained for 300 epochs with a patience of 100.

The performance was evaluated by measuring the inference time to ensure rapid detection for the control. mAP50 and mAP50-90 were calculated to verify high overlap between the ground truth and the predicted bounding boxes so they can be implemented as interactive objects in the control architecture. In addition, precision and recall were measured to ensure the detection of all classes. The precision score was important for indicating whether incorrect labels were used. Since bounding boxes are used as interactive buttons, an incorrect prediction could lead to the selection of the wrong task. Low recall would result in many pictograms not being detected, which would affect the control strategy's functionality. The F1 score was calculated for completeness and comparison to the metrics provided in the initial model selection.

The model was trained and evaluated on a computer running Windows 11 with an AMD Radeon RX 6750 XT graphics card and an AMD Ryzen 5 5600X six-core processor.

The evaluation was repeated for images in which different resolutions were applied in the data download. Additionally, a dataset of 90 images was created for a stress test of the model's performance. A measuring board was used to evaluate the model performance in connection with varying environmental conditions[3]. The measuring board displayed 24 pictograms, each having an edge length of 2 cm. The pictograms were spaced 1 cm apart. Two different background patterns were used. With both backgrounds, an occlusion of 0%, 25%, and 50% was simulated. Taking the images at each level of occlusion was repeated four times, applying the occluded area in each direction once. The setting was rotated in 45° steps within a range of ±90° to the upright position of the pictogram. No further digital augmentation was performed on the images. The dataset was annotated as stated above, but was not split into different groups due to its purpose.

The influence of varying distances was investigated separately. All parameters remained constant except for the distance between the camera and the measuring board. This distance varied from 30 cm to 200 cm in increments of 10 cm. Each object class had 16 instances. Both the stress test dataset and these images were not used in the model training.

F. INITIAL IMPLEMENTATION

An initial implementation was performed in which the model was integrated into a web server application that showed the camera scene and visualized the detected bounding boxes. The experiment was conducted to determine whether the inference time and precision of the bounding box location would meet the demands of the eye-tracking control. A Flask-based[4] web server was implemented to visualize camera data from an Intel RealSense D455 camera. The camera data was captured using the ROS2 Humble library provided by the manufacturer [36]. The outcome is shown in the supplementary materials video.

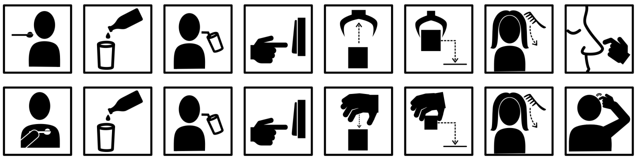

FIGURE 2. Upper row: Initial pictogram design, Lower row: Redesigned pictograms.

IV. RESULTS

A. TASK SELECTION

The tasks in this work were selected based on stakeholder interviews presented in previous works [10]. At least one task was selected for implementation from each category.

- Eating
- Filling a glass
- Drinking
- Switching a light switch
- Picking up an object
- Placing an object
- Brushing hair
- Scratching oneself

The tasks were chosen because they represent the challenges associated with complex tasks that include multiple objects as outlined in Section III-A. They were used to explore the shortcomings and requirements of the control in the implementation step.

B. PICTOGRAM DESIGN EVALUATION

After selecting the tasks, the initial design of the pictograms was created, as shown in the upper row of Figure 2. Eighteen fellow researchers were asked to identify the tasks represented by the pictograms to ensure understandability and collect feedback on the design. At this point, they were unaware of any answer options for the task description. The answers are summarized in Table 1. Most of the pictograms were interpreted correctly. Wrong answers were found in the pictograms "eat", "switch", "pick", and "scratch".

The pictograms were redesigned for easier identification and are shown in Figure 2 in the lower row. The new design resolved the following issues. The robotic gripper seen in the "pick" and "place" pictogram were changed to a hand, since people without robotics knowledge have trouble understanding what it depicts. It was decided that a grasping motion depicting a hand is easier to interpret. Similarly, the spoon was moved in front of the person in the pictogram to mimic the motion toward the mouth. In the scratching pictogram, the view was zoomed out to show the person's upper body and head, and lines were added to indicate movement. This approach has been found advantageous in other publications

[3] Measuring board and labeled data: https://app.roboflow.com/wlrisemanticlables/picotgram-stress-test/2

[4] https://flask.palletsprojects.com/en/stable/

TABLE 1. Overview of task interpretation of the pictograms. Researcher's CRR (Correct response rate) represents the rate the 18 researchers recognized the task from the pictogram without given answer options. Alternating answers are stated when incorrect guesses occurred. In the online study, the CRR was calculated based on 364 responses and evaluated the redesign pictograms.

	Research team		Online survey	
Task	CRR	Varying answers	CRR	Alternative answer
Eating	89%	Start of action, mouth	87%	Taking medication
Filling a glass	100%	-	89%	drinking
Drinking	100%	-	99%	-
Switch	83%	Ringing, selecting, dressing	92%	point at smth.
Picking up	83%	Release object, move up	52%	touch smth.
Placing	100%	-	97%	-
Brushing hair	100%	-	97%	-
Scratching oneself	72%	Tip of the nose, tap nose, smell	45%	Thinking, gesturing

to improve understandability for people with intellectual disabilities [37]. This increases the chance that the system will be accessible to a large group of people with varying disabilities.

The online survey aimed to investigate whether the pictograms were easy to understand. A total of 364 participants completed the online survey. Table 1 presents the correct response rate (CRR) for each pictogram. It describes how many participants have chosen the intended task description from the four available answers. If the CRR is lower than 95%, the second most frequently selected task is stated in the row "Alternative answers".

Participants had the opportunity to provide feedback. Of the 364 participants, 58 provided written comments containing actionable content, such as no emoticons or general statements indicating no issues. 25 of these responses concerned the selection options of the questions. Participants differentiated between cases in which their initial interpretation was not included among the predefined answer options and cases in which the pictogram design allowed for multiple plausible interpretations. Both response patterns had been anticipated during the questionnaire design phase. Due to the anticipated training approach for future users, the task description will be given with the pictogram to learn. Consequently, ambiguous initial interpretations are not expected to result in incorrect task selections during robot operation, although they may indicate the need for minor design refinements.

