Teaching Unknown Objects by Leveraging Human Gaze and Augmented Reality in Human-Robot Interaction

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Daniel Weber

aus Tübingen

Tübingen

2023

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	02.11.2023
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatterin:	Prof. Dr. Enkelejda Kasneci
2. Berichterstatter:	Jun.-Prof. Dr. Michael Krone

To my wife, my parents, my sister, and my grandmother, whose unwavering love and endless support are greater than my words could ever be.

Acknowledgments

In the course of my doctoral research in recent years, I have had the privilege of engaging in many exciting collaborations and fruitful discussions with numerous people, all of whom I have benefited from and whom I would like to express my sincere gratitude to. The successful completion of my research, culminating in this dissertation, would not have been possible without their support.

First of all, I would like to thank my two supervisors, Prof. Dr. Enkelejda Kasneci and Prof. Dr. Andreas Zell, for hosting me at the chairs Human-Computer Interaction and Cognitive Systems, respectively, and whose guidance has profoundly shaped both my research and my academic pursuits. The former, in particular, has been a consistent source of support, both in research and in general matters. Furthermore, special and sincere appreciation also go to Jun.-Prof. Dr. Michael Krone and Dr. Shahram Eivazi for their immediate willingness to evaluate my work. I am also very grateful to Vita Serbakova and Margot Reimold for always assisting me with the universities' bureaucracy, and to Uli Ulmer for providing all the exceptional technical support and consistently being reliably eager to help. In addition, I would also like to thank the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) for funding me and my project, as well as Michaela Bitzer, who has always been a reliable contact person at the university in these matters.

Throughout my journey, I have been fortunate to work with a number of former and present colleagues, whose experiences have greatly enriched my own. Foremost, I would like to express my heartfelt gratitude to Dr. Thiago Santini, Dr. David Geisler, Dr. Wolfgang Fuhl, and Dr. Nora Castner. I can't thank you enough for your warm and gracious welcome and for giving me an incredibly delightful start by taking me into your care. I deeply appreciate your patience in answering my countless questions and I truly enjoyed all our amusing conversations and funny moments both within our office and during the numerous "non-smoking smoking breaks". Special thanks also extend to the rest of my esteemed colleagues from both the Human-Computer Interaction team as well as the Cognitive Systems group, including Dr. Thomas Kübler, Dr. Efe Bozkir,

Acknowledgments

Dr. Nuri Benbarka, Dr. Julian Jordan, Dr. Benedikt Hosp, Dr. Hong Gao, Björn Severitt, Mario Laux, and many others, for their support and assistance whenever I was in need. This also applies in particular to Benjamin Kiefer, Martin Meßmer, Timon Höfer, and Valentin Bolz, who deserve special recognition. Your moral support and exhilarating gallows humor have been instrumental in making my academic journey so enjoyable and memorable. Especially the latter has been an invaluable presence in my life as a long-time and incredibly close friend.

Finally, and above all, I would like to thank my parents Brunhilde and Waldemar, my sister Marina, together with Sebastian and Lena, and my grandmother Elfriede, from the bottom of my heart. Your support, encouragement, and unconditional love mean more to me than you could possibly imagine. Likewise, I want to express my profound gratitude to my wife, Carmen, who has always believed in me, encouraged me, lifted me up, helped me through challenging times and accompanied me through all the ups and downs. Your support has been an immeasurable blessing in my life. Thank you for always being there for me – I am deeply indebted to you.

Daniel Weber

Abstract

Robots are becoming increasingly popular in a wide range of environments due to their exceptional work capacity, precision, efficiency, and scalability. This development has been further encouraged by advances in Artificial Intelligence (AI), particularly Machine Learning (ML). By employing sophisticated neural networks, robots are given the ability to detect and interact with objects in their vicinity. However, a significant drawback arises from the underlying dependency on extensive datasets and the availability of substantial amounts of training data for these object detection models. This issue becomes particularly problematic when the specific deployment location of the robot and the surroundings, including the objects within it, are not known in advance. The vast and ever-expanding array of objects makes it virtually impossible to comprehensively cover the entire spectrum of existing objects using preexisting datasets alone.

The goal of this dissertation was to teach a robot unknown objects in the context of Human-Robot Interaction (HRI) in order to liberate it from its data dependency, unleashing it from predefined scenarios. In this context, the combination of eye tracking and Augmented Reality (AR) created a powerful synergy that empowered the human teacher to seamlessly communicate with the robot and effortlessly point out objects by means of human gaze. This holistic approach led to the development of a multimodal HRI system that enabled the robot to identify and visually segment the Objects of Interest (OOIs) in three-dimensional space, even though they were initially unknown to it, and then examine them autonomously from different angles. Through the class information provided by the human, the robot was able to learn the objects and redetect them at a later stage. Due to the knowledge gained from this HRI based teaching process, the robot's object detection capabilities exhibited comparable performance to state-of-the-art object detectors trained on extensive datasets, without being restricted to predefined classes, showcasing its versatility and adaptability.

The research conducted within the scope of this dissertation made significant contributions at the intersection of ML, AR, eye tracking, and robotics. These findings not only

enhance the understanding of these fields, but also pave the way for further interdisciplinary research. The scientific articles included in this dissertation have been published at high-impact conferences in the fields of robotics, eye tracking, and HRI.

Zusammenfassung

Roboter finden aufgrund ihrer außergewöhnlichen Arbeitsleistung, Präzision, Effizienz und Skalierbarkeit immer mehr Verwendung in den verschiedensten Anwendungsbereichen. Diese Entwicklung wurde zusätzlich begünstigt durch Fortschritte in der Künstlichen Intelligenz (KI), insbesondere im Maschinellem Lernen (ML). Mit Hilfe moderner neuronaler Netze sind Roboter in der Lage, Objekte in ihrer Umgebung zu erkennen und mit ihnen zu interagieren. Ein erhebliches Manko besteht jedoch darin, dass das Training dieser Objekterkennungsmodelle, in aller Regel mit einer zugrundeliegenden Abhängig von umfangreichen Datensätzen und der Verfügbarkeit großer Datenmengen einhergeht. Dies ist insbesondere dann problematisch, wenn der konkrete Einsatzort des Roboters und die Umgebung, einschließlich der darin befindlichen Objekte, nicht im Voraus bekannt sind. Die breite und ständig wachsende Palette von Objekten macht es dabei praktisch unmöglich, das gesamte Spektrum an existierenden Objekten allein mit bereits zuvor erstellten Datensätzen vollständig abzudecken.

Das Ziel dieser Dissertation war es, einem Roboter unbekannte Objekte mit Hilfe von Human-Robot Interaction (HRI) beizubringen, um ihn von seiner Abhängigkeit von Daten sowie den Einschränkungen durch vordefinierte Szenarien zu befreien. Die Synergie von Eye Tracking und Augmented Reality (AR) ermöglichte es dem als Lehrer fungierenden Menschen, mit dem Roboter zu kommunizieren und ihn mittels des menschlichen Blickes auf Objekte hinzuweisen. Dieser holistische Ansatz ermöglichte die Konzeption eines multimodalen HRI-Systems, durch das der Roboter Objekte identifizieren und dreidimensional segmentieren konnte, obwohl sie ihm zu diesem Zeitpunkt noch unbekannt waren, um sie anschließend aus unterschiedlichen Blickwinkeln eigenständig zu inspizieren. Anhand der Klasseninformationen, die ihm der Mensch mitteilte, war der Roboter daraufhin in der Lage, die entsprechenden Objekte zu erlernen und später wiederzuerkennen. Mit dem Wissen, das dem Roboter durch diesen auf HRI basierenden Lehrvorgang beigebracht worden war, war dessen Fähigkeit Objekte zu erkennen vergleichbar mit den Fähigkeiten modernster Objektdetektoren, die auf umfangreichen Datensätzen trainiert

worden waren. Dabei war der Roboter jedoch nicht auf vordefinierte Klassen beschränkt, was seine Vielseitigkeit und Anpassungsfähigkeit unter Beweis stellte.

Die im Rahmen dieser Dissertation durchgeführte Forschung leistete bedeutende Beiträge an der Schnittstelle von Machine Learning (ML), AR, Eye Tracking und Robotik. Diese Erkenntnisse tragen nicht nur zum besseren Verständnis der genannten Felder bei, sondern ebnen auch den Weg für weitere interdisziplinäre Forschung. Die in dieser Dissertation enthalten wissenschaftlichen Artikel wurden auf hochrangigen Konferenzen in den Bereichen Robotik, Eye Tracking und HRI veröffentlicht.

Contents

List of Figures

List of Tables

List of Abbreviations

AGV	Automated Guided Vehicle
AI	Artificial Intelligence
AMR	Autonomous Mobile Robot
AR	Augmented Reality
DGA-GBVS	Dual Gaze-Assisted GBVS
FOV	Field of View
GA-GBVS	Gaze-Assisted GBVS
GBVS	Graph-Based Visual Saliency
HRC	Human-Robot Collaboration
HRI	Human-Robot Interaction
IoU	Intersection over Union
KNN	k-Nearest Neighbor
LLM	Large Language Model
mAP	mean average precision
mAR	mean average recall
ML	Machine Learning
OMD	Objects in Multiperspective Detail
OOI	Object of Interest
PbD	Programming by Demonstration
RANSAC	random sample consensus
RLHF	Reinforcement Learning from Human Feedback
ROI	Region of Interest

List of Abbreviations

ROS	Robot Operating System
SVM	Support Vector Machine
UAV	Unmanned Aerial Vehicle
UWP	Universal Windows Platform
VR	Virtual Reality

1 List of Publications

1.1 Publications Relevant to This Thesis

[1] **Daniel Weber**, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling Location Proposals of Unknown Objects through Gaze Information for Human-Robot Interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11086–11093. IEEE, 2020. doi:10.1109/IROS45743.2020.9340893.

[2] David Geisler, **Daniel Weber**, Nora Castner, and Enkelejda Kasneci. Exploiting the GBVS for Saliency aware Gaze Heatmaps. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5, 2020. doi:10.1145/3379156.3391367.

[3] **Daniel Weber**, Enkelejda Kasneci, and Andreas Zell. Exploiting Augmented Reality for Extrinsic Robot Calibration and Eye-based Human-Robot Collaboration. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 284–293. IEEE, 2022. doi:10.1109/HRI53351.2022.9889538.

[4] **Daniel Weber**, Wolfgang Fuhl, Andreas Zell, and Enkelejda Kasneci. Gaze-based Object Detection in the Wild. In *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*, pages 62–66. IEEE, 2022. doi:10.1109/IRC55401.2022.00017.

[5] **Daniel Weber**, Wolfgang Fuhl, Enkelejda Kasneci, and Andreas Zell. Multiperspective Teaching of Unknown Objects via Shared-gaze-based Multimodal Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 544–553, March 2023. doi:10.1145/3568162.3578627.

[6] **Daniel Weber**, Valentin Bolz, Andreas Zell, and Enkelejda Kasneci. Leveraging Saliency-Aware Gaze Heatmaps for Multiperspective Teaching of Unknown Objects. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. (Accepted for publication).

1.2 Further Publications

[7] Wolfgang Fuhl, **Daniel Weber**, and Shahram Eivazi. The Gaze and Mouse Signal as Additional Source for User Fingerprints in Browser Applications. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP*, pages 117–124. INSTICC, SciTePress, 2023. doi: 10.5220/0011607300003417.

[8] Wolfgang Fuhl, **Daniel Weber**, and Shahram Eivazi. Pistol: Pupil Invisible Supportive Tool to Extract Pupil, Iris, Eye Opening, Eye Movements, Pupil and Iris Gaze Vector, and 2D as well as 3D Gaze. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP*, pages 27–38. INSTICC, SciTePress, 2023. doi: 10.5220/0011607200003417.

[9] Wolfgang Fuhl, **Daniel Weber**, and Shahram Eivazi. GroupGazer: A Tool to Compute the Gaze per Participant in Groups with Integrated Calibration to Map the Gaze Online to a Screen or Beamer Projection. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP*, pages 109–116. INSTICC, SciTePress, 2023. doi: 10.5220/0011607000003417.

[10] Wolfgang Fuhl, Björn Severitt, Nora Castner, Babette Bühler, Johannes Meyer, **Daniel Weber**, Regine Lendway, Ruikun Hou, and Enkelejda Kasneci. Watch out for those bananas! Gaze Based Mario Kart Performance Classification. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–2. ACM, 2023. doi: 10.1145/3588015.3590136.

1.3 Scientific Contribution

This work explores a holistic perspective on Human-Robot Interaction (HRI), where interdisciplinary synergies from the research domains of Augmented Reality (AR), Machine Learning (ML), eye tracking and robotics empower a human to teach a robot novel objects, facilitating its adaptability to cope with unfamiliar environments. The most noteworthy contributions, as illustrated in Figure 1.1, encompass (1) the development of a natural and intuitive interaction channel between humans and robots utilizing AR and eye tracking, (2) a practical extrinsic robot calibration method, (3) multiple approaches for the identification of Regions of Interest (ROIs), (4) the segmentation of unknown objects through gaze-based HRI, and (5) robot learning by means of automatically recorded and labelled

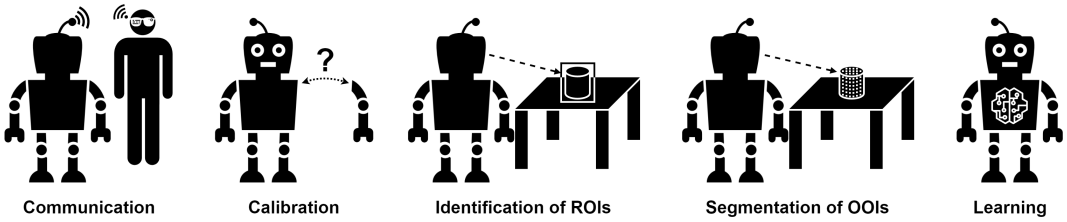


Figure 1.1: This thesis unites realms of research, that were previously running predominantly in parallel. The contributions cover a spectrum of domains including HRI, robotics, computer vision, machine learning, and various others. The figure provides an overview, visualizing some of the most significant contributions.

data. These contributions have been published at renowned conferences in the field of robotics, eye tracking, and HRI, and have paved the way for further research on robotic teaching.

The structure of this dissertation is as follows: Chapter 2 commences by emphasizing the differences between humans and robots and outlining how robots benefit humans. Following the various application areas of robots, their deficiencies and learning capabilities are explained. Subsequently, diverse approaches are described how humans can teach robots. These include, among others, the usage of AR and eye tracking, which in particular offer great potential for teaching unknown objects and which are further elaborated on in the context of HRI. Then, the general setting is presented, and the objectives are defined. At the end of the chapter, the utilized hardware is described and the fundamental terminology for subsequent evaluations is introduced. Chapter 3 presents the major contributions of the six scientific publications listed in Section 1.1. This encompasses the motivation, the methodology, as well as the main contributions and results of each paper. Finally, Chapter 4 concludes with a discussion of the findings, achievements and limitations of this work and provides an outlook on future research. The research conducted in the context of this dissertation was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy "Machine Learning: New Perspectives for Science" – EXC number 2064/1 – Project number 390727645.

2 Introduction

Humans and robots possess distinct capabilities and characteristics due to their inherent nature and design. These include, but are not limited to, emotional intelligence, intuition and creativity, learning and adaptation, physical and cognitive abilities, as well as physical and mental limitations.

Presumably one of the more obvious differences is intuition and creativity. Gut feelings, intuition, and a capacity for creativity allow humans to think outside the box. This quality facilitates the development of innovative ideas and solutions to problems. Robots, on the other hand, lack this genuine intuition and creativity as they are governed by logical patterns and unequivocal algorithms [11]. Although advances in robotics and Artificial Intelligence (AI) have enabled them to emulate certain aspects of intuition and creativity, they are still far from covering the entire spectrum of human intuition and creativity.

Closely related are the cognitive abilities. Humans are adept at interpreting ambiguous information, comprehending context, and making decisions based on multiple factors, which stem from their abilities in complex reasoning and critical thinking. While robots can accomplish certain cognitive tasks, such as identifying patterns or analyzing data, accurately and fast, their cognitive capabilities are limited and rarely extend beyond familiar circumstances [12].

Emotional intelligence empowers humans to express feelings and to convey, recognize, and understand emotions. In this way, we are able to empathize and connect with other human beings, develop social bonds, and build relationships. In contrast, robots are devoid of any concept of real emotion. Even though they can simulate or react to predefined emotions, their emotional abilities are exclusively artificial [13].

In terms of physicality, humans and robots differ in both capabilities and underlying limitations. Humans have versatile and adaptable physical abilities, such as dexterity, fine motor skills, and a broad portfolio of movements. However, they also experience physical and mental limitations owing to factors like fatigue, inconsistency regarding repetitions, and biological constraints. Robots, on the other hand, are proficient in precise

Figure 2.1: Robots excel over humans in terms of physical abilities, precision, and speed, rendering them exceptionally well-suited for tasks such as factory automation, as shown in the left image[1]. However, robots face limitations in their natural ability to autonomously adapt to unpredictable situations or operate in unfamiliar environments due to their lack of inherent abstraction capabilities. These can lead to failures under real-life conditions, as evident in the image[2] on the right, where an autonomous delivery vehicle got stuck due to an unexpected obstacle.

and repetitive actions, and are endowed with higher levels of strength and endurance in comparison to humans [14]. They are often engineered with specific physical characteristics required to accomplish a particular task. Apart from a dependency on power supply and maintenance, this tailored design and the absence of an intrinsic awareness of their surroundings often leaves them inflexible and inadequate to operate under unfamiliar conditions [15].

Some of these mentioned attributes render robots superior to humans for certain tasks, and others inferior (see Figure 2.1). Either way, this repertoire of characteristics can benefit humans, which is why robots continue to enter a growing array of application fields, serving as a valuable complement to human capabilities.

2.1 Applications Areas of Robots

Robots' excellent precision, working capacity, efficiency, and ability to operate in arduous and hazardous environments creates an increasing demand for robots in a variety of sectors. In the manufacturing industry, they are deployed extensively for tasks such as assembly [18, 19], painting [20, 21], welding [22, 23], packaging [24, 25], and quality

[1] © Haophuong21 / Wikimedia Commons / CC BY-SA 4.0 / Cropped from original. See [16].
[2] © Mbrickn / Wikimedia Commons / CC BY 4.0. See [17].

control [26, 27]. To this end, they are designed to perform these tasks repetitively, fast, and with high precision. Furthermore, in warehouses and distribution centers, robots are used to sort [28, 29], and Autonomous Mobile Robots (AMRs) and Automated Guided Vehicles (AGVs) are used to handle materials [30, 31, 32] to reduce human labor. Within the medical and healthcare sector, robots play an essential role conducting diagnoses [33, 34], assisting in surgeries [35, 36], and supporting rehabilitation [37, 38]. While robotic exoskeletons provide physiotherapy and mobility assistance for people with disabilities, surgical robots enable minimally invasive surgical operations. Agricultural robots, also known as agribots or agrobots, are utilized for agricultural purposes such as weeding [39, 40, 41], planting seeds [42, 43], disease and insect detection [44, 45, 46, 47], plant monitoring [48, 49, 50], spraying pesticides [51, 52, 53] and harvesting crops [54, 55, 56, 57]. They reduce the resources and labor costs required, protect humans and the environment from unnecessary use of chemical substances, and enable precision farming techniques. Furthermore, robots are deployed in space as part of space exploration missions to investigate distant planets, moons, and asteroids [58, 59, 60, 61]. Among others, their tasks include gathering data, conducting experiments and facilitating research. In defense and security, application fields of robots include surveillance [62, 63, 64], reconnaissance [65, 66, 67], support and rescue [68, 69, 70], as well as the detection and disposal of underground mines [71, 72, 73]. In these contexts, Unmanned Aerial Vehicles (UAVs) also play a crucial role because they can operate in impassable terrain: For example, they are highly suitable for search missions in maritime environments [74]. In the domestic and personal sphere, robots are primarily used as vacuum cleaners [75, 76, 77], lawn mowers [78, 79, 80] or social companions [81, 82].

The ever-expanding applications and benefits of robots have been substantially fertilized by the advances in the research field of AI, especially ML. These advancements have unlocked capabilities in robots for acquiring and learning certain behaviors to act in specific situations in a similar manner to humans or as desired by humans. However, the way robots learn fundamentally differs from that of humans. This foundational distinction expands upon the differences between humans and robots discussed earlier in this chapter and further emphasizes the unique nature of robot learning.

2.2 Robot Deficiencies and Learning

Humans possess a broad spectrum of learning styles, such as observational learning [83], conceptual learning [84] or learning by trial and error [85], feedback [86] or by means of a teacher [87]. This variety empowers them to link different concepts, understand and interpret complex contexts, as well as to perceive subtle clues and approach problems

Figure 2.2: In the left image[3], a consignment robot in a pharmacy retrieves packages of medicines from a storage shelf, and in the right image[4], a drone harvests apples. In both applications, the robot must detect the respective OOI prior to performing the action.

from unconventional angles. By extracting underlying principles, building contextual understanding and reusing past experiences, humans can adapt their knowledge to new circumstances and varying environments, respond to unexpected changes and apply their skills to other domains. The deficiency in the capacity for such generalizations and the lack of adaptability of behaviors and strategies are among the key shortcomings of robots. Robots suffer from an inability to think beyond established patterns, as their learning behavior primarily relies on programmed algorithms or machine learning techniques; the latter usually demands data-driven training to acquire new skills. Even in a familiar environment with a well-defined setting, a substantial amount of training data is required. On top of that, any change in circumstances may necessitate retraining. When robots are challenged with unpredictable or not predefined conditions that they have not been explicitly prepared for, their performance is negatively affected. However, deployed in close proximity to humans, the robot can be assisted by the human in dealing with such unfamiliar scenarios [12]. This implies that HRI holds the potential to mitigate the previously outlined disparities to some extent. Through interactive teaching by a human, knowledge can be imparted to the robot and obstacles can be collectively overcome.

In many of the robot applications described in Section 2.1, the identification of the relevant Objects of Interest (OOIs) plays a crucial role, as illustrated in Figure 2.2. The capability to detect them must be either acquired beforehand – often data-driven – or taught during runtime, for example by means of HRI or Human-Robot Collaboration (HRC). The former is the most popular and most widespread variant. Based on an available data set, the robot is trained and is then capable of interacting with the respective

[3] © UKM Elisabeth Deiters-Keul / Wikimedia Commons / CC BY-SA 3.0 / Cropped from original. See [88].
[4] © Tevel Aerobotics Technologies / Cropped from original. See [89].

objects, enabling potential subsequent actions like grasping. However, such an approach is based on two underlying assumptions. First and foremost, it is imperative that the future OOIs are known in advance. Secondly, the existence and accessibility of appropriate training data encompassing all OOIs is indispensable. These aspects are the crux of the issue at hand. The plethora of objects and thus potential interaction entities is – at least for all practical matters – basically unlimited. Consequently, the existence of appropriate data cannot be assumed without further ado. The problem that arises from such data dependency has already been hinted at above. If the setting is not precisely predefined and limited to fixed set of OOIs, this will impair the performance of the robot and thus prevent its successful deployment. Herein lies the necessity of the contributions of this work, as it becomes increasingly important to find alternative concepts that disengage from dataset dependency in robot learning. This dissertation revolves around one such an approach that entails the human teaching the robot the unknown objects, assisting it to cope in environments where novel objects are encountered.

2.3 Towards Teaching Robots

In general, robots can learn from humans in different ways. One possibility, for example, is through demonstration, which is particularly popular for teaching assembly tasks [90, 91, 92, 93]. Here, the robot typically observes the actions performed by a human and then replicates them through imitation. In a related framework of Programming by Demonstration (PbD), a human teacher assumes the role of manually guiding the robot [94, 95, 96], either by physically manipulating its limbs or utilizing an interface. The robot meticulously captures and stores the teacher's actions, enabling it to generalize and autonomously execute similar tasks in the future. Another teaching technique is Reinforcement Learning from Human Feedback (RLHF), which integrates valuable human insights into reinforcement learning [97, 98]. By incorporating evaluative feedback, such as rewards or punishments, provided by a human collaborator, the robot utilizes this information to refine its learning process. Through iteration, the robot progressively enhances its performance based on the guidance received through the human's assessments. Instructions in natural language offer another avenue for robots to learn from humans [99, 100]. The robot processes the verbal commands of a human teacher and can thus learn new tasks, such as grasping [101] or navigating [102], or improve its existing skills. Moreover, AR and Virtual Reality (VR) technologies offer immersive environments that facilitate the training of robots by humans [103, 104, 105, 106]. Within these augmented or virtual spaces, the human can interact with virtual objects or simulated events and the robot learns by assimilating the transferred information. Naturally, there are various other

methods beyond the scope of this dissertation that cannot all be discussed in detail. For more information and examples, the readers are referred to [107, 108, 109].

With regard to the intended teaching of novel objects, AR demonstrates significant potential. This technology stands out as a promising avenue to realize natural interaction and collaboration between human and robot.

2.4 Human-Robot Interaction Employing AR and Eye Tracking

The combination of robotics with augmented reality [110, 111, 112] and robotics with human eye tracking [113, 114, 115] led to the intensification of the interaction between humans and robots.

AR refers to the technology that integrates computer-generated content into the real world. By overlaying digital information, such as virtual objects, onto the physical world, the latter is augmented. According to Azuma [116], AR can be defined as a system that exhibits the following three properties: It combines real and virtual world, allows real-time interaction, and accurately registers virtual and real objects in 3D. The primary advantage of AR lies in its inherent ability to blend digital components into the user's individual perception of the physical world. Rather than merely displaying data, AR embeds immersive sensory modalities that are perceived as natural parts of the environment and enhance the user's overall experience. AR functions across a wide range of domains, such as medicine [117, 118, 119, 120], military [121, 122, 123], manufacturing [124, 125, 126], entertainment and games [127, 128, 129], education [130, 131, 132, 133], navigation and path planning [134, 135, 136], as well as tourism [137, 138, 139, 140]. Through integration of AR into HRI scenarios, the human gains awareness of the robot's state and intentions. AR offers a more immersive and contextually rich interaction by visualizing supplementary information of the robot that is not directly observable by the human [141]. In summary, AR bridges the digital world of the robot and the analog world of the human to facilitate communication between them.

Owing to the growing prevalence of AR glasses, such as the Microsoft HoloLens 2[5], the Magic Leap 2[6] or the recently announced Apple Vision Pro[7], eye tracking can be adroitly employed alongside AR. Eye tracking is the process of measuring a person's eye movements and focus. This technology enables the analysis and comprehension of eye behavior, such as the direction of the gaze [142, 143], the fixation duration on specific points [144], and gaze patterns [145]. It is applied in numerous fields and research areas,

[5] https://www.microsoft.com/en-us/hololens (accessed: June 16, 2023)
[6] https://www.magicleap.com/magic-leap-2 (accessed: June 16, 2023)
[7] https://www.apple.com/apple-vision-pro (accessed: June 16, 2023)

spanning psychology [146, 147, 148], marketing [149, 150, 151], driving [152, 153, 154, 155, 156, 157, 158, 159, 160], sports [161, 162, 163, 164], education [165, 166, 167, 168, 169, 170, 171, 172, 173], and medicine [174, 175, 176, 177, 178, 179, 180]. In the realm of robotics, especially with regard to HRI, eye tracking can be used, for example, to allow robots to proactively anticipate and perform tasks based on the eye movements of their human partner [181] or to empower people with disabilities to control assistive robots [182]. In this HRI context, eye tracking plays a pivotal role, as it enables the robot to perceive the focus of the human's attention. By tracking the user's eye movements, the system can ascertain where the user is looking, and subsequently, the robot can react in accordance with the interpreted user intentions. This seamless and natural interaction enables the human to effortlessly issue commands or express interest by simply looking at objects or specific locations of interest.

These benefits offered by AR and eye tracking were leveraged to close the existing gap of robots being confined to predefined scenarios, mentioned in the concluding remarks of Section 2.2. The interdisciplinary contributions of this dissertation furthered this holistic perspective, ultimately fulfilling the objectives detailed in the following.

2.5 Setting and Objectives

As briefly covered in Section 2.2, the central goal of this dissertation was to develop a framework for teaching a robot unknown objects, with a strong focus on flexibility in non-predefined scenarios. For this purpose, the human directs the robot's attention towards the OOI by looking at it and pointing it out via gaze. Subsequently, after the robot has segmented the object visually, it records it from slightly different angles and finally learns and later redetects it based on the class information provided by the human during the teaching process. The general setting is visualized in Figure 2.3.

In order to accomplish this overall goal, a detailed series of intermediate steps had to be completed. More specifically, this thesis addresses each of the following challenges and provides methods to approach them in realistic settings:

C1. The human gaze has to be mapped from the human's frame of reference to the robot's frame of reference to convey the information regarding where the human is looking and directing their gaze.

C2. The identification and localization of the OOI on the robot side must be accomplished. To this end, the robot needs to visually segment the object, even though the object is not yet known to it at this point. This segmentation process allows the robot to distinguish and delimit the object from its surroundings.

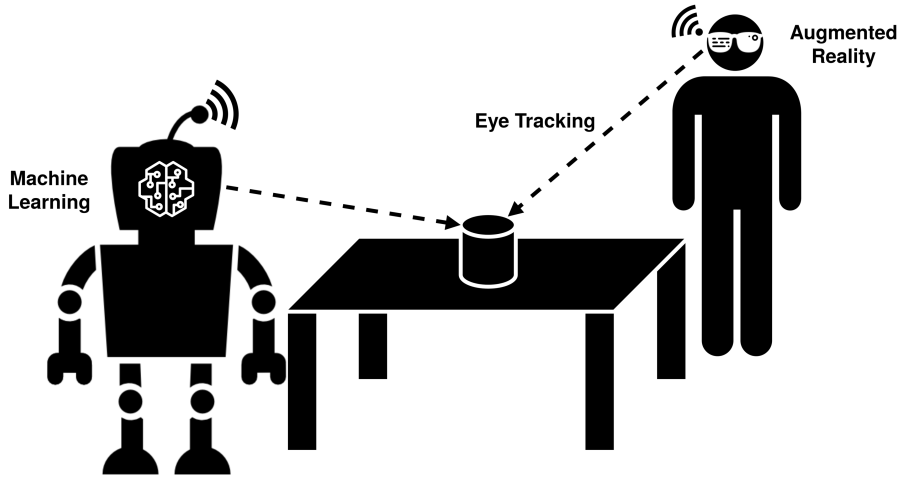


Figure 2.3: The human and the robot both stand in front of a table. The human selects the OOI using gaze. The robot must identify the object correctly and then learn it by means of the class information which the human provides. The communication takes place through an AR interface and Wi-Fi.

C3. In order for the human and the robot to communicate with each other and exchange data at all, a bidirectional connection needs to be established between them using an AR interface. The implementation of this dedicated interface is essential, as it serves the dual purpose of enabling communication and tracking the human gaze.

C4. Prior to the human teaching the robot, the human must ascertain whether both, human and robot, are sharing their attention on the same object. Meaning, the human needs to receive feedback regarding the segmented OOI in order to verify and intervene if necessary.

C5. As part of the teaching process, the human must impart the class information to the robot.

C6. To facilitate the robot's learning process, it is crucial to develop a procedure by which the robot can examine the object from various perspectives, as previously described, and generate autonomously labeled training data. Subsequently, this data can be utilized to train the robot using ML models.

All of the above intermediate challenges were successfully completed through pronounced interdisciplinary efforts and all contributed to the higher holistic objective. Their

solutions will be explained in more detail in Chapter 3 and Chapter 4. Upon completion, the robot's capability to redetect taught objects could be assessed and evaluated.

2.6 Hardware and Evaluation Fundamentals

2.6.1 Hardware

The HRI teaching task described above requires a variety of different hardware. This hardware will be described in more detail below, as it reoccurs throughout the dissertation.

In the first part of Chapter 3, for the general investigation of the potential of eye tracking in determining ROIs, standard eye trackers, decoupled from any AR functionality, are used. In the AR-related part of Chapter 3, from Subsection 3.2.2 and on, the hardware remains the same and is pictured in Figure 2.4. The mobile robot employed throughout

Figure 2.4: The left image depicts the Scitos G5 robot, and the right image[8] illustrates a person interacting with a virtual object while wearing the HoloLens 2. Both devices were used throughout this dissertation.

[8]The hummingbird was generated with DALL-E 3.

is the Scitos G5 developed by MetraLabs[9], with the robotic middleware suite Robot Operating System (ROS) [183] as framework to control it. Additionally, the robot is further equipped with a Kinova Jaco2[10] robot arm. The human partner wears a pair of AR glasses, specifically the HoloLens 2. This AR device manufactured by Microsoft is equipped with a built-in eye tracker. The eventual user interface developed for the HoloLens 2, as a crucial component of this work, is implemented within the game development environment Unity[11]. The actual communication between the HoloLens 2 and ROS is handled by the Universal Windows Platform (UWP) version of ROS# [184], which is a collection of open-source software libraries and tools designed to facilitate communication and data transfer between Unity applications and ROS.

2.6.2 Evaluation Fundamentals and Terminology

The proficiency of the teaching pipeline introduced in this dissertation is determined by the extent to which the robot was able to redetect the objects it was taught by the human. For better understanding of this pipeline and its evaluation, an overview of the common terminology frequently appearing in the subsequent chapters is provided here.

Among these terms is Intersection over Union (IoU). The IoU, also known as the Jaccard index, is a measure for the similarity of two sets. It is defined by the intersection of the two sets divided by their union. In the research field of computer vision, the IoU is widely applied to compare and quantify the similarity of two bounding boxes in order to assess the accuracy of object detection algorithms and models. It can assume values ranging from 0 to 1, where 0 signifies that the two bounding boxes do not overlap, and 1 indicates that they are identical in terms of position and size. Thus, a high IoU score suggests a better alignment or larger overlap. Commonly, values of 0.5 and above are considered acceptable, albeit slightly generous, an IoU of 0.9 is rather strict, and 0.7 a reasonable compromise in between.

Further relevant terms revolve around the MS COCO metrics [185], which are common evaluation metrics in object detection. The central focus lies on the average precision and the average recall across a variety of IoU thresholds. In this context, the precision is defined as the ratio of correct predicted bounding boxes to the total number of predictions, while recall represents the fraction of correct predictions among the relevant bounding boxes. Meaning, precision and recall specify how many of the obtained predictions are correct and how many of the relevant items were detected, respectively. An object detection model can then be evaluated for different thresholds of the model's

[9] https://www.metralabs.com/mobiler-roboter-scitos-g5 (accessed: June 16, 2023)
[10] https://www.kinovarobotics.com/product/gen2-robots (accessed: July 10, 2023)
[11] https://unity.com (accessed: June 16, 2023)

confidence scores, resulting in pairs of precision and recall values. Based on these pairs, a precision-recall curve can be constructed. The average precision then results from the area under this curve. In practice, however, MS COCO determines the average precision by averaging 101 interpolated precision values at equidistant recall values using a step size of 0.01 between 0 and 1, denoted by $[0 : 0.01 : 1]$. The abbreviations $AP_{50} = AP^{IoU=0.5}$ and $AP_{75} = AP^{IoU=0.75}$ refer to the average precision values at the IoU thresholds of 0.5 and 0.7, respectively. Furthermore, $AP = AP^{IoU=0.5:0.05:0.95}$ is the average precision averaged across all IoU thresholds in $[0.5 : 0.05 : 0.95]$. Analogously, this notation extends to the average recall. In this case, the maximum recall is ascertained allowing 1, 10, and 100 detections per image, respectively, averaged over IoUs, and is represented by the abbreviations $AR_1 = AR^{max=1}$, $AR_{10} = AR^{max=10}$, and $AR_{100} = AR^{max=100}$. Each of the aforementioned metrics are calculated independently for each individual class. Additionally, the mean across all given classes is denoted by the mean average precision (mAP) and the mean average recall (mAR).

3 Major Contributions

In this chapter, the relevant research contributions towards the objectives and setting introduced in Chapter 2 are further detailed. First, the motivation sets up the respective research question. Then, the methodology shows the processes involved with addressing the research question and subquestions, followed by the work's main contributions and results. An overview of the subsequent papers published at high-impact conferences in the fields of robotics, eye tracking and HRI is listed in Chapter 1.

This research has an impact on multiple fields, which are illustrated in Figure 3.1. These fields interlock and combine to form a system that enables two-way, human-robot interaction, ultimately the human to teach the robot unknown objects in a natural and feasible way.

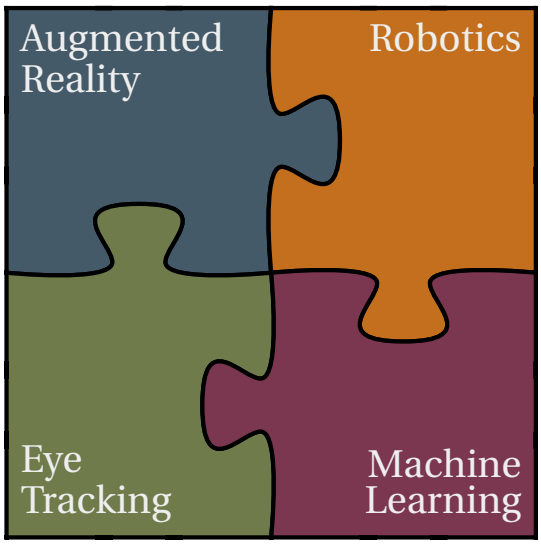


Figure 3.1: The distinct research fields of augmented reality, eye tracking, machine learning, and robotics are intertwined as essential parts of a larger HRI system.

This chapter is divided into two parts. The first part of this chapter, Section 3.1, investigates the general potential of the human gaze in determining ROIs. These defined ROIs based on human gaze offer great potential towards the goal of teaching a robot unknown objects, since the robot needs to identify which object a human is looking at in the first place, that is, what belongs to the object and what does not. In this part, the major focus lies on the domains eye tracking and machine learning. The accompanying publications are included in Chapter A.

The aforementioned findings are applied in the second part of the chapter, Section 3.2. Here, the other two domains from Figure 3.1, Augmented Reality and Robotics, become involved. All parts together finally empower the robot's ability to perceive and learn unknown objects by collaborative interaction with the human. The accompanying publications are included in Chapter B.

3.1 Investigating the Potential of Gaze in Determining Regions of Interest

The first subsection Subsection 3.1.1 examines how gaze can be potentially used for unknown object detection without considering the stimulus (image of the observed scene). The second subsection Subsection 3.1.2 introduces the concept of saliency-aware gaze heatmaps.

3.1.1 Gaze-based Object Detection in the Wild

> **Daniel Weber**, Wolfgang Fuhl, Andreas Zell, and Enkelejda Kasneci. Gaze-based Object Detection in the Wild. In *2022 Sixth IEEE International Conference on Robotic Computing (IRC '22)*.

Motivation

One of the first challenges before a human can teach a robot unknown objects in a HRI setting, as described in Section 2.5, is to let the robot perceive and locate the object the human is looking at. This semantic awareness applies not only to learning, but to generally all desired interactions of a robot with an unknown object of interest. However, detecting an object that is not known is a non-trivial problem. At the same time, the human gaze can assist and provide additional information. The main motivation of this work was to investigate how far the information capacity of the gaze spans, and whether it is possible to detect objects without any context from the scene image. Thereby, object detection was

conceived in the same way as face detection, where the primary task is to detect whether a face is present or not and, if so, to determine its position. Usually, no classification takes place. As this is not possible without a scene image, also in relation to objects, we have limited ourselves as well to the binary detection task and subsequent localization.

Principal Methodology

The foundation for the investigations was a self-reported data set. This data set comprised of multiple participants, who were instructed to move freely inside and outside the venue while wearing a head-mounted eye tracker. The gaze and the Field of View (FOV) were recorded without any specifications of how long the participants should look at the encountered objects. Afterward, the OOIs were labelled.

For the object detection task, the gaze data was divided into ranges of gaze points using temporal windows and then classified whether an object has been observed. If the classification resulted in the detection of an object of interest, a regression of the bounding box parameters followed. As input feature to the ML models, the spatial distribution of the gaze points in the respective time window encoded as 2D and 3D heatmaps was used.

Main Findings

From the methodological point of view, the extension of the 2D heatmap encoding into the three-dimensional space was essential. As a result, in combination with a k-Nearest Neighbor (KNN) approach, a classification accuracy of 92 % was achieved. Overall, multiple different ML models were investigated as a proof of concept. In addition to KNNs, this also included Bagged Trees, Support Vector Machines (SVMs), and Gaussian Process. In a detailed evaluation, the performance of these models were analyzed using different time window sizes and grid sizes for the heatmap features. Apart from the high classification accuracy, the regression of the bounding box parameters yielded an average absolute error for the position of around 6 %. However, the determination of the bounding box dimensions proved to be more difficult than the position. The best average absolute error of the height and width of the bounding boxes ranged between 10 % and 15 %.

Besides the classification accuracy and the average absolute error of the bounding box parameters, the speed and resource consumption of the object detection using the gaze heatmap features were investigated. For speed, the execution time for 1000 different inputs with a batch size of one was measured. Both heatmap features required only a fraction of the time needed by conventional image-based object detectors. In numbers, this translates to 8 to 58 seconds and 64 to 611 seconds for the 2D and 3D heatmap features, respectively. In contrast, only the smallest of the conventional image-based object

detectors was able to stay with 164 seconds below 3 minutes, whereas the majority took significantly more than 10 minutes. For instance, the very popular Faster R-CNN [186] with a ResNet-50-FPN backbone [187] needed the most time with over 8705 seconds for the 1000 different inputs. This made it several orders of magnitude slower than the models that used the proposed heatmap features.