The option to respond with free text was not available. The goal was to elicit feedback about the tasks, not to collect assumptions about what the pictogram might indicate. Consequently, participants provided design improvements for the pictograms. The pictograms that received the most feedback were "scratch" with ten comments, "pick" and "place" with five comments each, and "drink" and "fill glass" with four comments each. These pictograms resembled the pictograms with the lowest CRR. Six answers were given regarding general design. For the learning process of future users, it was important to understand that some pictograms were ambiguous in their meaning until users viewed their counterparts. This was the case with the "fill glass" pictogram, which had "drink" as a second answer option and was followed by the "drink" pictogram. The same was true for the "pick" and "place" pictograms. Another design suggestion was to replace the spoon in the "eat" pictogram with a knife and fork. The use as robot control was intentionally omitted to avoid biasing participants' perception and interpretation of the pictograms. We agree that a knife and a fork would represent the task "eat" more detailed. However, in future use, this could cause unease by indicating that the robot will approach the user with a fork and knife. The task "scratch" had the worst CRR and the most feedback. This was due to the provided answer options. Many participants chose the option "thinking" instead of "scratch" and provided feedback that matched their selection.

TABLE 2. Model performance after training with 300 epochs. The parameters indicate performance for varying image sizes and the stress test conducted to evaluate real-world application.

Parameter	1280 px	640 px	Stress test
mAP50	0.949	0.845	0.59
mAP50-90	0.852	0.726	0.496
Precision	0.966	0.904	0.802
Recall	0.919	0.739	0.414
F1 score	0.942	0.813	0.538
Inference	420.4 ms	62.5 ms	429.2 ms

Lastly, a participant asked if "scratching" could be considered a task. There is no item in the IFC catalog that handles scratching as an individual task. However, the classification scheme allows for the interpretation that it is classified under "Looking after one's health," an overarching category that includes multiple tasks which can partially be achieved with a robot. This suggests a potential discrepancy in how daily tasks are defined within assistive robotics.

C. PERFORMANCE EVALUATION

The dataset was validated with image sizes of 1280 and 640 pixels to compare the impact of compression on performance. The dataset was evaluated using the code provided in this repository[5]. Table 2 shows the performance for both validations with image sizes of 1280 and 640 pixels. Due to compression, the model's performance with 640 px images was considerably worse, but required less inference. The initial implementation into the interface revealed that the pictogram size must be significantly larger relative to the camera scene to correctly detect the pictograms when the image size is set to 640 px. This procedure is demonstrated in the video provided in the supplementary materials. Pictograms were more easily detected when they had a size of 2.5 by 2.5 cm at a distance of one meter or less. Inference was sufficient to provide a pleasant control experience. To ensure functionality across the full range of a robot's workspace and detect smaller pictogram sizes, we recommend using images with 1280 px.

[5] Evaluation code: https://github.com/AnkeLinus/PictogramFidMarkerSet

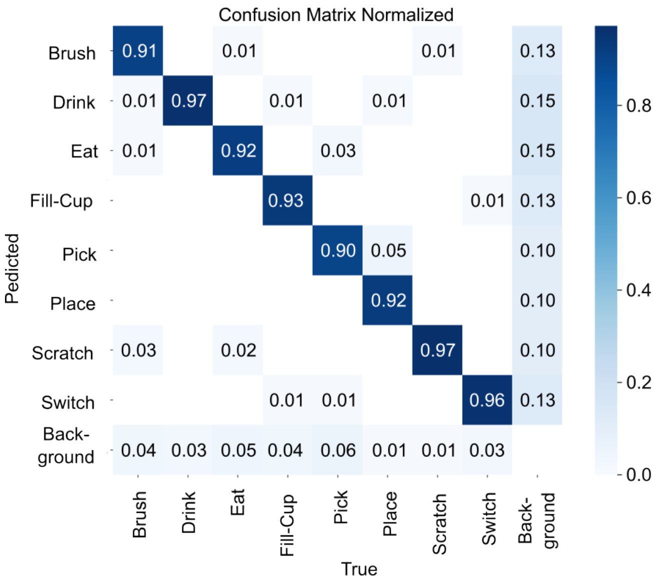

FIGURE 3. Normalized confusion matrix for test image size of 1280 px.

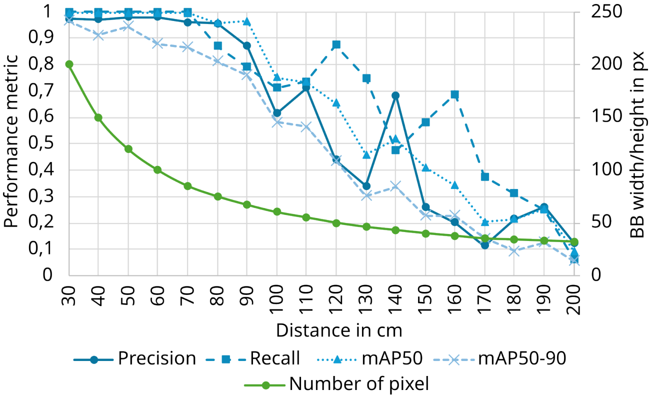

FIGURE 4. Model performance with varying distance between pictograms and camera.

Figure 3 shows the resulting normalized confusion matrix. Each row corresponds to the ground-truth class, while each column represents the predicted class. Correct classifications are shown along the diagonal, whereas off-diagonal entries indicate misclassifications. Due to the small size of the pictograms relative to the overall image size, visual artifacts such as holes or reflections may occasionally be misinterpreted as valid pictograms, leading to 10% to 15% false positives per class. Misclassifications occurred across all classes as well as in the background category. Since the average correct detection per class is over 90% and the initial implementation showed that repositioning the camera corrects the initial incorrect detection, we consider this model sufficiently precise for developing a prototype control.