Regarding resource efficiency, the difference was even more pronounced. While the use of the heatmap features only occupied a few hundred kilobytes to a few megabytes of memory for a single input, the smallest comparison model YOLOv5n [188] allocated around 270 MB. The most resources were claimed by Faster R-CNN, which allocated over 1.7 GB of memory. Therefore, the conventional image-based object detectors were also several orders of magnitude more inefficient in terms of memory.

The dataset created in the context of this work was unique at the time of publication, hence it was made publicly available to the research community[12].

Overall, the work has shown that the gaze carries important information that can be useful and harnessed for object detection even if the class of the OOI is unknown. Consequently, combined with a scene image, it can unfold even more impact, which will be elaborated on in the following section.

3.1.2 Exploiting the GBVS for Saliency aware Gaze Heatmaps

> David Geisler, **Daniel Weber**, Nora Castner, and Enkelejda Kasneci. Exploiting the GBVS for Saliency aware Gaze Heatmaps. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '20)*.

Motivation

Although, gaze can assist locating unknown OOIs, the use of pure gaze data in the form of heatmaps – or similar representations – has its limitations. Firstly, humans can sometimes unconsciously fixate on objects, which is why objects do not necessarily have to be observed entirely, as the human already perceives them beforehand. Secondly, the eye tracking signal is always subject to some degree of error, so the resulting estimated gaze is never perfectly accurate. For this reason, this work combined visual information and gaze signal. Algorithms such as Graph-Based Visual Saliency (GBVS) [189] extract visually attentive areas of the stimuli, which are likely to attract the attention of a human observer. The resulting saliency maps indicate regions of particular interest. In combination with the recorded human gaze, deficiencies on both ends can be compensated, leading to better overall estimation of the observed ROIs.

[12] https://cloud.cs.uni-tuebingen.de/index.php/s/QPzJC48xDGsjnZK

Principal Methodology

Fusing gaze and visual information was accomplished by incorporating the gaze into the salience maps determined by the GBVS. The original GBVS algorithm is composed of three main steps. Namely, 1.) the extraction of feature maps, comprising low-level features such as luminance, orientation, or color, based on the input image, 2.) the ensuing computation of activation maps, and finally 3.) their normalization and the aggregation of all the activation maps obtained from all the different feature maps.

The first and last steps have been retained in this work as in the original. In the second step, the idea is to weight a pixel more saliently the more it differs from its surroundings. A graph is constructed that connects each pixel to all other pixels in the feature map generated in the first step. In other words, the fully-connected graph represents the image, specifically the feature map, whereby the nodes can be interpreted as the pixels. The weight assigned to an edge of the graph is the product of the spatial and visual differences of the two pixels connected by the respective edge. Finally, the desired activation map can be treated as a state vector within a Markov chain operating on this graph. More precisely, the activation map is derived from the state of equilibrium (or stationary distribution), with the edge weights defining the Markov transition matrix. The stationary distribution is given as a solution to an eigenvector problem, that is, as the eigenvector of the transition matrix corresponding to the eigenvalue of $\mathbf{1}$ (identity). In practice, this eigenvector is typically determined by iteratively multiplying the Markov transition matrix by a probability vector with an initially uniform distribution. However, rather than choosing the initial activation map to be uniformly distributed, it is initialized by means of the observed gaze points, namely the measured visual activation ascertained from the gaze signal. This tailored initialization allows for the integration of eye tracking data in the saliency calculation process.

Main Findings

Due to the challenging nature of conducting a quantitative evaluation within this particular context, the evaluation primarily relied on experimental demonstration. Three different types of stimulus were analyzed: A painting, a short video, and a text-rich website. The salience maps obtained using the aforementioned method were visualized by overlaying them in the form of a heatmap onto the respective stimulus. In this way, saliency enhanced gaze heatmaps were examined alongside the standard gaze heatmaps and the fixation maps, which visually depict the areas where the observer's eyes were fixated or focused.

In the case of the website stimulus, it was immediately apparent that the approach

described takes direct account of the displayed text as opposed to the pure gaze heatmap. The gaze signal was guided to the individual lines and letters, making the entire heatmap appear sharper and more meaningful. As a result, it was easier to judge whether certain regions attract the intended level of visual attention and to assess the ease of visual accessibility for observers in perceiving the presented information. This finding is particularly relevant in sectors such as web design and advertising.

A more challenging stimulus was the painting, where the foreground contrasted less distinctly from the background. Also in this context, the strength of the proposed method emerged. The saliency-aware gaze heatmap exhibited a better balance between areas that were more difficult to perceive due to their complexity and thus were observed for longer periods of time, and areas that were less complex and hence observed only briefly. In contrast, the conventional gaze heatmap did not consider the region's accessibility to the observer and misleadingly suggested higher levels of interest in certain areas than they truly deserved, simultaneously undermining the prominence of other genuinely relevant regions.

In summary, the findings revealed that the saliency-aware gaze heatmaps effectively guide the eye-tracking signal towards salient regions, producing a more accurate attention pattern.

3.2 Perceiving and Multiperspective Teaching of Unknown Objects

This section elaborates on the intermediate steps that were necessary to achieve the main objective stated in Section 2.5. All the following papers worked towards this goal, and their respective motivations must therefore always be considered in this overall context. The first two subsections, Subsection 3.2.1 and Subsection 3.2.2, focus on the perception and localization of unknown OOIs. The latter two subsections, Subsection 3.2.3 and Subsection 3.2.4, deal with the learning process in which the human teaches the robot unknown objects within a HRI scenario.

3.2.1 Distilling Location Proposals of Unknown Objects through Gaze Information for Human-Robot Interaction

Daniel Weber, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling Location Proposals of Unknown Objects through Gaze Information for Human-Robot Interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '20)*.

Motivation

This work marked the beginning of the HRI teaching pipeline, in which now a real robot was deployed. However, instead of using AR glasses, as in the later course of the research, the gaze was still estimated and recorded with a regular head-mounted eye tracker.

Prior to learning unknown objects later in the project, the robot must first understand which (unknown) object the current OOI is. That means that the localization problem must be solved beforehand. This time, unlike in Section 3.1, not directly on the basis of the gaze data and the corresponding stimulus from the FOV of the human, but rather from the robot's perspective. In order to ensure further processing later on, it was primarily a matter of narrowing down the ROI by means of a bounding box around the OOI. The straightforward utilization of modern neural network-based object detectors was thereby precluded due to their reliance on pre-existing training data, which cannot be assumed in general and excludes the detection of objects not contained therein. Moreover, it would contradict the fact that the OOI are truly unknown.

Principal Methodology

Consequently, the approach presented in this work was algorithm-based. It can be divided into three building blocks.

The first block aims at mapping the human's gaze into the robot's frame of reference. This can be achieved either by directly locating the robot in the camera frame of the human's eye tracker or vice versa, or alternatively by indirect co-location. For the latter, at least four common points must be known in the respective FOVs of the robot and the human. If this is the case, the respective position of the human and the robot can be determined by means of trilateration and accordingly also the mutual position. To ensure the presence of common anchor points, fiducial markers were used, as their detection is robust and efficient. In practice, the transformation from the human reference frame to the one of the robot is achieved by finding a homography that performs a perspective transformation between the image plane of the eye tracker and the image plane of the robot's body camera. The homography can then be used to translate the human gaze points to the corresponding coordinates within the robot's camera image.

The second block requires the robot to predict candidate bounding boxes that are likely to contain an object. This task was facilitated by utilizing location proposal methods, which were applied to the image captured by the robot's scene camera. Such methods are commonly employed to effectively reduce the search space, thereby accelerating the detection process and diminishing the computational costs involved. In the present case, selective search [190] was resorted to due to the fact that it is class-independent. This

property makes it suitable for unknown OOI, just like in our case, where the class is not known in advance. Typically, the output of such location proposal methods consists of thousands of bounding box candidates, which is why the cardinality of the output set was reduced in the third block.

In the third and final building block, the robot's set of proposed candidate locations was distilled using gaze information from the human partner, previously mapped from the human's frame of reference to that of the robot. The intention was to significantly reduce the number of candidates while simultaneously increasing their relevance. As a filtering mechanism, the requirement that the human gaze must fall within the bounding box associated with the respective OOI was leveraged. This ensured that the resulting subset contained only bounding boxes that had an intersection with the object tagged by the gaze point. The distillation was tailored in such a way that the hierarchical order of the proposed bounding boxes was preserved. The order of the proposals hints at the likelihood of them containing an object.

Main Findings

A qualitative analysis showed that eye tracking, marker recognition, and gaze mapping operated in real time. Therefore, the proposed method proved to be suitable for real-time HRI. One prerequisite, however, was that a sufficient number of fiducial markers were visible to reliably estimate the mutual position of human and robot and to map the gaze as accurately as possible. As long as this premise was fulfilled, the human was not constrained in his movements and was able to move freely.

In the scope of a quantitative analysis, the position indices of the bounding boxes were evaluated object-wise before and after distillation. The position index denoted the position of a bounding box in terms of the hierarchical order in which it appeared in the set of location proposals. The smaller the position index of a bounding box pertaining to the respective OOI, the faster it can be found by iteration and the fewer communication with the robot is necessary to select it. However, a low position index is not the sole significant factor. The quality of the bounding box, that is, its accuracy, is equally crucial. Ideally, a bounding box perfectly enclosing the OOI appears in the first position set of proposals. Regarding the accuracy of the boxes, the IoU, also sometimes called the Jaccard index, was assessed as a performance metric. This metric quantified the extent to which the bounding boxes align with the ground truth. The output of the state-of-the-art object detector FCOS [191], which was pre-trained on the MS COCO dataset [185], served as a ground truth. In general, an object is considered as correctly detected if the IoU of the corresponding bounding box exceeds a threshold value of 0.5 [192, 193]. Once the IoU

threshold reaches 0.7, the detection can be deemed reasonably good.

The experiments have revealed that the overwhelming majority of boxes within the full set of location proposals possess an IoU value of less than 0.1, rendering them irrelevant. Instead of having to search through over 2000 proposals, the distillation method significantly reduced this number to an average of about 126. At the same time, the position index of the best existing box was improved from 1315 to 61.5. This trend was also evident with regard to the precision, that is, the fraction of the relevant boxes (boxes with an IoU of at least 0.7) among the boxes distilled by the method. The distillation process increased the precision from less than 2 % to almost 40 %. Among the first three bounding boxes after distillation, at least one box consistently exhibited an IoU value of 0.7 or higher, with an average IoU of over 0.81, which equates an accuracy of almost 90 % compared to the best box in the full set of proposals. Considering the first 15 boxes, the accuracy was even further enhanced to almost 98 %.

Overall, the conducted proof of concept demonstrated functionality and validity in the sense that the distillation process increased the precision by a factor of approximately 21 and was able to locate objects comparably well as the neural network-based object detector FCOS, although pre-training was completely dispensed with. Moreover, the proposed method had the capability to detect objects that FCOS was not specifically trained on and were therefore undetectable within the FCOS framework by nature.

3.2.2 Exploiting Augmented Reality for Extrinsic Robot Calibration and Eye-based Human-Robot Collaboration

Daniel Weber, Enkelejda Kasneci, and Andreas Zell. Exploiting Augmented Reality for Extrinsic Robot Calibration and Eye-based Human-Robot Collaboration. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI '22)*.

Motivation

In order to make the interaction between human and robot as pleasant as possible, it is crucial to establish a communication way between the two parties that is feasible and natural. For this reason, AR in the form of the HoloLens 2, a pair of AR glasses manufactured by Microsoft, was integrated into this work. In the previous work, outlined in Subsection 3.2.1, successful transmission of gaze data from the human to the robot was accomplished. However, it necessitated an overlapping FOV with fiducial markers within it, and even more pertinent, the communication was limited to a unidirectional exchange. Both of these problems ought to be solved in this work by means of AR. The

latter is especially important because another aspect needed to be addressed. Whereas up to this point the localization of objects was restricted to the 2D images of the robot camera, the robot and the human actually operate in three-dimensional space. Therefore, in this work, the 3D positions of the unknown OOIs should be determined. The AR-based two-way communication was then intended to enable the robot to provide the human with feedback regarding the detected OOIs.

Principal Methodology

The HoloLens 2 constitutes the junction between the digital world of the robot and the analog world of the human. All interactions take place wirelessly via an implemented AR interface. The interface offers gesture and speech navigation capabilities, enabling users to issue control commands to the robot or access its camera stream, among other functionalities. Conversely, the robot can superimpose detected objects directly in the human's FOV.

In order for the two interaction participants to be aware of each other's position, a calibration must first be carried out. Simultaneously, this serves to determine the transformation between the robot and its body camera. During this procedure, virtual counterparts of the robot and its camera are positioned according to the physical instances by means of QR codes. The QR codes, however, are only needed during the calibration and do not impose any burden during runtime. After calibration, the HoloLens 2 acts as a bridge between human and robot, ensuring that the robot maintains awareness of the human's position at all times. In this way, the AR interface facilitates real-time provision of human gaze information to the robot, unaffected by the movements of either the human or the robot, and without imposing any restrictions on their FOV.

The localization strategy of the OOIs in three-dimensional space is based on a sequence of well-known computer vision techniques enhanced by gaze data. The starting point forms the point cloud that originates from the robot's body camera. Due to the preceding calibration, the position of the camera and thus the point cloud is known to both the human and the robot. The point cloud first undergoes a pass through filter, followed by a voxel grid filter, aimed at diminishing complexity by reducing the number of points within the cloud. These two processing stages increase the computing time significantly. Due to the extrinsic camera calibration, the orientation of the table on which the objects are placed is known, enabling the identification and extraction of the corresponding plane using random sample consensus (RANSAC) [194]. Finally, by conducting Euclidean clustering on the remaining points of the point cloud and incorporating the

gaze information to identify the points belonging to the OOI, a segmented representation of the OOI is obtained.

Main Findings

The calibration method described above demonstrated high practicality in the conducted trails, especially due to its efficiency and minimal time expenditure. Depending on the experience of the user, a single calibration cycle typically took only between 15 and 40 seconds. Hence, the method highlighted its suitability for spontaneous recalibrations, enabling system modifications, such as adjustments to the camera's tilt, during runtime. Quantitatively, it proved to be reliable and accurate, deviating from the reference method by only a few millimeters on average. However, due to the absence of a real ground truth, it could not be conclusively decided which method was more accurate, as the deviations from each other were within the margin of error. In terms of speed, the presented AR-based calibration was clearly superior.

The permanent bidirectional communication channel introduced by the AR interface enabled continuous real-time segmentation of the respective OOI observed by the human. A 3D bounding box derived from the segmented point cloud can finally be overlaid in the human's FOV to indicate the specific object that the robot believes the human is focusing on. A comprehensive visualization is shown in Figure 3.2.

The quality of the segmented objects was assessed in two different ways: In 2D and in 3D. The neural network-based 3D object detectors VoteNet [195] and Frustum ConvNet [196], which were originally intended as baselines, only achieved a mAP of 27.8 % and less than 1 % respectively for the test objects and were therefore discarded. Note that in 3D, an object is typically already considered to be detected at an IoU threshold of 0.25 [197, 198]. Remarkably, even with a threshold twice as large, the method proposed in this work achieved a flawless recall rate of 100 %. Moreover, the mean 3D IoU of all test objects reached a value of almost 0.7, further emphasizing the quality of the results.

In addition to the evaluation in 3D, a 2D assessment was carried out to mitigate potential susceptibility to bias due to self-labeled 3D ground truth. The 3D bounding boxes were projected onto the 2D image plane of the robot's camera and then compared to the output of FCOS, which served as the ground truth. The achieved mean IoU of 0.81 considerably surpassed the 2D detection threshold of 0.5.

In summary, all test objects were successfully detected, and the system exhibited intuitive access to natural communication and HRI.

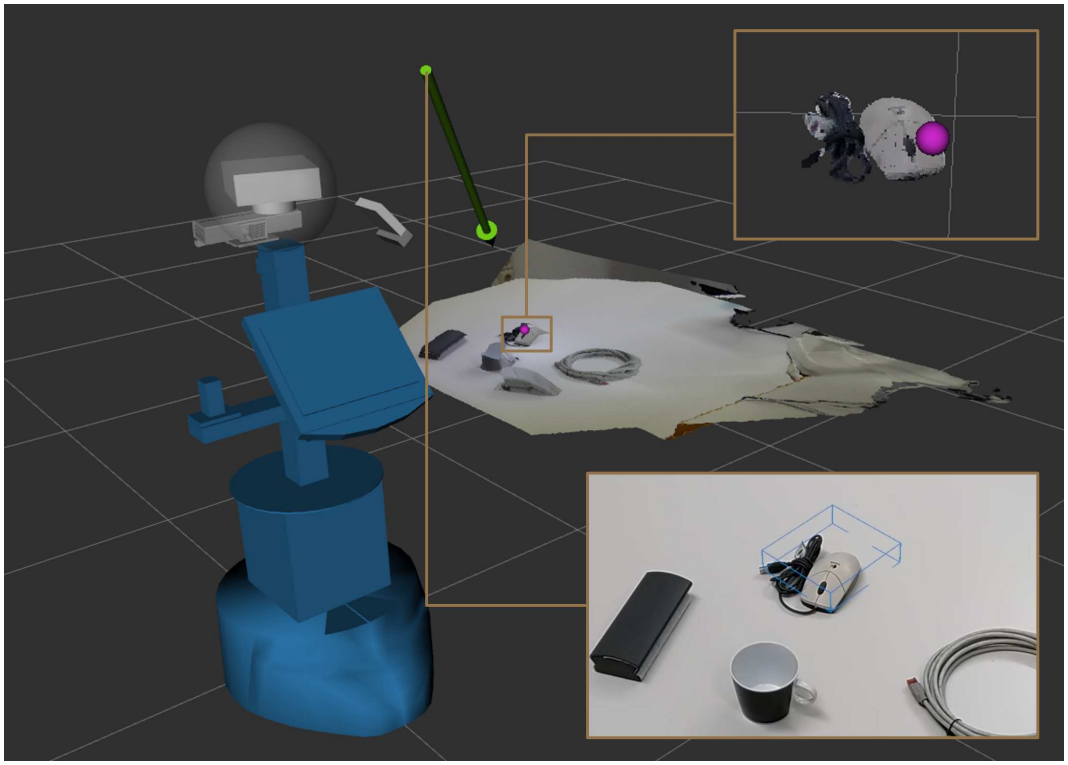


Figure 3.2: Visualization of the robot and the human observing a scene. The human's gaze vector is represented by the green arrow. The intersection of the gaze ray with the environment is depicted as a purple sphere. The top right shows the object segmented by the robot and the bottom right shows the feedback (blue bounding box) provided to the human, displayed in the human's FOV.

3.2.3 Multiperspective Teaching of Unknown Objects via Shared-gaze-based Multimodal Human-Robot Interaction

Daniel Weber, Wolfgang Fuhl, Enkelejda Kasneci, and Andreas Zell. Multiperspective Teaching of Unknown Objects via Shared-gaze-based Multimodal Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23)*.

Motivation

After the previous work in Subsection 3.2.2 had enabled the robot to segment unknown OOI by means of human gaze, this work aimed to further enhance the robot's capabilities

by addressing the key aspect of learning. In order to enable the robot to detect the objects independently and without help, the human should teach the robot within the context of a HRC scenario. The human should act in the pivot role of the teacher and, after visually indicating the object to the robot by gaze pointing and then verbally providing the corresponding class name, impart the knowledge to facilitate the robot's understanding of the OOI. In pursuit of this, the robot was supposed to capture various images of the OOI from multiple different angles. These images were then to be automatically labeled and utilized for training purposes. Eventually, the robot's proficiency in detecting the recently learned objects was tested.

Principal Methodology

The segmentation principles were borrowed from the work presented in the previous Subsection 3.2.2. The AR interface introduced therein was further extended to support multimodal HRI and the transmission of class information. To this end, once the human and robot have directed their attention towards the same object, the bounding box of the OOI is displayed through the HoloLens 2. The human user can then conveniently approve the bounding box using speech commands or gestures. Subsequently, the class name can be specified and submitted to the robot via a virtual keyboard or again via speech recognition.

For robust training purposes, the robot necessitates an abundant amount of data. Consequently, it autonomously generates a comprehensive dataset comprising images recorded from multiple perspectives. For this purpose, the robot utilized a second camera attached to its arm to perform a circular movement around the OOI and examines it from all sides. Thereby, he proceeds according to Algorithm 1. The camera mentioned therein always refers to the one attached to the robot arm and not to the body camera used for segmentation.

The core idea resolves around the autonomous labeling of each image with the ROI as the robot moves the camera around the object. The ROI is determined by transforming the segmented 3D point cloud into the respective current camera frame, which changes dynamically as the robot arm moves. The 3D points are then projected onto the 2D image plane of the camera using the intrinsic camera parameters, and the 2D bounding box is derived from the boundary points. Finally, the robot stores the ROIs, RGB and depth images as well as the intrinsic and extrinsic camera parameters. The extrinsic parameters specify the camera's position in relation to the object. The entire progress of this multi-perspective acquisition of training data is visually reported to the human by means of the AR interface.

Algorithm 1 Multiperspective Recording

Require: Class c, point cloud pc_{3D}, and bounding box box of segmented OOI

```
p ← calcCircularPath(box)
w ← getReachableWaypoints(p)
t ← calcTrajectory(w)
while moving along trajectory t do
    pc'_3D ← transform(pc_3D)          ▷ Transform 3D point cloud to camera frame
    pc_2D ← project3DToPixel(pc'_3D)   ▷ Project onto image plane of camera
    roi ← min/max(pc_2D)               ▷ Calculate 2D bounding box
    Store in folder c: RGB image, depth image, camera parameters, roi
end while
```

In terms of the object detection architecture, the robot was equipped with state-of-the-art models such as Faster R-CNN [186], which were fed with the obtained RGB images. Although, in principle, numerous datasets are available in the field of computer vision, it cannot be presupposed that they contain a particular OOI. Nevertheless, this offers an opportunity to build on. Therefore, the training of the robot's object detector was rooted in the extension of existing knowledge ad hoc in the situation through teaching assuming a general awareness of objectness in the form of pretraining on irrelevant objects. By applying a transfer learning approach, only the classification and regression heads were reinitialized and trained, while the feature layers remained frozen. This training strategy allowed successful training even on non-high-end hardware, such as the robot, while mitigating the risk of overfitting.

Main Findings

As part of the evaluation, the robot was taught ten different classes, each with two objects, using the teaching pipeline described above. Per teaching run, that is, per object, the robot acquired a large amount of labelled multiperspective training data in a short period of time (about 1 minute). After training, the robot's gained knowledge was examined on a separate test set, which comprised two other objects from each of the ten classes, distinct from the objects used during training.

In total, several different object detectors were tested, including Faster R-CNN and FCOS [191], among others. It showed that the robot was able to generalize from the training objects to unseen objects of the same respective classes. A thorough evaluation using Faster R-CNN revealed that the robot successfully detected the majority of objects, achieving an impressive mAP_{50} of almost 70 %. In contrast, the same object detector, when trained on the entire MS COCO [185] dataset, achieved an mAP_{50} of over 80 % on

the six classes that were also part of MS COCO. However, its performance on all ten test classes was only slightly around 50 %, which is considerably less compared to the results obtained through the multiperspective teaching pipeline.

The complete training data generated by the robot consists of more than 3100 viewpoints and has a considerable density of information as it encompasses various essential components mentioned above. Therefore, especially due to the inclusion of the camera poses, it becomes particularly appealing for other prominent research areas, such as Neural Radiance Fields [199, 200, 201, 202]. For this reason and in order to ensure reproducibility, this data, along with the validation and test set, were collected into the Objects in Multiperspective Detail (OMD) dataset, which was made publicly available to the research community[13]. In addition, the complete code base of the HRI system, featuring the AR interface, ROS nodes, and learning policy, was made publicly available as well[14].

All in all, the introduced novel teaching pipeline employing multimodal HRI demonstrated its practical efficacy as an intuitive and natural method for teaching the robot new, yet unknown objects using few instances. Furthermore, it enabled the robot to detect classes that lack a dedicated training dataset.

3.2.4 Leveraging Saliency-Aware Gaze Heatmaps for Multiperspective Teaching of Unknown Objects

> **Daniel Weber**, Valentin Bolz, Andreas Zell, and Enkelejda Kasneci. Leveraging Saliency-Aware Gaze Heatmaps for Multiperspective Teaching of Unknown Objects. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '23)*. (Accepted for publication).

Motivation

Although the robot was successfully taught through HRI in the previous work from Subsection 3.2.3, this approach exhibited a drawback. The outcome heavily relied on the quality of the segmentation discussed in Subsection 3.2.2, as it was instrumental in determining the ROIs. Furthermore, even minor inaccuracies in the transformation calibration between the robot, arm, and wrist camera can propagate throughout the system and accumulate to discrepancies in the result. A more direct and robust approach with fewer system components was therefore intended to reduce the complexity and thus the susceptibility to errors of the entire system. Rather than relying solely on the segmentation by

[13] https://cloud.cs.uni-tuebingen.de/index.php/s/2oRPs2o3FZkdBHW
[14] https://github.com/dnlwbr/Multiperspective-Teaching

HRC, a more straightforward approach involved the human observer devoting a slightly longer duration of time focusing the OOI and then to consider a series of the resulting gaze points.

Principal Methodology

Albeit, building upon the preceding achievements, the underlying paradigm that the OOIs need to be fully identified prior to the robot's data acquisition was abandoned.

The human looks at the object and initiates the teaching procedure by means of speech recognition. For a duration of 10 seconds, the gaze data is logged while the AR interface provides an audible countdown. Based on the gathered gaze points, the robot approximates the position and size of the OOI and records it from different angles as before. Instead of the segmented point cloud being transformed into the camera frame and then mapped onto the 2D image plane, it is the gaze points that undergo this procedure. Theoretically, the bounding box labels could be determined directly from the edges of the transformed and projected gaze points. However, as mentioned above, even slight imprecision in the hardware calibration or in the tracking of the human's position or gaze can cause an offset in the ensuing ROI. To compensate for these inaccuracies, the gaze points were refined and guided towards the OOI using saliency. This step reflected the findings of Subsection 3.1.2 in the form of the Gaze-Assisted GBVS (GA-GBVS), which is additionally extended to the Dual Gaze-Assisted GBVS (DGA-GBVS). For each perspective captured by the robot, feature maps are extracted from the corresponding RGB image. These feature maps are then used along with the gaze data to generate activation maps, which are eventually normalized and merged into a saliency-aware gaze heatmap.

By means of Otsu's method[203], a threshold value is set to delimit relevant points and to binarize the heatmap. This also sharpens the edge of the heatmap points, which then define the ROI. Note that relying solely on saliency maps, without considering gaze points, is insufficient for reliably determining the ROI. This is because there may be other salient areas within the stimulus that are not part of the intended OOI.

Finally, the robot can proceed analogously to Subsection 3.2.3, leveraging the acquired and labeled perspectives to learn the taught objects.

Main Findings

In order to compare the approach with the results published in [5] and discussed in Subsection 3.2.3, the evaluation was carried out accordingly. Exactly the same ten objects were taught to the robot via the new HRI pipeline, whose detection capability was then evaluated on the OMD dataset. Faster R-CNN served as the backbone.

In general, the described method consistently outperformed the previous approach in almost all classes. This basically applied to all common object detector metrics. Regarding the mean average precision, the previous values for mAP_{50}, mAP_{75}, and mAP were improved from 66.9 %, 31.4 %, and 33.6 % to 73.6 %, 38.1 %, and 39.5 %, respectively. A model trained on the full MS COCO dataset, as in Subsection 3.2.3, was again able to perform quite well on the known classes, but overall, including the unknown classes, only achieved a mAP_{50}, mAP_{75}, and mAP of 50.7 %, 33.3 %, and 30.5 %, respectively. Furthermore, this model was outperformed, even for some of the known classes, by the method presented in this work.

A similar pattern emerged with regard to the mean average recall. The mAR_1, mAR_{10}, and mAR_{100} were increased from 43.7 %, 50.1 %, and 50.4 % to 50.1 %, 54.4 %, and 54.4 %, respectively, in comparison to the scheme presented in Subsection 3.2.3 and thereby also surpassed the model trained on the entire MS COCO set. The latter only achieved a recall of 35.2 % in each of the three metrics. Likewise, this model was exceeded even for some of the known classes by the new proposed teaching pipeline.

Additionally, the curves of precision and recall as functions of IoU, generally revealed superior values compared to the previous approach from Subsection 3.2.3, especially at high IoU thresholds. This improvement indicated more accurate bounding boxes, enhancing the learning and detection capabilities of the entire HRI system.

4 Discussion & Outlook

This chapter discusses the publications presented in the previous Chapter 3 and aligns them with the objectives of the thesis outlined in Section 2.5. The central aspect is the development of a HRI teaching pipeline, as elaborated therein. The findings regarding the potential of gaze in determining ROIs, based on the publications [2] and [4] from the list in Chapter 1, are discussed in Section 4.1. Section 4.2 delves into the multiperspective teaching approach for unknown objects by means of HRI, utilizing AR and eye-tracking techniques. Here the insights of [1], [3], [5], and [6] are examined. Finally, a conclusion is drawn, providing a concise summary of the findings, which is followed by an outlook of possible future steps and directions to consider.

4.1 The Potential of Gaze in Determining Regions of Interest

In the general setting, without the robot and AR, the focus lies not on the teaching per se, but on the preceding identification of the OOIs. The results from publication [4] have demonstrated that the gaze by itself can already yield information about the position of objects and their approximate size up to a certain degree. The image of the scene does not necessarily have to be included, but due to the omission of image information, time windows with a fixed size were used. Especially the classification of the time windows, that is, whether an object was within focus or not, was convincing with 92 % for the best combination of parameters. Conversely, it was also found that without using the scenery, the estimation of the size of the bounding box belonging to the OOIs was quite challenging and produced mixed results. However, this is not particularly surprising, as humans do not observe objects completely in every detail in order to perceive them appropriately. Furthermore, since the subjects were allowed to move freely, different gaze points from subsequent timestamps could belong to the same point in the environment and vice versa. This further complicated the bounding box regression.

A clearer picture emerges with regard to the estimation time and the required re-

sources. The introduced heatmap feature, resulted in a tremendous speed boost in comparison to standard object detectors such as FCOS [191] or RetinaNet [204], while requiring only a fraction of the resources. This discrepancy can be attributed to the notable difference in size between heatmaps and images when employed as input features for machine learning models. This disparity can be attributed to the significant difference in sizes between heatmaps and images when used as input features for the ML models. Depending on the number of grid cells into which the FOV was divided to build the heatmap input features, even the three-dimensional variant had a relatively small size, containing only $50^3 = 125\,000$ values. They are, therefore, easier to process compared to images, which, at the resolution of $1088 \times 1080 \times 3$, had to handle a much larger number of 3 525 120 values. The low hardware requirements are particularly attractive for the operation on mobile robots, as their hardware is limited and not arbitrarily expandable. Even though this is a really powerful advantage, in practice one would rather seek to make use of any information available and look for a balanced combination to also enhance the regression of the bounding box parameters. To exploit the full potential of the robot, visual information or even depth data could therefore be incorporated in addition to the heatmap features. Such a combination remains the subject of future research.

The experiments in [2] revealed that gaze signal can be effectively combined with saliency maps. These maps identify and highlight salient areas and thus process visual information. As described in Subsection 3.1.2, the joint fixation-saliency maps were superior to the standard fixation heatmaps. The eye-tracking signals could be combined with the GBVS algorithm and inaccuracies in the gaze data could thus be corrected. This observation was true both for simple stimuli like a website, where the text prominently contrasts with the background, and for more intricate stimuli such as paintings. The downside of this approach, which is also its main limitation, is the runtime and memory consumption. As the size of the input image increases, the size of the Markov transition matrix grows quadratically. The initialization of this matrix has a complexity of $\mathcal{O}\left(n^2\right)$. The combined effect of these two factors restricts the GBVS to extremely low input resolutions, resulting in a reduced level of acuity. Consequently, the original implementation of the GBVS algorithm sets the internal resolution to a maximum edge length of 32 pixels [205]. Nevertheless, by sparsifying the matrix through the omission of very small values, the complexity can be reduced to $\mathcal{O}(n)$. Moreover, in the eventual application within the HRI teaching scenario, the blurring effect is less relevant compared to the original work. This is because the goal is not to generate a comprehensive saliency map, but rather to refine the gaze map in light of the stimulus.

With the outcome of these general experiments, a step was made towards the objectives outlined in Section 2.5. The observations, which highlight the benefit of gaze data

in determining ROIs and that saliency can mitigate inaccuracies in gaze signals, hold particular relevance to the challenges C2 and C6.

4.2 Multiperspective Teaching of Unknown Objects via Human-Robot Interaction

While the previous studies were detached from AR and HRI, in [1] the robot came into play. The main takeaway was that the presented approach successfully enabled the robot to detect unknown objects, that the human was looking at, within the 2D image of its body camera. Remarkably, this method performed comparably well to the state-of-the-art object detector FCOS [191] without requiring any pretraining, meaning that the deployment was not tied to predefined objects. By leveraging gaze information, the precision of the selective search algorithm witnessed a substantial increase by a factor of over 20. It should be noted, that although the devised method yielded a distilled and this more relevant output compared to the original, it was not univocal, as it still consisted of a set of multiple proposals. However, this characteristic is not actually a disadvantage, but rather presents opportunities as it allows the human to select the best and most appropriate bounding box in cases where the first one was not suitable. This is especially helpful when the OOI is difficult to distinguish from the background or when it comprises varying colors that do not clearly suggest to the robot whether it represents a single object or multiple distinct objects. Furthermore, the possibility of making decisions through HRI closely resembles the interaction and learning process between humans, rendering it a natural approach. Eventually, the challenges C1 (gaze mapping) and C2 (unknown object localization) could be solved within the realm of an ordinary tracker without any AR functionality. Nevertheless, the former relied on the detection of fiducial markers in order to map the gaze from the human's frame of reference to that of the robot. Due to the fact that the mapping process required a sufficient number of adequately sized markers to be present in both the robot's and the human's field of vision, the human's mobility was restricted, and the system became more cumbersome and error-prone. The latter was in turn not suitable for unknown object localization in three-dimensional space.

Both mentioned problems were further improved and fully resolved by the publication [3] from the list in Chapter 1. By using the HoloLens 2, the fiducial markers became obsolete. The position and orientation of the device and thus of the human could be tracked in real time without restricting the movements of the human or the robot in any way. Since the robot and the human were thus constantly aware of each other's position, the gaze vector and gaze point could be directly transformed from one coordinate system

to the other. This solved C1 entirely. The necessary information was exchanged through the dedicated communication channel provided by the specifically implemented AR interface. Furthermore, this interface was responsible for the registration of the human gaze data. Consequently, C3 could be marked as completed. Additionally, the involvement of speech and gesture control, which was absent in the preceding work [1], allowed for a more natural and intuitive interaction. This enhancement significantly improved the user-friendliness.

Another milestone was that the unknown object localization was lifted from 2D to 3D. This was indispensable for the later teaching process. The possibility of the human using gaze to guide the robot's attention towards the OOI, which the robot then segments in real time, also laid the foundation for the ensuing research. Gaze pointing offers an intuitive and less ambiguous alternative to pointing with a finger and, unlike speech, it can be utilized prior to the robot knowing the object. With this, C2 could ultimately be deemed resolved. Moreover, due to the synergy of the AR and segmentation components, a virtual three-dimensional bounding box around the object segmented by the robot can be visualized within the field of view of the HoloLens 2. Through this feedback mechanism, the robot can provide direct indications to the human regarding its estimation of the human's attentional focus. Consequently, the person can immediately identify whether the robot's assessment aligns with his or her own perception. Hence, C4 was also solved within the scope of publication [3].

Even though the presented algorithm-based approach clearly outperformed the state-of-the-art neural network-based 3D object detectors, it has a limitation to be considered. The segmentation is not able to distinguish objects that are close to each other. As the semantic understanding of objects for the localization task is not yet evolved, situations may arise where objects positioned in close proximity to each other are interpreted as a single large object. However, this could be addressed by means of additional HRI in which the human advises the robot of the approximate size of the OOI.

For the follow-up work, C5 and C6 remained to be eliminated. These were addressed in [5] from the list in Chapter 1, focusing on the teaching aspect of the pipeline. With regard to C5, the previously implemented AR interface was extended to enable the human to convey the class information to the robot, thereby enhancing the collaboration between the human and the robot. The virtual bounding box derived from the robot's segmentation feedback serves two purposes. Firstly, it verifies the robot's attention, and secondly, once the shared attention of the human and the robot is ensured, the human can select the OOI encompassed by the virtual bounding box. This interactive selection mechanism appears intuitive and natural, as it can be conducted either by gestures or by speech. The same modalities can then be used to determine the class of the OOI. For this purpose, the

AR interface implements a virtual keyboard, solving C5.

Last, C6 is addressed by means of the robotic arm technique and the transfer learning approach described in Subsection 3.2.3. The autonomous and efficient way for the robot to examine the OOI in more detail and acquire labelled training data completed the teaching pipeline. As a result, the robot was able to successfully redetect the objects it had been taught through HRI, fulfilling the overarching objective outlined in 2.5. In particular, it must be emphasized that the objects used in the test phase were distinct from the objects, that the robot was taught with. The robot was thus able to generalize to unseen object entities of the previously learned classes. This reveals the capabilities of the robot's object detection system, representing a big step towards the deployment of robots in unfamiliar and non-predefined environments.

Naturally, the results achieved through HRI teaching could hardly match those attained by training on extensive datasets containing thousands or even millions of images. At least in the assessment of the individual classes. Assuming the initial issue of not having adequate datasets for all classes, the strengths of teaching through HRI became apparent. Considering all classes, rather than only those contained within the corresponding datasets, the evaluations revealed that the teaching approach outperformed the baseline with an mAP of 33.6 % compared to 30.5 %. This highlights the teaching approach's enhanced flexibility and adaptability. Consequently, in practice, one would neither want to forego prior knowledge, if it is available, nor the flexibility of HRI teaching. Hence, a combination of pretraining and HRI teaching emerges as the best strategy as prior knowledge can be used, but still be expanded when unknown objects occur.

Either way, the performance of supervised machine learning methods relies on the quality of the input data during training. In the current form of the HRI teaching pipeline, the output of the object segmentation from publication [3] described in Subsection 3.2.2 is thus a carrying factor. When exposed to challenging sensor data, such as with extremely dark or highly reflective object surfaces, the depth determination might be inaccurate, leading to insufficient segmentation of the OOI. This in turn can affect the labels of the training data gathered by the robot and hinder the robot's learning progress. While this issue does not fully impede the robot's ability to learn unknown objects, it can be alleviated by increasing the number of objects from the respective class during the training phase.

Alternatively, in [6], an alternative approach was attempted to determine the labels on the image data collected by the robot. Hence, challenge C2 was solved in a different way, not with upstream segmentation, but based on a series of gaze points collected over a period of time. In combination with the saliency of the individual training images, the ROIs could be calculated and used the corresponding labels. By encoding the three-dimensional gaze points as saliency-aware 2D gaze heatmap, the human gaze could be

aligned with the OOI and imprecision in the signal could be rectified. Compared to the segmentation in publication [3], an even greater amount of gaze information from the human has been incorporated. As a result, the estimated ROIs are more accurate and the labels of higher quality. This is also reflected in the robot's detection performance. In fact, the results have improved compared to the preceding HRI variant in [5], for example, the mAP has further increased from 33.6 % to 39.5 %. Even in terms of the class-specific analysis, it was possible to keep up with the model that was trained on the entire MS COCO dataset. Furthermore, one of the limitations of the segmentation in [3], wherein objects in close proximity were erroneously interpreted as a single object, is remedied. Although the advantages outweigh the disadvantages, there is also a small drawback. Specifically, after the robot has examined the OOI, an intermediate processing step is necessary, because due to the runtime of the GA-GBVS and DGA-GBVS techniques already mentioned earlier, the labels cannot be produced during the recording phase itself. In practice, however, this is to a certain extent negligible, as the calculations can take place in parallel with the HRI teaching of other new objects.

4.3 Conclusion and Outlook

In conclusion, the presented interdisciplinary research broke new ground by fusing fields that were previously running predominantly in parallel. More specifically, this dissertation marks a significant step towards teaching robots their unknown environment, particularly when the required training data is limited or unavailable. Instead of relying entirely on data based pre-training, HRI was consulted to foster the robot's object detection abilities. By engaging in such collaborative settings, the robot could successfully be taught unknown objects within its environment by its human peer. The robot became capable of independently detecting these objects without further external assistance, enhancing its adaptability to non-predefined scenarios. Along this line, a variety of innovative steps were solved at the intersection of robotics, human eye tracking, AR and ML. This includes, in addition to successfully teaching the robot, natural and human-like interaction by means of eyes and speech. The novel AR-based extrinsic calibration required for this purpose is characterized by its speed, ease of use, as well as competitive accuracy in comparison to classical approaches. To summarize, it can be concluded that all challenges, initially outlined at the beginning of this dissertation, were successfully addressed, resulting in the attainment of the overarching goal of teaching unknown objects through HRI. This accomplishment not only validated the intended outcome, but also aligned with the broader desire of closing the yawning gap of data dependency in robot learning.