In addition to the standard performance evaluation, we conducted a stress test to examine the system's performance boundaries more closely. The background, occlusion, and rotation were systematically varied to test how the model would behave in real-world applications. Table 2 shows the overall performance decrease under these conditions in the "Stress test" column. Occlusion was found to have the highest impact on the performance. While the mAP50 was 0.958 (P = 0.945, R = 0.9) without any occlusion, it deteriorated to 0.339 (P = 0.346, R = 0.343) with 50% occlusion. Since this decrease dominated the results, the dataset was split further to investigate the impact of the background changes and rotation without the influence of occlusion. The background changes also impacted the performance of the image. Precision, recall, mAP50, and mAP50-90 were lower for the patterned background condition. This effect was visible in all occlusion conditions. As an example, mAP50 was 0.985 with 0% occlusion, and a blank background, but degraded to 0.932 with the patterned background. With 50% occlusion, mAP50 decreased from 0.432 to 0.295 when comparing the blank and patterned backgrounds. Rotating the images revealed no significant differences or trends in any metric. Performance varied for mAP50 in a range of 0.05.

For real-world application, these findings suggest being cautious with occluded pictograms, as it can happen in everyday life. Ideally, caregivers can help reposition the object. Regarding robot control, it is important to return objects to their original location with their initial orientation to reduce the occurrence of occluded pictograms. The decline in performance in patterned environments highlights shortcomings that can be anticipated in cluttered scenes, suggesting that there should be additional tests in real-world settings.

Lastly, the impact of varying distances was systematically tested. The performance metrics are displayed in Figure 4. The metrics indicate stable performance up to a distance of 70 cm with 2 cm pictograms. For example, this would correspond to a bounding box width and height of 85 px and a camera resolution of 4032 px x 3024 px. Between 80 cm and 120 cm (75 px to 50 px size) the model's performance is visibly unstable. Recall and precision fluctuate significantly, and mAP50 decreases substantially. Under these test conditions, the model's performance was found insufficient under these test conditions with distances greater than 130 cm. The resolution of the images is less than 46 px, which severely limits a stable detection. It has to be ensured that the detection is robust within the robot's maximum reach. For example, the Kinova Jaco arm has a maximum reach of 90 cm. Due to variations in the user's head position relative to the robot's base, the maximum distance between the user's eyes and the desired object including the pictogram can exceed this distance. In such cases, the unstable distance can be reached, but due to the scalable size of the pictograms, this issue can be resolved by printing them larger than 2 cm enhancing the pixel count within the pictogram.

V. DISCUSSION

Object size, geometry, gaze location, and user intention influence the occurrence of unintended task selections and misinterpretations of gaze behavior in eye-tracking controls [10]. The presented approach reduces these influences by

TABLE 3. Advantages and disadvantages of key contributions.

Key contribution	Advantage	Disadvantage
Usability and accessibility	Easy pictogram application, extensible framework, customization pictogram size	Pending real-world user study, handling of complex tasks
Interpretability	Easy-to-understand, no text	Generalized task taxonomy, tasks outside the robot's ability
Scalability	Scalable to new tasks	Lack of clear object anchors
Robustness	Sufficient model performance, sufficient differentiation	Accuracy errors

providing pictograms as visual guidance and selectable objects. Depending on the arrangement of the objects, object detection can generate overlapping bounding boxes, resulting in unintentional selection (see Figure 5-A). For slim objects, such as forks, object segmentation could lead to regions that are too small, causing the gaze to be localized outside the segmented object due to accuracy errors. These issues can be resolved with the presented pictogram approach, as shown in Figure 5-B. Since the pictograms depict specific tasks, the user can explicitly select their desired task, thereby gaining more control over the system. This, in turn, may lead to an increase in their confidence in the system's capabilities. Furthermore, because of their rectangular shape, pictograms do not display narrow object features, which makes them easier to select by gaze. Figures 5-C and 5-D show examples of the bounding box overlap in real-world scenes. Applying the pictograms creates wider gaps between the bounding boxes, reducing the error source for misinterpreted gaze fixations.

A. REFLECTION ON KEY CONTRIBUTIONS

The reflection on the key contribution revolves around the question of whether the developed dataset and model meet the initial demands. In Table 3 the advantages and disadvantages of this approach are stated and will be discussed below in detail.

1) Usability and accessibility

In research, reducing the workload by increasing independence is often equated with completely transferring care assistance to robots. Caregivers, family members, and potential users indicated that they prefer to use time freed from care duties for social interaction. Therefore, this approach assumes that users are in contact with a caregiver or family member for assistance with applying the pictograms and other tasks at least part of the day. This control strategy aims to facilitate the implementation of daily task approaches and generate a system capable of performing a multitude of tasks that fulfills the preferences and aims of potential users. The application of the pictograms by caregivers and family members is straightforward because they do not require alignment or focus on the scene for detection, as shown by the rotation variation in the performance evaluation. Implementing these pictograms takes only a few seconds. Additionally, the selection of objects can be limited to reduce the risk of injury when handling knives or breakable objects. Larger pictograms improve detection and selection for people with ocular disabilities, such as eye tremors. They can be scaled up during the printing process. The area of the virtual interactive button can therefore be scaled larger, making it easier to stay within the bounding box boundaries. Due to their quadratic shape, the overlap of

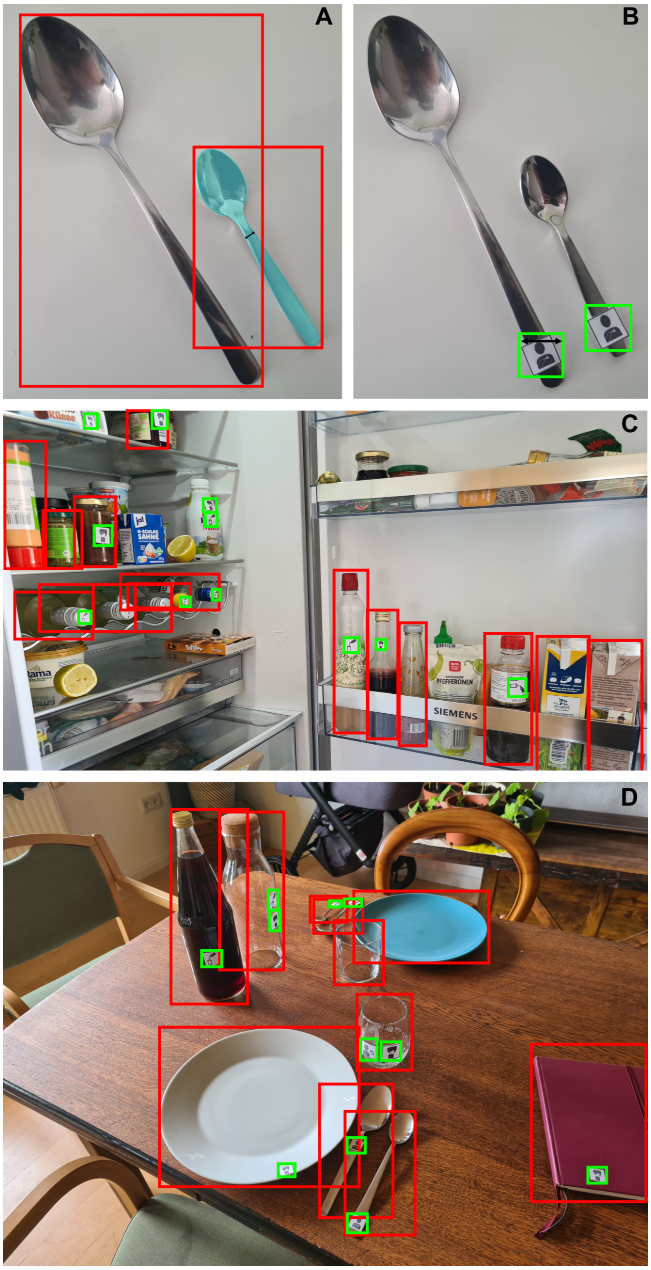

FIGURE 5. A: Overlap of bounding boxes (red) and small regions in object segmentation (blue highlighted area) are shown. B: Due to the size of the pictograms, no overlap occurs, and the region size is wider, as indicated by the black arrows. C and D: Examples of in-the-wild scenes from the dataset showcasing overlaps of object bounding boxes (red) and solved issue with pictogram bounding boxes (green).

bounding boxes is still strongly limited. On large pictograms with edge lengths exceeding 10 cm the desired object for interaction can be placed upon the pictogram. In this scenario, the occlusion has to be controlled to ensure stable prediction of the pictograms.

Lastly, the question arises as to whether users will perceive this approach. No studies involving people with disabilities have been conducted at this point, due to ongoing work on integrating this strategy into robot control. Considering the findings of others such as Nunez Sardinha et al. or Kudo et al., we tried to narrow down this gap. Nunez Sardinha et al. presented a similar approach based on a diegetic GUI that evaluated user perception of an eye-tracking control with 3D-printed fiducial markers [21]. Although the control strategy differs from this approach, participants quickly learned to interact with the system, which required minimal workload. Since pictograms are easier to interpret than fiducial markers, similar results can likely be found with our approach, though this must be verified in the future. Additionally, design variations proposed by Kudo et al. and shown to make pictograms easier for people with cognitive limitations to understand were applied.

2) Interpretability

An easy interpretation of the pictograms improves the usability by reducing frustration in learning and distinguishing the pictograms. The online study revealed that most of the pictograms are easily understandable. Presenting the pictograms alongside a description of the task would have reduced the difficulty some participants experienced in distinguishing between tasks. This assumption can be derived from the feedback evaluation, which shows that some participants' initial guesses were not represented among the provided answer options.

We learned two lessons from this survey. First, the survey indicated that participants might confuse tasks with each other. However, the limitations of what a robotic arm can perform significantly restrict the available options. This was the case in the "scratching" task in which 45% of participants selected "scratching" and 46% selected "thinking". This limits the system's scalability to the context of robotic arm use. In adjacent robot applications, the approach has to be reevaluated. The second lesson learned was that closely related tasks reduce interpretability. For example, eating was confused with taking medication and balancing an object. When integrating such tasks into the model, their pictograms must be clearly distinguishable. Additionally, a sufficiently generalized task taxonomy should be defined to avoid overly fine-grained task distinctions.

Additional improvements to the pictograms were considered to make them easier to understand but ultimately rejected. We discussed whether pictograms needed to be more unique to improve user interpretation and model detection performance. The utilization of diverse symbols in pictograms has been posited as a method to enhance detection performance. However, given humans' aptitude for pattern recognition, this proposal has been dismissed. Another improvement that was considered involved adding text to the pictograms to clarify their meaning. This idea was rejected because it is uncertain whether users are able to read and understand the presented language, which would require datasets in multiple languages. Because the pictograms are small and users can print them themselves, it is not guaranteed that the writing will be large enough to be readable.

3) Scalability

Some tasks lack a clear object anchor. In our set, two of the eight tasks ("scratch" and "place") cannot be directly associated with a specific object, and the task "switch" is limited applicable in public environments. While reusable pictogram stickers can be attached to objects such as plates, they cannot be placed directly on food. Therefore, this approach should primarily be understood as a high-level task selection tool for a broad set of complex activities, which can then be refined through additional gaze-based input and control modules of the community. Such controls for robotic feeding tasks are a well-researched field. Multiple authors have presented suitable solutions [38]–[40]. Similar approaches exist for all other tasks represented in this paper, such as drinking [41] and picking up and placing objects [42], [43]. Alternatively, intent inference and prediction methods can specify the objects and preparation steps with which the user wants to interact in a "prepare meal" task, as demonstrated by Koochaki et al. or Huang and Mutlu [6], [44]. Similar pictogram anchors must be used for other tasks, such as scratching, since the pictogram cannot be attached to the itchy area. However, the pictogram can be placed within the user's visual range and selected at any time.

Daily tasks include personal hygiene, eating, drinking, moving objects, and leisure and work activities. Often, performing daily tasks requires more than a robotic arm, such as the assistance of caregivers or additional assistive devices. In tasks related to personal hygiene, human assistance is often preferred. This is due to the potential confidentiality concerns associated with the use of robotic cameras. Petrich et al. summarize the life domains of functioning resulting from the WHODAS 2.0 test, which resulted in 25 tasks. The study identifies twelve tasks that are suitable for robotic-arm assistance and can be broken down into smaller tasks [45]. If the set of pictograms is expanded to include these twelve tasks, it is assumed that most users will be able to learn and distinguish the pictograms. Furthermore, the pictograms can be simplified to avoid overwhelming details and specify differences within these task groups, while the model's performance should not deteriorate.