Despite this, there are still several opportunities for potential enhancements and

future research. For instance, as of now, no form of optimization has taken place with regard to the object detection backbone. Naturally, the hyperparameters of the neural network models could be fine-tuned. However, this is also accompanied by an increased risk of overfitting to a specific setting or environment. This, in turn, would negate the advantage of modularity inherent in the presented system, which in principle works with any object detector, and would limit the flexibility of the entire system. It would therefore be more beneficial to develop a model that takes all robot sensors into account, rather than solely relying on the RGB images. Especially, the depth data could contain valuable additional information that can be leveraged. By incorporating the full range of sensor data, such a model would unlock enhanced capabilities and maximize the system's potential.

Another direction to consider is the further intensification of the interaction between the human and the robot. The potential that HRI offers has not yet been fully exhausted, and humans continue to hold significant value in providing further support. Especially within situations where an object class has been detected incorrectly, the human can play an even more pivotal role in instructing and correcting the robot. In this regard, new opportunities also arise due to the ascent of increasingly powerful Large Language Models (LLMs). Prominent exemplars such as GPT-4 [206] and PaLM 2 [207], used in ChatGPT and Google Bard, respectively, along with Meta's LLaMA [208] hold the potential to advance communication, streamlining the conveyance of intricate scenarios and subject matters to the robot. Large visual models or large multimodal models, like SAM [209], may also, in the future, facilitate tasks such as identifying the ROI of the unknown OOIs or even serve as a backbone for the teaching process itself. For the latter, it is important to note that the comprehension of objects in these models is rooted in extensive amounts of textual and image data. As a result, training these large-scale models, comprising hundreds of billions of parameters, necessitates supercomputers equipped with exceptionally high-performance hardware. Therefore, training these models directly on the limited hardware of a robot is currently entirely precluded. For the former, even with a robust generalization due to the substantial volume of training data, it must be ensured that totally unfamiliar objects, not previously included in the training data, are detectable.

Furthermore, the presented system still has to prove its operational capability in alternative contexts outside the office environment. Thus far, the used objects have exhibited a considerable degree of diversity and the tests encompassed multifaceted conditions, while the HRI teaching process has been conducted within the confines of controlled office scenarios, rather than real-world environments where the robot would be confronted with a plethora of disturbing and irrelevant objects within its FOV.

Even though there is still further research required to refine the system's general

applicability, it has already indicated some promising aptitude for certain applications. For instance, it was possible to design a prototypical extension that enabled the robot to acoustically name the classes of objects that the human was looking at. This might already be advantageous in domestic settings for handicapped or elderly people with speech deficits or limited mobility, in order to point out an object of their desire to other people in the vicinity.

In all, the emergence of the presented research findings in robotics, HRI, AR, ML, and eye tracking will prospectively fuel the steady expansion of potential applications and highlight the growing demand for HRI based teaching.

A Investigating the Potential of Gaze in Determining Regions of Interest

The following publications are enclosed in this chapter:

[4] **Daniel Weber***, Wolfgang Fuhl*, Andreas Zell, and Enkelejda Kasneci. Gaze-based Object Detection in the Wild. In *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*, pages 62–66. IEEE, 2022. doi:10.1109/IRC55401.2022.00017.

[2] David Geisler, **Daniel Weber**, Nora Castner, and Enkelejda Kasneci. Exploiting the GBVS for Saliency aware Gaze Heatmaps. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5, 2020. doi:10.1145/3379156.3391367.

* equal contribution

A.1 Gaze-based Object Detection in the Wild

A.1.1 Abstract

In human-robot collaboration, one challenging task is to teach a robot new yet unknown objects enabling it to interact with them. Thereby, gaze can contain valuable information. We investigate if it is possible to detect objects (object or no object) merely from gaze data and determine their bounding box parameters. For this purpose, we explore different sizes of temporal windows, which serve as a basis for the computation of heatmaps, i.e., the spatial distribution of the gaze data. Additionally, we analyze different grid sizes of these heatmaps, and demonstrate the functionality in a proof of concept using different machine learning techniques. Our method is characterized by its speed and resource efficiency compared to conventional object detectors. In order to generate the required data, we conducted a study with five subjects who could move freely and thus, turn towards arbitrary objects. This way, we chose a scenario for our data collection that is as realistic as possible. Since the subjects move while facing objects, the heatmaps also contain gaze data trajectories, complicating the detection and parameter regression. We make our data set publicly available to the research community for download.

A.1.2 Introduction

Recent research has shown that eye tracking has becoming increasingly relevant for a variety of applications. These include even dynamic real-world scenarios, such as driving [210], medicine [211], and sports [212]. Especially the combination with computer vision problems [213], has in turn great potential for the employment of eye tracking in other fields, such as robotics [1]. In the field of robotics, the focus is often on the interaction with the environment, for example, detecting and grasping objects [214]. In such settings, however, the interaction entities are often unknown due to the enormous amount of potentially existing objects. For this purpose, a semantic understanding of scenes must be present. In conveying this understanding, humans can play an important role and provide assistance to the robot. One modality that has proven to be particularly suitable and helpful for such human-robot collaboration (HRC) settings is the human gaze [3]. Gaze allows objects to be intuitively selected by the human and communicated (e.g., gaze pointing) to the interaction partner (e.g., robot). An additional advantage of the gaze

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

modality is that it is far more unambiguous than gestures and, unlike speech, can also be used effortlessly in the case of unknown objects whose class name may not be known at all.

In this work, we address the problem of unknown object detection in real-world scenarios based on gaze. This is an essential challenge for HRC, as an example. After all, if the robot could detect an unknown object by the fact that the human is looking at it, this paves the way for further interaction possibilities. We refer to object detection in a similar manner to face detection. In face detection, the task is to estimate whether there is a face or not. In our task, the challenge is to find out whether the current gaze pattern belongs to a perceived object or not. While there is work investigating unknown object detection on static imagery, there is little research addressing unknown object detection on videos and settings in the wild. Along this line, [215] used fixations to infer the saliency of objects. A gaze map was used by [216], who combined it with candidate regions to segment objects. In the work by [217], gaze points were grouped into clusters to determine whether a cluster belonged to an object of interest and whether it was looked at intentionally or unintentionally. However, all these related works used multiple gaze points on one image, which is only possible if the stimulus (image of the observed scene) is static or if, for instance, eye tracking data from multiple people is used, as in [213]. Contrary to all aforementioned related works, we present a method capable of using gaze data from a single person in dynamic scenes, i.e., with non-static stimuli, to detect unknown objects.

Our way to meet this challenge is by considering and analyzing gaze data across multiple frames and constructing a heatmap from it. In contrast, [1] significantly reduced the amount of candidate bounding boxes of unknown objects on a static image using only one gaze point. In another recent work in a HRC scenario, [3] achieved segmentation of unknown objects and calculated corresponding bounding boxes in 3D space in real time. Although only one gaze point was required here, the scene image including depth information was needed. Some other approaches dispense with the gaze altogether, but focus rather on single-class images [218], or use additional information, e.g., from a depth sensor [219]. While robots typically have many sensors, they often have limited computing power. Additionally, there is often only one object of interest at a time, obviating the need to detect all objects at once. By completely omitting image data and employing gaze data instead, we can accomplish the task of detecting unknown objects of interest and still saving large amounts of required computer resources.

In this work, we build on existing work and pave the way for successful human-robot interaction through the following main contributions:

- We present a method for detecting unknown objects in a scene without stimulus,

based solely on gaze information.

- We only use heatmaps instead of scene images, enabling thus for a significantly faster approach than image-based object detection, while at the same time requiring considerably less computational resources.

- We make our unique data set, which contains both gaze data and bounding boxes of the observed objects, publicly available to the research community for download at https://cloud.cs.uni-tuebingen.de/index.php/s/QPzJC48xDGsjnZK.

A.1.3 Method

In this work, we follow two goals. First, we classify which gaze points or ranges of gaze points belong to an object, and we assign temporal windows to the gaze points, which belong to an annotated bounding box. This creates a classification problem in which the gaze points windows with an associated bounding box are assigned to class one and gaze points windows without a bounding box are assigned to class zero.

The second goal is to regress the bounding box parameters on the gaze points. These parameters are the width and height, as well as the x and y position. For this task, we also assigned the gaze points to temporal windows. For the regression, we used only temporal windows with associated bounding box, since all others have no parameters for the regression.

We decided to use a spatial distribution as a feature since this worked best in our initial evaluations. This spatial distribution is a heatmap as previously proposed by [220] to classify gaze position data. To create such a heatmap, the gaze position data of a temporal window are used, and the individual gaze positions are assigned to cells in the heatmap (grid). Each time window results in one heatmap. After the assignment, the heatmap is divided by the sum over all values to obtain a distribution. As an extension to the approach in [220], we extended the 2D heatmap to 3D. This was possible because the software used for gaze determination generates 3D gaze points [221] based on a k-nearest neighbor regression. In the case of the 3D heatmap, a cell is assigned to each gaze point based on its spatial position with the difference to the 2D heatmap that the depth or distance of the gaze points is additionally considered along the z-axis. The assignment procedure is illustrated in Figure A.1.1.

A formal description of the generation of the heatmap in 3D is given in Equation A.1.1.

$$\text{heat}\left(\left\lfloor \frac{p_x}{R_x} \cdot G_x \right\rceil, \left\lfloor \frac{p_y}{R_y} \cdot G_y \right\rceil, \left\lfloor \frac{p_z}{R_z} \cdot G_z \right\rceil\right) \mathrel{+}= 1. \tag{A.1.1}$$

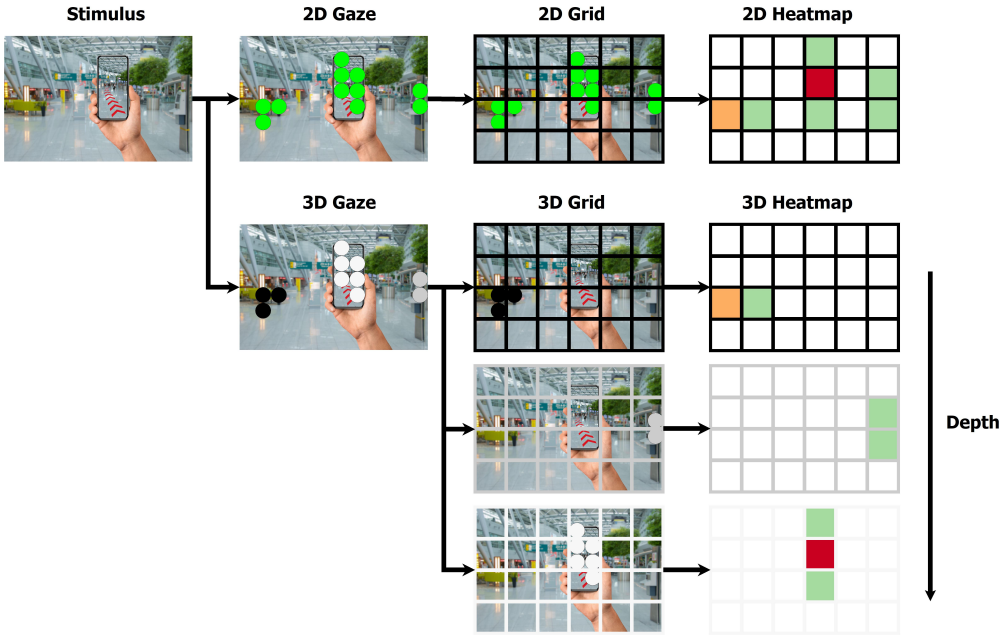


Figure A.1.1: Creation of a 2D or 3D heatmap based on the gaze information and the stimulus resolution.

The gaze positions in x, y, and z coordinates in an Euclidean coordinate system are denoted by p_x, p_y, and p_z, respectively. The constants R_x, R_y, and R_z represent the maximum resolution of the stimulus in x and y direction and the maximum depth supported by the software Pistol [221]. By dividing the gaze points by the maximum resolution, these ranges are normalized between 0 and 1. Subsequently, these values are multiplied by the number of grid cells (G_x, G_y, and G_z) and rounded to the nearest integers, denoted by "$\lfloor\cdot\rceil$". These new values correspond to the index in the heatmap and the selected cell is incremented by one, denoted by "+=". In the case of a 2D heatmap, the cell for depth (z coordinate) is fixed at one.

Equation A.1.2 describes the normalization of the heatmap in 3D and 2D since for the 2D case there would be only one depth.

$$\text{heat}(x,y,z) = \frac{\text{heat}(x,y,z)}{\sum_{i=1}^{G_x}\sum_{j=1}^{G_y}\sum_{k=1}^{G_z}\text{heat}(i,j,k)}. \tag{A.1.2}$$

The variables x, y, and z are the indexes to the heatmap corresponding to the x-axis, y-axis, and z-axis. Finally, the one-dimensional vector resulting from the flattening of the heat map can be used as an input feature for various machine learning techniques.

A.1.4 Study Design & Data Acquisition

In this section, we describe the dataset we used. In order to evaluate our approach, a dataset was required which contains not only eye tracking information but also, in addition to the gaze points, the bounding boxes of the objects that the participants were looking at. Since, to the best of our knowledge, no such dataset exists or is publicly available, we collected a novel data set. At the beginning, a calibration was performed with each participant, following the procedure described in [221]. Subsequently, the subjects were allowed to move freely around the site. In this course, they should look at arbitrary objects they encountered. There was no specification as to how long they were supposed to look at the objects. To evaluate gaze accuracy, the participants were asked to look at the calibration marker again at the end of each recording. All recordings were conducted with the Pupil Invisible eye tracker, a head-mounted eye tracker developed by Pupil Labs, whose scene camera provides RGB images with a resolution of 1088 × 1080. Each participant captured three recordings (each recording was about five minutes long, including calibration and evaluation), resulting in 14 valid videos in total. This led to a total length of about one hour of recording, consisting of 102 620 frames of which 27 946 contained objects.

Finally, we labeled the obtained data with DarkLabel [222]. Figure A.1.2 shows individual example moments from the recordings. Due to the errors related to the gaze estimation, the gaze points are, especially for small objects, not always on the labeled object, even though the participant was actually looking at it. In fact, even for a human, it is not always easy to determine the target object, and sometimes only possible considering the context and the observation of an image sequence. This demonstrates quite clearly the difficulties and challenges associated with this task. Our final, publicly available

Figure A.1.2: The images, some of them zoomed in, show exemplary moments of our data, where the objects that were consciously observed are labeled with a bounding box.

dataset only contains the gaze information and bounding boxes, yet no stimuli-related information.

A.1.5 Evaluation

In this section, we evaluate the classification of the gaze points with respect to the affiliation to an object, and we try to extract the position and the size of the object from those. To this end, we applied a variety of different, well-established machine learning methods and list here a selection comprising the best of them. In the classification experiments, we always specify the mean accuracy of a 5-fold cross validation. For the regression experiments, the mean error as a percentage of the image resolution from a 5-fold cross validation is given. We evaluated different heatmap grid sizes as well as different time window sizes. We conducted our evaluations on a computer system with Windows 10, an AMD Ryzen 9 3950X 16-core processor with 3.50 GHz, and 64 GB DDR4 Ram. All machine learning methods were implemented on the Matlab version 2021b and for reproducibility we restrict ourselves to Matlab's default parameters.

The assignment of classes (object or no object) to time windows was done based on the presence of an annotated object in the time window. This means that if there was an annotated object in the time window, the class was set to one, and zero otherwise. In the regression, only time windows with an existing annotated object were used. Here, the parameters of the annotated object closest to the central timestamp of the time window were chosen. This was assigned because, in most cases, our subjects moved while looking at an object. Thus, there are usually different positions and sizes of bounding boxes in a time window.

Figure A.1.3 and Table A.1.1 show a summary of the results of our classification exper-

Table A.1.1: Best and worst classification results of the 2D and 3D heatmap features. The mean is denoted by μ and the standard deviation by σ.

Feature	ML	Accuracy		$\mu \pm \sigma$
		Worst	Best	
2D heatmap	KNN	68	88	83.2 ± 3.8
	Bagged Trees	79	89	86.3 ± 1.9
	Gaussian SVM	73	84	79.4 ± 2.6
3D heatmap	KNN	73	92	87.8 ± 3.3
	Bagged Trees	80	89	86.8 ± 1.3
	Gaussian SVM	72	83	76.5 ± 3.6

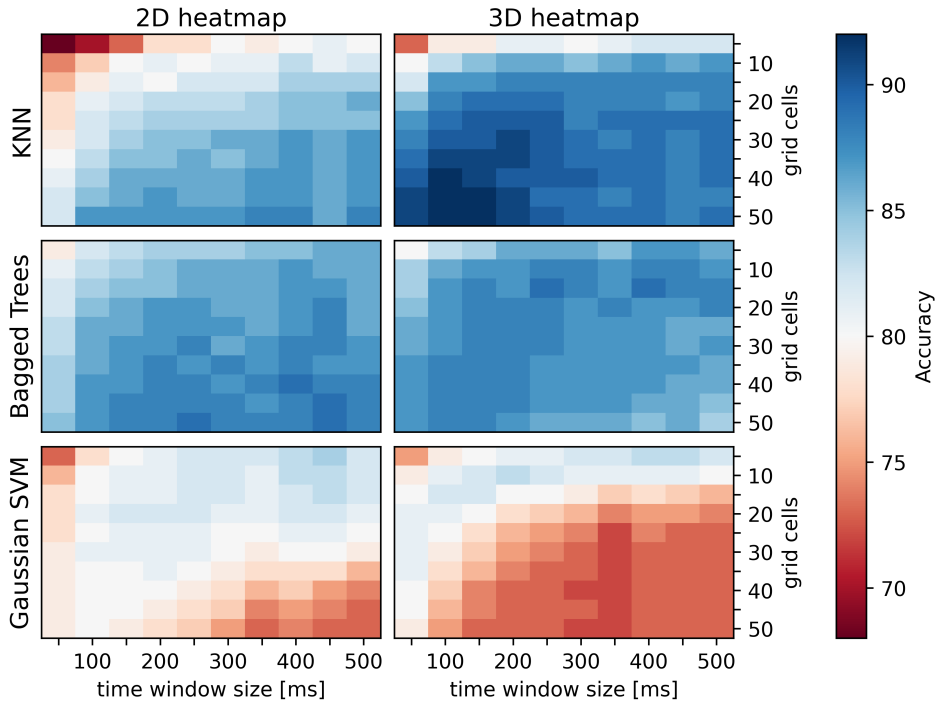


Figure A.1.3: Classification results of the 2D and 3D heatmap features for different time window sizes (in ms), number of grid cells, and machine learning methods illustrated in a heatmap. The results are the average accuracy of a 5-fold cross validation.

iment. Comparing the results of the three methods (KNN, bagged trees, and Gaussian SVM) for the 2D heatmap feature, the approach based on bagged trees achieves the best results. Looking at the progression over the grid and time window size, we can see that the KNN and the bagged trees perform best with a high number of grid cells and large time windows. In contrast, the Gaussian SVM performs best at a small number of grid cells but still large time windows. Moving on to the 3D heatmaps, the accuracy of the KNN method improves by 4 percent to 92 percent, which is also significantly better than the bagged trees.

The best results of our regression experiment are shown in Table A.1.2. Looking at the individual methods (Gaussian process regression, bagged trees, and Gaussian SVM), we see that all methods perform similarly well. As expected, based on the spatial heatmap feature, the position estimation is the most accurate. In contrast, the regression of the bounding box size, using only gaze data an no stimuli, is even more difficult than the position estimation and therefore less accurate. Comparing the results for the 2D and

Table A.1.2: Best regression error results as the average absolute error of a 5-fold cross validation in percentage. The columns X and Y denote the position of the bounding box, W is the width, and H is the height of the bounding box.

Feature	ML	Error			
		X	Y	W	H
2D heatmap	Gaussian Process	6.1	6.8	12.2	15.1
	Bagged Trees	6.4	6.9	12.0	14.3
	Gaussian SVM	6.4	6.9	13.4	15.5
3D heatmap	Gaussian Process	5.8	6.0	9.9	11.6
	Bagged Trees	6.4	6.7	10.5	12.3
	Gaussian SVM	6.2	6.2	11.0	12.9

the 3D heatmap feature, the position results remain about the same, with some overall improvement. In terms of bounding box size, the best results improve significantly for all of the three methods. All in all, the Gaussian process method combined with the 3D heatmap feature performs best.

Figure A.1.4 shows a qualitative extract of the Gaussian process regression in comparison to the ground truth. Naturally, the position is more accurate than the bounding box size, since humans tend not to observe the entire object when looking at it. Overall, however, both can be determined quite well.

Hereafter, we will investigate the runtime and memory requirements. It should be borne in mind that classical object detectors pursue a slightly different goal than we do. Whereas in their case all objects are to be detected, we are primarily interested in the existence of an object of interest, that is, the one that the human is looking at. Since classical object detectors only use scene images and do not obtain information about human gaze behavior, they cannot know whether a human is looking at an object, nor which object. Thus, it would be a matter of chance whether the statement is correct.

With the regression task, the detection of all objects would be possible. Here, however, we encounter a different real-world problem, outside of laboratory conditions, which also makes our method so appealing. Since we are in a wild world, the objects of interest are extremely diverse and their number tremendous. The vast majority of objects in our dataset, such as doorknobs, light switches, and fire extinguishers, are simply not part of any publicly available data sets, such as Microsoft COCO [185] or ImageNet [223], that are typically used for training. Since the methods differ too much in this respect, we need a benchmark that covers more the commonalities. Therefore, in the remainder of this section, we will establish a baseline comparison in terms of speed and computing resources. As a baseline,

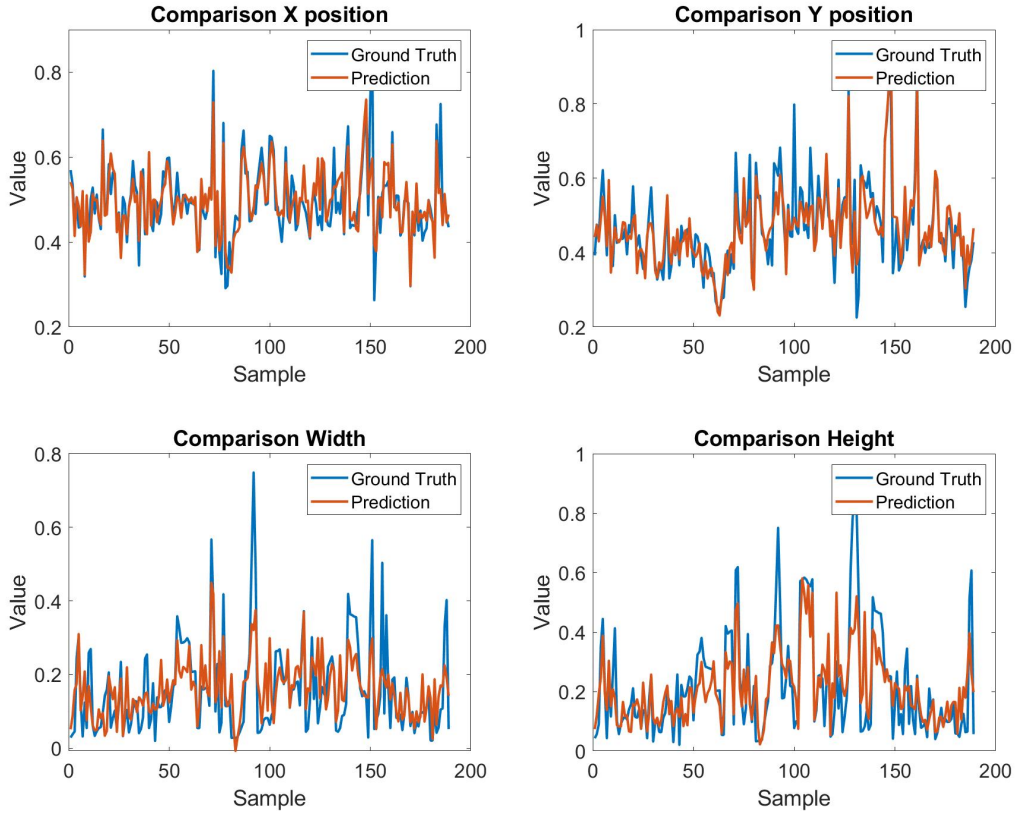


Figure A.1.4: Qualitative evaluation of the bounding box parameter regression. The results are from the Gaussian Process Regression with a time window size of 100, a grid cell number of 15 and the 3D heatmap feature.

we use state-of-the-art object detectors. These include Faster R-CNN [186], FCOS [191], and RetinaNet [204], each with a ResNet-50-FPN backbone [187], SSDlite320 [224] and Faster R-CNN both with a MobileNetV3 Large backbone [225], as well as SSD300 [224] with a VGG16 backbone [226]. These are supplemented by various YOLOv5 [188] variants. In order to test the speed, we measured the runtime of all methods on the CPU for 1000 individual predictions, i.e. 1000 different inputs with a batch size of one. The resource consumption was determined by measuring the amount of memory required for a single input. For our method with the heatmap input features, we used a time window size of 250 ms. For the classic object detectors, the $1088 \times 1080 \times 3$ RGB images were used as input. The summary of the results are shown in Table A.1.3.

The fastest are the Gaussian SVM and the KNN with the 2D heatmap feature. The Bagged Trees are slower, but the runtime increases proportionally less as the number

Table A.1.3: Comparison of the required resources for the different input features. The time column indicates the execution time for 1000 different inputs at a batch size of one in seconds. The memory column specifies the required memory of a single input in kilobytes. For the 2D and 3D heatmap features, the results shown are from a time window size of 250 ms and a grid cell number of 30.

Feature	ML	Time [s]	Memory [KB]
2D heatmap	KNN	10.8	424
	Bagged Trees	57.8	1134
	Gaussian SVM	8.6	406
3D heatmap	KNN	276.9	3045
	Bagged Trees	64.7	1467
	Gaussian SVM	610.6	3650
RGB Image	F. R-CNN [186] (RN50)	8 705.3	1 745 456
	F. R-CNN [186] (MN)	1 205.6	545 400
	FCOS [191]	4 723.2	995 416
	RetinaNet [204]	5 184.5	1 390 580
	SSD300 [224]	900.8	529 744
	SSDlite320 [224]	163.7	293 788
	YOLOv5n [188]	200.6	270 168
	YOLOv5s [188]	486.3	312 104
	YOLOv5m [188]	1 127.6	421 904
	YOLOv5l [188]	2 174.5	622 536
	YOLOv5x [188]	3 677.9	940 508

of grid cells increases. Consequently, the runtime for the 3D heatmap feature is in the range of one minute for the 1000 predictions while the runtime for KNN and Gaussian SVM increases considerably from a few seconds to several minutes. Nonetheless, it is immediately apparent that the runtime is in general significantly lower compared to the object detectors using the RGB images as input features. While only the smaller models like YOLOv5n and SSDlite remain under three minutes, the other models are much slower. In particular, the computation time required by the popular Faster R-CNN (RN50) exceeds that of the Bagged Trees by a factor of over 100.

A similar picture emerges with respect to the RAM allocated for one single prediction. The memory requirements of the bagged trees are larger for small inputs, but do not increase as much in proportion to the number of grid cells as for the KNN and the Gaussian SVM. Overall, the heatmap features require only a few 100 KB to a few MB. This is substantially less than the most frugal neural network YOLOv5n, which needs around

270 MB. Faster R-CNN with the ResNet-50 backbone requires the most memory with over 1.7 GB. Again, the factor is more than 100 times larger than for the Gaussian SVM with the maximum number of 50 grid cells. Compared to the Bagged Trees, it even exceeds 860 times.

In summary, our method is several orders of magnitude faster than conventional object detectors while requiring only a fraction of their resources.

A.1.6 Conclusion

In this work, we addressed object detection in the wild by means of gaze data. Our results show that it is possible to detect objects and determine their bounding box based solely on gaze information. Additionally, we have used a variety of machine learning methods to show that they work for solving such challenges. Besides, the functionality of several machine learning methods proves that our heatmap feature, which we have extended to 3D, can be used efficiently for this problem. In comparison to classical object detectors that use image input features, we have shown that object detection by means of our heatmap features is significantly faster while only requiring a fraction of the computational resources. This is of major relevance due to the fact that robots usually have only limited computing capacity at their disposal and cannot be equipped with powerful graphics units as they consume a lot of power.

However, a significant amount of work remains for the future as we plan to extend our proof of concept to a real robot by making the gaze of the human collaborator accessible to it. Our approach can serve as a foundation for future applications in the field of human-machine interaction and HRC, where robots can learn new objects from humans through instant knowledge sharing. Hence, we hope our methods and dataset can help to advance researchers in this challenging context.

A.2 Exploiting the GBVS for Saliency aware Gaze Heatmaps

(a) Fixation sequence on the painting *An Unexpected Visitor* from Ilya Repin.

(b) Regular gaussian like fixation heatmap.

(c) GBVS attention map with incorporated gaze signal.

Figure A.2.1: (a) shows the sequential fixation signal, where the size of the circles encodes the fixation time. (b) shows the corresponding gaussian like fixation heatmap. (c) shows the output of the proposed approach, where the tracked fixations are incorporated into the GBVS attention map calculation.

A.2.1 Abstract

Analyzing visual perception in scene images is dominated by two different approaches: 1.) Eye Tracking, which allows us to measure the visual focus directly by mapping a detected fixation to a scene image, and 2.) Saliency maps, which predict the perceivability of a scene region by assessing the emitted visual stimulus with respect to the retinal feature extraction. One of the best-known algorithms for calculating saliency maps is GBVS. In this work, we propose a novel visualization method by generating a joint fixation-saliency heatmap. By incorporating a tracked gaze signal into the GBVS, the proposed method equilibrates the fixation frequency and duration to the scene stimulus, and thus visualizes the rate of the extracted visual stimulus by the spectator.

A.2.2 Introduction

Our eyes move around to perceive and understand the scene in order to compensate for our limited- but clearest- foveal vision. When viewing a scene, we frequently focus our attention, known as a fixation, before shifting to another area with a rapid eye movement known as a saccade. The selectivity of the focused scene locations is a highly optimized and developed process, mainly driven by two factors: 1.) visual scene features, extracted

by the retina (bottom-up), and 2.) the interpretation of the extracted features regarding their semantic value by higher cognitive processes, and subsequent identification of the next fixation target (top-down) [227, 228, 229]. Modeling and understanding this reciprocal process is a long-term core topic in cognitive psychology and the computer vision community [230].

While retinal feature extraction can be modeled using saliency maps such as GBVS [189], the selectivity of visual attention can be measured using eye tracking. Saliency maps are predominantly bio-physiologically inspired algorithms to predict potential fixation targets [231]. Therefore, regions emitting a strong, recognizable visual stimulus are identified and emphasized by replicating the retinal visual stimulus processing. Eye tracking, on the other hand, often tracks the pupil center and extrapolates the line of sight to a scene image. A subsequent typical visualization is to illustrate the extracted fixations as a heatmap overlay on the scene image. This is used to investigate the visual attention on the scene, and to identify areas of particular interest.

However, fixation heatmaps are subject to some limitations. For instance, slight shifts in the eye-tracking signal make it difficult to identify the scene parts that attracted the visual attention and which information of the scene was actually perceptible. In addition, depending on the implementation, long or frequent fixations on the same scene region may lead to a high density in the fixation heatmap. Hence, it is assumed that these regions are particularly relevant to the spectator since a comparatively high amount of visual information was extracted. However, frequent or long fixations may also be caused by difficult scene conditions such as low contrasts. Thus, the visual information may be harder to extract, and therefore requires longer or more frequent fixations to be perceived.

In this work, we propose to incorporate detected fixations from an eye tracking signal into the calculation of the GBVS attention map. The resulting heatmap equilibrates the measured visual attention to the retinal-perceivable stimulus, and thus visualizes the density of perceived information in the scene more accurately as pure fixation or saliency heatmaps.

Structure of the Paper: Section A.2.3 gives a short introduction into state-of-the-art eye tracking visualization and saliency methods. Section A.2.4 contains a comprehensive description of how the proposed approach takes place in the GBVS algorithm. Section A.2.5 shows the exemplary application of the proposed visualization to different types of stimuli. The final sections A.2.6 state the limitations of the presented approach and the final remarks.

A.2.3 Related Work

The eye tracking community is a research powerhouse. Continuous improvements in tracking accuracy, precision, and availability over the last decades made eye tracking to one of the most eminent sensors in numerous research fields: Psychology, HCI, medicine, neuroscience, marketing, and many more. In particular, the success of recent years in the field of vision-based eye tracking has boosted the technology in terms of affordability, convenience, and usability for a broad community [232, 233, 234, 235]. However, the ability to conduct comprehensive eye-tracking studies led to increasing demand for sophisticated methods for visualization and qualitative evaluation of the acquired data [236].

An initial exploratory step in eye tracking studies is often to examine the spatial location, duration, and frequency of fixations as a heatmap over the stimulus [237]. This can be efficiently calculated over a large amount of data and gives a first impression of the distribution of visual attention on the stimulus [238, 239]. But, in order to gain deeper insights into the data, an extensive repertoire of different visualization techniques is available, such as various saccade metrics [240, 241, 242], AOI hierarchies [243, 244], or extensive interactive visualizations including stimulus and time domains [245, 246, 247, 248], etc. A comprehensive overview can be found in the survey of [249] and [250].

Saliency maps assess the stimulus by modeling the retinal signal processing to determine whether a scene area is particularly prominent in its immediate neighborhood, and therefore more likely to be perceived. The stimulus is evaluated by its intensity and its opponent color spaces: Driven by the neuronal circuit of the photoreceptors. Additionally, further feature spaces can be formed, such as edge orientation and difference formation of sequential images [251, 189, 252, 228, 253, 254]. While these bottom-up approaches mainly reproduce the feature extraction of the retina perception, newer deep-learning-based approaches show great success in modeling the whole processes, from the retinal feature extraction up to the semantic interpretation of the visual cortex and higher cognitive levels. They include the recognition and evaluation of abstract forms regarding their object-relatedness and semantic relevance [255, 256, 257, 258, 259, 260, 261, 262, 263].

The proposed approach combines the worlds of fixation heatmaps and salience maps, as a novel visualization technique. The resulting heatmap provides insights into the extracted information rate of the scene and extends the existing visualization techniques towards a more stimulus-driven paradigm.

A.2.4 Method

Similar to most saliency map approaches, the GBVS algorithm is divided into 3 consecutive steps [189]:

1. Extraction of a feature map M_t on a given image I_t.
2. Calculation of an activation map A_t based on M_t.
3. Normalization and combination of the activation map A_t.

Our approach amends step (2) by injecting the gaze signal g_t into the calculation of the activation map A_t. Steps 1 and 3 remain unchanged to the GBVS publication [189] and not further discussed here. The subscript t indicates the time domain since the gaze signal is given as a time series of consecutive fixation points. However, it also simplifies the handling with dynamic stimuli, such as videos. In the following, we assume that for each t exists a corresponding gaze signal g_t, as well as a stimulus I_t, respectively a feature map M_t.

The GBVS interprets the activation map as a state vector of a Markov model. The transition between two states is defined by a dissimilarity score over the feature map M_t. Thus, a random walk over the Markov model empowers those states that are dissimilar in the respective feature map. Analogous to the original GBVS, the dissimilarity between the two states i and j in the feature map M_t is defined as follow:

$$d_t(i,j) = \left| \log \frac{M_t(i)}{M_t(j)} \right|, \tag{A.2.1}$$

where $M_t(i)$ is the i-th value of the corresponding feature map M_t. The transition weight $w_t(i,j)$ between the two states i and j is defined as the product of their dissimilarity score $d_t(i,j)$ and a distance weight $F_w(i,j)$:

$$w_t(i,j) = d_t(i,j) \cdot F(i,j). \tag{A.2.2}$$

The distance weight adds a local sensitivity to the dissimilarity score. Thus, states that are dissimilar to their immediate neighborhood are emphasized while the impact of the dissimilarity score is attenuated with increasing distance. $F(i,j)$ is defined as an exponentially weighted square distance between the states i and j in their spatial dimension in the input image I_t:

$$F(i,j) = \exp\left(-\frac{\big(x(i)-x(j)\big)^2 + \big(y(i)-y(j)\big)^2}{2 \cdot \sigma} \right), \tag{A.2.3}$$

where $x(i)$ and $y(i)$ is the x- and y-coordinate of the i-th state in the respective input image I_t. The free parameter σ controls the shape of the exponential distance weight. The larger σ is chosen, the more weight is given to the dissimilarities of more remote states.

The final Markov transition matrix T_t is then assembled as follows:

$$T_t = \begin{pmatrix} 1 & w_t(0,1) & \dots & w_t(0,n) \\ w_t(1,0) & 1 & \ddots & w_t(1,n) \\ \vdots & \ddots & \ddots & \vdots \\ w_t(n,0) & w_t(n,1) & \dots & 1 \end{pmatrix}, \tag{A.2.4}$$

where n is the number of elements in the feature map M_t respective the input image I_t.

The activation map A_t is then calculated by k repeated multiplication with the transition matrix T_t:

$$A_t^{(k)} = T_t \cdot A_t^{(k-1)}. \tag{A.2.5}$$

Incorporate Gaze: Up to this step, the procedure follows the original GBVS algorithm. However, instead of initializing $A_t^{(0)}$ equally distributed, the gaze position is encoded as initial activation map:

$$A_t^{(0)} = q \cdot A_{t-1}^{(k)} + (1-q) \cdot \left(F(0,g_t), \dots, F(n,g_t)\right), \tag{A.2.6}$$

where $F(i,g_t)$ is the exponential weighted square distance between the recorded gaze position g_t and the spatial location of the i-th element in the activation map. In other words, the activation map is initialized by the measured visual activation from the eye tracking signal. Additionally, parameter $q \in [0,1]$ controls the influence of the previously calculated activation map A_{t-1} into the initialization of $A_t^{(0)}$. Thus, for $q > 0$, $A_t^{(0)}$ encodes the recently measured visual attention, but also the history of previous predicted attention areas. This smooths the resulting activation map $A_t^{(k)}$ in the temporal domain, and makes noise in the gaze signal less significant. However, it also poses the risk to generate a distorted activation map. For instance, on a dynamic stimulus: the previous predicted attentive area in frame I_{t-1} is located somewhere in frame I_t. Yet, $A_t^{(0)}$ provides values at this area and the Markov model will adapt it to the next salient region – which may not have ben actually focused on. Nevertheless, this effect only occurs if the content of the scene changes significantly, for instance on scene cuts in movies, or opening a new web page while browsing.

When generating static heat maps (such as Figure A.2.1), it is common to ignore the temporal domain completely. In this case, q is set to zero. The overall heatmap $\mathbf{A}^{(k)}$ is then the weighted sum over $A_t^{(k)}$:

$$\mathbf{A}^{(k)} = \sum_t A_t^{(k)} \cdot b_t, \tag{A.2.7}$$

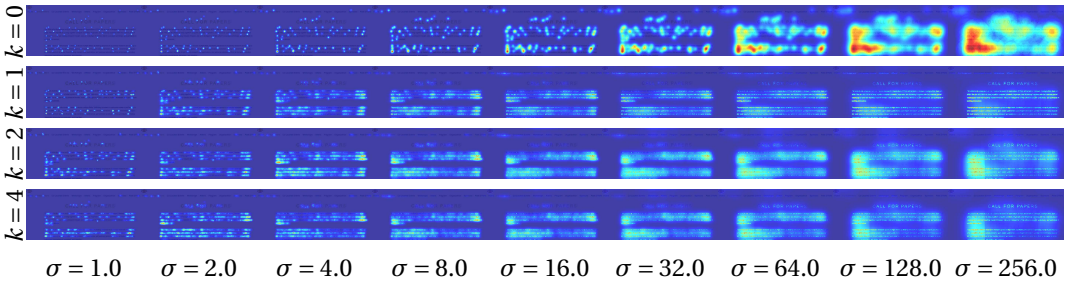


Figure A.2.2: Influence of the parameters σ (horizontal) and k (vertical) on the adaptation of the fixations heatmap to the text stimulus. The input is the same as used in figure A.2.3. For large k, the injected visual attention is increasingly distributed across the entire stimulus. The parameter σ should be chosen depending on the desired level of detail of the visualization. Using a text stimulus, it is usually reasonable to choose a high degree of detail (here $\sigma \leq 8$), in order to visualize the perception rate of single words or lines.

where the weighting b_t, for instance, can be chosen in relation to the fixation time.

Parameters: On regular gaussian like gaze heatmaps, σ models the area of visual attention (foveal perception) and/or the expected noise of the eye tracking signal, and thus controls the acuity of the resulting heatmap. In the proposed approach, σ controls the distribution of visual attention deduced from the fixation signal. But also how far the Markov model may adopt this distribution to the underlying stimulus in each iteration. The number of iterations is controlled by the parameter k. Whereby for $k = 0$, $A_t^{(0)} = A_t^{(k)}$ corresponds to a regular gaussian like fixation heatmap of a single fixation point (respectively $\mathbf{A}^{(0)}$ overall fixation points). Figure A.2.2 shows how the initial gaze heatmap $A_t^{(0)}$ is gradually distorted to the stimulus for each additional iteration over equation A.2.5.

Implementation Details: The main limitation of GBVS is runtime and memory consumption. The transition matrix T_t grows in quadratic size with the input size n, and thus quickly exceeds the available memory (e.g. $> 9.4 \cdot 10^{10}$ elements on a VGA resolution). Additionally, the initialization of T_t requires a runtime complexity of $\mathcal{O}\left(n^2\right)$. Both together, limit the GBVS to very low input resolutions, which leads to a loss of acuity. Thus, the standard parametrization of the GBVS toolbox limits the internal resolution to an edge length of 32px [205].