4) Robustness

In HRI, robustness refers to a system's ability to maintain safe, reliable, and effective interactions with humans, even when behavior, environmental conditions, or sensor input are uncertain or variable. In this context, the proposed approach promotes robustness through reliable task selection, which enables effective interaction. Intention inference could be enhanced by first selecting the desired task and then specifying the relevant time window and objects for intent inference. However, several factors in this approach can reduce robustness, including hallucinations or misclassifications by the object detection model and unintentional gaze behavior. To mitigate errors in object detection, we diversified the appearance of the pictograms and applied data augmentation strategies. The likelihood of unintentional task selection due to unintentional gaze behavior is reduced by the pictograms' small size, as shown in Figure 5. This modification substantially decreases the spatial region in which a gaze overlap results in a valid selection. Together, these measures improve the system's overall robustness in realistic interaction scenarios.

However, as demonstrated in the stress test, the minimum size of the pictogram is constrained by the camera's resolution and the eye tracker's accuracy. If the pictogram is too small relative to the image, recall and precision will decrease because the pictogram will no longer be reliably detected. Additionally, due to accuracy errors exceeding the size of the bounding box, the gaze might be tracked outside the interactive button area. The robot's maximum reach limits the distance between the user and reachable objects, thereby limiting the minimum pictogram size. However, the selection process is limited not only by accuracy, but also by the overlap of the bounding box with the physical pictogram. In the stress test, mAP50-90 drop indicated that this overlap is reduced with increasing occlusion and distance. We therefore emphasize the importance of adapting pictogram size to user preferences while ensuring a minimum size of 2 cm to maintain stable detection performance and to avoid areas that are too small for robust gaze tracking.

VI. GAZE CONTROL AND ROBOT IMPLEMENTATION

The model and approach presented are implemented in an eye-tracking-driven control of robotic arms, as discussed in a separate work [46]. The scope of this paper is to present the control strategy in detail. The usage of the presented strategy becomes clearer when the implementation is outlined.

In this control, the tasks are selected by gazing at the pictograms as indicated in Section III-A. Overlapping the gaze with the detected bounding box initiates the task execution process. Each pictogram is connected to information that defines the task, the need for a second object, and the pictogram's location in the room. A second camera mounted on the robot detects objects in the scene. Pictograms fixed to the objects were used to increase the number of features found in a feature-matching approach. This approach finds the same object in both scenes, thereby transferring knowledge about the selected task and object to the robotic system.

During a brainstorming session, other applications of this approach were discussed. It was found that it may also be transferred to other robotic applications, such as training humanoid robots, providing a quick and simple way to label objects with corresponding tasks, and facilitating communication between humans and robots during social interactions. Screens displaying GUIs for robotic arms on electric wheelchairs are not necessary with this approach since users can select their desired task in real-world scenes with eye-tracking glasses, making the setup of the system more lightweight and less power-consuming for everyday use. This approach has the potential to facilitate the integration of artificial feedback methods, as employed in augmented and mixed reality approaches [12], [47]. In these contexts, the data, when coupled with pictograms, can be utilized to articulate the selected task or visually highlight the selected task pictogram, further improving the user's understanding of the system.

VII. LIMITATIONS

This work mentions limitations in the design process of the pictograms. Since the anticipated user group was not involved in the design process, the pictograms could be made more understandable by a larger peer group of people with disabilities. In the online survey, participants were not asked directly about physical impairments, as this information could not be verified and could potentially lead to bias in the data evaluation process.

This approach used a YOLOv12n model. Future research must evaluate whether more lightweight models or a pruning strategy can yield better results. Object detection exhibits certain variance in accuracy, as evidenced by the mAP50-90 score. If necessary, integrating fiducial markers into the pictogram itself, such as TopoTags [21], [48], [49] can enhance accuracy. Due to limited resolution related to the pictogram size, future work will investigate the potential application of such fiducial markers.

VIII. CONCLUSIONS

The aim of this work was twofold. First, a control input strategy was created to solve the ongoing challenge of correctly selecting between multiple tasks in eye-tracking-driven selection processes. Secondly, we developed an easy-to-learn, scalable, and robust control strategy for both users and robots. We presented a pictogram-based approach in which easy-to-learn pictograms serve as fiducial markers that enable the robot to locate the desired object and interpret the desired task within a scene. Eight pictograms representing daily tasks were designed. A pilot dataset containing 1,008 images from in-the-wild scenes was created and augmented to enhance its size. A YOLOv12n model, trained on this data, was used to evaluate the dataset performance and to estimate adequacy for prototyping. Precision (=0.97) and recall (=0.92) indicated sufficiently accurate detection of small pictograms in domestic home and workplace-related scenes. A stress test and online survey for evaluating the interpretation of the pictograms were conducted to show limitations of this approach in real-world application. These values indicated adequate performance for implementation into the setup, which produced promising results. Concerns regarding usability, robustness, and scalability were discussed. Key findings included the fact that scalability and robustness are influenced by the model's performance. While integrating a multitude of tasks may reduce model performance, the effect can be mitigated through varying the design of pictograms and using reasonable object anchors. Restricting the task set appropriately to subtasks that can be achieved with the robotic arm makes it feasible for the user to learn the meaning of the pictograms. We anticipate that this strategy will improve usability and task selection by defining the interaction interval and the involved objects, thereby reducing ambiguity in real-world scenes for task intent inference and explicit control inputs.

ACKNOWLEDGMENT

The authors acknowledge the use of ChatGPT (OpenAI) for assistance with code debugging during the development of this work.

IX. APPENDIX

The online survey consisted of 9 questions, with questions 1 to 8 featuring a pictogram and providing answer options in the form of radio buttons. Each pictogram was introduced with the question: "Which daily activity depicts this pictogram?". Question 9 asked for the participant's feedback. In Table 4, questions 1 to 8 are stated with the available answers.