On closer examination, however, it is apparent that most values in T_t are extremely small and have no significant impact to the resulting activation map $A_t^{(k)}$. Thus, after applying a threshold l, the transition matrix T_t becomes predominantly sparse. Furthermore, assuming that $M_t \in [0,1]$, the elements of T_t, which potentially exceed the threshold

l can be determined in relation to σ:

$$l < F(i,j), \tag{A.2.8}$$

and resolves to:

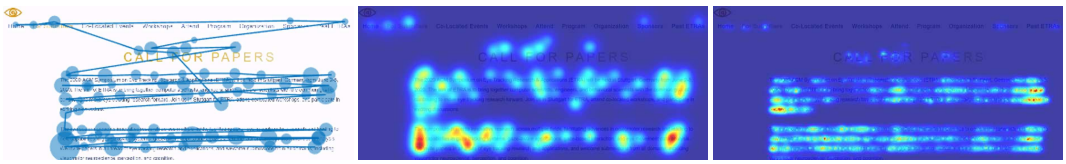
$$\sqrt{-2\cdot\sigma\cdot\log(l)} \geq \sqrt{\left(x(i)-x(j)\right)^2+\left(y(i)-y(j)\right)^2}, \tag{A.2.9}$$

where the right term is the euclidean distance between the the i-th and j-th element in the feature map M_t. Thus, initializing T_t only requires the calculation of $2\cdot\sqrt{-2\cdot\sigma\cdot\log(l)}$ elements per row, since all other elements are not exceeding the threshold l. This reduces the actual runtime from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. Similar considerations can be made for the initialization of $A_t^{(0)}$ (although this is not a bottleneck). However, due to the sparseness of T_t and $A_t^{(0)}$, solving equation (A.2.5) is much faster [264].

A.2.5 Experimental Demonstration

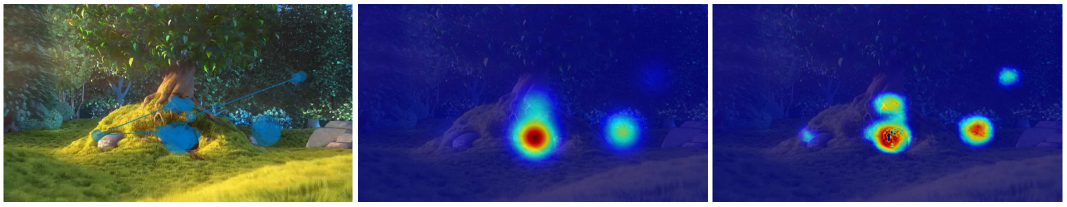
Figures A.2.1, A.2.3, and A.2.4 demonstrate the application of the proposed visualization on different stimulus types: the *An Unexpected Visitor* painting from Ilya Repin, the *Call for Papers* website from ETRA 2020 as text, and a short video snippet of *Big Buck Bunny* from the Peach open movie project [265]. The gaze signal was recorded by a Tobii Pro Spectrum at 1200Hz. The fixation locations and duration were extracted using the fixation filter I-VT provided by Tobii Pro Lab and default parametrization [266]. All stimuli were presented as full screen on the Monitor at 1920 × 1080 pixels.

On the text stimulus, it is recognizable how the Markov model depicts the measured visual attention to paragraphs, lines, down to single words and characters. Therefore, the acuity of the heatmap is increased, and consequently, interpretations about the perception



Input fixation map & stimuli | Gaze Heatmap ($\sigma = 32.0$) | GBVS Heatmap ($\sigma = 4.0$, $k = 2$)

Figure A.2.3: The left image shows the recorded fixation sequence on the ETRA 2020s *Call for Papers* website. The middle image shows a regular gaussian fixation heatmap ($\mathbf{A}^{(0)}$). The right image shows the output of the proposed method ($\mathbf{A}^{(k=2)}$).



Input fixation map & stimuli | Gaze Heatmap ($\sigma = 16.0$) | GBVS Heatmap ($\sigma = 8.0,\ k = 2$)

Figure A.2.4: The left image shows the recorded fixation sequence on a short snippet of the video clip *Big Buck Bunny* from the Peach open movie project [265]. The middle image shows a regular gaussian fixation heatmap ($\mathbf{A}^{(0)}$). The right image shows the output of the proposed method ($\mathbf{A}^{(k=2)}$).

rate to text passages are simplified. For instance, in the field of web design and advertising, the proposed model can help to analyze whether a certain area attracts the desired level of visual attention and whether the presented information was easily visual accessible to the spectator.

However, in this context, text reading is a relatively unambiguous challenge, since the text is very salient to its background. At the same time, the text is often the only element that attracts the visual attention of the reader. The strength of the proposed method of visual attention visualization is particularly evident in more complex stimuli as shown in figures A.2.1 and A.2.4. Considering the *An Unexpected Visitor* painting, the fixations are mainly on the faces in the scene, but also on some miscellaneous areas, such as hands, the paintings in the background, or feet. However, the regular fixation heatmap has a particularly pronounced fixation cluster on the face of the woman in the background. This can be attributed to the fact that this face is particularly difficult to perceive due to its low contrast. Yet, the long and frequent fixations in this area lead to a suppression of all other fixations, which can lead to the interpretation that this area was of higher interest for the spectator. The GBVS generated fixation heatmap incorporates not only the fixation duration and frequency but also how accessible the stimulus in the region is to the observer. The result is a much more balanced fixation heatmap, where all the fixated heads are clearly pronounced.

A.2.6 Final Remarks

The proposed method extends the well-known GBVS saliency algorithm by incorporating the measured visual attention. The resulting heatmap visualizes a predicted perception rate of scene areas for an individual or multiple spectators. However, as the most bottom-

up saliency algorithm, GBVS uses exclusively intrinsic scene features to predict whether certain scene content is attractive for fixation. It turns out, this is very accurate for a free viewing scenario. Yet, various tasks may require the spectator to direct their visual attention to less saliency scene areas. The proposed algorithm might distort these fixation points to a close salient region and thus weigh the perception rate based on the wrong stimuli. This limitation can be compensated by using a small σ and high-resolution scene images but requires a high accuracy of the fixation point.

In practice, however, it has been shown that the proposed visualization generates more intuitive heatmaps than pure fixation heatmaps. Thus, the presented visualization provides an ingenious overview of the scene areas with a distinctive high rate of visual awareness.

B Perceiving and Multiperspective Teaching of Unknown Objects

The following publications are enclosed in this chapter:

[1] **Daniel Weber**, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling Location Proposals of Unknown Objects through Gaze Information for Human-Robot Interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11086–11093. IEEE, 2020. doi:10.1109/IROS45743.2020.9340893.

[3] **Daniel Weber**, Enkelejda Kasneci, and Andreas Zell. Exploiting Augmented Reality for Extrinsic Robot Calibration and Eye-based Human-Robot Collaboration. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 284–293. IEEE, 2022. doi:10.1109/HRI53351.2022.9889538.

[5] **Daniel Weber**, Wolfgang Fuhl, Enkelejda Kasneci, and Andreas Zell. Multiperspective Teaching of Unknown Objects via Shared-gaze-based Multimodal Human-Robot Interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 544–553, March 2023. doi:10.1145/3568162.3578627.

[6] **Daniel Weber**, Valentin Bolz, Andreas Zell, and Enkelejda Kasneci. Leveraging Saliency-Aware Gaze Heatmaps for Multiperspective Teaching of Unknown Objects. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. (Accepted for publication).

The publications templates have been slightly adapted to match the formatting of this dissertation. The ultimate versions

B.1 Distilling Location Proposals of Unknown Objects through Gaze Information for Human-Robot Interaction

B.1.1 Abstract

Successful and meaningful human-robot interaction requires robots to have knowledge about the interaction context – e.g., which objects should be interacted with. Unfortunately, the corpora of interactive objects is – for all practical purposes – infinite. This fact hinders the deployment of robots with pre-trained object-detection neural networks other than in pre-defined scenarios. A more flexible alternative to pre-training is to let a human teach the robot about new objects after deployment. However, doing so manually presents significant usability issues as the user must manipulate the object and communicate the object's boundaries to the robot. In this work, we propose streamlining this process by using automatic object location proposal methods in combination with human gaze to distill pertinent object location proposals. Experiments show that the proposed method 1) increased the precision by a factor of approximately 21 compared to location proposal alone, 2) is able to locate objects sufficiently similar to a state-of-the-art pre-trained deep-learning method (FCOS) without any training, and 3) detected objects that were completely missed by FCOS. Furthermore, the method is able to locate objects for which FCOS was not trained on, which are undetectable for FCOS by definition.

B.1.2 Introduction

In today's modern world, interaction between human and machines is omnipresent, e.g. in the figure of Alexa and Siri. Moreover, the significant progress in augmented reality is also pushing the boundaries of the cooperation between human and robots. For instance, this emerging kind of human-robot interaction (HRI) has already helped to optimize manufacturing steps in production as well as been applied in factories [267] and for assembly guidance [268]. The great majority of such technological developments has been strongly fueled by machine learning methods, such as neural networks. The collection of huge databases allows us to train and continuously improve (deep) neural networks in order to fulfill challenging tasks. However, these use cases typically operate under the assumption that there are sufficient data sets for training available. But what if the training data is biased (e.g., geographically [269]) or not labeled, for example in many production processes – such as, the assembly of a recently developed electric engine of a car? Furthermore, in some application scenarios such as search and rescue work with unmanned aerial vehicles (UAVs) or the classification of medical images, there might be

Figure B.1.1: Without specialized pre-training, the robot does not know the objects in front of it. Nonetheless, through our proposed approach, the robot is capable of detecting these unknown objects based on gaze-based human-robot interaction without any training instances.

very few or no available training instances. For example for UAVs, aerial footage is simply difficult to obtain [270], whereas for medical images, storage is often prohibited due to patient privacy [271].

In addition, labeling data is a costly process due to the amount of human effort involved. Drawing a high quality bounding box in an image, including quality and coverage verification, can take a human from 7 up to 42 seconds per object [272, 273]. With multiple objects in a scene this can quickly add up to prohibitive amounts.

In this paper we address this challenge by connecting findings from two research areas: eye tracking and robotics. On the human side, we use the gaze modality to enable the exchange of information for a specific problem on the robot side, namely the detection of unknown objects. Our goal is to enable the deployment of a robot in a non-predefined scenario and to explain an interaction context to the robot, e.g., the class of an object after detection. That is, rather than using a neural network for object detection, we resort to the human gaze and want the robot to detect which object the human is looking at, even though its class is not yet known (see Figure B.1.1). Moreover, interaction requires online operation – in contrast to post processing. To the best of our knowledge, this is the first work to combine the well known technique of selective search [190], which outputs thousands of class-independent object location proposals, with human gaze information

to separate useful and useless areas of interest in a scene image. Thus, the proposed approach enables us to detect and process objects in an image without training but still in an efficient way. In summary, our most important contributions are:

1. A novel method towards the deployment of robots in non-predefined scenarios.

2. We are the first to connect eye tracking and robotics to detect unknown objects without the usage of neural networks, alleviating training-data dependency.

3. As a proof of concept we conduct an experiment and demonstrate the validity and feasibility of our method.

B.1.3 Related Work

Mapping gaze data from a head-mounted eye tracker with moving point of view, i.e. coordinate system, to a known reference frame is a well-known and open problem in current research. Most works, such as [181, 274], that were confronted with this issue solved it by using fiducial markers. Even though [275] additionally tested feature matching and achieved reasonable results, markers provided better stability and reliability at significantly less computational cost in all of their test cases. Apart from that, their purpose was to match a picture of an image displayed on a screen to a planar reference image, which was very similar to the one displayed on the screen. As described in [276], feature matching reaches its limit when applied to a three-dimensional target object. Accordingly, in our case it is more difficult to find and match features than with a simple painting or a poster, especially when the viewing perspectives differ significantly. In [277] the authors succeeded in mapping the gaze by utilizing velocity features. However, this was limited to the user looking at one of several pre-defined key points.

In recent years, object recognition has been one of the most intensively researched areas in computer vision. The availability of better hardware led to the emergence of deep neural networks as a go-to solution for object detection. YOLO [278], Mask R-CNN [279], SSD [224] and FCOS [191] are great examples of the extensive use of neural networks that constantly have been pushing the boundaries of object detection. These networks typically rely on fully supervised learning methods and the existence of large annotated data sets, such as PASCAL VOC [192], Microsoft COCO [185] and Imagenet [223]. This means that they do not generalize well and lack reliability on unknown domains [280].

Moreover, with increasing climate-related public awareness, there has been some research focusing on energy efficiency of neural networks [281] and its environmental impact [282]. [283] analyzed the power consumption of popular image classification

models. Consequently, we follow the recommendation of [282] and prioritize a simple non-deep-learning approach instead.

A few works have already investigated the combination of eye tracking and computer vision tasks. The authors of [284] performed gaze guided object recognition by matching features around human fixations to features from known objects in a database. After a database was created, it was possible to classify an image, but not to determine the position of the object within the image. [285] concentrated on annotating images with bounding boxes. They utilized fixation points to extend existing training data with gaze information. Subsequently, a model was trained that predicted bounding boxes from the fixations while viewing an image. A strategy for superpixel segmentation with eye tracking data was proposed by [215]. Just like the previous method, training data was already required right from the start. In addition, both methods require multiple gaze points. In contrast, our method is able to operate with as few as one gaze point, thus being applicable in an online fashion.

In this paper, we build on existing work and benefit from collaborative working with a robot. In this context, eye tracking can play an important role and connect humans and robots in a natural and intuitive manner, offering an additional communication channel available even when traditional channels, such as speech and gestures [286], might not be available for HRI – e.g., during microsurgery [287]. We use the human gaze to enable a robot to interact with its unknown environment by letting it recognize objects we are looking at. Thereby, we bridge the gap between existing approaches for object detection and data independence with eye tracking.

B.1.4 Method

In this work, we propose finding pertinent and accurate location proposals of unknown objects through gaze information. This process can be thought of as three building blocks: 1) estimating the human partner's gaze in the robot's frame of reference, 2) generating location proposals for unknown objects, and 3) distilling the location proposals using the gaze information. Throughout this section, we assume the robot to be equipped with at least one camera.

Gaze Estimation

The most straightforward and inexpensive way of estimating the partner's gaze in the robot's frame of reference is by estimating the gaze directly through the robot's sensors – e.g., through appearance or model-based remote gaze estimation methods [288, 289].

However, this poses a key limitation as the partner must be facing the robot, severely limiting the perspectives from which gaze-based HRI can happen.

This limitation can be alleviated through multiple remote eye trackers distributed around the environment or the usage of a head-mounted eye tracker. However, in both cases, it is necessary to map the estimated gaze from the eye tracker frame of reference to the robot's. This transformation can be achieved in multiple ways, for example by 1) directly finding the eye tracker's pose in the robot's camera or vice versa, or 2) indirect co-location, by finding at least four corresponding points in images of the eye tracker's and robot's cameras[1].

In this work, we favor the usage of a head-mounted eye tracker due to the reduced costs (i.e., only a single eye tracker is required) and user constraints. Moreover, we employ fiducial markers [290] for co-location as these provide a robust and inexpensive solution to the gaze mapping issue that can be employed in traditional HRI scenarios such as in factories, care facilities, or individual homes.

Unknown Object Location Proposal

Location (or region) proposal methods consist of determining candidate object locations (e.g., bounding boxes, or segmentation masks) that *might* contain an object. This task can be realized, for example, through segmentation [291], randomly-sampled boxes classification [292], jumping windows [293], and selective search [190]. Such methods are typically used as an alternative to exhaustive search for object detection to reduce the search space, speeding up the detection and reducing the associated computing costs.

The cardinality of the proposed locations set is, naturally, image-dependent but tends to be in the order of thousands. Normally, each location proposal is run through a pre-trained classifier to detect whether an object is present in it. However, many of these methods, such as the ones proposed by [190, 292], have a particularly interesting property: The proposed locations are *class-independent*. In other words, within the proposed regions there are objects that a computer vision system might not have been trained to identify – i.e., unknown objects. This begs the question: Can we identify pertinent locations from the set of proposals for interaction or to teach a robot about new objects in a natural way?

[1] By finding the plane defined by the these four points, one can estimate the pose of each camera relative to the plane and, thus, the pose of one camera relative to the other.

Distillation Through Gaze Information

In this work, we approach the task of identifying location proposals that are pertinent for a human-robot interaction from the full set of class-independent proposals by using gaze information from the human partner. This distillation process can be activated through multimodal interactions – e.g., through touch or voice. Nevertheless, we also envision an automatic approach in which the robot notices the human's gaze continously attending to a region where no known object has been identified yet.

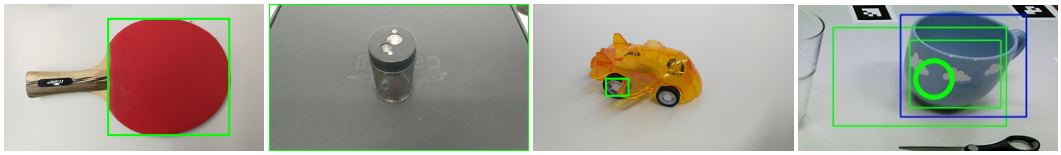
In order to obtain an initial set of candidate bounding boxes, we resort to selective search [190]. Selective search uses the segmentation method from Felzenszwalb and Huttenlocher [294] to analyze the intensity of the pixels of the image and perform segmentation. The segmented parts and groups of adjacent segments are then used to calculate and propose regions of interest. In other words, this algorithm-based approach combines the high recall of exhaustive search with the image guided sampling process of segmentation and outputs bounding boxes in a hierarchical order. The benefits here are two-fold: the method can capture all possible object locations and the region proposals are guided by the structure of the image, such as color, texture, size and shape, leading to a reduced number of proposed locations. In this paper, we will refer to the position with respect to the order in which the boxes appear in the output set of region proposals as position *index*.

Although the number of bounding boxes is reduced in comparison to an exhaustive search approach, this does not effect the high recall we need to ensure that we can find a suitable box for each object. Moreover, it is possible to further distill the regions into a smaller and more-pertinent set of proposals: Since we know that the gaze coordinate has to lie within the searched bounding box, we can employ this information as a filtering mechanism. Let $(x_1^{(i)}, y_1^{(i)}) \in \mathbb{N}^2$ be the lower left and $(x_2^{(i)}, y_2^{(i)}) \in \mathbb{N}^2$ be the upper right corners of the bounding box $B_i \in B$, where B is the full set of class-independent proposals. By tracking our gaze point $g = (x, y) \in \mathbb{N}^2$, we can distill a smaller subset $B_g \subset B$ of pertinent proposals from B:

$$B_g := \left\{ B_i \in B \,\middle|\, x_1^{(i)} \leq x \leq x_2^{(i)},\ y_1^{(i)} \leq y \leq y_2^{(i)} \right\}.$$

This subset B_g contains only bounding boxes that have an intersection with the object marked by the gaze point. As we will see later, to achieve satisfactory results, we are dependent on a high gaze-tracking accuracy and a robust gaze mapping.

Note that getting multiple (but hierarchically-sorted) bounding boxes proposals is not a disadvantage but an advantage in our use case. As previously mentioned in [190], an object can consist of different colors, multiple objects can have the same color, or the



(a) table tennis racket (b) sharpener (c) toy (d) cup

Figure B.1.2: Objects can vary in shape and size, have different backgrounds and can consist of multiple colors. This may cause errors regarding the detection. The green boxes in the figure indicate proposed regions. In (a) the red part is proposed earlier, meaning the corresponding bounding box has a lower position index than the whole racket. (d) shows the first three proposals we receive for the blue cup. The first two (green) are not as accurate as the third (blue). Through interaction it is possible to communicate the preferred bounding box.

object could be indistinguishable from its background. In Figure B.1.2 one can see that this could lead to problems if the detection fails in terms that the only proposed bounding box is not correct or the object is not detected at all.

Moreover, we strive for a more human-like learning process, in the sense of an interaction between robot and human, similar to that of a human with another human. Multiple proposals also mean that we can decide to choose the second or third proposed and more accurate box instead of the first one (see Figure B.1.2d). Interaction between robot and human makes these decisions possible and brings us closer to a natural learning process.

B.1.5 Experimental Setup

In order to showcase a working proof of concept of the proposed application, we collected a session for a participant (one of the system's designers) with the whole system working in real-time[2]. This session serves as basis for our initial evaluation of the system.

On a table, we placed different objects, including partially overlapping objects to some extent. To have a wide appearance range, we selected objects with distinct sizes, colors, and shapes. In Figure B.1.3, one can see the robot and his view in front of the table with all objects he is supposed to detect. For the sake of simplicity and for later evaluation, we have used ordinary office and household items that are all part of the Microsoft COCO data set [185].

As hardware, we used the first generation of Pupil Core [295], a head mounted eye tracker developed by Pupil Labs. Although Pupil Labs provides a software solution called *Pupil Capture* and *Pupil Player*, we decided to utilize *EyeRecToo* [233], an open-source

[2]Eye tracking and gaze mapping working at about 30 frames per second.

Figure B.1.3: With a Microsoft Kinect v2 the robot sees different objects on a table: Keyboard, scissors, cups, bottle, fork, knife, spoon, mouse and a small toy car.

software for real-time pervasive head-mounted eye-tracking. The main reasons were the calibration method *CalibMe* [232], the robust detection of ArUco markers, and slippage robustness [234]. *EyeRecToo*'s pupil tracking pipeline was set to use *PuRe* [296] / *PuReST* [297]. Our robot counterpart is a Scitos G5 from MetraLabs [298] equipped with a Microsoft Kinect for XBox One. We accessed the RGB channels of the Kinect v2 using *ROS* [183], *libfreenect2* [299], and *iai_kinect2* [300]. For the implementation, we make extensive use of the *OpenCV* [301] library.

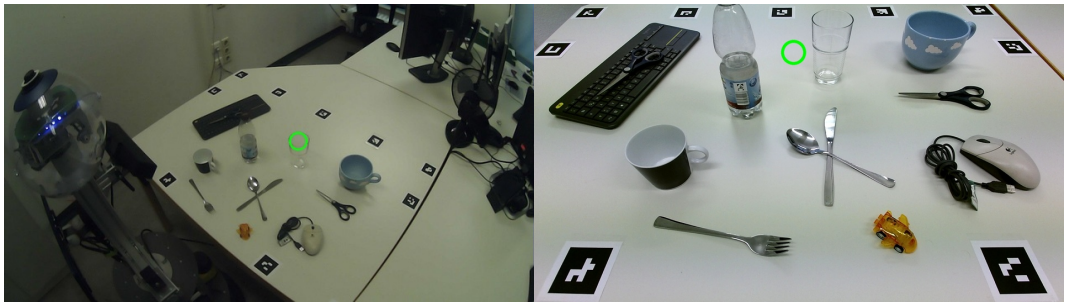
B.1.6 Evaluation

To establish reference ground-truth values for the object locations, we have employed the Fully Convolutional One-Stage Object Detector (FCOS) [191] trained on Microsoft COCO [185], using the ResNeXt-64x4d-101 backbone with deformable convolutions. This serves

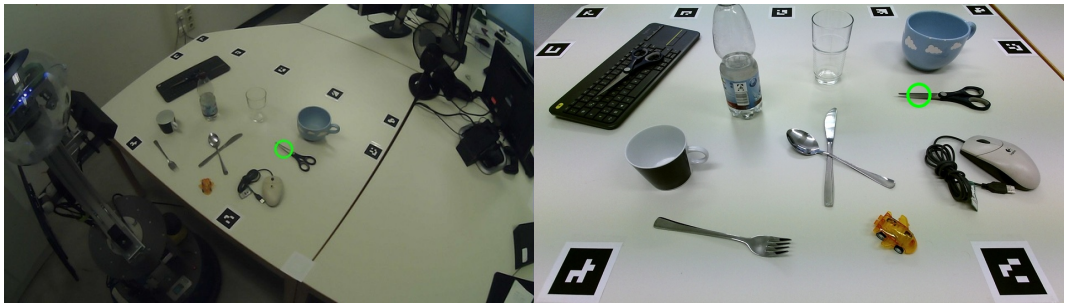
as a baseline representing a state-of-the-art object detection for supervised learning. Given an image viewed from the robot's perpective, the output of FCOS is shown in Figure B.1.5a. It is worth noting that the bottle was detected twice; in this case, we opted to ignore the smaller inaccurate bounding box. Moreover, neither the knife that overlaps with the spoon nor the scissor placed on the keyboard are recognized by FCOS, despite all of these classes being present in the training data. Thus, we discuss these separately.

Qualitative Analysis

Both eye tracking and marker detection work in real time, as well as the subsequent gaze mapping. Therefore, our method is suitable for real-time human-robot interaction. As long as the accuracy in all three steps is high enough, the robot knows at any time where we are looking at. The human is even unrestricted in his movements. Figure B.1.4 shows two attempts of pointing out an object to the robot. One was successful and the other one failed. Although the human from whom the gaze point in Figure B.1.4a originated



(a) Failed



(b) Successful

Figure B.1.4: A failed and a successful attempt of mapping the human gaze (left) on the robot's view (right).

actually looked at the glass and his gaze was tracked correctly, the gaze point in the robot's view is not on the glass, i.e. the mapping procedure was problematic in this case. This exemplifies that enough markers have to be detected to guarantee reliable mapping and usability. This could be ensured, for example, by using more accurate markers such as infrared tokens. In addition, the tracking of the human gaze must work reliably to achieve satisfactory usability. Therefore, we have carefully calibrated the eye tracker to achieve the desired accuracy. During interaction, however, the device is likely to slip [302] such that slippage robustness is paramount.

In contrast to the gaze mapping, the region proposal achieves real-time operation only at lower frame rates. The calculation of all the 2198 region proposals on our picture of the robot's view with a resolution of 1900x1080 took about 2.7 seconds with the "quality" method. Nonetheless, this is not a problem, as region proposal is not required for each frame but only sporadically. Once a correct bounding box for the intended object has been found, it can be tracked with well-known tracking algorithms like KCF [303] or CSRT [304].

Quantitative Analysis

To evaluate the efficiency of our method we compare the position indices of each bounding box within the complete hierarchical set of region proposals from the selective search algorithm with the indices we have distilled. Of course these boxes should not only be easy and fast to find but have to be accurate as well. For this reason, we need to investigate the similarity of the proposed boxes w.r.t. the ground truth. As measurement for accuracy, we calculate the Jaccard index $J(B_1, B_2)$, also known as Intersection over Union (IoU). This means that the closer the Jaccard index is to 1, the greater the similarity between the boxes. For object detection, if the Jaccard index is more than 0.5, a detection is typically considered correct [192]. Nevertheless, in general, a higher value is desirable. [193] provides a comparison of different values of the Jaccard index and describes 0.5 as very loose, 0.9 as very strict and 0.7 as reasonable compromise in between. Therefore, we set 0.7 as threshold and characterize bounding boxes with at least this value as "sufficient". This allows us to analyze whether the selective search algorithm is a good choice and provides region proposals that are accurate enough, i.e. sufficient, for our use case.

In Table B.1.1 the Jaccard index between the boxes predicted by FCOS and the best box in our set of proposals is listed for each item. Note that the knife and scissor placed on the keyboard are omitted from Table B.1.1 because they are not recognized by FCOS, which means we do not have any reference values for these items. We will discuss these items separately at the end of this section. Besides, depending on whether we want to

Table B.1.1: Comparison between the full and our distilled set of bounding boxes.

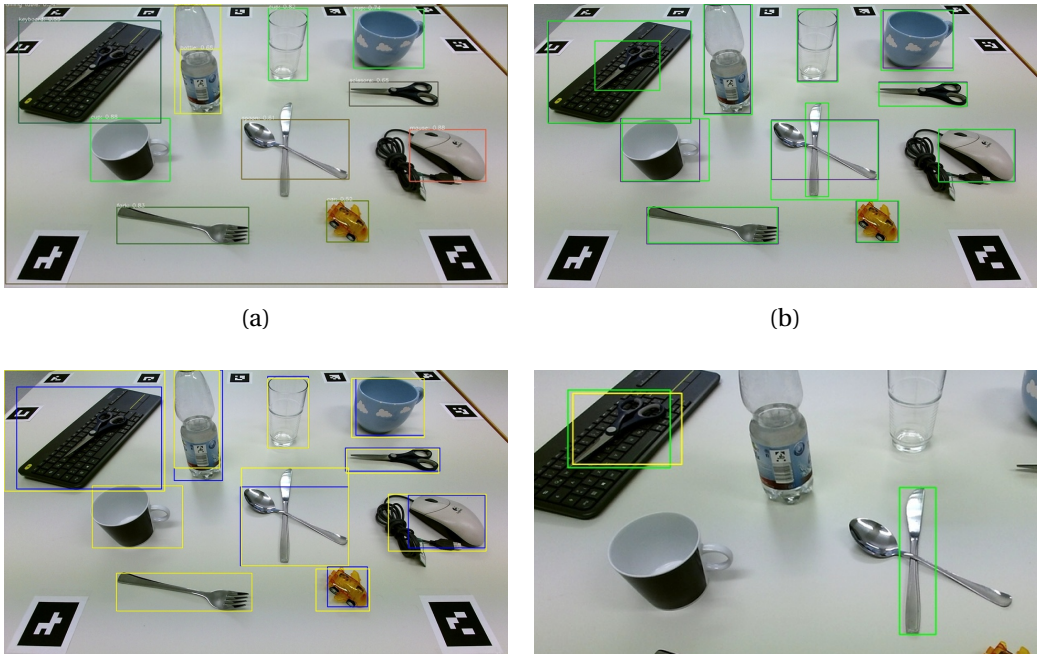
Item	FCOS	Best total		#Boxes	First sufficient			Best among first 15			Recall		Precision		F_1 score	
	Confidence	Index	IoU	Dist.	Index	IoU	Acc.[1]	Index	IoU	Acc.	Full	Dist.	Full	Dist.	Full	Dist.
Bottle	0.69	**292**[2]	0.943	98	1	0.851	90.24 %	**12**	0.943	100 %	1	1	0.012	0.265	0.023	0.419
Cup (black)	0.88	1748	0.863	221	**3**	0.828	95.94 %	**3**	0.828	95.94 %	1	1	0.029	0.29	0.057	0.449
Cup (blue)	0.74	1199	0.918	110	3	0.870	94.77 %	15	0.899	97.93 %	1	1	0.014	0.273	0.027	0.429
Fork	0.83	1873	0.945	34	**1**	0.939	99.37 %	**1**	0.939	99.37 %	1	1	0.013	0.824	0.025	0.903
Glass	0.82	1429	0.935	110	3	0.896	95.83 %	4	0.922	98.61 %	1	1	0.020	0.391	0.038	0.562
Keyboard	0.55	1839	0.988	189	1	0.751	76.01 %	9	0.981	99.29 %	1	1	0.046	0.529	0.087	0.692
Mouse	0.88	883	0.968	231	1	0.714	73.76 %	11	0.955	98.66 %	1	0.96	0.011	0.104	0.023	0.188
Scissor	0.68	1137	0.954	89	2	0.880	92.24 %	9	0.898	94.13 %	1	1	0.018	0.449	0.036	0.620
Spoon	0.61	670	0.727	118	**3**	0.720	99.04 %	**3**	0.720	99.04 %	1	1	0.002	0.034	0.004	0.066
Toy car	0.52	2079	0.946	68	3	0.712	75.26 %	5	0.909	96.09 %	1	1	0.021	0.662	0.040	0.797
Ø	0.72	1314.9	0.919	126.8	2.1	0.816	89.25 %	7.2	0.899	97.91 %	1	0.996	0.018	0.382	0.036	0.512

[1] Accuracy compared to the best box in the full set (ratio of the two Jaccard indices).
[2] Bold indices within the same line indicate identical boxes.

consider the mouse cable or not, the values in Table B.1.1 naturally change. Even though we can distill boxes for both cases, here we stick to the output of FCOS and only consider the mouse without cable. One must keep in mind, however, that the composition of the mouse from two sub objects leads to different predictions being made. For example, if the cable had not been bundled up, the relevant position index would indeed be smaller or the mean Jaccard index would be larger. In our test, the use of a second gaze point allowed the delimitation to boxes containing the cable and the mouse body alone. Compared to regular object detection, we are not bound to fixed ideas of objects but can vary the object's bounding box depending on the situation. It is worth noting that this is a good example where our method surpasses all pre-trained object detectors in terms of flexibility. In this particular case, there is not only one correct box, but two. *Gaze allows us to resolve such ambiguities.*

We can observe that the Jaccard indices with few exceptions are all above 0.85 and the vast majority is even above 0.9 (see column 4). This is highlighted visually in Figure B.1.5b, which shows the bounding box detected by FCOS and by the proposed approach. With a mean value of the Jaccard indices of 0.919, the proposed boxes are highly relevant. This value can also be used as upper bound for the accuracy of our fast distillation. Also worth noting are the massively high indices of the respective best box in each category. Whereas the position index of the best box of the bottle in the full set is the lowest at 292, the mean value of the position index is about 1315. Through distillation, we managed to improve that value to an average of 61.5. Figure B.1.6 illustrates that the vast majority of the boxes in the full set of proposals has an Jaccard index below 0.1 and is therefore irrelevant.

As Table B.1.1 and Figure B.1.5c show, we do not have to find exactly the best boxes. With our fast distillation method, we were able to provide a proposal among the first 15 boxes of each distilled subset with at least 94 % accuracy to the best. That is, with an average accuracy of even 97.91 %, we were almost as accurate as the best possible box,



(a) (b)

(c) (d)

Figure B.1.5: (a) shows the objects detected with FCOS. The confidence of the prediction can also be seen in Table B.1.1. (b) shows a comparison of the total best bounding box (green) with the ground truth given by FCOS (purple). (c) shows a comparison of the best bounding box among the first 15 proposals (blue) with earlier sufficient boxes (yellow). Note that these boxes are identical for the cup and the fork. The knife and scissors on the keyboard have been omitted as they are handled separately. (d) shows the best bounding boxes distilled for the knife and the scissor on the keyboard.

with a much smaller position index. Therefore, we need much less communication with the robot to reach the desired box.

In addition, we have considered earlier sufficient boxes in the sense of boxes with a Jaccard index of at least 0.7. Figure B.1.5c shows these bounding boxes along with the best boxes among the first 15 described above. Their position index is of course lower and, in our particular case, never higher than three. As highlighted in Table B.1.1, it is often sensible to fall back to earlier boxes. For instance, in the case of the blue cup, it is possible to reduce the position index from 15 to 3 while reducing the Jaccard index only by 0.03. In contrast, the toy car's position index is acceptable either way, and a significant drop in accuracy results if the position index is lowered from 5 to 3. In general, the average accuracy is about ten percent lower compared to the best possible box, but, as previously

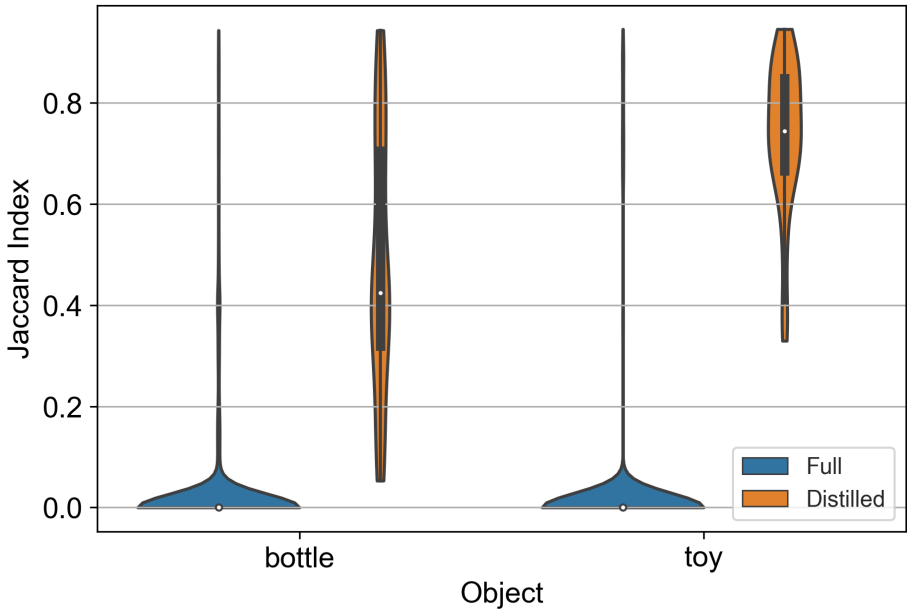


Figure B.1.6: The violin plot shows the distribution of the Jaccard indices for the full and the distilled set of bounding boxes using the example of the bottle and the toy.

mentioned, it is found early since it is one of the first three proposals.

Considering that the sufficient boxes (Jaccard index > 0.7) are the relevant ones, we define a) *recall* as the ratio between relevant boxes retrieved by our method and all relevant boxes, as well as b) *precision* as the ratio of boxes retrieved by our method that are relevant. The per-object recall and precision are reported in Table B.1.1 together with the F_1 score $\left(2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}\right)$, which is the harmonic mean of precision and recall. Whereas the recall remained virtually the same, the precision increased significantly due to the distillation using the proposed method. To be more specific, while on average not even 2 % of the boxes in the full set could be considered as sufficient, almost 40 % of the distilled boxes have a Jaccard index of at least 0.7. Moreover, the resulting mean F_1 value is about 14 times higher for the distilled sets compared to the full set.

Finally, we would like to discuss the objects that were not recognized by FCOS. The knife and the spoon lying on top of each other was a difficult task. Both FCOS and our proposed method struggled on this part. Although FCOS was not able to detect the knife at all, we at least managed to get a sufficient box with a high position index of 75. Figure B.1.5d shows our best possible result. However, we needed several attempts to match the human gaze point with the knife since the knife's width is relatively small.

Detecting the dark blue scissor on the black keyboard was on the other hand quite

easy in terms of mapping the gaze point on the object. Even though it was generally an even harder task with respect to the color of the background, unlike FCOS, the proposed method was able to find one sufficient and one best bounding box with the position index of 45 and 99, respectively. These bounding boxes were also relatively late to reach but far earlier than 635 and 1251, their indexes in the full set.

Limitations

Our method is based on a high accuracy in each partial step. This is an issue if we have either bad gaze tracking or mapping, which could result by too small or too few markers, low-quality hardware, or external disturbances. Even if the mapping part is accurate, a sloppy gaze estimation can lead to a gaze point that does not overlap with the object. With an inaccurate gaze point in the robot's view, accurate bounding box proposals are difficult and sometime impossible to distill. In this case, we have to repeat pointing out to the object.

Furthermore, with one exception (scissors on the keyboard), the proof of concept was carried out on a plain white table. Although we would expect more candidate boxes in less homogeneous settings, our experiments suggest that there would still be highly relevant boxes due to the high recall that is in the nature of the method.

B.1.7 Conclusion

In this work, we have proposed and evaluated a novel method that enables the deployment of robots in non-predefined scenarios. The proposed method combines automatic object location proposals with human gaze to distill pertinent location proposals. Just by looking at an object and some human robot communication, we can find a bounding box with a Jaccard index of almost 0.9 compared to the ground truth. These boxes can then be used to quickly extend the robot's object detection neural network.

Out of thousands of possible region proposals, we successfully distilled useful object-independent bounding boxes, increasing the precision of the location proposals by over 21 times with virtually no recall loss. Despite challenging scenarios, our method was consistently applicable and does not need any training at all. Relative to a state-of-the-art object detector (FCOS) trained on the Microsoft COCO data set, we achieved an average Jaccard index of almost 0.9 for at least one box out of the first 15 proposals. Looking only at the first sufficient box of each object, we observed an average accuracy of 89.25 % compared to the respective best possible box in the full set of proposals.

Since our gaze method significantly improved the position index of important bounding boxes compared to the large initial number of region proposals, it enables concentrat-

ing exclusively on relevant boxes. This allows the robot to find the intended object more quickly and to generally reduce the necessary communication, improving the human-robot interaction. In addition, we could find bounding boxes to objects that could not even been detected by FCOS.

In summary, our proposed method is therefore a broadly applicable and natural way to achieve unknown-object detection by a robot in HRI scenarios. However, a significant amount of work remains for future work as we plan to extend our proof of concept by evaluating our method with more participants and additionally investigate the impact of imperfect labels on the training of neural networks.

Acknowledgment

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

B.2 Exploiting Augmented Reality for Extrinsic Robot Calibration and Eye-based Human-Robot Collaboration

B.2.1 Abstract

For sensible human-robot interaction, it is crucial for the robot to have an awareness of its physical surroundings. In practical applications, however, the environment is manifold and possible objects for interaction are innumerable. Due to this fact, the use of robots in variable situations surrounded by unknown interaction entities is challenging and the inclusion of pre-trained object-detection neural networks not always feasible. In this work, we propose deploying augmented reality and eye tracking to flexibilize robots in non-predefined scenarios. To this end, we present and evaluate a method for extrinsic calibration of robot sensors, specifically a camera in our case, that is both fast and user-friendly, achieving competitive accuracy compared to classical approaches. By incorporating human gaze into the robot's segmentation process, we enable the 3D detection and localization of unknown objects without any training. Such an approach can facilitate interaction with objects for which training data is not available. At the same time, a visualization of the resulting 3D bounding boxes in the human's augmented reality leads to exceedingly direct feedback, providing insight into the robot's state of knowledge. Our approach thus opens the door to additional interaction possibilities, such as the subsequent initialization of actions like grasping.

B.2.2 Introduction

More and more robots are being used in environments within a close proximity to humans. The possible applications of robots are diverse and possible interactions with humans are multifaceted. Whether as a tour guide in museums [305] or as an assistant in supermarkets [306], each interaction scenario involving robots has its own challenges. Furthermore, successful technical advances in augmented reality (AR) have promoted the interaction and collaboration between humans and robots. Consequently, AR has found application in factories [267] and in imitating assembly processes that a human demonstrates [307].

The long list of possible use cases results in at least as many tasks that need to be solved. Among these tasks, the conveyance of the interaction context, such as the specification of an object to interact with, is particularly challenging. Many tasks, especially object

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

detection, can be accomplished through the benefit of machine learning methods, such as neural networks. While advances in machine learning have had a major impact on the development of human-robot interaction, there are also some drawbacks. Typically, many of these approaches require a sufficient amount of available training data, which cannot always be guaranteed. This data dependency ties the deployment of robots to predefined scenarios and limits interaction with the environment, e.g. with unknown objects that cannot be detected. For example, if a supermarket changes its assortment of products, the robot can usually only interact with the new items if it has learned them beforehand. Our goal is to enable data-independent object detection for cases where no training data is available.

Another even more fundamental problem is the calibration of the robot. In order for a robot to perceive a scene, its sensors, such as a fixed, but adjustable camera, must be properly targeted and its position relative to the robot base must be known. Therefore, the scene or the purpose of the operation needs to be identified in advance, at least to a certain degree. In addition, calibration of extrinsic robot parameters is often laborious [308] since, in most cases, either the existence of a second sensor in the form of a laser scanner or another camera is assumed, or expensive external tools are used. Both make subsequent adjustments in response to changing circumstances difficult. On top of that, the authors of [309] noted that robots in public attract the curiosity of people, especially children. In particular, children tend to touch the robot or exhibit abusive behavior when unobserved. This, in turn, can often lead to misalignments of the robot's sensors and require frequent recalibrations. A less time-consuming calibration method is beneficial in this case.

In this work, we attempt to fill this gap at the intersection of research fields of human-robot interaction, eye tracking, and augmented reality. More specifically, we aim at a flexible deployment of robots, detached from predefined scenarios by leveraging collaboration with humans instead of training data. Our contribution with this work is twofold:

On the one hand, we present a convenient method for determining the transformations between the robot and a sensor, in our case a camera, as well as between the human and the robot. With our method, time does not have to be spent repeatedly for each calibration run, but only once during the initial setup. Subsequent calibrations can then be performed in a matter of seconds, making the method particularly suitable for situations where frequent recalibrations are required. The calibration can be executed at any time during runtime and allows both the human and the robot to move freely.

On the other hand, after utilizing said calibration, we fuse existing point cloud clustering methods with eye-tracking information to showcase the 3D detection of unknown

objects. More precisely, the robot and the human collaborate so that the robot detects which object the human is looking at without knowing the interaction context in advance. Based on our calibration, we can establish a connection for continuous exchange of interaction information. The human continually provides the robot with gaze data and the robot responds with the bounding box of the target object. The human's perception is augmented by integrating the robot's feedback directly into the human's reality. All of this without training and in an online fashion, not after the fact.

In summary, our most important contributions are as follows:

1. We show and evaluate a calibration method via an augmented reality interface that is suitable for the deployment of robots in ever-changing scenarios and allows the robot's capabilities to be further extended by providing it with a new, additional real-time information channel — the human gaze.

2. We are the first to use augmented reality in a human-robot collaboration scenario to segment unknown objects in three-dimensional space without the use of neural networks. We also provide direct feedback to the human, enabling subsequent interactions.

The remaining part of this paper is structured as follows. After a discussion of the related work, in Section B.2.4 we describe and formalize our approach in detail. Our results and the limitations of our approach are discussed in Section B.2.5. Section B.2.6 concludes this work and gives an outlook on our future activities.

B.2.3 Related Work

Employing gaze information to achieve human-robot interaction with unknown objects requires significant multidisciplinary efforts, which we will discuss in this section. From how 1) robots collaborate with humans, to 2) augmented reality in robotics, 3) robot calibration and 4) 3D object detection, to 5) mapping human gaze to a known frame of reference and 6) previous applications of eye tracking in the context of computer vision.

Collaborative Settings

In recent years, scenarios in which humans and robots work together side by side have gained attention. Interaction with robots invites interesting possibilities for beneficial collaboration in human everyday life [310]. In [311] a system was presented, that enables a robot to perform cooperative search with a human teammate, where the robot assists the human teammate in navigation to the search target. Collaboration between human

and robot is also widespread in industrial environments, such as in assembly tasks [312], surface finishing applications [313] or welding work [314]. In addition to the application in industry, robots have more and more of a social purpose. Due to the lack of medical personnel and rising costs in the health sector, social robots are increasingly being used in the health care system [315]. They are typically deployed for surgical assistance [316], rehabilitation [37], elderly care [317], and as companion robots [318].

Augmented Reality in Robotics

With the increasing availability of various augmented reality glasses, the impact of AR in research and industry has also grown. In [319], head orientation and pointing gestures were used to control an industrial robot arm for pick-and-place tasks. However, the arm was fixed in the room to facilitate coordinated transformation by means of a marker attached to the wall and the set of interaction objects was fixed. An AR device was also used by [320] in a multimodal communication setup to help a robot decide which object a human pointed to using gestures, gaze, and speech. In this setup, again, the objects were predefined and their positions were additionally measured accordingly in advance. The authors of [321] visualized sensor data from a robot using AR glasses. All sensors, though, were already calibrated, which additionally allowed for the utilization of a localization algorithm. Following on from this, the same authors recently used a deep learning-based approach in [322] to determine the mutual position of the robot and AR device. Nevertheless, this approach was not suitable for real time scenarios due to the limited computational capacity of the AR glasses. Within a manipulation frame, [323] used pre-trained 2D object detectors to determine 3D bounding boxes. This required a fiducial marker to be in the field of view at all times and was limited to a single object per pass. Such problems of ambiguity we will solve with gaze.

Extrinsic Robot Calibration

Modern robots are usually equipped with a large number of sensors, most frequently RGB-D cameras. Ensuring their operability requires the most accurate calibration of extrinsic parameters, i.e. their position on the robot base. A classical approach to this is the use of calibration patterns. By observing the pattern, [324] determined the mutual position between a camera and a 2D laser range finder. With only one image, but several markers, [325] succeeded in calibrating a camera with respect to a second camera or a laser scanner. In both cases, however, the existence of a second sensor was a prerequisite and a common field of view of these two was mandatory.

In [326], a framework for parameter estimation using a motion capture system was

built. While such systems, including Vicon [327] or OptiTrack [328] can be very accurate, they require careful calibration beforehand. In addition, they are time-consuming to set up and expensive due to the amount of hardware involved, such as multiple cameras. We try to close this gap with a fast and universally applicable method.

3D Object Detection

Due to the higher level of difficulty, many 3D object detectors are inspired by detection in 2D. This includes the projection of the point cloud into bird's eye view [329] or cropping on frustums based on 2D bounding boxes [197], [196]. Few also operate on the point clouds directly [195]. What they all have in common, however, is that they rely heavily on the availability of training data and focus predominantly on road scenes or furniture pieces. An approach to instance segmentation of unseen objects was proposed by [330]. While they did not need real world images, they had to generate a large amount of synthetic data for which 3D CAD models were required. As an alternative to neural networks, [219] used a saliency-driven approach to detect unknown objects. Nonetheless, the results were influenced to some extent by a parameter that depended on the size of the objects, and, due to the long calculation time, the system was not suitable for real-time applications.

Gaze Mapping

Mapping gaze data from a moving eye tracker to another coordinated frame is still an unsolved challenge and thus ongoing research [274]. One possible solution to this challenge is feature matching. For example, [275] achieved promising results with such a method, however it reaches its limit with diverging camera perspectives. The authors also found that better robustness at less computational cost was achieved with fiducial markers [275]. Such markers were used in recent works by [274] and [1], among others. One disadvantage of this approach is that fiducial markers have to be in the field of view of both cameras, restricting thus movements. With our AR-based approach, we overcome this problem and ensure stable gaze mapping despite free movement and thus independent of the field of view.

Eye Tracking and Computer Vision

Although not yet very popular, there are some works that have tried to solve computer vision problems with eye tracking. In [284] for example, features in the neighborhood of human fixations were matched to features of known objects to determine the class of the respective object. Statements about its position could not be made in this way. The

authors of [215] reduced the number of superpixels for salient object detection with gaze data. In contrast to our approach, however, this required both multiple gaze points and training data. With only one gaze point, [1] managed to drastically reduce the number of candidate bounding boxes of a region proposal method, but this method is only applicable in 2D.

In this work, we build on existing research to improve human-robot interaction. While speech and gesture are popular channels for communication, gaze is challenging [331] and often neglected. In the following, we link eye tracking and augmented reality to address classical calibration problems as well as data-independent 3D object detection in a collaborative manner.

B.2.4 Methods

In this work, we propose finding 3D positions of unknown objects by incorporating human gaze into the robot's segmentation process. For this purpose, we first introduce the interface used to communicate with the robot. Subsequently, we present an extrinsic robot calibration method, which is particularly characterized by its flexibility and ease of execution. In our case, we calibrate a camera's position relative to the robot's base, but in principle the method can be applied to any sensors. Finally, we explain the segmentation process that applies said methods.

Augmented Reality Interface

All interaction with the robot is guided via an augmented reality interface and serves as a two-way communication channel between human and robot. In this way, we can, for example, control the movement of the robot, access the robot's camera feed, or perform the extrinsic calibration between robot and its camera. In addition, we can provide the robot with the human gaze data and display the results of the object detection. We use the HoloLens 2 from Microsoft, a head-mounted pair of mixed reality glasses with a built-in eye tracker. For the development of AR applications, Microsoft provides an open-source cross-platform toolkit called Mixed Reality Toolkit (MRTK). The creation and development of our interface takes place in the game development environment Unity. We use the versions MRTK 2.7.2 and Unity 2019.4.29. For the actual communication between the HoloLens' Universal Windows Platform (UWP) and the robot operating system (ROS), we resort to the UWP version of ROS# [184], a set of open source software libraries and tools for communicating with ROS from Unity applications. On system startup, the robot launches ROS#'s file_server package as well as rosbridge_server from the rosbridge_suite. As soon as the AR interface is started on the HoloLens, it immediately establishes a

connection with the robot via Wi-Fi. Thereupon, ROS# uses the rosbridge protocol to send JSON based commands via WebSockets, enabling the deployment of custom publishers and subscribers. During runtime, the menu of our interface can be opened by looking at the user's palm. Created virtual objects can then be selected by voice or gestures. For example, menu buttons can be simply pressed with a finger or other virtual objects can be selected by looking at them and pinching the thumb and index finger together or saying "select".

Calibration & Gaze Estimation

The incorporation of the human gaze into the robot's world requires the estimated gaze to be mapped from the reference frame of the human, provided by the HoloLens, into the robot's frame of reference. For this purpose, the transformation can be computed either directly, if the pose of one device in the frame of the other is known, or through indirect co-location by finding corresponding points in the image of the two associated cameras [1]. The former is often difficult to realize in practice, while the latter has some disadvantages, namely limiting the view of both participants to an overlapping field of view. Furthermore, for the robot to interact with objects in its field of view, the position and orientation of the robot's camera relative to its base must also be known. The solution comes in the form of augmented reality, which we can employ as a bridge. If we create virtual counterparts corresponding to the real poses of the respective frames, we become acquainted with the transformation between frames through the transformation between virtual elements. More precisely, we determine the mutual position of the robot and the robot's camera sensor by aligning them with the corresponding virtual objects and calculating the transformation occurring in between in the virtual space of the HoloLens. The authors of [307], [311] and [321] did something similar to align the coordinate systems of a robot and that of a HoloLens. However, in their case, all the necessary robot sensors had been calibrated beforehand. The advantage of our calibration method is that it splits the usual time-consuming extrinsic sensor calibration into multiple parts. In case of frequent calibrations, only the fast part needs to be repeated.

For us, the approach described means we can intertwine both of our problems: On the one hand, we can calibrate the position of the robot and the camera in relation to each other, and, on the other hand, we can establish a direct transformation between HoloLens and the robot, which means that the robot is aware of the gaze point at all times regardless of the field of view. An overview of the underlying pipeline is shown in Figure B.2.1.

We start by determining the poses of the two frames of interest. This is, in our case, the so called base_link on the robot side and the camera_base frame on the camera side. In

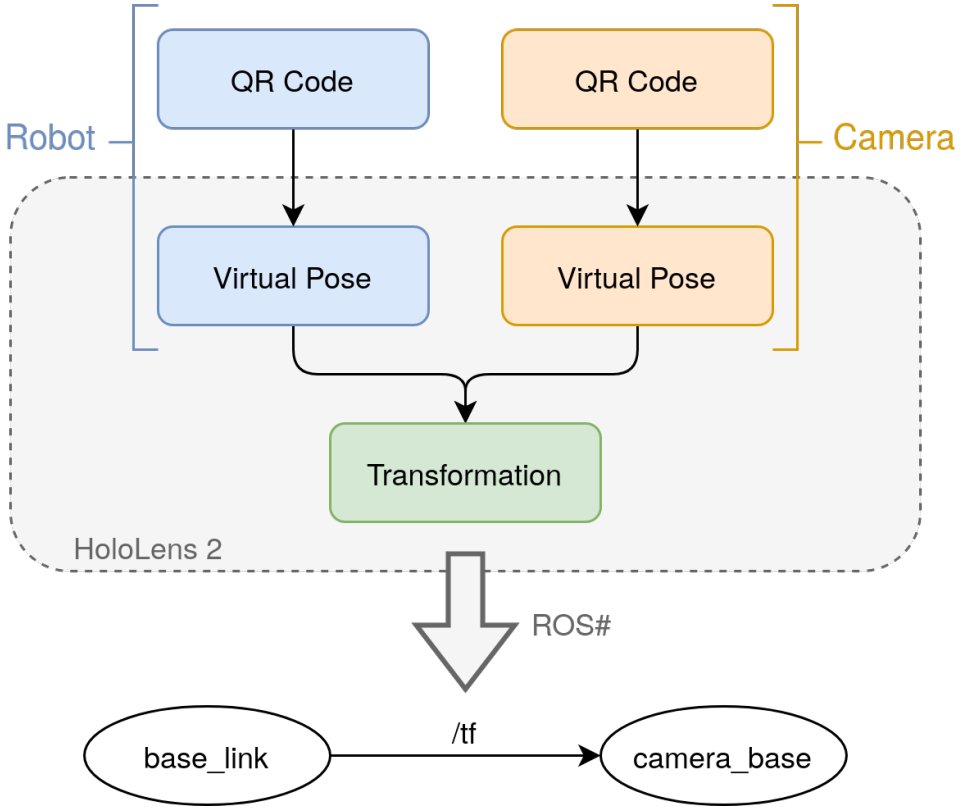


Figure B.2.1: The QR codes specify the position of the virtual versions of the robot and the camera. The intervening transformation can be determined in the virtual world of the HoloLens 2 and is then published via ROS#.

principle, however, any frame can be used whose origin is known relative to a point on the housing. To align real and virtual versions of the robot and its camera, we attach fiducial markers in the form of QR codes (see Figure B.2.2) as they allow for robust and inexpensive detection. The HoloLens 2 is moreover capable of detecting QR codes at the system level in the driver. However, we have to consider that there will be an offset between the pose of the markers and the actual frame. So let $\{b\}$, $\{m_b\}$, $\{c\}$ and $\{m_c\}$ be the coordinate frames of the robot's base (base_link), the QR code on the base, the camera (camera_base), and the marker on the camera, respectively. For two frames $f_1, f_2 \in \{\{b\},\{m_b\},\{c\},\{m_c\}\}$, let the transformation from f_1 to f_2 be denoted by ${}^{f_1}T_{f2} \in \mathrm{SE}(3)$. The connection between the frames can be illustrated by the following transformation graph:

$$\{m_b\} \xrightarrow{{}^{m_b}T_b} \{b\} \dashrightarrow^{{}^{b}T_c} \{c\} \xleftarrow{{}^{m_c}T_c} \{m_c\}.$$

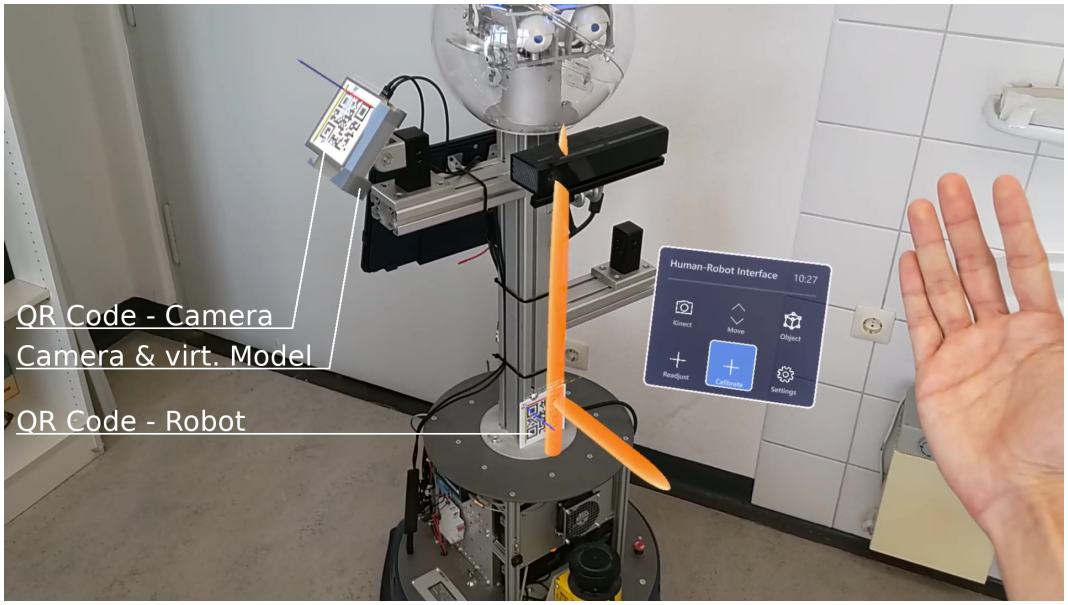


Figure B.2.2: The AR interface appears when looking at the open palm. The QR code on the camera positions the virtual camera model and the QR code on the robot's torso defines the robot's forward direction and center of rotation (orange).

The QR codes on the robot and camera can usually be attached to their housing so that they are either parallel or perpendicular to it. Thus, their orientations and, hence, the rotations to the corresponding frames are known. The same applies to the translation between $\{m_b\}$ and $\{b\}$, since the marker can be placed on the robot according to existing knowledge about other robot frames. If, contrary to expectations, this is not possible, we also managed to approximately estimate the center of rotation of the robot, i.e. where the base_link frame $\{b\}$ is located, as the geometric center of the virtual circle drawn by the marker on the camera as the robot rotates around its own axis. The translation from $\{m_c\}$ to $\{c\}$ can be determined with the help of manufacturer information about the dimensions of the camera. This means ${}^{m_b}T_b$ and ${}^{m_c}T_c$ are known.

We want to determine the transformation ${}^{b}T_c$. The idea is to add a frame $\{h\}$ corresponding to the coordinate system of the HoloLens to close the transformation graph:

$$\begin{array}{ccccccc} & & & {}^{h}T_{m_b} & & {}^{h}T_{m_c} & \\ & & & \{h\} & & & \\ & \swarrow & & & & \searrow & \\ \{m_b\} & \xrightarrow{{}^{m_b}T_b} & \{b\} & \dashrightarrow^{{}^{b}T_c} & \{c\} & \xleftarrow{{}^{m_c}T_c} & \{m_c\} \end{array}$$

After the two QR codes have been detected by the HoloLens, they can be selected via our AR interface and ${}^{h}T_{m_b}$ and ${}^{h}T_{m_c}$ can be estimated. Finally, the transformation ${}^{b}T_{c}$ from the base of the robot to the camera is given by the following equation:

$${}^{b}T_{c} = {}^{m_b}T_{b}^{-1}\,{}^{h}T_{m_b}^{-1}\,{}^{h}T_{m_c}\,{}^{m_c}T_{c}.$$

The result can be published from the HoloLens using ROS# to the transformation topic /tf, making it available to the robot.

Furthermore, we can use $\{h\}$ as a parent frame in which the robot's odometry frame is embedded. This gives us a reference point for the gaze information that we can access via MRTK. Associated with $\{h\}$, we can publish this data on a separate topic. This includes the gaze vector and the hit point of the eye gaze ray with the target.

It should be noted that the fiducial markers are only needed while performing the calibration. Once they have been detected and selected, the user is free from restrictions on the field of view. In addition, contrary to the usual procedure, we do not determine the calibration parameters externally and then store them in configuration files. This means that we can make changes to the camera, such as the tilt, even during runtime. This is a great advantage for use under changing scenarios.

Segmentation

We now address the problem of detecting unknown objects in the three-dimensional environment. We tackle this task by enhancing existing segmentation methods on the robot side with gaze information from the human collaborator. The segmentation process can be triggered either on demand by multimodal interaction, such as gestures or speech, or – empowered by the calibration method – continuously in real time. The assistance that the robot receives from the human should be limited solely to the provision of the gaze information. Apart from that, the segmentation should only take place on the robot's side. This makes sense due to the robot's higher resources and computing power compared to head-mounted devices like the HoloLens.

The segmentation process starts with a pass through filter where we assume that all relevant objects are between zero and three meters away from the camera, followed by a voxel grid filter with a leaf size of 0.03 along each axis that downsamples the point cloud we acquire from the robot's camera. This is not mandatory, but it reduces the computation time drastically and allows a segmentation in real time. In most cases, we can assume that the objects to be detected lie on a surface that is reasonably flat. This could be, for example, a table, a shelf, or the floor itself. We can take advantage of the parallelism between all these surfaces. Due to our calibration, we know the orientation of the camera

with respect to the robot standing on the floor. This means that we can transform the upward vector from the HoloLens world frame into the camera frame and thus obtain the normal vector of the surface, that is parallel to the floor and on which the objects are located, in the frame of the camera. We can then use RANSAC to search for the largest plane in the robot's field of view, namely the said surface, that is perpendicular to the given normal vector. Thereby, we set the maximum allowed deviation from the normal vector to 30 degrees. All points belonging to this plane are finally removed from the point cloud. In the next step, we let the gaze information flow in. Since the human is looking at the object of interest, we know at least one point on its surface. Starting from this point, we can cluster the point cloud using simple euclidean clustering. That is, we first use a k-d tree to find the point in the point cloud that is closest to the gaze point. Then we cluster the point cloud with respect to the Euclidean distance, a tolerance of 5 mm, and a minimum cluster size of 500. All points that belong to the same cluster as the nearest neighbor of the initial point result in the searched object. Note that without the gaze information we would not be able to distinguish between clusters belonging to objects, clusters of parts of the environment, or noise. *This subtle gaze interaction resolves ambiguities and brings us closer to a natural learning process.*

Finally, we do not only obtain an instance segmentation of an object, but we can also calculate a 3D bounding box from it. The box can be aligned properly in space again due to our calibration and the robot can share the result directly with the human via our AR interface. Thus, the bounding box can be displayed in the human's field of view, providing direct feedback and enabling a natural two-way communication component, as well as an initialization of further interactions of the robot with the object.

B.2.5 Evaluation

In our experiments, a Scitos G5 from MetraLabs [298] was employed as a robotic counterpart. It has been equipped with an Azure Kinect DK, whose relative position to the robot we want to calibrate. The camera also provides image data such as the point cloud on which we perform the object detection. All components communicate with each other using ROS [183].

Qualitative Analysis

One of the advantages of our method is already evident when performing a single calibration run. Whereas calibration methods based on data collection are time-consuming and difficult to automate [308], the entire procedure with our variant takes less than a minute. Depending on the user's experience, a single run usually takes only between 15

and 40 seconds. This is especially apparent when the camera needs to be adjusted more frequently, either because it has been unintentionally moved or because the setting has changed.

After calibration, the whole system, including gaze mapping and the object segmentation, runs in real time. Figure B.2.3 shows a visualization in RVIZ. The gaze ray vector as well as the coordinate of the hit point on the target are published with 59 Hz. Using the default configuration, the Azure Kinect provides the point cloud at 4 frames per second. Subsequently, segmentation reduces the rate of the outgoing segmented cloud and thus also that of the bounding box to 2 frames per second. Since the minimal fixation duration is, in most cases, at least 200 ms [332] and the recommended feedback delay time for manual pointing actions is approximately between 350 ms and 600 ms [333], an update every 0.5 seconds is sufficient. Consequently, our method is suitable for human-robot interaction in real time.

Some final example results of segmented objects and the respective bounding box can

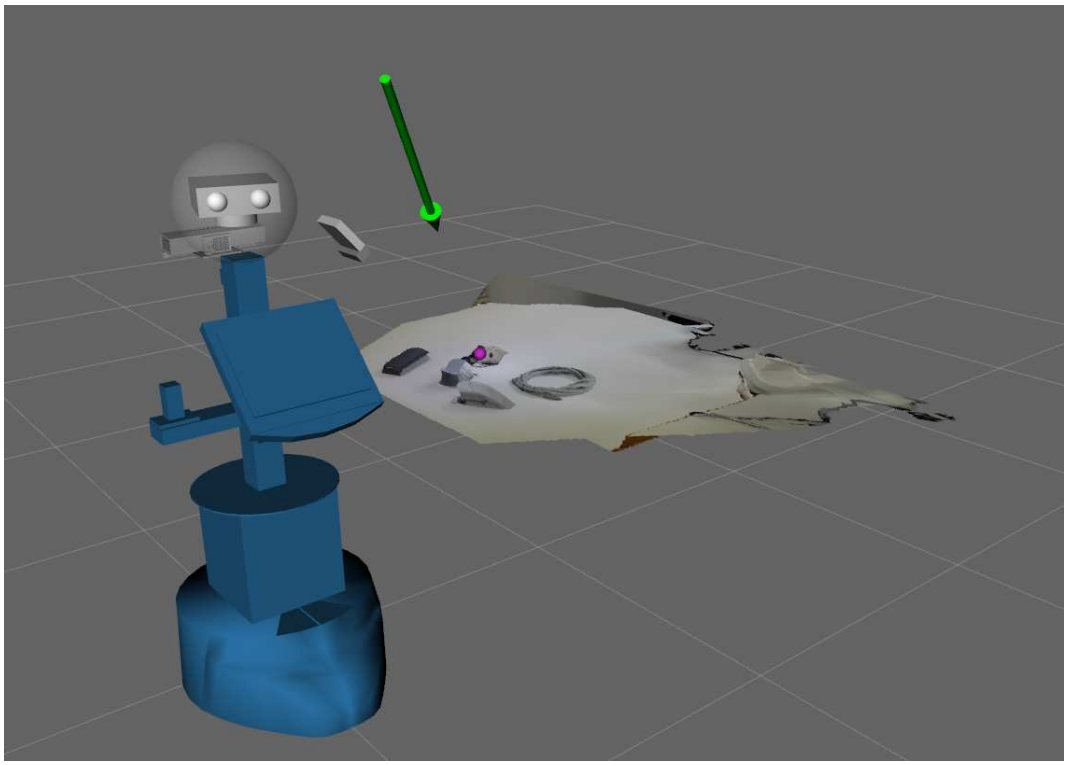


Figure B.2.3: The robot model with the camera positioned relative to it. The human gaze vector is shown as a green arrow and the gaze hit point as a purple sphere.

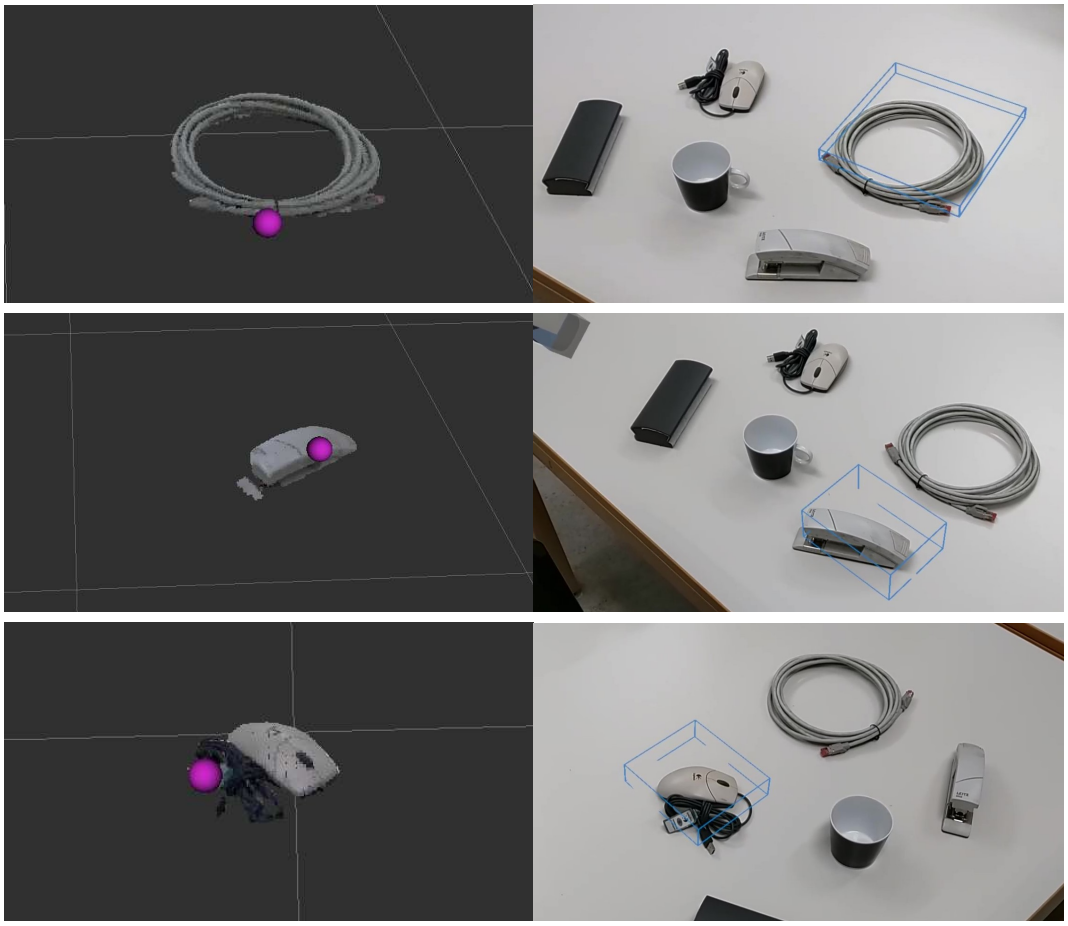


Figure B.2.4: The segmentation with the gaze point (left) and the resulting bounding box as seen from the human (right). The box is given in world coordinates, therefore tracking of already detected objects during movements of the robot is superfluous.

be seen in Figure B.2.4. For simplicity, we have chosen common household objects and office utensils, which we have placed on a table in front of the robot. In principle, both humans and robots can move freely around the table, since the position of both is known in the HoloLens based parent frame. However, to ensure that the robot's movements are tracked as precisely as possible, an additional localization procedure is required, which is beyond the scope of this work, as solving such a problem has already been extensively researched, and possible solutions can be found in the literature [334], [335]. Naturally, the current position can be manually repositioned at any time via our interface.

Quantitative Analysis

First we start with the evaluation of the calibration part. To establish a reference ground truth, we utilize the OptiTrack motion capture system [328]. We place multiple reflective markers on both the robot and the camera. These can be tracked by the Optitrack system with an accuracy of 1 mm. Given these point observations, we can calculate the robot and the camera poses with respect to the coordinate system of the motion capture system, and then compute the camera pose of interest relative to the coordinate system of the robot. Based on the deviations we have observed in several test trials, we estimate that this post-processing decreases the accuracy to about 3 mm. In this way, we determine the ground truth of the transformations from the robot frame to the camera frame for three different poses of the camera. Once horizontally, i.e. parallel to the floor, once vertically, i.e. perpendicular to the floor, and once in an inclined position at about 45 degrees. Without moving the camera in between, one of the system's designers performed the calibration 20 times per tilt using the method we presented in Section B.2.4. For each tilt, we evaluate the translation and rotation components separately.

Table B.2.1 shows the result of the translation part of our AR-based calibration compared to the calibration using OptiTrack. In the table, the translation in each direction is given with respect to the ROS coordinate system. The difference between the result of the OptiTrack system and the mean result from our calibration varies, but is not noticeably pronounced with respect to any direction. The largest difference is observed with 3 mm in the direction of the y-axis in the case of horizontal orientation. All other values do not differ at all or only 1 to 2 mm. Our analyses have shown that the same is true for the average of all individual differences to the ground truth.

Table B.2.1: The translation in meters determined by the calibration with OptiTrack as well as our AR interface.

	Vertical		Horizontal		Inclined	
Axis	OptiTrack	mARC[a]	OptiTrack	mARC	OptiTrack	mARC
x[b]	-0.081	-0.080	-0.079	-0.079	-0.077	-0.079
y	-0.295	-0.295	-0.324	-0.327	-0.331	-0.332
z	0.973	0.973	1.072	1.071	1.033	1.035
ØDist.[c]	0.003		0.004		0.003	

[a] Average value of our AR-based calibration
[b] Coordinate axes refer to the ROS coordinate system
[c] Average spatial distance of all runs calculated with the euclidean norm

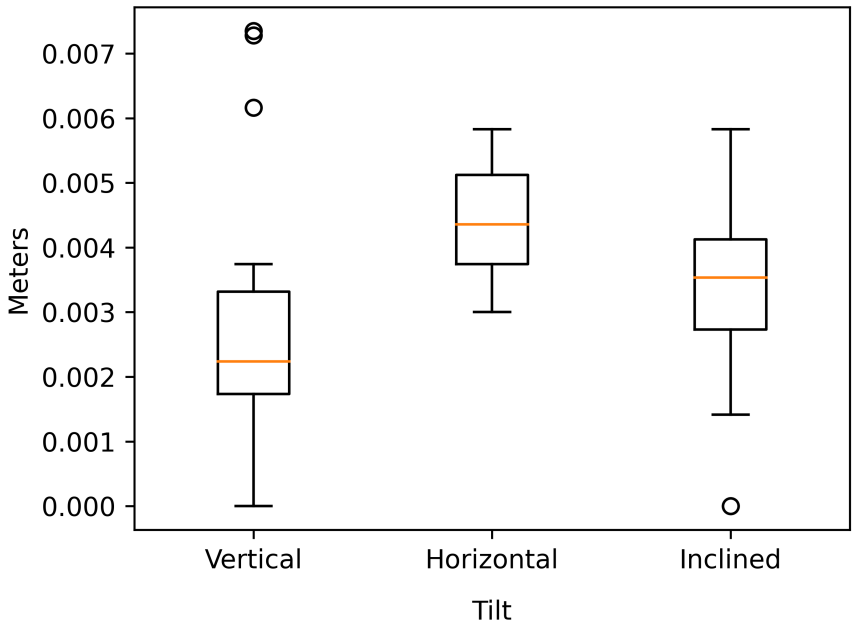


Figure B.2.5: The box plot represents the distribution of the translation errors with respect to the euclidean norm.

Since the deviation in individual directions is less relevant than the spatial distance, we also want to take this into account. We measured the euclidean distance of the translation of each individual calibration run from the ground truth translation. The results are reported in the last row of Table B.2.1. One can see that the spatial error does not exceed 4 mm on average. This is comparable to the accuracy of the extrinsic calibrations evaluated in [324] and [336]. The distribution of the individual euclidean distances to the ground truth are shown in the box plot in Figure B.2.5. Although the error in the vertical setting is generally the smallest, there were also some outliers. Basically, in all three scenarios the vast majority of errors were below 5 mm. The medians lie between 2 mm and 4.5 mm. Note that this is only slightly above the accuracy range of the reference ground truth estimated via OptiTrack.

We now examine the rotational error of the transformation. In general, each rotation can be expressed by an axis of rotation and an angle of rotation. This rotation angle can be considered a measurement of the similarity of two orientations. This means that for each rotation component determined by our calibration, we calculate the difference rotation, which transforms the obtained rotation into the ground truth rotation. The smaller the

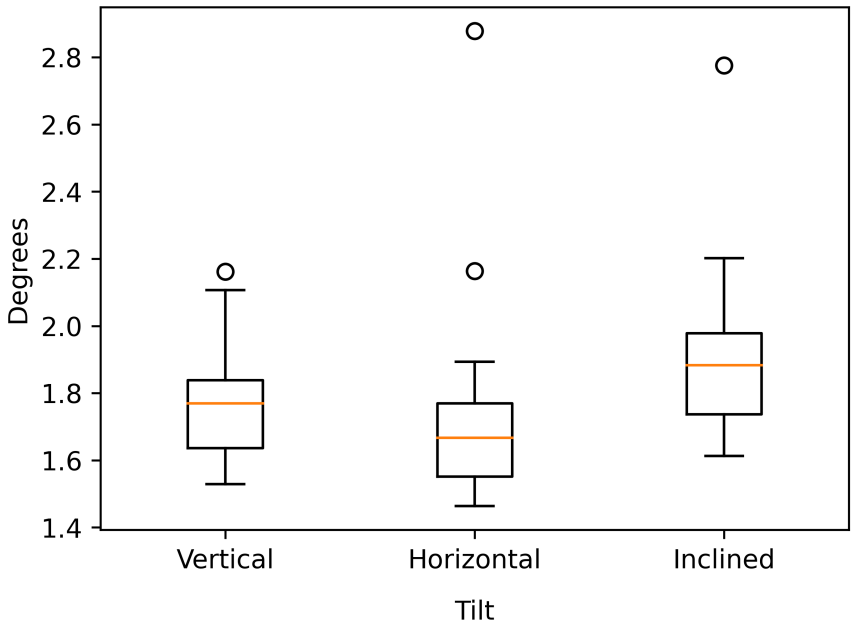


Figure B.2.6: The rotation errors displayed in a box blot.

angle of rotation, the more similar the two rotations. The angles of all difference rotations are plotted in Figure B.2.6. Although there are, again, a few outliers, most values are below 2 degrees with medians ranging from 1.6 to 1.9 degrees. The same applies to the average rotation error. Thus, the rotation error is of the same order of magnitude as that of classical approaches like [324]. All in all, the accuracy meets the requirements of most applications, including our gaze segmentation, while being flexible and fast.

Let us now have a closer look at the evaluation of the segmentation part. Although the performance of current 3D object detectors lags behind the state of the art in 2D object detection, there are 3D object detectors that promise good results on indoor datasets such as SUN RGB-D [198]. However, our experiments have shown that the claimed results are difficult to achieve in practical applications. One possible reason could be that, due to comparability, the evaluations are usually conducted on the same few categories [197], [196], [195]. As a result, performance on other classes is often significantly worse or remains unknown.

We trained several neural networks, such as VoteNet [195] and Frustum ConvNet [196] on the classes book, bottle, bowl, cup, keyboard, laptop, mouse, paper, plant, and telephone from the Sun RGB-D dataset. These objects were more appropriate for our

Table B.2.2: The IoU between the bounding boxes obtained by our method and the respective ground truth.

Class name	apple	backpack	book	bowl	clock	cup	keyboard	mouse	remote	tennis ball	mIoU
2D IoU	0.78	0.78	0.79	0.86	0.68	0.84	0.88	0.79	0.80	0.87	0.81
3D IoU	0.70	0.66	0.72	0.84	0.62	0.73	0.66	0.59	0.64	0.71	0.69

setup although smaller than the furniture used in the original papers. It turned out that all state-of-the-art networks performed very poorly on our set of objects and could not serve as reasonable reference ground truth. To put this in numbers: Whereas the mean average precision with a 3D Intersection over Union (IoU) threshold of 0.25 was only 27.8 % for Frustum ConvNet, this value was even less than 1 % for VoteNet. Thus, almost none of the available test objects were successfully detected by the neural networks and a meaningful comparison was therefore not possible. For this reason, we devised an alternative evaluation strategy and eventually conducted two different approaches. In the first one, we labeled the 3D bounding boxes of the objects in the acquired point cloud of the scene by hand and calculated the 3D IoU (with regard to the volume) for ten test objects. In the second one, we used a pretrained 2D object detector to avoid vulnerability regarding a bias in labeling. While modern 3D detectors are still far from being able to serve as ground truth, 2D detectors certainly are capable of doing so. Hence, we projected the points segmented by our method onto the 2D image plane and compared the resulting 2D bounding box with Faster R-CNN [186] (ResNet-101 backbone) trained on Microsoft COCO [185]. This dataset was also the criterion by which the ten test objects were selected. The results of both evaluations are shown in Table B.2.2. In the 2D case, all values are above 0.5 and thus all objects can be considered correctly detected [192], [193]. Furthermore, almost all values are even above 0.7 with a mean IoU of 0.81. In contrast, the 3D IoU values are naturally smaller. Nevertheless, all objects are again considered to be detected, using the usual 3D threshold of 0.25 as reference [197], [198]. The mean 3D IoU is 0.69. Figure B.2.7 shows the recall as a function of the IoU threshold at which a bounding box is classified as true positive. Note that even with a 3D IoU threshold of 0.5, which is twice as large as that used by the authors of VoteNet and others [196], [198], the recall is still 100 %.