TABLE 4. Overview of the answers available in each question and response distribution.

Shown Pictogram	Available answers	Number of responses
Eat	Eat	317
	Take medication	27
	Balance something	15
	Brush teeth	5
Fill a Glass	Pour a drink	326
	Drink	32
	Pour milk into coffee	4
	Water flowers	2
Drink	Drink	360
	Carry a glass	0
	Throw a can	0
	Blow bubbles	4
Switch	Operate a switch	337
	Point at something	21
	Call an elevator	3
	Show someone the way	3
Pick	Pick something up	190
	Take something out of a box	17
	Touch something	108
	Throw away trash	49
Place	Place something	354
	Build a tower out of blocks	3
	Create a route in the GPS	3
	Carry something fragile	4
Comb hair	Comb hair	355
	Massage head	1
	Put brush away	7
	Apply mascara	1
Scratch yourself	Scratch yourself	166
	Think	167
	Gesture	29
	Meditate	2

REFERENCES

[1] J. M. Beer, A. D. Fisk, and W. A. Rogers, "Toward a framework for levels of robot autonomy in human-robot interaction," *J. Hum.-Robot Interact.,* vol. 3, no. 2, pp. 74–99, Jul. 2014. DOI: 10.5898/*JHRI.3.2*.Beer.

[2] Y. Cao, S. Huang, S. L. Sørensen, Y. Yamakawa, and M. Ishikawa, "A Wearable Real-Time 2D/3D Eye-Gaze Interface to Realize Robot Assistance for Quadriplegics," *IEEE Access,* vol. 13, pp. 155259–155276, Aug. 2025, DOI: 10.1109/*ACCESS.2025*.3603617.

[3] MH. Farhadi, A. Rabiee, S. Ghafoori, A. Cetera, W. Xu, and R. Abiri, "Human-Centered Shared Autonomy for Motor Planning, Learning, and Control Applications," in *Bridging the Gap between Mind and Machine,* June 2026, DOI: 10.1007/*978-3-032-06713-5_14*.

[4] A. Fischer-Janzen, T. M. Wendt, and K. Van Laerhoven, "A scoping review of gaze and eye tracking-based control methods for assistive robotic arms," *Front. Robot. AI.,* vol. 11, Feb. 2024. DOI: 10.3389/*frobt.2024*.1326670.

[5] L. V. Herlant, R. M. Holladay, and S. S. Srinivasa, "Assistive teleoperation of robot arms via automatic time-optimal mode switching," in *ACM/IEEE HRI '16,* Christchurch, New Zealand, 2016, pp. 35–42, DOI: 10.1109/*HRI.2016*.7451731.

[6] C. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *ACM/IEEE HRI '16,* Christchurch, New Zealand, 2016, pp. 83–90, DOI: 10.1109/*HRI.2016*.7451737.

[7] H. Admoni and S. S. Srinivasa, "Predicting user intent through eye gaze for shared autonomy," in *AAAI '16,* Phoenix, Arizona, USA, 2016, pp. 585–590.

[8] Z. Zhang, B. Yang, X. Chen, W. Shi, H. Wang, W. Luo, and J. Huang, "Mindeye-Omniassist: A Gaze-Driven LLM-Enhanced Assistive Robot System for Implicit Intention Recognition and Task Execution," in *IEEE CBS '25,* Beijing, China, 2025, pp. 1-6, DOI: 10.1109/*CBS65871.2025*.11267731.

[9] A. Belardinelli, "Gaze-Based Intention Estimation: Principles, Methodologies, and Applications in HRI," *ACM J. Hum.-Robot Interact.,* vol. 13, no. 3, Sep. 2025. DOI: 10.1145/3656376.

[10] A. Fischer-Janzen, T. M. Wendt, D. Görlich, and K. Van Laerhoven, "Eye-Tracking-Driven Shared Control for Robotic Arms: Wizard of Oz Studies to Assess Design Choices," *J. Hum.-Robot Interact.,* Feb. 2026. DOI: 10.1145/3796524.

[11] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," in *IEEE/RSJ IROS '16,* Daejeon, South Korea, 2016, pp. 4193–4198, DOI: 10.1109/*IROS.2016*.7759617.

[12] R. R. Baptista, N. R. Gerszberg, R. V. Godoy, and G. J. G. Lahr, "MIHRaGe: A Mixed-Reality Interface for Human-Robot Interaction via Gaze-Oriented Control," *arXiv:2505.03929 [cs],* May 2025, DOI: 10.48550/*arXiv*.03929.

[13] N. Janaka, S. Zhao, and S. Sapkota. "Can Icons Outperform Text? Understanding the Role of Pictograms in OHMD Notifications," in *ACM CHI '23,* New York, NY, USA, 2023, pp. 1–23. DOI: 10.1145/*3544548*.3580891

[14] G. E. Lancioni, N. N. Singh, M. F. O'Reilly, J. Sigafoos, G. Alberti, O. Troccoli, I. Orlando, and C. Ricci. "Enabling people with intellectual and other disabilities to make verbal requests using cardboard chips with mini objects or pictures and a smartphone," *Front. Rehabil. Sci.,* vol. 28, no. 4, Sep. 2023 DOI: 10.3389/*fresc.2023*.1257493.

[15] K. Hirose, D. Chugo, S. Yokota, and K. Takase, "Camera-based localization for indoor service robots using pictographs," in *IEEE/ASME AIM '11,* Budapest, Hungary, 2011, pp. 924–929, DOI: 10.1109/*AIM.2011*.6027131.

[16] T. Nishiyama, H. Hoshino, K. Sawada, M. Yoneda, Y. Tokunaga, I. Takeuchi, S. Hattori, Y. Ichige, and A. Takanishi, "Development of User Interface for Humanoid Service Robot System," in *Cat. No.03CH37422,* Taipei, Taiwan, vol. 3, pp. 2979–2984, 2003, DOI: 10.1109/*ROBOT.2003*.1242048.

[17] B. Yang, J. Huang, M. Sun, J. Huo, X. Li, and C. Xiong, "Head-free, Human Gaze-driven Assistive Robotic System for Reaching and Grasping," in *IEEE CCC,* Shanghai, China, 2021, pp. 4138-4143.