Overall, our method hints at going far beyond the practical applicability of state-of-the-art neural network-based 3D object detectors, illustrating the importance of diverse solution strategies along with neural networks.

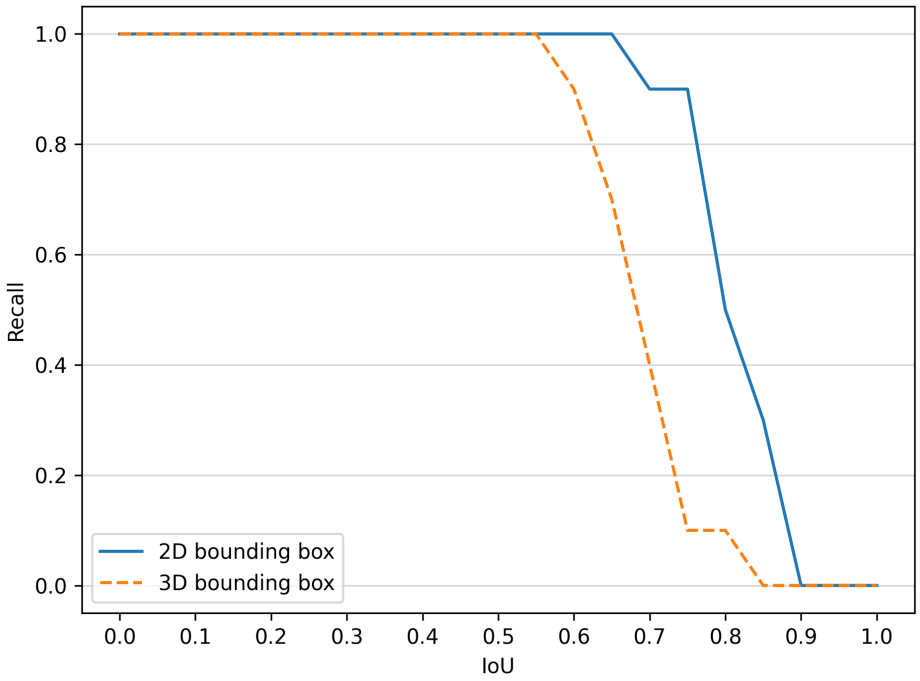


Figure B.2.7: The recall as a function of the IoU threshold at which the objects are considered to be detected.

Limitations

Although our method of calibration is remarkably fast and user-friendly, the initial setup takes some time. While less in-depth expert knowledge is required compared to other methods, care must be taken to ensure that the markers are positioned accurately and that the distances to the corresponding frames can be determined. However, since this is a one-time step, this time expenditure is not of any significance compared to the time saved in each subsequent calibration run.

Furthermore, as with any other existing method, our segmentation and the calculated bounding box strongly depend on the quality of the original point cloud provided by the depth sensor. In this regard, the perspective of the robot's camera on the object also plays a role and whether the depth sensor can correctly determine the distance at the edges of the objects. However, the fact that the image quality has an influence on the result is in the nature of things and could be resolved by using multiple cameras or additional angles.

For objects that are too close to each other, it is not possible to keep them apart by extracting euclidean clusters. In this case, one could, for instance, resort to a min-cut

based segmentation algorithm, also generally suitable, since a point must be given in the center of the object, which can be provided by the gaze point. In our tests, min-cut segmentation indicated promising results, but also required the approximate size of the respective object as an additional input argument.

B.2.6 Conclusion

In this work, we presented a novel method that allows for the deployment of robots under changing and non-predefined conditions. In this course, we combined robotics, augmented reality, and eye tracking to improve human-robot collaboration. Merely by receiving gaze information from its human partner, the robot was capable of detecting and segmenting unknown objects.

While most existing methods for extrinsic robot calibration are time consuming and often quite complicated to conduct, we have developed a method that is user-friendly, customizable at runtime, and takes only a few seconds to complete. At the same time, our evaluation has shown that we still achieve competitive accuracy compared to classical methods.

In addition, we bridge the two worlds of human and robot through the use of head mounted augmented reality glasses, giving the robot access to another persistent information channel — the human gaze. Just by having a human look at an object, the robot was able to segment objects it has not seen before and calculate associated three-dimensional bounding boxes. This goes beyond the capabilities of some state-of-the-art 3D object detectors and we found that our method works in situations where current existing neural networks have failed. Through direct feedback in the augmented human reality, the human is continuously informed about the results and the initialization of further interactions between the robot and the object is possible. This could be especially relevant for physically disabled people who are limited to movements in the head or neck area, in combination with a robotic arm that helps them grasp or manipulate objects.

In summary, our proposed method is versatile and facilitates general human-robot collaboration, as well as unknown object detection in the context of such scenarios in particular. However, there remains a significant amount of future work as we seek to investigate our segmentation in more challenging scenarios and realize a subsequent interaction between the robot and the objects.

B.3 Multiperspective Teaching of Unknown Objects via Shared-gaze-based Multimodal Human-Robot Interaction

B.3.1 Abstract

For successful deployment of robots in multifaceted situations, an understanding of the robot for its environment is indispensable. With advancing performance of state-of-the-art object detectors, the capability of robots to detect objects within their interaction domain is also enhancing. However, it binds the robot to a few trained classes and prevents it from adapting to unfamiliar surroundings beyond predefined scenarios. In such scenarios, humans could assist robots amidst the overwhelming number of interaction entities and impart the requisite expertise by acting as teachers. We propose a novel pipeline that effectively harnesses human gaze and augmented reality in a human-robot collaboration context to teach a robot novel objects in its surrounding environment. By intertwining gaze (to guide the robot's attention to an object of interest) with augmented reality (to convey the respective class information) we enable the robot to quickly acquire a significant amount of automatically labeled training data on its own. Training in a transfer learning fashion, we demonstrate the robot's capability to detect recently learned objects and evaluate the influence of different machine learning models and learning procedures as well as the amount of training data involved. Our multimodal approach proves to be an efficient and natural way to teach the robot novel objects based on a few instances and allows it to detect classes for which no training dataset is available. In addition, we make our dataset publicly available to the research community, which consists of RGB and depth data, intrinsic and extrinsic camera parameters, along with regions of interest.

B.3.2 Introduction

As technology progressed, more and more robots were developed for the industrial sector, and their fields of application became diversified. Numerous industries, including automotive, electronics, rubber and plastics, cosmetics, pharmaceutical, and food and beverage, benefit from their superior precision, efficiency, working capacity and tolerance to arduous and hazardous environments [337]. In the immediate environment of humans, robots are also increasingly employed in the form of service assistants, for instance in supermarkets such as Walmart [338, 339] or as tour guides in museums [305, 340]. This success has also been fueled by recent advances in machine learning, particularly in computer vision, which allows robots to understand their environment and detect objects

and people within it. However, a core assumption is almost always that a large amount of training data with high quality labels exist. Many large car manufacturers or companies such as Google, Tesla, and Uber have therefore established their own image and video databases, in most cases by outsourcing to crowdsourcing platforms such as Amazon Mechanical Turk [341]. Regarding service robots operating in warehouses or office environments, there are often no publicly accessible datasets that are tailored to the respective environment, comprise all relevant object classes, and are fully labeled. Consequently, state-of-the-art object detectors perform excellently given the existence of sufficient training data, but are limited to deployment in previously specified scenarios predefined by the training data [1]. As soon as an object is not included, it exceeds the capabilities of the object detector and pushes the robot to the limits of its possibilities. In fact, the proportion of objects covered by data sets is vanishingly small compared to the quantity of objects existing in practice. For example, ImageNet [223], one of the largest publicly available image datasets, contains just 1000 classes, while the number of classes existing in the real world obviously far exceeds this number. This fact hinders the deployment of robots in unknown environments and the dynamic adaptation to unfamiliar conditions.

In this work, we aim to make mobile robots not only more capable in terms of the tasks they have to accomplish, but also alleviate data dependency, in the sense that we extend the robot's basic knowledge by building on an existing general understanding of objects and adding new classes. More specifically, we teach a robot novel, unknown objects to enable it to redetect said objects within the environment in which the learning process took place. To this end, we employ fluent and intelligent human-robot interaction, at the intersection of research fields of computer vision, eye tracking, and augmented reality (AR). By means of the latter, we realize a multimodal communication channel, using human gaze to direct attention to an object, and speech or gestures to convey the relevant class information to the robot. Subsequently, the robot visually segments the object of interest and takes a series of images of the object from slightly different angles. The data obtained in this user-friendly and convenient process is rich in information and encompasses extrinsic and intrinsic camera parameters, as well as regions of interest, in addition to RGB and depth images. In conjunction with the class information provided by the human, the robot learns the respective object accordingly.

In summary, our main contributions are as follows:

1. We propose a novel pipeline to teach a robot new, yet unknown objects.

2. Towards this goal, we combine gaze and augmented reality in a human-robot interaction scenario to enable a feasible and swift acquisition of large amounts of labeled training data.

3. We evaluate the learning process in detail with multiple models, different learning methods, and with varying amounts of data.

4. We present Objects in Multiperspective Detail (OMD), a versatile dataset, and make it publicly available to the research community under https://cloud.cs.uni-tuebingen.de/index.php/s/2oRPs2o3FZkdBHW.

5. We make our system with our entire code base publicly available to the research community under https://github.com/dnlwbr/Multiperspective-Teaching.

B.3.3 Related Work

Engaging multimodal human-robot interaction to teach a robot unknown objects requires significant multidisciplinary efforts, which we will discuss in this section. From 1) augmented reality in robotics, and 2) previous applications of eye tracking in the context of computer vision, to 3) unknown object detection, to 4) how robots learn and 5) collaborate with humans.

Augmented Reality in Robotics

With the increasing popularity of augmented reality devices, industrial applications and research also expanded [342]. Especially the combination of AR, with robotic-assisted surgeries showed potential [343], as the human remains in control via AR, but can take advantage of the precision and consistency of the robots [344]. In terms of robot control and path planning, [319] controlled an industrial robot arm in pick-and-place tasks using an AR device and [345] designed an AR interface to plan, preview and execute the trajectory of a robot arm. In multi-agent systems, AR has also been used for visual feedback [346] and remote control [347] of robot swarms. Enhancing the perception of the real world through AR has thus proven to be an appealing way to communicate with robots.

Eye Tracking and Computer Vision

Eye tracking has also become an important tool in both research and industry [348]. For this reason, [348] developed an eye tracking software that works on mobile devices such as mobile phones or tablets and does not require any sensors other than a camera. The authors of [349], analyzed the mobile 3D eye tracking data using computer vision (and augmented reality). This involved tracking 3D markers and aligning them with virtual proxies. In [284], the class of objects was identified by matching the features of known

objects with features in the neighborhood of human fixations. However, this required the object to appear in the database previously created specifically for this purpose, and it was also not possible to make statements about the position of the objects. Nevertheless, eye tracking data can help to improve the performance of segmentation algorithms [350]. Thus, in [285], fixations were used to train a model to annotate object locations, and in [3] gaze was incorporated into the segmentation process of a point cloud. Here in turn, in both cases it was not possible to make a statement about the object classes.

Unknown Object Detection

The problem of unknown object detection was also addressed using eye tracking in [1]. In this work, the number of candidate bounding boxes of a region proposal method was significantly reduced, but a classification was not possible. Using heat maps instead of scene frames, [4] categorized video segments based on whether a person was looking at an object and determined the parameters of the associated bounding box. The detected objects were all unknown, but again the classes were not determined. The authors of [351] addressed the problem of unknown object detection using a one-class support vector machine. Since the learning process was incremental, multiple robots were involved, connected to each other via a cloud-based station where all the processing took place. In this approach, unknown objects were only identified as unknown without adding the class information to the learning process. The purpose was to filter the unknown objects from the classified objects and forward only known objects in order to avoid sending incorrect information to the robots. In another recent work, [352] proposed to exploit additional predictions of semantic segmentation models and quantify the uncertainty of the proposed segmentations. Again, the classification task was binary and only the categories known and unknown were determined without eventually learning the objects.

Teaching of Robots and Machines

Robots are employed in a wide range of applications, especially in industry. These include teaching assembly tasks [353], where the robot learns from human demonstration: First the robot observes the human, then it imitates the human's movements. A different way of learning through user interaction is proposed in [354]. In this approach, natural language is used to explain to the robot, which tasks it should perform. One such typical task is grasping objects. Solving this challenge is part of current research such as [355], where an object detection approach was used to learn good grasping poses. Data driven approaches, as stated in [356], are often addressed by providing training data in form of labeled examples, by trial-and-error, or through human demonstration. This means that

communication between human and robot is also important here for a flexible learning progress without prior offline data generation.

Joint Attention in Collaborative Settings

Teaching in collaborative scenarios between human and robot has been investigated by [357] and [358], among others. In [357], natural language context for one-shot learning of visual objects has been used to enable a robot to recognize a described object. In this proof of concept, however, the objects had to be unambiguously distinguishable by color or spatial relationships, and the component parts also had to be uniquely describable by linguistic expressions. In [358], a teaching system for object categorization was proposed. This system allowed the user to visualize the intermediate states of categorization, that is, to which category the robot would assign an object. Through interaction, the categorization could be improved and corrected, but all objects had to be marked with AR markers in order to be recognized at all. In combination with picking tasks, [359] and [360] also taught a robot new objects. In both cases, however, exactly the same objects of each class were used for both training and testing, which positively biases the results. We use several different objects per class, which is more in line with real-world conditions.

In this work, we combine the different research areas of eye tracking and AR for a simple and natural interaction between robot and human to jointly solve a computer vision problem.

B.3.4 Method

The goal of this work is to teach a robot unknown objects in its environment, in such a way that the robot is later capable of detecting these objects in this same environment. To this end, we will explain below 1) how we communicate with the robot through the modalities vision, gaze, speech, and gestures 2) how the robot identifies the object of interest, 3) how the human teaches the robot, and 4) how the robot eventually manages to learn.

Augmented Reality Interface

In order to enable the human to teach the robot anything, a communication channel is mandatory. As suggested by [3], we meet this need in the form of an augmented reality interface. The entire communication between human and robot takes place via this interface, that is, the robot can be controlled, the gaze information of the human can be transmitted and the respective class information of the objects can be conveyed. On the human side, we deploy the HoloLens 2, which is a pair of head-mounted augmented

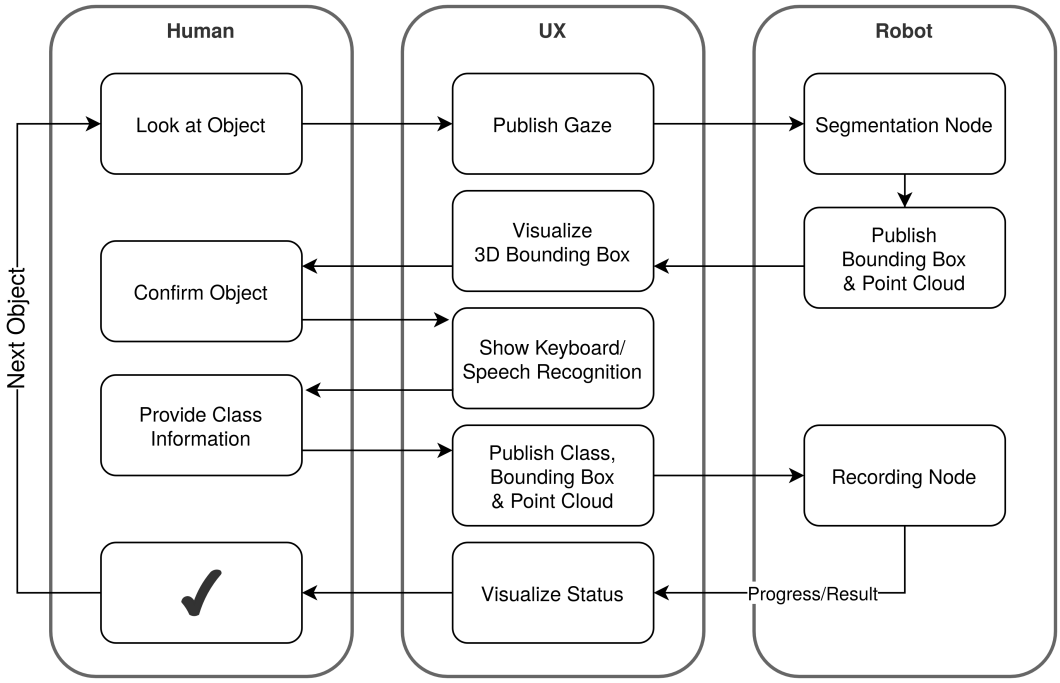


Figure B.3.1: Overview of the entire teaching pipeline. The AR user interface (UX) acts as a bridge between human and robot.

reality glasses manufactured by Microsoft with a built-in eye tracker. We developed the interface, i.e. the HoloLens application, using the game engine and real-time development platform Unity, version 2019.4.36. In addition, we used assets from the Microsoft Reality Toolkit, MRTK 2.7.2, which Microsoft supplies specifically for this purpose. The data interconnection between the Universal Windows Platform (UWP) app on the HoloLens and the robot operating system, ROS [183], takes place via ROS# [184]. The open-source software library ROS# exchanges JSON based commands with ROS through the rosbridge_suite from within Unity applications. Both the HoloLens and the robot are continuously connected with each other via WiFi, and data, such as the human's gaze information, can be sent and received in real time. Finally, gestures and speech serve to operate the AR interface and to interact with the robot. Figure B.3.1 illustrates how the AR interface acts as a bridge between human and robot and shows the interplay of the individual components of our teaching pipeline, which we will describe in more detail below.

Identifying the Unknown Object of Interest

In order for the robot to learn a new object, it has to identify it as such in the first place. This is quite a fundamental problem, as it is, in a sense, a chicken-and-egg problem. For the robot, it is difficult to detect the object of interest as it does not know it at this point and it is yet to be taught. Therefore, since the robot must identify the object before it has learned it, the deployment of neural networks is not possible at this point, and determining where the object begins and where it ends is not trivial. Instead, we want to incorporate the human's gaze information to help the robot locate the target object. This means that the human looks at the object, whereupon the robot can distinguish it from the rest of the environment. For this purpose, we take the approach of [3] as a basis, who segmented observed objects using human gaze and the point cloud obtained from the depth sensor of the robot's scene camera. Thereby, a calibration determines the respective position of the robot and HoloLens, and the HoloLens' motion sensors ensure that the mutual position is tracked during human movements. In addition, the gaze point, that is, the point at which the human is looking, is continuously tracked by the HoloLens and published as a point in 3D space through our user interface using ROS#. Thus, the corresponding ROS topic can be subscribed by the ROS system of the robot, which means that the gaze point is known to the robot at all times and can be used for segmentation. In the first instance of the segmentation described in [3] a pass through filter and a voxel grid filter are applied to reduce the size of the point cloud. Subsequently, the ground is extracted using RANSAC and eventually the object is isolated by means of the gaze point and Euclidean clustering. We adapt this method with slight adjustments in the last step. Instead of assigning only the cluster closest to the gaze point to the object, we consider all clusters within a certain distance and with a certain size. We set the threshold for the maximum distance between cluster and gaze point to 2 cm and the minimum cluster size to five points. This way it is possible to even segment very flat objects that do not protrude far from the ground. Depending on which object the person is currently looking at, the respective object is then segmented in real time. All the mentioned point cloud processing is accomplished using the open-source Point Cloud Library (PCL) [361].

Owing to the calibration carried out at the beginning, the position of the object of interest is known both from the location of the human wearing the HoloLens as well as from the location of the robot. The former allows the robot to display its feedback regarding the segmented object as a 3D bounding box on the HoloLens, namely in the human's field of view, using a subscriber, attached to a virtual bounding box, that updates the position of the box depending on the segmentation results. We can take advantage of the latter during the teaching process described below.

Teaching through Joint Attention

The attention of the robot and the human is now jointly directed at one and the same object. The next step is to confirm to the robot that the framed object is the object of interest and to provide a class information. By performing a pinching gesture with the index finger and thumb (within the field of view of the HoloLens) during the fixation of the object with the human eyes, we can select the object via the AR interface. Alternatively, it is also possible to just say “select”. Thereupon, a virtual keyboard appears in the human’s field of vision, on which he can now enter the class name of the object. Again, it is alternatively possible to simply resort to speech. Figure B.3.2 shows the implemented keyboard from the human perspective.

Our next goal is to have the robot autonomously capture images of the object of interest, which it can later use as training data. In order to get as many images from multiple angles as possible, we attach a second camera to the wrist of the robot’s arm. The robot is now supposed to move this camera in a circle around the object. This means that

Figure B.3.2: In the human’s field of view, a bounding box of the object segmented by the robot is displayed. After the selection it is possible to specify the class name of the respective object using a virtual keyboard or speech.

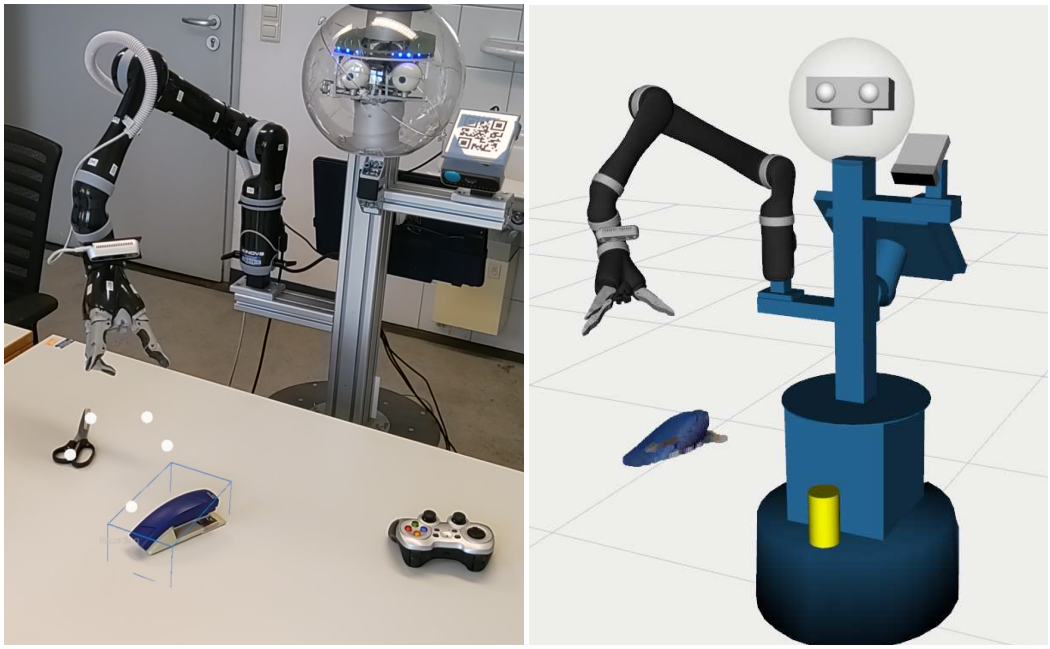


Figure B.3.3: The left side shows the recording process from the human's augmented view and the right side is a visualization in RVIZ with the point cloud of the segmented object. The point cloud is used both to display the bounding box for the human and to label the images captured by the camera on the wrist of the robotic arm.

once the human has transmitted the class name to the robot by means of a ROS action, it calculates a circular trajectory of reachable points. Due to physical limitations, such as the length of the arm, this is usually a partial segment of the circle. During the movement, the camera is aligned in such a way that it points at a 45 degree angle to the center of the previously determined 3D bounding box. As the distance between the center and the camera, we use twice the length of the diagonal of the bounding box with a minimum safety distance of twice the distance between the camera and the robot arm end effector. The recording process is illustrated in Figure B.3.3.

By virtue of the calibration described in the previous section, the robot is not only aware of the position of the object of interest, but also capable of computing the transformation to all of its coordinate systems, namely the robot frames. This includes the camera on the robot's arm. The essential aspect in this step is to continuously transform the point cloud of the segmented object into the coordinate system of the camera. By projecting the point cloud onto the 2D image plane of the camera, we can derive the region of interest from the boundary points. Thus, for each captured image, we can additionally store a 2D

bounding box calculated in this way. Overall, the stored synchronized data are RGB and depth images as well as the regions of interest with the 2D bounding boxes. In addition, we also store the positions of the camera to the captured object. Eventually, the robot is able to automatically produce hundreds of labeled training images in a remarkably brief period of time. More precisely, teaching one object takes about one minute and yields about 300 images. The progress and completion of the recording process is in turn transmitted to the HoloLens via the ROS action, visualizing to the human when the robot is ready for a new object.

Transfer Learning

Following the teaching part, the learning process of the robot now ensues. To enable the robot to independently detect the previously seen objects in the future, it must use the information at its disposal in the form of the training data it has created itself. In other words, the robot, or rather its neural network based object detectors, will be trained on the RGB images obtained. This will be accomplished by means of transfer learning. Hence, we assume some prior awareness of objectness, since our method should be seen as an extension rather than a replacement for the training with common large datasets, such as ImageNet [223], PASCAL VOC [192] or MS COCO [185]. Such an approach is realistic, since most of the existing objects are not part of these datasets. We aim to extend this state of knowledge with our method. Consequently, we resort to state-of-the-art object detectors, such as Faster R-CNN [186] and FCOS [191]. Starting from one of these pretrained models respectively, we delete the last classification layers, and then reinitialize them with the appropriate number of output neurons for our use case. Finally, we retrain said layers with our data, freezing all other neurons from the preceding feature layers. By freezing the feature layers responsible for the general comprehension of objects and fine-tuning only the last few layers, we prevent overfitting [362, 363]. In fact, since we only retrain the classification heads of the models, even training on the robot itself is possible without relying on a high-performance GPU. Naturally, the training may take longer. In Section B.3.6, we will evaluate the performance of the aforementioned models, among others.

B.3.5 Dataset: Objects in Multiperspective Detail

The method introduced in the previous section allowed us to create a new type of dataset. While we publish the validation and test set mainly for the sake of reproducibility of our results, we think that our training set might be especially interesting for further purposes. In the following, we would like to explain the training data in more detail and provide

some statistics. We will elaborate on the validation and test set within the scope of our evaluation in Section B.3.6.

Our training set is particularly characterized by the fact that it supplies many details about individual objects. While most object detection datasets often consist of many images with different objects, our data depicts the objects from many different angles. This is especially attractive for objects that appear different from the front and back, such as a gamepad.

The set consists of 3113 perspectives in total and contains the classes fork, frisbee, gamepad, hole puncher, knife, scissors, shuttlecock, stapler, table tennis ball and toothbrush. For each class, there are two different entities in the set, differing in color, shape, or both. Figure B.3.4 shows some sample images and Figure B.3.5 illustrates the distribution of the viewpoints. The objects are each placed individually on a table and for every camera perspective multiple pieces of information are included. For each RGB image, alongside the region of interest, there is a corresponding depth image that is aligned to the color image. All RGB images and depth images have a resolution of 1920 × 1080. In

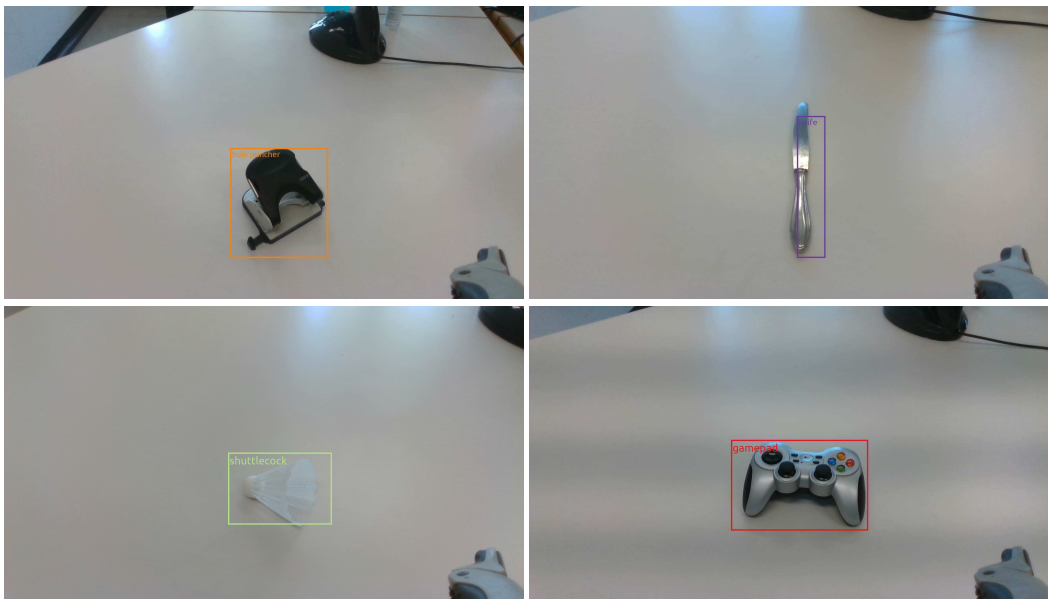


Figure B.3.4: Sample images of a hole puncher, a knife, a shuttlecock and a gamepad from our dataset. The quality of the bounding boxes may vary depending on the point of view and may sometimes be slightly too large, too small or offset. In all images, however, the majority of the box always covers the respective object. The objects contrast differently with the background in terms of flatness and color.

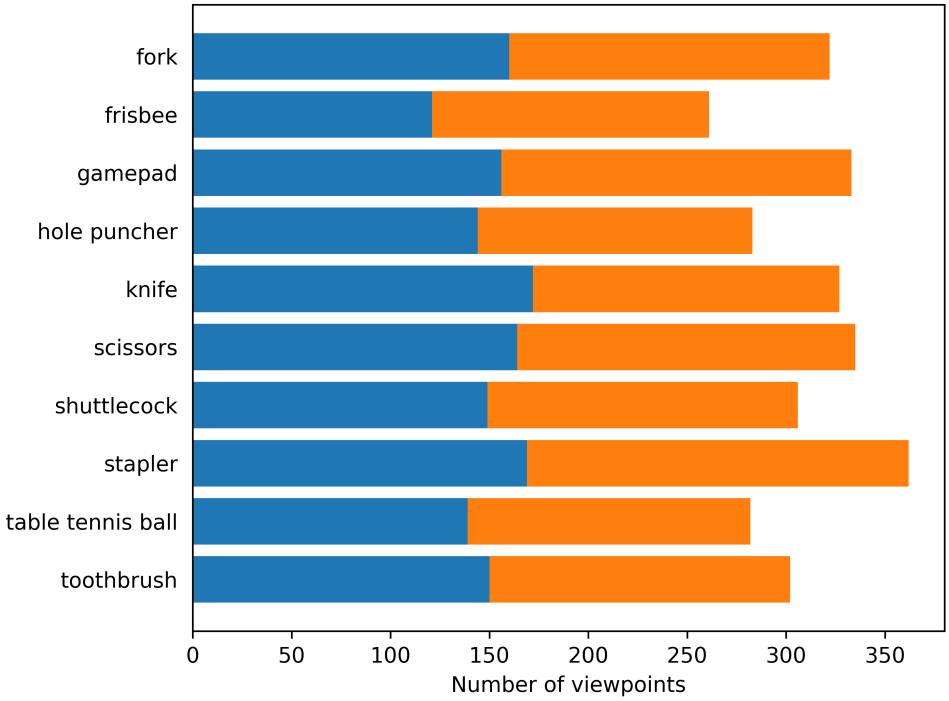


Figure B.3.5: Distribution of the viewpoints across the categories. The colors indicate the two different items within the classes.

addition, all camera poses are available. They are specified independently of the robot as a transformation, composed of translation vector and rotation quaternion, from the coordinate system of the camera to the one of the object. The last component is the meta-information about the camera with the intrinsic parameters of the camera calibration.

Altogether, we believe that the high information density in our data is also interesting for other research areas where camera positions are crucial, such as Neural Radiance Fields [199, 200, 201, 202]. There, either synthetic data must be used or, given real data, the camera positions (and depths) can only be roughly approximated via structure from motion. For this reason, we make our dataset, Objects in Multiperspective Detail (OMD), publicly available to the research community.

B.3.6 Evaluation

For all our experiments, alongside the aforementioned HoloLens 2 worn by the human, we employed a Scitos G5 from MetraLabs [298] as robot. The body camera through which the robot observes the scene and performs the segmentation is an Azure Kinect DK

from Microsoft. The robot arm that was additionally installed on the Scitos is a Kinova Jaco2 [364] with 6 DoF. The camera attached to the wrist of the arm in order to take pictures of the objects is an Intel RealSense D435.

For training we use our own training set, which we have explained in detail in Section B.3.5. Since the objective of this work is to teach the robot its environment, we also had to record a validation and test set located in the environment where the learning process took place. As mentioned earlier, alongside the training set, we will also publish the validation and test set to ensure the reproducibility of our results. The validation and test set consist of 1051 and 1410 regular images, respectively, which were manually labeled by hand using DarkLabel [222]. The classes represented therein are the same ten as in the training set. For each class, four distinct objects of the respective category were available. The validation set contains the same two objects of each class as the training set, whereas the test set contains the other two. That is, the objects differ in shape, color, or both from those used in training. The objects were photographed randomly grouped (within the set) in the robot's office environment. We made sure to create challenging scenarios as well, such as items being stacked or the toothbrush still being in its packaging, as shown in Figure B.3.6.

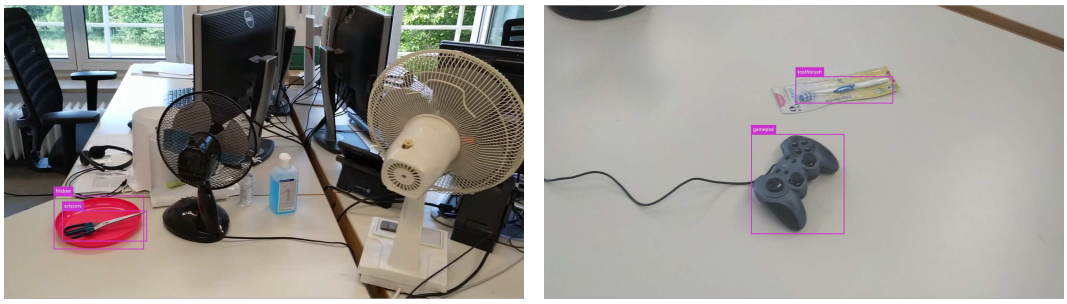


Figure B.3.6: Sample images from the test set. The set of objects is disjoint with the ones from the training set (see gamepad). The set is also diverse in terms of the clutter of the background and the distances to the objects.

In the following, we evaluate our learning pipeline using several state-of-the-art object detectors. Consequently, the object detectors Faster R-CNN [186], RetinaNet [204], FCOS [191], and SSD300 [224] serve as a foundation. We complement these with various backbones, such as ResNet-50-FPN [187], VGG16 [226] and MobileNetV3 Large [365, 225]. All backbones were trained on ImageNet and can be left as is, since we deliberately picked object classes for our evaluation that had no intersection at all with this dataset. The reason behind our choice of the ten test objects was as follows. On the one hand, they must not appear in ImageNet due to the backbones, but on the other hand, at least a part

of them ought to be in MS COCO so that we have a comparison later on. Furthermore, within each class there had to be several different looking objects of that class. All of this together limited the selection accordingly, especially since we tried to avoid perishable classes like food. Hence, in terms of the actual object detectors and to ensure that our objects are indeed unknown to the models, we had to train them on a subset of MS COCO. More precisely, we extracted the classes fork, frisbee, knife, scissors, sports ball, and toothbrush from the dataset using the tool Fiftyone [366] and then trained the above mentioned detectors on the remaining part. In doing so, we followed the respective training recipe of the original implementation and, for consistency, adhered thereto in all of our subsequent experiments in our own training pipeline. The only exception in our transfer learning approach was the type of data augmentation applied. In this case, we used random photometric distortion, random zoom out, random cropping, and random horizontal flipping for all models (not just SSD300) to prevent overfitting. Subsequently, we trained using the method described in Section B.3.4.

In all our experiments, we evaluate according to the MS COCO metric [185], namely the average precision for varying intersection-over-union thresholds (IoU). In this context, we use the abbreviations AP = $AP^{IoU=0.5:0.05:0.95}$, AP_{50} = $AP^{IoU=0.5}$, and AP_{75} = $AP^{IoU=0.75}$ within a class and mAP as the average over all categories. Analogously, this applies to the average recall, where we consider the maximum recall given 1, 10, and 100 detections per image, respectively, and use the abbreviations AR_1 = $AR^{max=1}$, AR_{10} = $AR^{max=10}$ and AR_{100} = $AR^{max=100}$. Again, mAR denotes averaged over all categories. Unless otherwise stated, average precision and average recall refer to AP and AR, respectively.

A comparison of all tested models is provided in Table B.3.1. Faster R-CNN with the ResNet-50 backbone generally performed best in terms of average precision. With a MobileNetV3 backbone, the performance was significantly worse in terms of both precision and recall. FCOS and RetinaNet are slightly behind Faster R-CNN in terms of

Table B.3.1: Comparison of all machine learning models trained in a transfer learning fashion. The best values are highlighted in bold.

Model	Backbone	mAP	mAP_{50}	mAP_{75}	mAR_1	mAR_{10}	mAR_{100}
Faster R-CNN [186]	ResNet-50 [187]	**33.6**	**66.9**	31.4	43.7	50.1	50.4
Faster R-CNN [186]	MobileNetV3 [365, 225]	15.5	38.1	6.1	23.7	27.4	27.7
Faster R-CNN [186]	MobileNetV3 [365, 225]†	13.0	38.4	3.1	22.2	25.5	25.5
FCOS [191]	ResNet-50 [187]	30.6	47.6	**35.9**	44.7	53.8	55.0
RetinaNet [204]	ResNet-50 [187]	31.2	52.4	34.3	**46.2**	**57.6**	**59.1**
SSD300 [224]	VGG16 [226]	8.0	19.1	5.0	21.0	31.6	34.0

† Tuned for mobile use cases

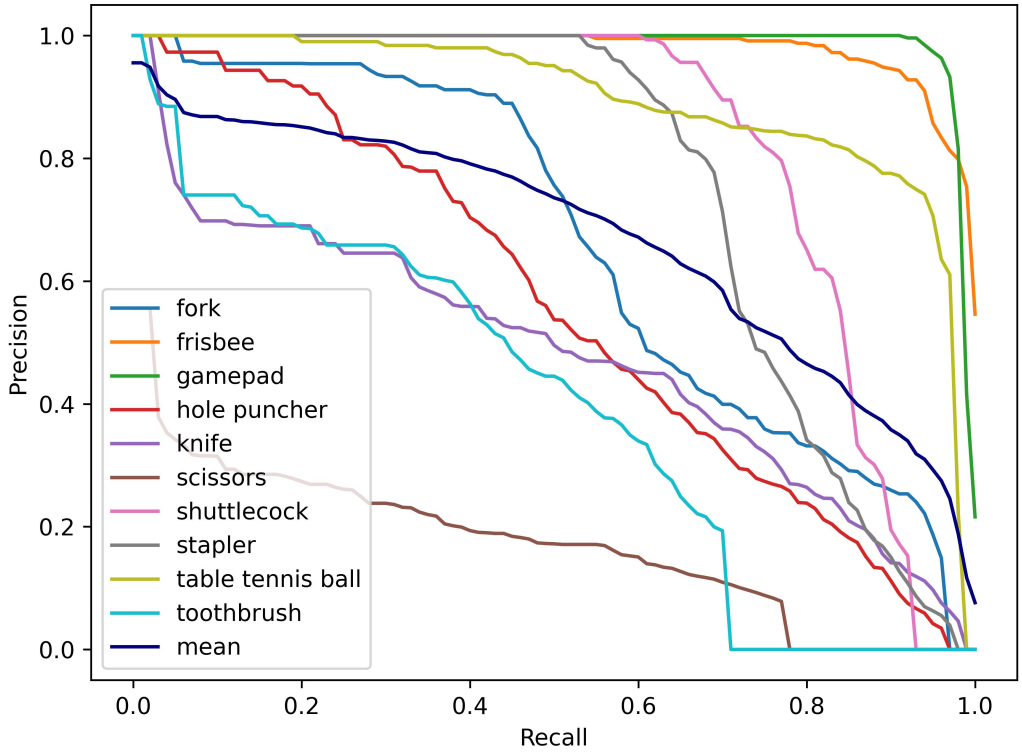


Figure B.3.7: Precision-recall curve of Faster R-CNN at an IoU of 0.5. Above this value, objects can be considered as detected [192, 193].

the mean average accuracy. The latter has the best recall values, while FCOS has the best mean average precision at an intersection-over-union of 0.75. SSD300 clearly lags behind all other models in terms of precision.

As we proceed, we will continue with Faster R-CNN for further analysis, since MS COCO [185] considers mAP as the single most important metric. In general, the model seems to detect the objects quite well, but some classes cause more difficulties than others. This also becomes apparent by looking at the curve in Figure B.3.7. Even at low recall values, the precision of the scissors is below 0.5. The class toothbrush also decreases early. On the other hand, the precision for the classes gamepad and frisbee is consistently excellent, even for a high recall.

In Table B.3.2, we compare different training variants. If we ignore the baseline variant (COCO) for a moment, we find that the transfer learning method, in which only the last layers had been trained (TL-F), has the best AP for the majority of classes. In particular,

Table B.3.2: Comparison of the average precision (AP) for different training types of Faster R-CNN on our test set. Namely, apart from the backbone, trained from scratch (S) or trained in the sense of transfer learning with frozen non-classification layers (TL-F) or completely unfrozen (TL-U), respectively. All three on the data collected by the robot. The best values are highlighted in bold. The last column (COCO) serves as an orientation and reports the results of Faster R-CNN trained on the entire MS COCO training set.

Class	S	TL-U	TL-F	COCO
fork	**21.7**	**21.7**	18.6	63.8
frisbee	19.5	47.6	**58.8**	65.9
gamepad	38.4	24.8	**62.6**	-
hole puncher	**42.5**	26.3	23.0	-
knife	20.4	19.1	**27.6**	50.0
scissors	6.5	**7.3**	5.1	70.9
shuttlecock	24.9	27.8	**51.5**	-
stapler	29.3	24.8	**38.9**	-
table tennis ball	3.4	27.6	**44.0**	17.3
toothbrush	**21.6**	8.8	6.2	37.4
mAP_{50}	62.8	**68.8**	66.9	84.4
mAP_{75}	9.8	7.5	**31.4**	55.4
mAP	22.8	23.6	**33.6**	50.9

mAP_{75} is significantly higher. It is worth mentioning that the baseline values (COCO) are naturally superior. The difference between the full MS COCO training set and the part we used for pretraining remains still 25 713 objects in 14 296 images, which is five times as many images as we used. In addition, our images are distributed among all ten classes, while MS COCO does not contain four of them and the baseline thus does not recognize them at all. This demonstrates the strength of our pipeline, which is designed to enable the learning of additional, as yet unknown classes, that is, to extend existing knowledge. The comparison with the baseline, which is the ideal case, namely 1) a suitable data set exists 2) it is accessible and 3) the object is part of it, serves primarily to better classify our results into the overall picture. It is not intended to outperform a model trained on such a large data set, but rather to determine an upper bound and test how close we can get with our method. Taking this into account, it is remarkable how well our pipeline has learned especially the classes gamepad or shuttlecock, whose AP is even higher than the mAP of the baseline. Furthermore, since the table tennis ball occurs in MS COCO only as a subset of the class sports ball, we can see how our system becomes more attentive to table tennis

Table B.3.3: Results of Faster R-CNN trained via TL-F on different sized subsets of our dataset. The best values are highlighted in bold.

	25%	50%	75%	100%
mAP	31.5	32.9	31.7	**33.6**
mAP_{50}	64.7	66.7	66.7	**66.9**
mAP_{75}	28.6	28.4	26.0	**31.4**
mAR_1	41.1	42.9	41.4	**43.7**
mAR_{10}	46.9	49.9	46.9	**50.1**
mAR_{100}	47.2	50.2	47.1	**50.4**

balls. In contrast, the class frisbee does not quite reach the baseline as the corresponding MS COCO class contains exclusively frisbees. In the case of the category hole puncher, the results are satisfying even without pre-existing basic knowledge (S).

Considering the amount of time needed for the respective training, major differences become apparent. Training of the entire MS COCO training set lasted the longest, at more than two days on two NVIDIA RTX A4000 deployed in parallel. The other three variants were trained on our dataset on only one of the GPUs and took 4 hours for the entire model (S, TL-U) and 2.5 hours for the freezed variant (TL-F), respectively. As mentioned above, although for consistency reasons we trained 26 epochs as in the original recipe, the weights with the best validation accuracy that we eventually used for testing were often reached earlier. For Faster R-CNN trained via TL-F, this was even the case after three epochs (starting at mAP = 0.0 before training), which corresponds to a training time of about 40 minutes on our Scitos equipped with an NVIDIA GeForce GTX 1050 Ti. This makes the entire pipeline also suitable for stand-alone learning directly on the robot.

Finally, we analyze the influence of the amount of images used for training. Table B.3.3 lists the results of Faster R-CNN trained using TL-F for varying dataset sizes. The images were removed from the sequence of perspectives with equidistant spacing. Apart from the case where 75% of the data is used, the tendency emerges that as the number of images increases, so does the average precision and average recall. The best result is obtained using all the data.

B.3.7 Limitations & Discussion

Similar to all supervised machine learning methods, we depend on the quality of the training data. In our case, this can vary depending on the preceding segmentation of the point cloud. This in turn is naturally dependent on the quality of the data obtained by

the depth sensor of the robot's scene camera. Especially with very dark or glossy surfaces, we noticed that the depth sensor had problems determining the depth accurately. As a result, the accuracy of the bounding boxes suffers, which eventually has an impact on performance. However, this problem can be compensated with an even larger number of objects and our tests have shown, moreover, that the robot is still capable of detecting the learned objects in its environment despite such difficulties.

One further point is that while our approach generalizes well even to other objects of the learned classes, our tests were inferior on popular datasets such as MS COCO. This is due to the diversity of the images and the versatile situations depicted in them. For instance, fruits such as bananas and apples can be found in their natural form as well as cut into small pieces in a fruit salad. This task can only be solved with an enormous amount of training data. Our method, on the other hand, although it cannot achieve the performance of training on large datasets, is primarily designed to teach the robot objects for which data does not yet exist. In such scenarios, a semantic scene understanding is necessary and humans can assist in gaining this understanding by means of our method. While an extension of existing datasets would also be conceivable, our method, in contrast, does not require tremendous labeling resources and can be used spontaneously in the respective situation. Our evaluations (Table B.3.2) show that our method is capable of learning unknown objects that are not detected by the baseline. It can therefore be used more flexibly without the need to know the situation in advance and rely on the existence of appropriate datasets. Moreover, teaching through two-way interaction is extremely natural, especially since the AR system enables real-time communication between human and robot by directly connecting both worlds, the analog world of the human and the digital world of the robot, so that the human can supply information (gaze, class) to the robot, thus initiating the recording process, and the robot can in turn communicate feedback visually. Pointing to objects using gaze completes the interplay, as it is intuitive, less ambiguous than pointing with a finger and, unlike speech, can be applied before the object is known to the robot. All in all, we therefore consider our approach to be less of a replacement and more of a supplement.

B.3.8 Conclusion

In this work, we presented a novel pipeline towards the deployment of robots in non-predefined scenarios. To this end, we leveraged human gaze and augmented reality in the interaction between robot and human to successfully teach the robot new, yet unknown objects in its environment.

In order for robots equipped with machine learning based object detectors to detect

their environment and the objects contained therein, a lot of training data is usually required. In practice, however, under unpredictable conditions and due to the wide range of existing objects, the availability of a suitable data set can not always be guaranteed. Our approach can complement popular datasets in exactly such situations and produce large amounts of automatically labeled, non-synthetic training data in a user-friendly manner and in a short period of time. On the basis of such data, we have trained state-of-the-art object detectors in several different ways and shown that it is possible to learn and detect new objects in this manner. In fact, the training can even take place standalone on the robot due to transfer learning, without the need for tremendous computational resources. Further, with a few instances, it was possible for the robot to generalize to unseen objects in the given class and to detect classes that could not be detected by the baseline due to an unsuitable underlying training dataset. This makes our teaching pipeline a valuable extension to training exclusively on standard datasets. Overall, our approach is supremely natural and intuitive by virtue of its multimodality, including AR and the shared gaze of human and robot. The dataset we have recorded in the course of our evaluation is also made publicly available and is characterized by a high level of information density owing to the many different perspectives on the respective object and the data gathered in this process. As a result, it has the potential to be relevant for a variety of purposes, aside from ours.

However, a significant amount of work remains for the future as we plan to investigate the usability of our system in a user study as well as to extend our approach to enable the robot to successfully detect objects outside of its trained environment and to further leverage the acquired knowledge through active learning in another human-robot interaction scenario.

Acknowledgments

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. We also thank Julia Dietl for her valuable efforts in labeling.

B.4 Leveraging Saliency-Aware Gaze Heatmaps for Multiperspective Teaching of Unknown Objects

B.4.1 Abstract

As robots become increasingly prevalent amidst diverse environments, their ability to adapt to novel scenarios and objects is essential. Advances in modern object detection have also paved the way for robots to identify interaction entities within their immediate vicinity. One drawback is that the robot's operational domain must be known at the time of training, which hinders the robot's ability to adapt to unexpected environments outside the preselected classes. However, when encountering such challenges a human can provide support to a robot by teaching it about the new, yet unknown objects on an ad hoc basis. In this work, we merge augmented reality and human gaze in the context of multimodal human-robot interaction to compose saliency-aware gaze heatmaps leveraged by a robot to learn emerging objects of interest. Our results show that our proposed method exceeds the capabilities of the current state of the art and outperforms it in terms of commonly used object detection metrics.

B.4.2 Introduction

With all the advancements in technology, the industrial sector witnessed a proliferation of robots and an expansion of their areas of application, spanning a wide range of industries and use cases. From automotive and electronics over food and beverage to rubber and plastics, cosmetics, and pharmaceutical industries, an array of domains profit from the exceptional precision, work capacity, efficiency and high tolerance in demanding and hazardous environments of these robots [337]. Robots are also becoming more prevalent in close proximity to humans, serving as tour guides in museums [305, 340] or as service assistants in places like supermarkets, such as Walmart [338, 339]. The recent advancements in machine learning, especially in computer vision, have been a major catalyst for this success, as they enabled robots to identify objects and individuals within their environment with great accuracy and speed. Nevertheless, a fundamental premise underlying most cases is the existence of abundant training data with high quality labels. In practice, the assumption that a fully labelled data set is available, is suited to the field of application, and encompasses all relevant objects, is not inevitably valid. While state-of-the-art object detectors can achieve outstanding performance, given sufficient training data, their deployment is restricted to predefined scenarios imposed by the available training data [1]. An unknown object that was not initially already part of the training

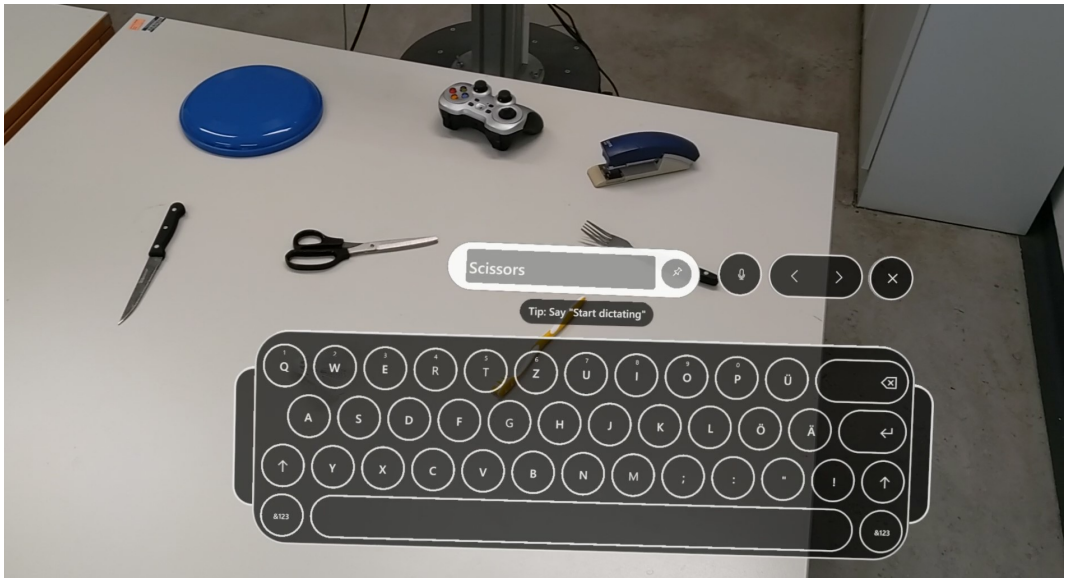


Figure B.4.1: The augmented reality interface through which the human can teach the robot the class of an object using a virtual keyboard or speech.

set, but appears in front of a robot, cannot be detected by definition. As a result, object detectors and, consequently, robots reach their capability limits. For the latter, however, scene understanding is indispensable for all kinds of interaction.

In this work, we aim at minimizing such constraints towards the deployment of robots in unfamiliar environments by promoting the adaptation to new conditions involving unknown objects, while simultaneously reducing data dependency. Within a human-robot interaction setting, we teach a robot novel objects using both multimodal and natural communication channels, such as augmented reality (see Figure B.4.1), speech, and human gaze. The human looks at an object and provides the class information, whereupon the robot, equipped with an object detector, learns it. To this end, the robot autonomously acquires a series of images using a robot arm and determines the area of interest based on the gaze data. Following up on the work of [5], where the robot had to rely on point cloud segmentation, we streamline the teaching process by using saliency aware gaze heatmaps. The core idea is to leverage the Graph-Based Visual Saliency (GBVS) algorithm [189] in combination with gaze [2] to refine the human gaze points and guide the robot's attention to the salient parts of the images that are of interest. As a result, the teaching process becomes more lightweight, yet more efficient.

Overall, the contribution of our work is twofold:

1. In a human-robot interaction setting, we utilize augmented reality to encode 3D gaze points of a human as saliency-aware 2D gaze heatmaps, applying a single and a dual gaze-assisted approach.

2. Having taught a robot unknown objects, our evaluations demonstrate that we outperform the current state of the art with regard to the common object detection metrics.

B.4.3 Related Work

Teaching a robot to detect unknown objects through multimodal human-robot interaction is a complex task that necessitates substantial collaboration between different research fields, such as computer vision, eye tracking, and robotics.

In [351], the authors approached the task of detecting unknown objects by employing a one-class support vector machine. As the learning process was carried out incrementally, a series of robots were operating simultaneously, connected to a cloud-based system where all the computations were performed. In this approach, the classification of unknown objects was limited to the property of being unknown and the learning process did not incorporate any specific class information. By separating known and unknown items and exclusively relying on the familiar objects, the amount of incorrect data transmitted to the robots ought to be reduced. The awareness that an object is unknown was also the basis in [367]. This work addressed the problem that datasets often contain unlabelled objects that are misinterpreted by object detectors as known classes. The authors identified such out-of-distribution objects in videos to develop an unknown-aware object detector. This model, however, used that information solely to reduce the number of false positives. The unknown objects were neither learned nor classified. Another recent work, as presented in [352], proposed to assess the uncertainty of the predictions of semantic segmentation models to binarily classify whether an object was known or unknown. Again, no attempt was made to learn the individual categories.

In the field of eye tracking, the problem of unknown object detection was investigated in [1]. This proof of concept aimed to decrease the number of candidate bounding boxes generated by a region proposal method, without being able to classify them. The same authors ascertained in [4] whether an object was being observed by a person and determined the corresponding bounding box parameters using simple heatmaps instead of a scene image. All the objects detected in the video segments were unknown, but their categories were not identified. In [368], several egocentric videos of museum visitors we taken to train a model to identify, among 15 different objects, which specific object a subject was currently looking at. However, the experimental setting was fixed and the

set of 15 objects were specified in advance. Furthermore, the model was only capable of identifying the attended object, and none of the other known objects in the scene. Hence, the model did not have the ability to detect objects in a general sense. A similar objective was analysed in [369]. In a virtual reality study the set of objects that a user's attention was directed towards, was narrowed down based on the combination of head pose and linguistic description. Nevertheless, neither was the class of the attended object learned, nor could the object itself be detected at a later point in time.

Several researchers, including [357] and [358], have also investigated the potential of collaborations involving both humans and robots for teaching purposes. The work presented in [357] utilizes a natural language context to enable a robot to recognize objects through one-shot learning based on visual descriptions. Nevertheless, in this proof of concept, the objects were required to be easily distinguishable through color or spatial relationships, and their component parts needed to be precisely describable using linguistic expressions to avoid ambiguity. The authors of [358] introduced a teaching system for object categorization that provided users with the ability to visualize the intermediate categorization stages, meaning the class into which the robot would categorize an object. While the system allowed for interactive improvement and correction of the categorization, it required all objects to be marked with fiducial markers to be recognized at all. Previous works, such as [359] and [360], have taught robots new objects in conjunction with picking tasks. However, both studies used identical objects for both training and testing, which creates a positive bias in the results. In order to reflect the real-world conditions more accurately, we use distinct objects per class in our experiments.

B.4.4 Method

The objective of this work is to teach a robot unknown objects within its environment when training data does not yet exist or may not be available. In other words, the human looks at an object that is unknown to the robot and provides the respective class information. The robot then autonomously records the object from several different viewpoints and labels the data. Finally, the robot uses the obtained data to learn the respective class. Building upon the work of [5], we effectively extend their method, resulting in a significant improvement of the achieved results.

The interaction between the human and the robot requires a way of communication, which we meet by means of the augmented reality interface introduced in [5]. This user interface operates on a HoloLens 2 worn by the human and enables real-time transmission of the human's gaze data to the robot. In the course of the teaching process, however, we eliminate the assumption that the robot must identify the object prior to recording

the data. Instead of a segmentation that provides a 3D bounding box with the size and position of the object, our method is based on successive gaze points of a given time interval. The human selects the object of interest by saying “select” and then observes it for a few seconds (10 seconds in our experiments), seeking to capture the entire surface of the object as completely as possible. A visualization is shown in Figure B.4.2. Based on the resulting gaze points the position and size of the object is approximated in the world space. Thereby, the median of all gaze points serves as the center of a 3D bounding box and three times the interquartile range as its size, component-wise. One limitation of [5] is the dependence on the segmentation presented in [3] and, consequently, on the depth sensor data of the robot’s body camera, which leads to problems with very flat, dark or glossy objects. By directly determining the position of the object using the gaze points, these types of objects do not pose a challenge for our approach. Moreover, since we will later also refer to the gaze data to automatically label the recorded perspectives with a bounding box each, we achieve better overall results with less error proneness due to fewer system components.

After the human has looked at the object for the specified time interval, a virtual

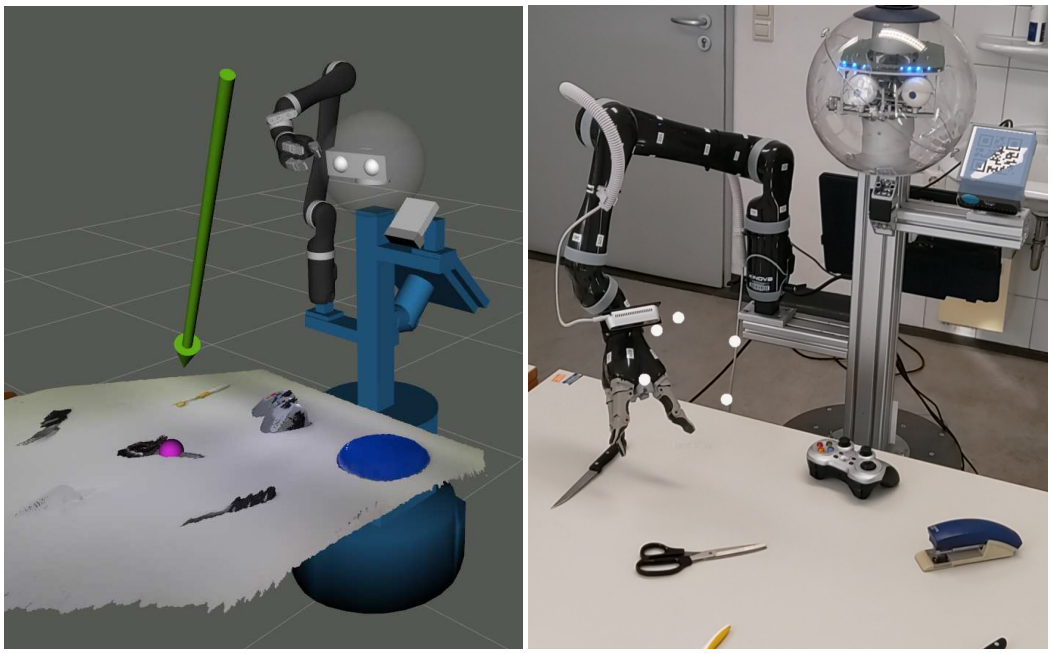


Figure B.4.2: On left side, the human gaze ray (green arrow) and the ensuing gaze point (purple sphere) are visualized in RVIZ. The right side shows the recording process as seen from the augmented view of the human.

keyboard automatically appears in the human's field of view. Using this keyboard or alternatively by speech, the class of the object can now be specified (see Figure B.4.1). Once the human has submitted the input, it is transmitted to the robot, which then calculates a circular path around the object from which it can be recorded by means of a robot arm. Due to physical constraints, such as the robot arm length, the final reachable trajectory of the arm is usually a sub-segment of the circle. The robot then automatically brings its arm to the start of the trajectory and moves along it. Using a camera, which is attached to the wrist of the robot arm, the robot records the object from various angles. This process is illustrated in Figure B.4.2. For each perspective, the gaze points gathered at the beginning are continuously mapped from the world space to the respective coordinate system of the moving camera. Subsequently, these transformed 3D gaze points are projected onto the 2D image plane of the camera. In principle, a 2D area of interest could now already be determined for each image from the boundary points. However, in practice, many small inaccuracies are involved. These include among others the eye tracking and gaze determination, the calibration of the robot with body, arm and camera, as well as the transformation between the individual robot frames. Although these are negligible individually, they add up in combination to an offset that corrupts the result. This offset can also be seen in Figure B.4.4b.

In order to determine the region of interest more precisely, we incorporate saliency in addition to the gaze data. A saliency map is an image that highlights the most visually prominent regions in an input image that are likely to attract the attention of a human observer. Inspired by [2], we leverage the Graph-Based Visual Saliency (GBVS) algorithm [189] for saliency-aware gaze heatmaps. GBVS refers to a computational method to determine saliency maps based on principles of graph theory. It uses the visual and spatial relationships between image pixels to estimate their degree of importance. The GBVS algorithm can be structured into three steps:

1. Extraction of a feature map $\mathcal{M} \in \mathbb{R}^{m \times n}$ based on a given Image $\mathcal{I} \in \mathbb{R}^{m \times n}$.

2. Generation of an activation map $\mathcal{A} \in \mathbb{R}^{m \times n}$ based on $\mathcal{M}$.

3. Normalization of the activation map $\mathcal{A}$ (and combination with the activation maps of all other feature maps).

In the first step, low-level visual features such as color, luminance, and orientation are extracted from the image $\mathcal{I}$. The creation of such feature maps $\mathcal{M}$ is a well-known task and can be found in the literature [228, 370], therefore we refer the readers to these sources for further details.

The second step is to calculate the saliency of each feature map based on the concept that a pixel is more salient if it is different from its surroundings. This step is modeled on [2], except that we omit the temporal domain as we merge the series of successive gaze points into a single heat map per perspective. The idea is to construct a graph representation of the image, where each node represents an image pixel, and the edges between nodes reflect the relationships between the pixels in terms of their visual dissimilarity or spatial proximity. The activation map $\mathcal{A}$ can then be interpreted as a state vector of a Markov chain on this graph. So let the visual dissimilarity of two nodes $s_i = (p,q)$ and $s_j = (u,v)$, where $p,u \in \{1,2,\dots,m\} =: [m]$, $q,v \in [n]$ and $i,j \in [mn]$, be defined as

$$d(s_i,s_j) = \left|\log \frac{\mathcal{M}(s_i)}{\mathcal{M}(s_j)}\right|. \tag{B.4.1}$$

Further, let $\pi_1 \colon [m]\times[n] \to [m], (x,y) \mapsto x$ and $\pi_2 : [m]\times[n] \to [n], (x,y) \mapsto y$ be the projection onto the first and second coordinate, respectively. Then $F \colon ([m]\times[n])^2 \to \mathbb{R}$ given by

$$F(s_i,s_j) = \exp\left(-\frac{\hat{F}(s_i,s_j)}{2\sigma^2}\right), \tag{B.4.2}$$

where

$$\hat{F}(s_i,s_j) = \big(\pi_1(s_i)-\pi_1(s_j)\big)^2 + \big(\pi_2(s_i)-\pi_2(s_j)\big)^2, \tag{B.4.3}$$

is the exponential weighted square distance of the two nodes s_i and s_j. The variable σ is a free parameter that is usually set to one tenth to one fifth of the mean value between the width and height of the feature map [189].

We now consider a fully-connected, directed graph that links all nodes of the lattice $\mathcal{M}$. The weight assigned to the directed edge from node s_i to node s_j is the product of the visual dissimilarity d in the domain of $\mathcal{M}$ and the spatial proximity F:

$$w(s_i,s_j) = d(s_i,s_j)\cdot F(s_i,s_j). \tag{B.4.4}$$

By normalizing the weights of each node's outbound edges to 1, we can define a Markov chain. This allows us to establish an equivalence between nodes and states, and between edge weights and transition probabilities. The Markov transition matrix between the mn states is

$$\mathcal{T} = (t_{i,j})_{1\le i,j\le mn} \in \mathbb{R}^{mn\times mn}, \tag{B.4.5}$$

where $t_{ij} = w(s_i,s_j)$. The final activation $\mathcal{A}$ results from the equilibrium distribution, that is, $\hat{\mathcal{A}}\mathcal{T} = \hat{\mathcal{A}}$, where $\hat{\mathcal{A}} \in \mathbb{R}^{1\times mn}$ is the flattened version of $\mathcal{A}$. This boils down to an eigenvector problem. In practice, a common method for determining the equilibrium

distribution involves repeatedly multiplying the Markov transition matrix with a vector $v \in \mathbb{R}^{1\times mn}$ that is initially uniformly distributed. Consequently, given $\lim_{k\to\infty} v\mathcal{T}^k = \hat{\mathcal{A}}$, $\hat{\mathcal{A}}$ can be estimated as $\hat{\mathcal{A}} = v\mathcal{T}^k$, using a sufficiently large $k \in \mathbb{N}$.

Motivated by [2], we incorporate gaze into this step. Rather than distributing the initial (flattened) activation map v uniformly, we initialize it based on the gaze points:

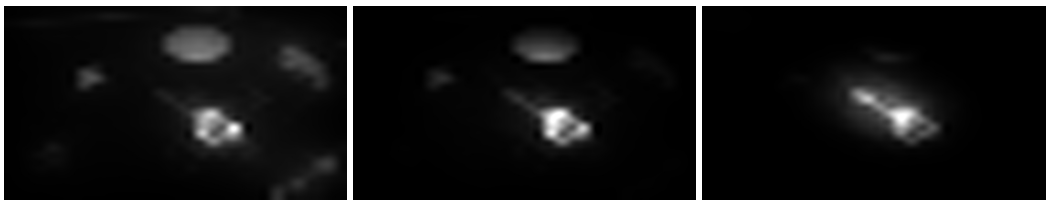
$$v = \sum_t \frac{v_t}{\|v_t\|_1}, \quad v_t = \big(F(s_0, g_t) \ \dots \ F(s_{nm}, g_t)\big), \tag{B.4.6}$$

where g_t is the recorded gaze position at time t. In this way, v is influenced by the weighted distance of the pixels to the individual gaze points. Since the equilibrium state remains the same regardless of the initialization, we do not perform the multiplication with the transition matrix multiple times, but only once, that is $k = 1$. The greater the value of k, the more significant the impact of saliency, while the influence of the gaze decreases proportionally. Depending on the quality of the gaze data, this value can be adjusted accordingly.

The next and last GBVS step is to "normalize" and combine the activation maps. The aim is to further accentuate salient areas, in order to produce an informative saliency map that is not overly uniform. This can be achieved with an analogous Markovian approach as in step 2, selecting the weights of the transition matrix $\mathcal{T}'$ as follows:

$$w'(s_i, s_j) = \mathcal{A}(s_j) \cdot F(s_i, s_j). \tag{B.4.7}$$

While the authors of [2] only included gaze data in the second step, we would like to point out that this is additionally possible in this third step, if the influence of the gaze needs to be increased further. In that case, the initial normalized activation map v' has to be initialized according to (B.4.6). Under certain circumstances, such as when an input image contains multiple objects, our experiments have shown that this Dual Gaze-Assisted GBVS



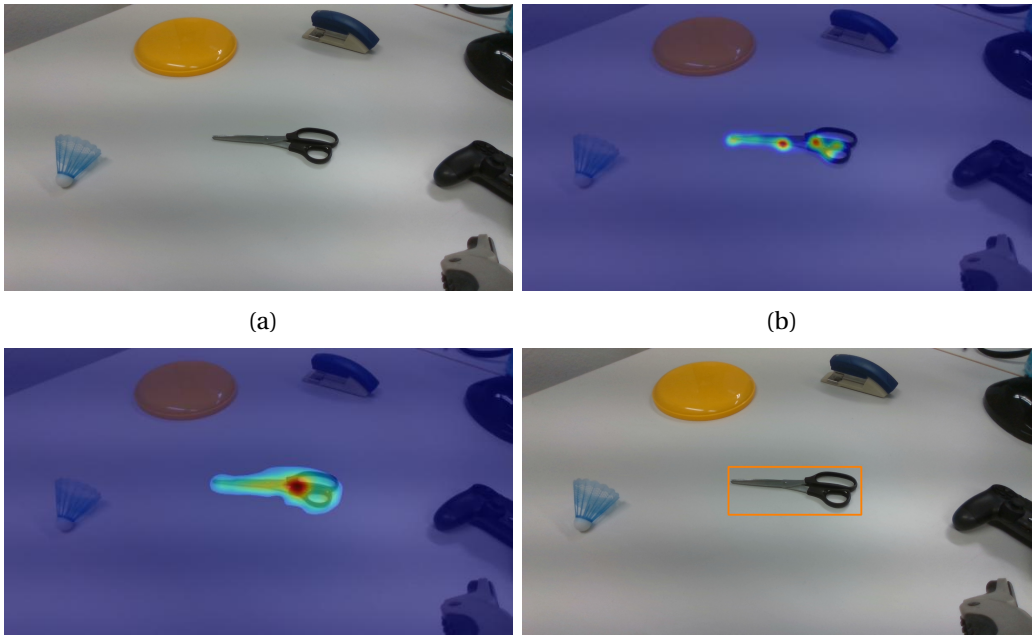
(a) (b) (c)

Figure B.4.3: A comparison of the saliency maps obtained by (a) GBVS, (b) Gaze-Assisted GBVS, and (c) Dual Gaze-Assisted GBVS.

approach (DGA-GBVS), where gaze data is taken into account in the normalization step as well, can be beneficial and may lead to improved performance. Nevertheless, when it comes to the overall learning process, the consideration of gaze data only in the second step (GA-GBVS) outperforms the method proposed in [5] even more apparent.

The finalization is done by combining the activation maps of all extracted feature maps into one single saliency map. This can be achieved by summing up and then normalizing the outcome to the image value range. The ultimate resulting saliency-aware gaze heatmap contains less noise than the standard GBVS saliency map and more intensely concentrates attention on the area of interest. A comparison of the final saliency maps is shown in Figure B.4.3.

Owing to less noise, a bounding box can now be determined on the basis of the boundary points, whereby a threshold value for the points under consideration is set beforehand using Otsu's binarization [203]. An example with the intermediate stages of



(a) (b)

(c) (d)

Figure B.4.4: The intermediate stages of the bounding box determination. (a) Various images of the object are taken by means of the robot arm. (b) The 3D gaze points are mapped to each image obtained. (c) The heatmap is refined using GBVS in combination with gaze. (d) Eventually, after Otsu's binarisation [203], the boundary points lead to the desired bounding box.

the entire bounding box estimation is shown in Figure B.4.4.

In the final step, each image taken with the robot arm can now be automatically labeled with a bounding box and the robot can be trained using the transfer-learning approach proposed in [5]. This means that we follow the same training routine and also pretrain the robot on a subset of MS COCO that does not contain any objects that will be taught later in the course of the evaluation.

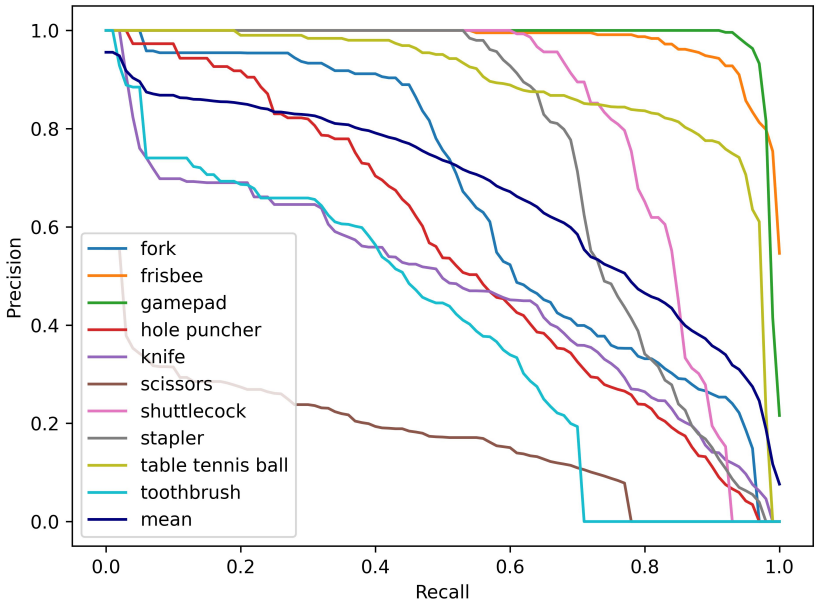
B.4.5 Evaluation

Our evaluation setup comprises of a Scitos G5 robot manufactured by MetraLabs [298], which serves as the base platform. Attached to this platform is a Kinova Jaco2 [364] robotic arm featuring six degrees of freedom. The camera, which is attached to the wrist of the arm and is guided around the objects by the robot to capture images from multiple perspectives, is an Intel RealSense D435 [371].

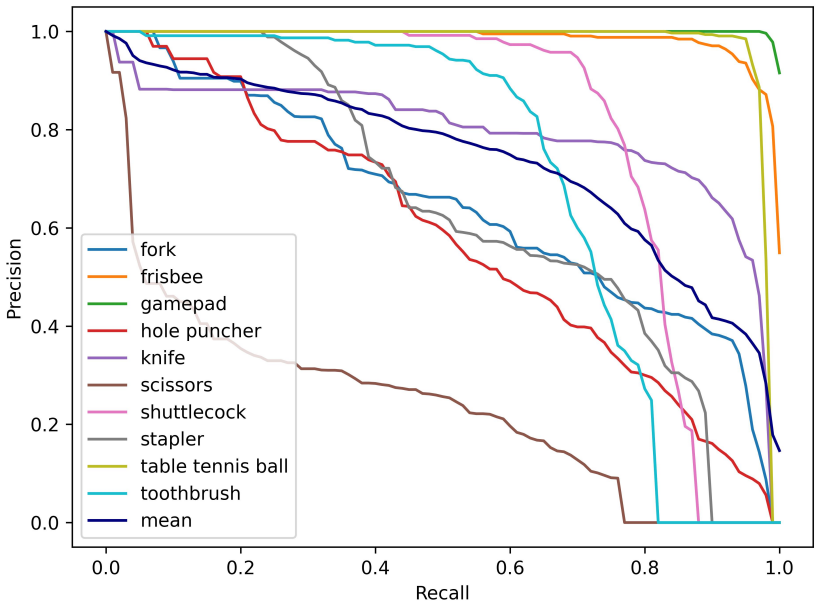
For the comparability of our results, we evaluate on the publicly available Objects in Multiperspective Detail (OMD) dataset [5] and use the approach presented there as a baseline. The dataset contains ten classes, which we teach to the robot following the methodology described in detail in Section B.4.4. For this purpose, conforming to [5], we place the two training items of each class individually on the table in front of the robot, which differ in shape, color or both from the two items in the test set. To replicate their results, we employ Faster R-CNN [186] with the ResNet-50-FPN [187] backbone as underlying machine learning model and we also pretrained the backbone on Imagenet [223] and Faster R-CNN on a subset of MS COCO [185], excluding the classes from the OMD dataset, which we intend to teach.

Upon completion of the training, we evaluate based on the MS COCO metrics [185]. That means we examine the average precision and the average recall for a variety of intersection-over-union (IoU) thresholds. To simplify the discussion, we adopt the following abbreviations: AP denotes the average precision across all IoU thresholds from 0.5 to 0.95, with a step size of 0.05. The abbreviations AP_{50} and AP_{75} correspond to the average precision at IoU thresholds of 0.5 and 0.75, respectively; All are calculated separately for each individual class. The analog notation applies to the average recall, where we use the abbreviations AR_1, AR_{10}, and AR_{100} to refer to the average recall when allowing up to 1, 10, and 100 detections per image, respectively. Additionally, we use mAP and mAR to represent the mean average precision and the mean average recall over all classes.

Figure B.4.5 compares the precision-recall curves of the baseline method and our approach at an IoU of 0.5. This is the threshold above which objects can be considered as detected [192, 193]. In general, the mean curve shows a tendency towards better precision



(a) Baseline



(b) GA-GBVS

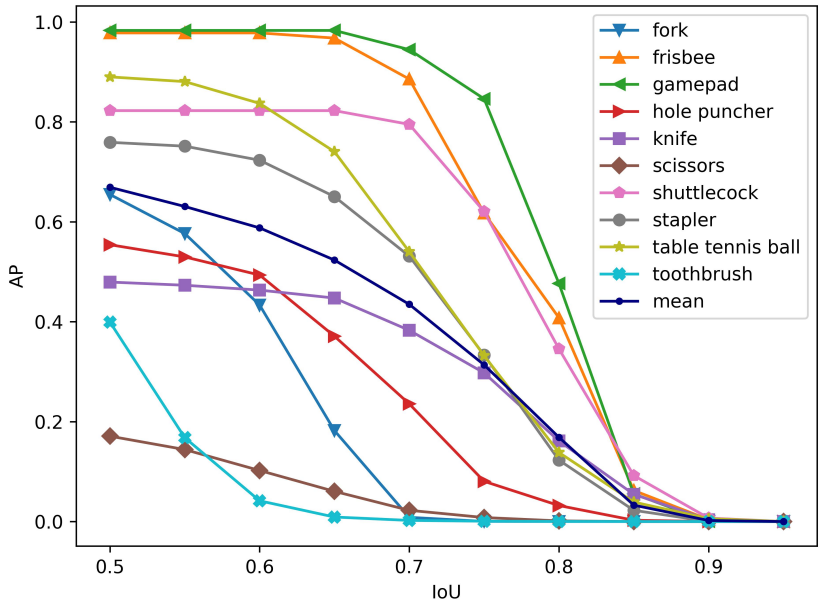
Figure B.4.5: Precision-recall curves at an IoU of 0.5 of (a) the baseline [5] and (b) our teaching approach using GA-GBVS.

at higher recall applying our method. Although the stapler class has deteriorated slightly, the overall improvement is particularly apparent for the knife, the toothbrush, and the table tennis ball. While, for example, in the latter the curve of the baseline decreases continuously beginning with a recall value of 0.3, our method still reaches a precision of almost 1 at a recall value of over 0.9. The knife and toothbrush reveal an even more pronounced improvement over the baseline, as both curves in Figure B.4.5(b) are much more concave and lie further above the corresponding baseline curves in Figure B.4.5(a).

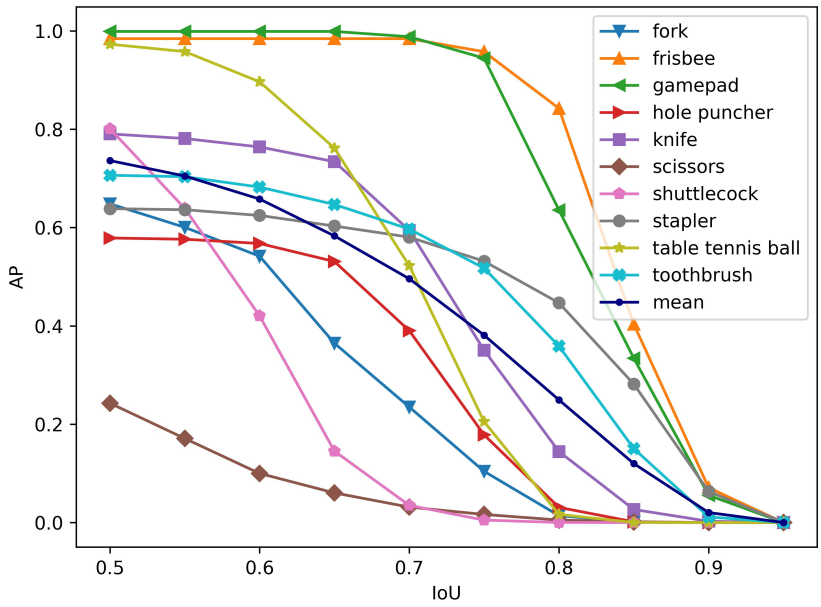
The same picture emerges when the average accuracy is considered as a function of the IoU. A comparison can be seen in Figure B.4.6. The results of the toothbrush show the most improvement. Whereas with the baseline the average precision drops steeply right at the beginning, our method demonstrates a slower, more gradual decline, remaining almost constant at first. In contrast, the shuttlecock has deteriorated, however, the overall result across all classes is still better. This is supported by the curve representing the mean of all classes, which proves that our method outperforms the baseline in terms of mean average accuracy at every IoU value. Specifically, our method achieves $mAP_{60} \approx 0.71$ and $mAP_{80} \approx 0.25$, respectively, while the baseline achieves only around 0.63 and 0.17 at the same IoU values.

Figure B.4.7 displays the recall as a function of the IoU. This comparison also shows that our method outperforms the baseline in terms of recall averaged across all classes. Although the mAR_{50} is approximately the same for both methods, the baseline values decrease faster as the IoU threshold increases. As an example, our method yields an mAR_{80} of approximately 0.39, while the baselines value is already 0.32 at the same IoU threshold.