[18] R. M. Aronson and H. Admoni. "Eye Gaze for Assistive Manipulation," in *ACM/IEEE HRI '20,* New York, NY, USA, 2020, pp. 552–554. DOI: 10.1145/*3371382*.3377434.

[19] L. Wöhle and M. Gebhard, "Towards Robust Robot Control in Cartesian Space Using an Infrastructureless Head- and Eye-Gaze Interface," *MDPI Sensors,* vol. 21, no. 5, Mar. 2021. DOI: 10.3390/s21051798.

[20] B. Yang, J. Huang, X. Chen, X. Li, and Y. Hasegawa, "Natural Grasp Intention Recognition Based on Gaze in Human–Robot Interaction," *IEEE JBHI,* vol. 27, no. 4, pp. 2059-2070, April 2023, DOI: 10.1109/*JBHI.2023*.3238406.

[21] E. Nunez Sardinha, M. Munera, N. Zook, D. Western, and V. Ruiz Garate, "Diegetic Graphical User Interfaces and Intuitive Control of Assistive Robots via Eye-gaze," in *IEEE/RSJ IROS '25,* Hangzhou, China, 2025, pp. 17383–17390, DOI: 10.1109/*IROS.60139.2025*.11247404.

[22] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: a review," *J. Hum.-Robot Interact.*, vol. 6, no. 1, pp. 25-63, May 2017. DOI: 10.5898/*JHRI.6.1*.Admoni.

[23] R. M. Aronson, N. Almutlak, and H. Admoni, "Inferring Goals with Gaze during Teleoperated Manipulation," in *IEEE/RSJ IROS '21,* Prague, Czech Republic, 2021, pp. 7307–7314, DOI: 10.1109/*IROS51168.2021*.9636551.

[24] C. Staub, S. Can, B. Jensen, A. Knoll, and S. Kohlbecher, "Human-computer interfaces for interaction with surgical tools in robotic surgery," in *IEEE RAS & EMBS BioRob,* Rome, Italy, 2012, pp. 81–86, DOI: 10.1109/*BioRob.2012*.6290850.

[25] Z. Wu, B. Ai, T. Silver, and T. Bhattacharjee, "SAVOR: Skill Affordance Learning from Visuo-Haptic Perception for Robot-Assisted Bite Acquisition," *arxiv.2506.02353 [cs],* Sep. 2025, DOI: 10.48550/*arXiv.2506*.02353.

[26] M. Chi, Y. Liu, Q. Zhang, and C. Zeng: "A demonstration trajectory segmentation approach for wheelchair-mounted assistive robots," *CAAI Trans. Intell. Technol.,* vol. 10, no. 3, pp. 738–754, June 2025. DOI: 10.1049/*cit2*.12358

[27] A. Nanavati, V. Ranganeni, and M. Cakmak, "Physically Assistive Robots: A Systematic Review of Mobile and Manipulator Robots That Physically Assist People with Disabilities," *ARCRAS,* vol. 7, pp. 123–147, July 2024. DOI: 10.1146/*annurev-control-062823*-024352.

[28] C. A. Stanger, C. Anglin, W. S. Harwin, and D. P. Romilly, "Devices for assisting manipulation: a summary of user task priorities," *IEEE TNSRE,* vol. 2, no. 4, pp. 256–265, Dec. 1994. DOI: 10.1109/*86*.340872.

[29] S. W. Brose, D. J. Weber, B. A. Salatin, G. G. Grindle, H. Wang, J. J. Vazquez, and R. A. Cooper, "The Role of Assistive Robotics in the Lives of Persons with Disability," *IEEE Am. J. Phys. Med. Rehabil,* vol. 89, no. 6, pp. 509–21, June 2010. DOI: 10.1097/*PHM*.0b013e3181cf569b.

[30] Z. Bien, D. Kim, M. Chung, D. Kwon, and P. Chang, "Development of a wheelchair-based rehabilitation robotic system (KARES II) with various human-robot interaction interfaces for the disabled," in *IEEE/ASME AIM '03,* Kobe, Japan, 2003, pp. 902–907, DOI: 10.1109.*AIM.2003*.1225462.

[31] T. Bhattacharjee, E. K. Gordon, R. Scalise, M. E. Cabrera, A. Caspi, M. Cakmak, and S. S. Srinivasa, "Is More Autonomy Always Better? Exploring Preferences of Users with Mobility Impairments in Robot-assisted Feeding," in *ACM/IEEE HRI '20,* Cambridge, United Kingdom, 2020, pp. 181–190.

[32] Roboflow and Supervision, "Computer Vision Model Leaderboard," Roboflow. 27. Dec. 2025. [Online]. Available: https://leaderboard.roboflow.com/

[33] Y. Tian, Q. Ye, and D. Doermann, "YOLOv12: Attention-Centric Real-Time Object Detectors," in *NeurIPS,* Dec. 2025, DOI: 10.48550/*arXiv*.12524.

[34] B. Dwyer, J. Nelson, T. Hansen et al., "Roboflow (Version 1.0) [Software]," Roboflow. 27. Dec. 2025. [Online]. Available: https://roboflow.com.

[35] K-M. Ponomarjova, A. Fischer-Janzen, T. M. Wendt, and K. Van Laerhoven, "WLRI-AD: assistive device dataset for daily living automation," *Pers. Ubiquit. Comput.,* vol. 30, no. 2. Jan. 2026. DOI: 10.1007/s00779-026-01854-2.

[36] Intel Corporation, "ROS Wrapper for Intel(R) RealSense(TM) Cameras," Git Hub. 27. Dec. 2025. [Online]. Available: https://github.com/IntelRealSense/realsense-ros

[37] M. Kudo, "Graphic design of pictograms focusing on the comprehension of people with intellectual disabilities–The next step in standardization: Pictogram design and evaluation methods," *Visible Language,* vol. 56, no. 3, pp. 58-85, Dec. 2022.

[38] G. Canal, G. Alenyà, and C. Torras, "Personalization Framework for Adaptive Robotic Feeding Assistance," in *ICSR '16,* vol. 9979, 2016, pp. 22-31. DOI: 10.1007/*978-3-319-47437-3*_3.

[39] R. Feng, Y. Kim, G. Lee, E. K. Gordon, M. Schmittle, S. Kumar, T. Bhattacharjee, and S. S. Srinivasa, "Robot-Assisted Feeding: Generalizing Skewering Strategies across Food Items on a Realistic Plate," *arXiv:1906.02350 [cs],* Sep. 2019, DOI: 10.48550/*arXiv*.02350.

[40] T. Bhattacharjee, G. Lee, H. Song, and S. S. Srinivasa, "Towards Robotic Feeding: Role of Haptics in Fork-Based Food Manipulation," *IEEE RA-L,* vol. 4, no. 2, pp. 1485-1492, April 2019. DOI: 10.1109.*LRA.2019*.2894592.

[41] M. Pascher, A. Baumeister, S. Schneegass, B. Klein, and J. Gerken, "Recommendations for the Development of a Robotic Drinking and Eating Aid - An Ethnographic Study," in *INTERACT 2021,* Bari, Italy, 2021, pp. 331–351.

[42] H. Wang, X. Dong, Z. Chen, and B. E. Shi, "Hybrid gaze/EEG brain computer interface for robot arm control on a pick and place task," in *IEEE EMBC,* Milan, Italy, 2015, pp. 1476–1479, DOI: 10.1109.*EMBC.2015*.7318649.

[43] A. Shafti, P. Orlov, and A. A. Faisal, "Gaze-based, Context-aware Robotic System for Assisted Reaching and Grasping," in *IEEE ICRA '19,* Montreal, QC, Canada, 2019, pp. 863–869, DOI: 10.1109/*ICRA.2019*.8793804.

[44] F. Koochaki and L. Najafizadeh, "A Data-Driven Framework for Intention Prediction via Eye Movement With Applications to Assistive Systems," *IEEE TNSRE,* vol. 29, pp. 974–984, May 2021, DOI: 10.1109/*TNSRE.2021*.3083815.

[45] L. Petrich, J. Jin, M. Dehghan, and M. Jagersand, "A Quantitative Analysis of Activities of Daily Living: Insights into Improving Functional Independence with Assistive Robotics," *IEEE ICRA '22,* Philadelphia, PA, USA, 2022, pp. 6999–7006, DOI: 10.1109/*ICRA46639.2022*.9811960.

[46] A. Fischer-Janzen, T. M. Wendt, and K. Van Laerhoven"Eye-Tracking-Driven Control in Daily Task Assistance for Assistive Robotic Arms," *arxiv.2601.17404 [cs],* Jan. 2026, DOI: 10.48550/*arXiv*.2601.17404.

[47] K. Park, S. Choi, H. Moon, J. Lee, Y. Ghasemi, and H. Jeong, "Indirect Robot Manipulation using Eye Gazing and Head Movement For Future of Work in Mixed Reality," in *IEEE VRW,* Christchurch, New Zealand, 2022, pp. 585–590, DOI: 10.1109/*VRW55335.2022*.00107.

[48] R. M. Aronson, T. Santini, T. C. Kübler, E. Kasneci, S. Srinivasa, and H. Admoni, "Eye-Hand Behavior in Human-Robot Shared Manipulation," in *ACM/IEEE HRI '18,* Chicago, IL, USA, 2018, pp. 4–13.

[49] G. Yu, Y. Hu, and J. Dai. "TopoTag: A Robust and Scalable Topological Fiducial Marker System," in *IEEE TVCG '21,* vol. 27, no. 9, Sept. 2021, pp. 3769–3780, DOI: 10.1109/*TVCG.2020*.2988466.

ANKE FISCHER-JANZEN received the M.Sc. degree in electrical engineering from the Karlsruhe Institute of Technology, Karlsruhe, in 2019 and is currently working as doctorate student for a doctor of engineering in Information Technology from University of Siegen, Siegen, Germany, since 2021. She worked in the Ubiquitous Computing group with Kristof Van Laerhoven and in various projects in the Work-Life Robotics Institute with Thomas M. Wendt.

From 2020 to 2026, she was a Research staff member with the Work-Life Robotics Institute at Offenburg University of Applied Sciences. Her research interests include assistive robotics and human-robot interaction with focus on the control of articulated robotic arms by gaze, and published multiple articles in various conferences and journals. She is Reviewer in journals such as *Assistive Technologies* and conferences such as *CHI* and *HRI*.

THOMAS M. WENDT received the M.Sc. degree in mechatronics from De Montfort University in Leicester, U.K., in 1999, and the Ph.D. degree from the Albert-Ludwig-University of Freiburg, Germany, in 2009, as part of Prof. Leonhard Reindl's Department Of Microsystems (IMTEK). Before his current role, he was a Professor for electrical engineering and Professor for robotics, additively manufactured electronics and wireless technologies with the Offenburg University of Applied Sciences. He has currently an Endowed Professorship for Cobotics and Socio-Technological Systems and is head of the Work-Life Robotics Institute with the Offenburg University of Applied Sciences, Germany. His research interests include robotics and safety, focusing on the use of AI, humanoid robotics, sensing, and smart grippers. He is an active member of the robotics community, contributing as an reviewer and editor for leading journals and conferences, and holds six patents.

KRISTOF VAN LAERHOVEN received the M.Sc. degree in computer science from the University of Brussels, Belgium, in 1999, and the Ph.D. degree from the University of Lancaster, U.K., in 2006, as part of Hans Gellersen's Embedded Interactive Systems Group. He worked on projects with the Luc Steels' AI Laboratory, University of Brussels. Before his current role, he was a Professor in embedded systems with the University of Freiburg. He led an Emmy Noether Junior Research Group in embedded sensing systems with the Technical University of Darmstadt, funded by German Research Foundation (DFG). He is currently a Full Professor in ubiquitous computing with the University of Siegen, Germany. His research interests include wearable and distributed sensing systems, particularly in machine learning, for recognizing human activities, attention, and stress patterns. He is also an active member of the ubiquitous computing community, contributing as an editor for leading journals and conferences.