In Table B.4.1 an extract of the detailed average precision values is listed. The results of both of our teaching variants, GA-GBVS and DGA-GBVS, are compared with the baseline approach [1]. Furthermore, in accordance with their evaluation methodology, we also compare our method with a model trained on the entire MS COCO dataset, that is, including the classes present in the OMD dataset. However, the authors have pointed out that this comparison has to be treated with caution and is only conditionally meaningful, as it involves fundamentally disparate starting points. On the one hand, the model that has been trained on the entire MS COCO dataset naturally has an advantage, as it has seen far more and also more diverse images. On the other hand, only six out of the ten OMD classes are part of MS COCO, which means that more images are distributed among fewer classes and, in addition, the four other classes cannot be detected at all. This lack of coverage entails that there is no flexibility to detect unknown objects. Nevertheless, this is exactly the task. The intention is to make it possible to detect all objects, irrespective of whether a corresponding data set is available, and to teach the unknown objects

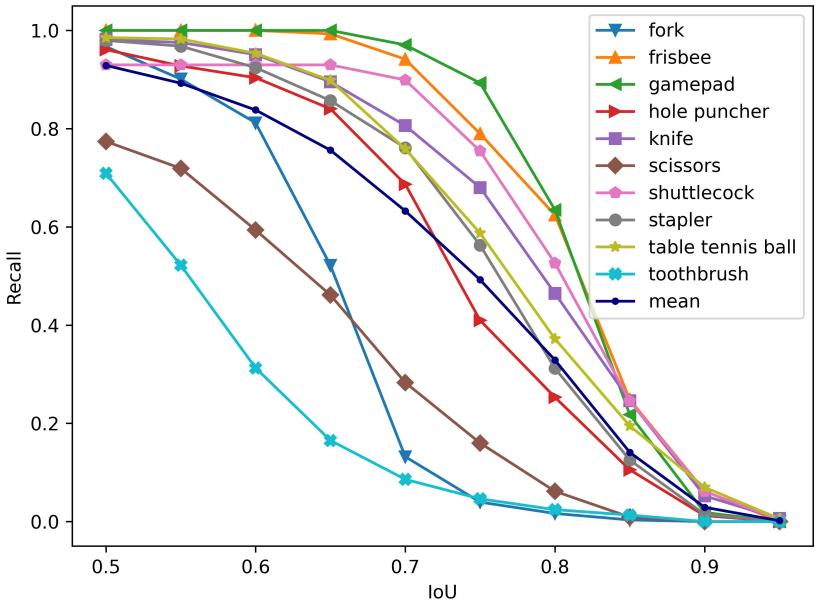


(a) Baseline

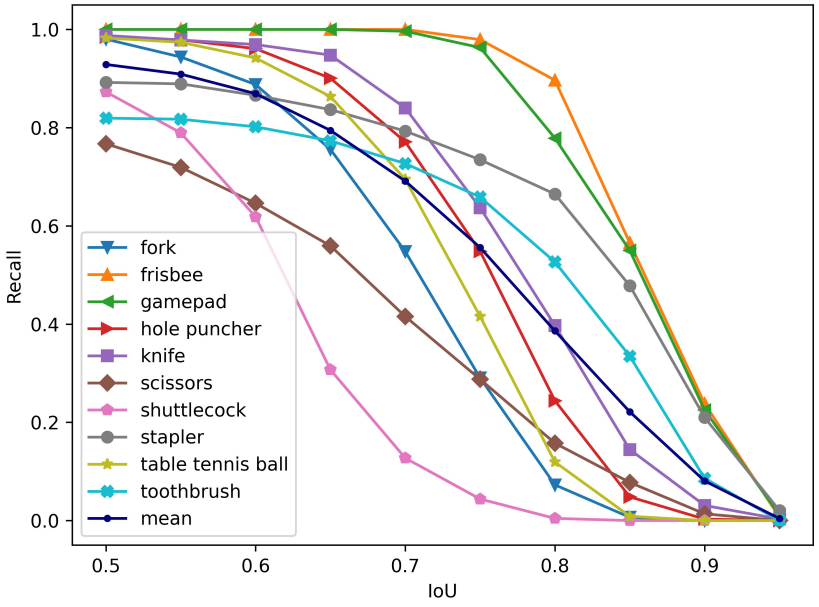


(b) GA-GBVS

Figure B.4.6: AP-IoU curves of (a) the baseline [5] and (b) our teaching approach using GA-GBVS.



(a) Baseline



(b) GA-GBVS

Figure B.4.7: Recall-IoU curves of (a) the baseline [5] and (b) our teaching approach using GA-GBVS.

Table B.4.1: The average precision on the OMD test set for the different training methods. The best values are printed in bold.

Class	AP_{50}				AP_{75}				AP			
	[5]	GA†	DGA‡	COCO	[5]	GA	DGA	COCO	[5]	GA	DGA	COCO
fork	65.5	64.8	68.7	**98.3**	0.0	10.4	2.3	**77.6**	18.6	25.1	21.8	**63.8**
frisbee	97.8	**98.4**	97.6	87.1	61.7	**95.8**	23.0	79.8	58.8	**72.0**	46.3	65.9
gamepad	98.3	**99.9**	99.5	0.0	84.6	**94.4**	32.3	0.0	62.6	**69.5**	49.0	0.0
hole puncher	55.4	57.9	**59.5**	0.0	8.0	**17.8**	0.1	0.0	23.0	**28.5**	18.0	0.0
knife	47.9	79.0	78.9	**87.9**	29.8	35.1	48.3	**50.7**	27.6	41.9	44.7	**50.0**
scissors	17.1	24.3	50.1	**94.9**	0.8	1.6	24.0	**87.1**	5.1	6.3	27.4	**70.9**
shuttlecock	**82.3**	80.1	32.8	0.0	**62.1**	0.5	0.0	0.0	**51.5**	20.5	5.2	0.0
stapler	**75.9**	63.8	61.6	0.0	33.3	**53.2**	52.3	0.0	38.9	**44.1**	41.0	0.0
table tennis ball	89.0	**97.3**	91.0	67.4	**33.2**	20.5	27.3	0.8	**44.0**	43.3	38.1	17.3
toothbrush	39.9	70.6	66.0	**70.9**	0.1	**51.7**	13.6	36.6	6.2	**43.7**	27.7	37.4
mean	66.9	**73.6**	70.5	50.7	31.4	**38.1**	22.3	33.3	33.6	**39.5**	31.9	30.5

† Gaze-Assisted GBVS (GA-GBVS)
‡ Dual Gaze-Assisted GBVS (DGA-GBVS)

ad hoc if necessary, expanding the present knowledge of the robot. In fact, this is precisely the strength of our method and what differentiates the starting points. Still, we want to consult COCO as an additional comparison to the baseline [5] to get a kind of "upper bound" and to assess which values are realistic in case a suitable dataset exists, is available and contains the relevant objects. The values for mAP and mAR, however, refer to all ten classes to reflect the result on the overall task.

Our method outperforms the state-of-the-art approach proposed by [5] across nearly all classes. For example, while the average accuracy at an IoU of 0.5, the threshold at which an object is considered to be detected [192, 193], slightly deteriorated for the shuttlecock with GA-GBVS compared to [5], the toothbrush showed significant improvements. In terms of the scissors, DGA-GBVS is superior to the more basic GA-GBVS approach. It is remarkable that our training method utilizing GA-GBVS has surpassed the AP values of MS COCO in three classes, which is especially noteworthy given that MS COCO considers AP to be the primary metric for evaluation [185]. Overall, our approach utilizing GA-GBVS outperforms all other alternatives across all IoU values in terms of the mean average precision (mAP). Moreover, a central aspect to emphasize is that teaching the robot via GA-GBVS from scratch yields also promising performance, as indicated by our achieved mAP_{50}=71.9, mAP_{75}=32.2, and mAP=34.6 scores. That means even without any pretraining our method is superior to the baseline.

With respect to the average recall, our method can also prevail. The results in Table B.4.2 demonstrate that we not only obtain a higher AR than the baseline for almost

Table B.4.2: The average recall on the OMD test set for the different training methods. The best values are printed in bold.

Class	AR_1				AR_{10}				AR_{100}			
	[5]	GA†	DGA‡	COCO	[5]	GA	DGA	COCO	[5]	GA	DGA	COCO
fork	28.4	40.4	34.3	**72.3**	33.9	44.9	36.9	**72.3**	33.9	44.9	36.9	**72.3**
frisbee	66.3	**76.7**	53.7	71.6	66.6	**76.9**	53.9	71.6	66.6	**76.9**	53.9	71.6
gamepad	67.3	**75.2**	57.9	0.0	67.3	**75.2**	58.1	0.0	67.3	**75.2**	58.1	0.0
hole puncher	34.2	**42.5**	26.7	0.0	49.2	**54.0**	37.2	0.0	51.0	**54.4**	37.6	0.0
knife	41.4	57.8	59.8	**63.2**	60.4	59.4	62.0	**63.4**	60.6	59.4	62.0	**63.4**
scissors	23.4	20.5	44.2	**74.8**	30.6	36.4	51.5	**75.4**	30.6	36.4	51.5	**75.4**
shuttlecock	**61.0**	27.1	9.6	0.0	**62.1**	27.6	14.8	0.0	**62.1**	27.6	15.0	0.0
stapler	45.6	**57.6**	53.8	0.0	54.0	**63.8**	60.3	0.0	55.0	**63.8**	60.3	0.0
table tennis ball	**56.1**	49.8	49.0	23.8	**58.1**	50.0	51.9	23.8	**58.1**	50.0	51.9	23.8
toothbrush	13.6	**54.0**	35.4	45.9	18.8	**55.4**	37.7	45.9	18.8	**55.4**	37.7	45.9
mean	43.7	**50.1**	42.4	35.2	50.1	**54.4**	46.4	35.2	50.4	**54.4**	46.5	35.2

† Gaze-Assisted GBVS (GA-GBVS)
‡ Dual Gaze-Assisted GBVS (DGA-GBVS)

every class regardless of the number of detections per image, but also partially exceed the "upper bound" of MS COCO. Eventually, our approach based on GA-GBVS also outperforms all other alternatives in terms of the mean average recall (mAR).

B.4.6 Limitations

While our approach does partially mitigate inaccuracies arising from factors such as gaze tracking, robot-to-arm calibration, and human-to-robot transformations, the extent of the ensuing deviations varies depending on the arm's position. This means that the offset between the projected gaze points on the recorded images is also varying, which in turn causes the quality of the final bounding boxes to differ slightly. Nonetheless, as our results support, the quality is sufficient for the model to learn the objects successfully.

B.4.7 Conclusion

In this paper, we introduced a novel technique enabling robots to adapt to unfamiliar environments, along with the objects contained therein, beyond the predefined ones. Rather than having the potential interaction entities dictated by fixed datasets, we expand existing knowledge in an ad hoc manner. Such approaches are inevitable in the long run, as the world is multifaceted and dynamic and the number of existing objects – at least for all practical purposes – is infinite, precluding the compilation of comprehensive datasets that cover them all. The existence of a suitable dataset that contains the required

classes and is also available can therefore not be inevitably assumed. For this reason, we tasked the human with the role of a teacher, educating the robot about unknown objects of interest. To this end, we transformed human gaze data acquired by means of an augmented reality interface into saliency-aware gaze heatmaps. This process involved two different gaze-assisted approaches, which eventually allowed the robot to precisely perceive the region of interest. Based on the class name provided by the human, the robot was capable of learning new objects in a flexible way. The results of our evaluation have shown that our proposed method is superior to the current state of the art in terms of commonly used object recognition metrics. This remains the case even if we omit the pretraining used by the baseline. Therefore, our findings suggest that our approach has the potential to significantly enhance the adaptability of robots to novel scenarios and objects. Altogether, we hope that our approach will offer new avenues for future research in human-robot interaction, leading to more dynamic and versatile robotic systems in non-predefined scenarios. Despite the progress made, a considerable amount of work remains to be addressed as we intend to conduct a user study to asses the usability of our system as well as to further broaden the interaction between human and robot in an active learning context.

Acknowledgment

This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645.

Bibliography

[1] Daniel Weber, Thiago Santini, Andreas Zell, and Enkelejda Kasneci. Distilling location proposals of unknown objects through gaze information for human-robot interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11086–11093. IEEE, 2020. doi:10.1109/IROS45743.2020.9340893.

[2] David Geisler, Daniel Weber, Nora Castner, and Enkelejda Kasneci. Exploiting the gbvs for saliency aware gaze heatmaps. In *ACM Symposium on Eye Tracking Research and Applications*, pages 1–5, 2020. doi:10.1145/3379156.3391367.

[3] Daniel Weber, Enkelejda Kasneci, and Andreas Zell. Exploiting augmented reality for extrinsic robot calibration and eye-based human-robot collaboration. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 284–293. IEEE, 2022. doi:10.1109/HRI53351.2022.9889538.

[4] Daniel Weber, Wolfgang Fuhl, Andreas Zell, and Enkelejda Kasneci. Gaze-based object detection in the wild. In *2022 Sixth IEEE International Conference on Robotic Computing (IRC)*, pages 62–66. IEEE, 2022. doi:10.1109/IRC55401.2022.00017.

[5] Daniel Weber, Wolfgang Fuhl, Enkelejda Kasneci, and Andreas Zell. Multiperspective teaching of unknown objects via shared-gaze-based multimodal human-robot interaction. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 544–553, March 2023. doi:10.1145/3568162.3578627.

[6] Daniel Weber, Valentin Bolz, Andreas Zell, and Enkelejda Kasneci. Leveraging saliency-aware gaze heatmaps for multiperspective teaching of unknown objects. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023. (Accepted for publication).

[7] Wolfgang Fuhl, Daniel Weber, and Shahram Eivazi. The gaze and mouse signal as additional source for user fingerprints in browser applications. In *Proceedings of*

the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP, pages 117–124. INSTICC, SciTePress, 2023. ISBN 978-989-758-634-7. doi:10.5220/0011607300003417. URL https://doi.org/10.5220/0011607300003417.

[8] Wolfgang Fuhl, Daniel Weber, and Shahram Eivazi. Pistol: Pupil invisible supportive tool to extract pupil, iris, eye opening, eye movements, pupil and iris gaze vector, and 2d as well as 3d gaze. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP*, pages 27–38. INSTICC, SciTePress, 2023. ISBN 978-989-758-634-7. doi:10.5220/0011607200003417. URL https://doi.org/10.5220/0011607200003417.

[9] Wolfgang Fuhl, Daniel Weber, and Shahram Eivazi. Groupgazer: A tool to compute the gaze per participant in groups with integrated calibration to map the gaze online to a screen or beamer projection. In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP*, pages 109–116. INSTICC, SciTePress, 2023. ISBN 978-989-758-634-7. doi:10.5220/0011607000003417. URL https://doi.org/10.5220/0011607000003417.

[10] Wolfgang Fuhl, Björn Severitt, Nora Castner, Babette Bühler, Johannes Meyer, Daniel Weber, Regine Lendway, Ruikun Hou, and Enkelejda Kasneci. Watch out for those bananas! gaze based mario kart performance classification. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–2. ACM, 2023. ISBN 9798400701504. doi:10.1145/3588015.3590136. URL https://doi.org/10.1145/3588015.3590136.

[11] Evana Gizzi, Mateo Guaman Castro, and Jivko Sinapov. Creative problem solving by robots using action primitive discovery. In *2019 Joint IEEE 9th International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pages 228–233. IEEE, 2019.

[12] John Carff, Matthew Johnson, Eman M El-Sheikh, and Jerry E Pratt. Human-robot team navigation in visually complex environments. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3043–3050. IEEE, 2009.

[13] Diana Löffler, Nina Schmidt, and Robert Tscharn. Multimodal expression of artificial emotion in social robots using color, motion and sound. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, pages 334–343, 2018.

[14] Tobias Werner, Dominik Riedelbauch, and Dominik Henrich. Design and evaluation of a multi-agent software architecture for risk-minimized path planning in human-robot workcells. In *Tagungsband des 2. Kongresses Montage Handhabung Industrieroboter*, pages 103–112. Springer, 2017.

[15] Adrian Boteanu, David Kent, Anahita Mohseni-Kabir, Charles Rich, and Sonia Chernova. Towards robot adaptability in new situations. In *2015 AAAI fall symposium series*, 2015.

[16] Haophuong21. https://commons.wikimedia.org/wiki/File:Robot-cong-nghiep-the-he-moi.jpg, 2021. Cropped from original. Licensed under CC BY-SA 4.0 (https://creativecommons.org/licenses/by-sa/4.0/deed.en). Accessed: 2023-11-30.

[17] Mbrickn (https://commons.wikimedia.org/wiki/User:Mbrickn). https://commons.wikimedia.org/wiki/File:Stuck_Starship_Robot.jpg, 2021. Licensed under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.en). Accessed: 2023-11-30.

[18] Jeremy A Marvel, Roger Bostelman, and Joe Falco. Multi-robot assembly strategies and metrics. *ACM Computing Surveys (CSUR)*, 51(1):1–32, 2018.

[19] Maria Kyrarini, Muhammad Abdul Haseeb, Danijela Ristić-Durrant, and Axel Gräser. Robot learning of industrial assembly task via human demonstrations. *Autonomous Robots*, 43:239–257, 2019.

[20] Ehsan Asadi, Bingbing Li, and I-Ming Chen. Pictobot: A cooperative painting robot for interior finishing of industrial developments. *IEEE Robotics & Automation Magazine*, 25(2):82–94, 2018.

[21] Binbin Zhang, Jun Wu, Liping Wang, and Zhenyang Yu. Accurate dynamic modeling and control parameters design of an industrial hybrid spray-painting robot. *Robotics and Computer-Integrated Manufacturing*, 63:101923, 2020.

[22] J Norberto Pires, Altino Loureiro, Tiago Godinho, Pedro Ferreira, Bruno Fernando, and Joel Morgado. Welding robots. *IEEE robotics & automation magazine*, 10(2): 45–55, 2003.

[23] J Norberto Pires, Altino Loureiro, and Gunnar Bölmsjo. *Welding robots: technology, system issues and application*. Springer Science & Business Media, 2006.

[24] Venketesh N Dubey and Jian S Dai. A packaging robot for complex cartons. *Industrial Robot: An International Journal*, 2006.

[25] Hyun Min Do, Chanhun Park, and Jin Ho Kyung. Dual arm robot for packaging and assembling of it products. In *2012 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 1067–1070. IEEE, 2012.

[26] Niko Herakovic. *Robot vision in industrial assembly and quality control processes*. INTECH Open Access Publisher London, UK, 2010.

[27] Andrzej Burghardt, Krzysztof Kurc, Dariusz Szybicki, Magdalena Muszyńska, and Tomasz Szczęch. Monitoring the parameters of the robot-operated quality control process. *Advances in Science and Technology. Research Journal*, 11(1), 2017.

[28] Meiyu Song and Shuo Xin. Robot autonomous sorting system for intelligent logistics. In *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pages 80–83. IEEE, 2021.

[29] Chang Liu, Ying Xu, Chunxia Tang, and Diansheng Chen. Rapid sorting robot system for e-commerce warehouse. In *2022 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1521–1525. IEEE, 2022.

[30] Maojia P Li, Prashant Sankaran, Michael E Kuhl, Raymond Ptucha, Amlan Ganguly, and Andres Kwasinski. Task selection by autonomous mobile robots in a warehouse using deep reinforcement learning. In *2019 Winter Simulation Conference (WSC)*, pages 680–689. IEEE, 2019.

[31] Lukas Polten and Simon Emde. Scheduling automated guided vehicles in very narrow aisle warehouses. *Omega*, 99:102204, 2021.

[32] Zheng Zhang, Juan Chen, and Qing Guo. Application of automated guided vehicles in smart automated warehouse systems: A survey. *CMES-COMPUTER MODELING IN ENGINEERING & SCIENCES*, 134(3):1529–1563, 2023.

[33] Purang Abolmaesumi, Septimiu E Salcudean, Wen-Hong Zhu, Simon P DiMaio, and Mohammad Reza Sirouspour. A user interface for robot-assisted diagnostic ultrasound. In *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164)*, volume 2, pages 1549–1554. IEEE, 2001.

[34] Siamak Najarian, Mehdi Fallahnezhad, and Ehsan Afshari. Advances in medical robotic systems with specific applications in surgery—a review. *Journal of medical engineering & technology*, 35(1):19–33, 2011.

[35] Jacob Rosen, Blake Hannaford, and Richard M Satava. *Surgical robotics: systems applications and visions*. Springer Science & Business Media, 2011.

[36] Brian S Peters, Priscila R Armijo, Crystal Krause, Songita A Choudhury, and Dmitry Oleynikov. Review of emerging surgical robotic technology. *Surgical endoscopy*, 32: 1636–1655, 2018.

[37] Hermano Igo Krebs, Jerome Joseph Palazzolo, Laura Dipietro, Mark Ferraro, Jennifer Krol, Keren Rannekleiv, Bruce T Volpe, and Neville Hogan. Rehabilitation robotics: Performance-based progressive robot-assisted therapy. *Autonomous robots*, 15(1): 7–20, 2003.

[38] Roberto Colombo and Vittorio Sanguineti. Rehabilitation robotics: technology and application. 2018.

[39] Frits K Van Evert, Gerie WAM Van Der Heijden, Lambertus AP Lotz, Gerrit Polder, Arjan Lamaker, Arjan De Jong, Marjolijn C Kuyper, Eltje JK Groendijk, Jacques J Neeteson, and Ton Van Der Zalm. A mobile field robot with vision-based detection of volunteer potato plants in a corn crop. *Weed Technology*, 20(4):853–861, 2006.

[40] Trygve Utstumo, Frode Urdal, Anders Brevik, Jarle Dørum, Jan Netland, Øyvind Overskeid, Therese W Berge, and Jan Tommy Gravdahl. Robotic in-row weed control in vegetables. *Computers and electronics in agriculture*, 154:36–45, 2018.

[41] S Gokul, R Dhiksith, S Ajith Sundaresh, and M Gopinath. Gesture controlled wireless agricultural weeding robot. In *2019 5th international conference on advanced computing & communication systems (ICACCS)*, pages 926–929. IEEE, 2019.

[42] Lin Haibo, Dong Shuliang, Liu Zunmin, and Yi Chuijie. Study and experiment on a wheat precision seeding robot. *Journal of Robotics*, 2015:12–12, 2015.

[43] Piyanun Ruangurai, Mongkol Ekpanyapong, Chatchai Pruetong, and Thaisiri Watewai. Automated three-wheel rice seeding robot operating in dry paddy fields. *Maejo International Journal of Science and Technology*, 9(3):403, 2015.

[44] Sai Kirthi Pilli, Bharathiraja Nallathambi, Smith Jessy George, and Vivek Diwanji. eagrobot—a robot for early crop disease detection using image processing. In *2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, pages 1684–1689. IEEE, 2015.

[45] Noa Schor, Avital Bechar, Timea Ignat, Aviv Dombrovsky, Yigal Elad, and Sigal Berman. Robotic disease detection in greenhouses: combined detection of powdery mildew and tomato spotted wilt virus. *IEEE Robotics and Automation Letters*, 1(1): 354–360, 2016.

[46] Noa Schor, Sigal Berman, Aviv Dombrovsky, Yigal Elad, Timea Ignat, and Avital Bechar. Development of a robotic detection system for greenhouse pepper plant diseases. *Precision agriculture*, 18:394–409, 2017.

[47] Beatriz Rey, Nuria Aleixos, Sergio Cubero, and José Blasco. Xf-rovim. a field robot to detect olive trees infected by xylella fastidiosa using proximal sensing. *Remote Sensing*, 11(3):221, 2019.

[48] Filipe Neves Dos Santos, Heber Miguel Placido Sobreira, Daniel Filipe Barros Campos, Raul Morais, Antonio Paulo Gomes Mendes Moreira, and Olga Maria Sousa Contente. Towards a reliable monitoring robot for mountain vineyards. In *2015 IEEE international conference on autonomous robot systems and competitions*, pages 37–43. IEEE, 2015.

[49] Marco Bietresato, Giovanni Carabin, Renato Vidoni, Alessandro Gasparetto, and Fabrizio Mazzetto. Evaluation of a lidar-based 3d-stereoscopic vision system for crop-monitoring applications. *Computers and Electronics in Agriculture*, 124:1–13, 2016.

[50] Renato Vidoni, Raimondo Gallo, Gianluca Ristorto, Giovanni Carabin, Fabrizio Mazzetto, Lorenzo Scalera, and Alessandro Gasparetto. Byelab: An agricultural mobile robot prototype for proximal sensing and precision farming. In *ASME International Mechanical Engineering Congress and Exposition*, volume 58370, page V04AT05A057. American Society of Mechanical Engineers, 2017.

[51] S Singh, TF Burks, and WS Lee. Autonomous robotic vehicle development for greenhouse spraying. *Transactions of the ASAE*, 48(6):2355–2361, 2005.

[52] Philip J Sammons, Tomonari Furukawa, and Andrew Bulgin. Autonomous pesticide spraying robot for use in a greenhouse. In *Australian Conference on Robotics and Automation*, volume 1, pages 1–9. Commonwealth Scientific and Industrial Research Organisation Canberra, Australia, 2005.

[53] Mohd Saiful Azimi Mahmud, Mohamad Shukri Zainal Abidin, Zaharuddin Mohamed, Muhammad Khairie Idham Abd Rahman, and Michihisa Iida. Multi-objective path planner for an agricultural mobile robot in a virtual greenhouse environment. *Computers and electronics in agriculture*, 157:488–499, 2019.

[54] Kanae Tanigaki, Tateshi Fujiura, Akira Akase, and Junichi Imagawa. Cherry-harvesting robot. *Computers and electronics in agriculture*, 63(1):65–72, 2008.

[55] Johan Baeten, Kevin Donné, Sven Boedrij, Wim Beckers, and Eric Claesen. Autonomous fruit picking machine: A robotic apple harvester. In *Field and service robotics: Results of the 6th international conference*, pages 531–539. Springer, 2008.

[56] DM Bulanon and T Kataoka. Fruit detection system and an end effector for robotic harvesting of fuji apples. *Agricultural Engineering International: CIGR Journal*, 12 (1), 2010.

[57] Qingchun Feng, Wei Zou, Pengfei Fan, Chunfeng Zhang, and Xiu Wang. Design and test of robotic harvesting system for cherry tomato. *International Journal of Agricultural and Biological Engineering*, 11(1):96–100, 2018.

[58] Gerd Hirzinger, Bernhard Brunner, Johannes Dietrich, and Johann Heindl. Rotex-the first remotely controlled robot in space. In *Proceedings of the 1994 IEEE international conference on robotics and automation*, pages 2604–2611. IEEE, 1994.

[59] Mark Yim, Kimon Roufas, David Duff, Ying Zhang, Craig Eldershaw, and Sam Homans. Modular reconfigurable robots in space applications. *Autonomous Robots*, 14:225–237, 2003.

[60] S Ali A Moosavian and Evangelos Papadopoulos. Free-flying robots in space: an overview of dynamics modeling, planning and control. *Robotica*, 25(5):537–547, 2007.

[61] Robert Bogue. Robots for space exploration. *Industrial Robot: An International Journal*, 2012.

[62] Paul E Rybski, Nikolaos P Papanikolopoulos, Sascha A Stoeter, Donald G Krantz, Kemal B Yesin, Maria Gini, Richard Voyles, Dean F Hougen, Brad Nelson, and Michael D Erickson. Enlisting rangers and scouts for reconnaissance and surveillance. *IEEE Robotics & Automation Magazine*, 7(4):14–24, 2000.

[63] Zubair Ghouse, Nishika Hiwrale, and Nihar Ranjan. Military robot for reconnaissance and surveillance using image processing. *International Research Journal of Engineering and Technology*, 4(5):2395–0072, 2017.

[64] Minal S Ghute, Kanchan P Kamble, and Mridul Korde. Design of military surveillance robot. In *2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC)*, pages 270–272. IEEE, 2018.

[65] Brian Chemel, Edward Mutschler, and Hagen Schempf. Cyclops: Miniature robotic reconnaissance system. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, volume 3, pages 2298–2302. IEEE, 1999.

[66] Danna Voth. A new generation of military robots. *IEEE Intelligent Systems*, 19(4): 2–3, 2004.

[67] Mrs NS Usha, S Priyadharshini, K Rohinee Shree, P Sabari Devi, and G Sangeetha. Military reconnaissance robot. *International Journal of Advanced Engineering Research and Science*, 4(2):237036, 2017.

[68] Byunghun Choi, Wonsuk Lee, Gyuhyun Park, Youngwoo Lee, Jihong Min, and Seongil Hong. Development and control of a military rescue robot for casualty extraction task. *Journal of Field Robotics*, 36(4):656–676, 2019.

[69] Rakshana Mohamed Ismail, Senthil Muthukumaraswamy, and A Sasikala. Military support and rescue robot. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 156–162. IEEE, 2020.

[70] Rakshana Ismail and Senthil Muthukumaraswamy. Military reconnaissance and rescue robot with real-time object detection. In *Intelligent Manufacturing and Energy Sustainability: Proceedings of ICIMES 2020*, pages 637–648. Springer, 2021.

[71] Kenzo Nonami, Qingjiu Huang, Daisuke Komizo, Yoichiro Fukao, Yoshitomo Asai, Yoshinori Shiraishi, Masaki Fujimoto, and Yosuke Ikedo. Development and control of mine detection robot comet-ii and comet-iii. *JSME International Journal Series C Mechanical Systems, Machine Elements and Manufacturing*, 46(3):881–890, 2003.

[72] Hajime Aoyama, Kazuyoshi Ishikawa, Junya Seki, Mitsuo Okamura, Saori Ishimura, and Yuichi Satsumi. Development of mine detection robot system. *International Journal of Advanced Robotic Systems*, 4(2):25, 2007.

[73] Waqar Farooq, Nehal Butt, Sameed Shukat, Nouman Ali Baig, and Sheikh Muhammad Ahmed. Wirelessly controlled mines detection robot. In *2016 International Conference on Intelligent Systems Engineering (ICISE)*, pages 55–62. IEEE, 2016.

[74] Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. Seadronessee: A maritime benchmark for detecting humans in open water. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2260–2270, 2022.

[75] Min-Chie Chiu, Long-Jyi Yeh, and YC Lin. The design and application of a robotic vacuum cleaner. *Journal of Information and Optimization Sciences*, 30(1):39–62, 2009.

[76] Florian Vaussard, Julia Fink, Valerie Bauwens, Philippe Rétornaz, David Hamel, Pierre Dillenbourg, and Francesco Mondada. Lessons learned from robotic vacuum cleaners entering the home ecosystem. *Robotics and Autonomous Systems*, 62(3): 376–391, 2014.

[77] Mun-Cheon Kang, Kwang-Shik Kim, Dong-Ki Noh, Jong-Woo Han, and Sung-Jea Ko. A robust obstacle detection method for robotic vacuum cleaners. *IEEE Transactions on Consumer Electronics*, 60(4):587–595, 2014.

[78] Rob Warren Hicks II and Ernest L Hall. Survey of robot lawn mowers. In *Intelligent Robots and Computer Vision XIX: Algorithms, Techniques, and Active Vision*, volume 4197, pages 262–269. SPIE, 2000.

[79] Guri B Verne. Adapting to a robot: Adapting gardening and the garden to fit a robot lawn mower. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 34–42, 2020.

[80] Juinne-Ching Liao, Shun-Hsing Chen, Zi-Yi Zhuang, Bo-Wei Wu, and Yu-Jen Chen. Designing and manufacturing of automatic robotic lawn mower. *Processes*, 9(2): 358, 2021.

[81] Lucile Bechade, Guillaume Dubuisson-Duplessis, Gabrielle Pittaro, Mélanie Garcia, and Laurence Devillers. Towards metrics of evaluation of pepper robot as a social companion for the elderly. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, pages 89–101. Springer, 2019.

[82] Moojan Ghafurian, Colin Ellard, and Kerstin Dautenhahn. Social companion robots to reduce isolation: A perception change due to covid-19. In *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part II 18*, pages 43–63. Springer, 2021.

[83] Albert Bandura. Observational learning. *The international encyclopedia of communication*, 2008.

[84] Jacob Feldman. The simplicity principle in human concept learning. *Current directions in psychological science*, 12(6):227–232, 2003.

[85] H Peyton Young. Learning by trial and error. *Games and economic behavior*, 65(2): 626–643, 2009.

[86] William K Balzer, Michael E Doherty, et al. Effects of cognitive feedback on performance. *Psychological bulletin*, 106(3):410, 1989.

[87] Eric A Hanushek and Steven G Rivkin. Teacher quality. *Handbook of the Economics of Education*, 2:1051–1078, 2006.

[88] UKM Elisabeth Deiters-Keul. https://commons.wikimedia.org/wiki/File:Kommissionierroboter_in_der_Apotheke_der_Universit%C3%A4tsklinik_M%C3%BCnster.jpg, 2015. Cropped from original. Licensed under CC BY-SA 3.0 (https://creativecommons.org/licenses/by-sa/3.0/deed.en). Accessed: 2023-11-30.

[89] Tevel Aerobotics Technologies. https://www.tevel-tech.com/. URL https://www.tevel-tech.com/wp-content/uploads/2022/10/Tevel-Aerobotics-FAR-1-1-1024x768.jpg. Cropped from original. Photo courtesy of Tevel Aerobotics Technologies. Accessed: 2023-08-07.

[90] Marjorie Skubic and Richard A Volz. Acquiring robust, force-based assembly skills from human demonstration. *IEEE Transactions on Robotics and Automation*, 16(6): 772–781, 2000.

[91] Rüdiger Dillmann. Teaching and learning of robot tasks via observation of human performance. *Robotics and Autonomous Systems*, 47(2-3):109–116, 2004.

[92] Weitian Wang, Rui Li, Yi Chen, Z Max Diekel, and Yunyi Jia. Facilitating human–robot collaborative tasks by teaching-learning-collaboration from human demonstrations. *IEEE Transactions on Automation Science and Engineering*, 16(2):640–653, 2018.

[93] Zuyuan Zhu and Huosheng Hu. Robot learning from demonstration in robotic assembly: A survey. *Robotics*, 7(2):17, 2018.

[94] Jacopo Aleotti, Stefano Caselli, and Monica Reggiani. Leveraging on a virtual environment for robot programming by demonstration. *Robotics and Autonomous Systems*, 47(2-3):153–161, 2004.

[95] Sylvain Calinon and Aude Billard. Active teaching in robot programming by demonstration. In *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*, pages 702–707. IEEE, 2007.

[96] Sonya Alexandrova, Maya Cakmak, Kaijen Hsiao, and Leila Takayama. Robot programming by demonstration with interactive action visualizations. In *Robotics: science and systems*, pages 1–9, 2014.

[97] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems*, 26, 2013.

[98] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[99] Stanislao Lauria, Guido Bugmann, Theocharis Kyriacou, Johan Bos, and A Klein. Training personal robots using natural language instruction. *IEEE Intelligent systems*, 16(5):38–45, 2001.

[100] MA Viraj J Muthugala and AG Buddhika P Jayasekara. A review of service robots coping with uncertain information in natural language instructions. *IEEE Access*, 6: 12913–12928, 2018.

[101] Lanbo She, Yu Cheng, Joyce Y Chai, Yunyi Jia, Shaohua Yang, and Ning Xi. Teaching robots new actions through natural language instructions. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 868–873. IEEE, 2014.

[102] Juan Fasola and Maja J Mataric. Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 143–150. IEEE, 2013.

[103] Thomas Pettersen, John Pretlove, Charlotte Skourup, Torbjorn Engedal, and T Lokstad. Augmented reality for programming industrial robots. In *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, pages 319–320. IEEE, 2003.

[104] Hangxin Liu, Yaofang Zhang, Wenwen Si, Xu Xie, Yixin Zhu, and Song-Chun Zhu. Interactive robot knowledge patching using augmented reality. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1947–1954. IEEE, 2018.

[105] F Tahriri, M Mousavi, HJ Yap, MD Siti Zawiah, and Z Taha. Optimizing the robot arm movement time using virtual reality robotic teaching system. *International Journal of Simulation Modelling*, 14(1):28–38, 2015.

[106] Vicente Román-Ibáñez, Francisco A Pujol-López, Higinio Mora-Mora, Maria Luisa Pertegal-Felices, and Antonio Jimeno-Morenilla. A low-cost immersive virtual reality system for teaching robotic manipulators programming. *Sustainability*, 10 (4):1102, 2018.

[107] J Gregory Trafton, Alan C Schultz, Dennis Perznowski, Magdalena D Bugajska, William Adams, Nicholas L Cassimatis, and Derek P Brock. Children and robots learning to play hide and seek. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 242–249, 2006.

[108] Andrea Lockerd and Cynthia Breazeal. Tutelage and socially guided robot learning. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, volume 4, pages 3475–3480. IEEE, 2004.

[109] Yang Xu, Chenguang Yang, Junpei Zhong, Hongbin Ma, Lijun Zhao, and Min Wang. Robot teaching by teleoperation based on visual interaction and neural network learning. In *2017 9th International Conference on Modelling, Identification and Control (ICMIC)*, pages 1068–1073. IEEE, 2017.

[110] Scott A Green, Mark Billinghurst, XiaoQi Chen, and J Geoffrey Chase. Human-robot collaboration: A literature review and augmented reality approach in design. *International journal of advanced robotic systems*, 5(1):1, 2008.

[111] Scott A Green, J Geoffrey Chase, XiaoQi Chen, and Mark Billinghurst. Evaluating the augmented reality human-robot collaboration system. *International journal of intelligent systems technologies and applications*, 8(1-4):130–143, 2010.

[112] Ryo Suzuki, Adnan Karim, Tian Xia, Hooman Hedayati, and Nicolai Marquardt. Augmented reality and robotics: A survey and taxonomy for ar-enhanced human-robot interaction and robotic interfaces. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–33, 2022.

[113] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017.

[114] Reuben M. Aronson, Thiago Santini, Thomas C. Kübler, Enkelejda Kasneci, Siddhartha Srinivasa, and Henny Admoni. Eye-hand behavior in human-robot

shared manipulation. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, HRI '18, pages 4–13, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-4953-6. doi:10.1145/3171221.3171287. URL http://doi.acm.org/10.1145/3171221.3171287.

[115] Lorenzo Scalera, Stefano Seriani, Paolo Gallina, Mattia Lentini, and Alessandro Gasparetto. Human–robot interaction through eye tracking for artistic drawing. *Robotics*, 10(2):54, 2021.

[116] Ronald T Azuma. A survey of augmented reality. *Presence: teleoperators & virtual environments*, 6(4):355–385, 1997.

[117] David W Roberts, John W Strohbehn, John F Hatch, William Murray, and Hans Kettenberger. A frameless stereotaxic integration of computerized tomographic imaging and the operating microscope. *Journal of neurosurgery*, 65(4):545–549, 1986.

[118] Michael Bajura, Henry Fuchs, and Ryutarou Ohbuchi. Merging virtual objects with the real world: Seeing ultrasound imagery within the patient. *ACM SIGGRAPH Computer Graphics*, 26(2):203–210, 1992.

[119] Tobias Sielhorst, Marco Feuerstein, and Nassir Navab. Advanced medical displays: A literature review of augmented reality. *Journal of Display Technology*, 4(4):451–467, 2008.

[120] Rong Wen, Wei-Liang Tay, Binh P Nguyen, Chin-Boon Chng, and Chee-Kong Chui. Hand gesture guided robot-assisted surgery based on a direct augmented reality interface. *Computer methods and programs in biomedicine*, 116(2):68–80, 2014.

[121] Simon Julier, Marco Lanzagorta, Yohan Baillot, Lawrence Rosenblum, Steven Feiner, Tobias Hollerer, and Sabrina Sestito. Information filtering for mobile augmented reality. In *Proceedings IEEE and ACM International Symposium on Augmented Reality (ISAR 2000)*, pages 3–11. IEEE, 2000.

[122] Mark A Livingston, Lawrence J Rosenblum, Simon J Julier, Dennis Brown, Yohan Baillot, II Swan, Joseph L Gabbard, Deborah Hix, et al. An augmented reality system for military operations in urban terrain. 2002.

[123] Eric Foxlin, Yury Altshuler, Leonid Naimark, and Michael Harrington. Flighttracker: A novel optical/inertial tracker for cockpit enhanced vision. In *Third IEEE and ACM*

International Symposium on Mixed and Augmented Reality, pages 212–221. IEEE, 2004.

[124] G Reinhart and C Patron. Integrating augmented reality in the assembly domain-fundamentals, benefits and applications. *CIRP Annals*, 52(1):5–8, 2003.

[125] Arthur Tang, Charles Owen, Frank Biocca, and Weimin Mou. Comparative effectiveness of augmented reality in object assembly. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 73–80, 2003.

[126] Ze-Hao Lai, Wenjin Tao, Ming C Leu, and Zhaozheng Yin. Smart augmented reality instructional system for mechanical assembly towards worker-centered intelligent manufacturing. *Journal of Manufacturing Systems*, 55:69–81, 2020.

[127] Rick Cavallaro. The foxtrax hockey puck tracking system. *IEEE Computer Graphics and Applications*, 17(2):6–12, 1997.

[128] Ronald Azuma, Howard Neely, Mike Daily, and Jon Leonard. Performance analysis of an outdoor augmented reality tracking system that relies upon a few mobile beacons. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 101–104. IEEE, 2006.

[129] Rick Cavallaro, Maria Hybinette, Marvin White, and Tucker Balch. Augmenting live broadcast sports with 3d tracking information. *IEEE MultiMedia*, 18(4):38, 2011.

[130] Steve Chi-Yin Yuen, Gallayanee Yaoyuneyong, and Erik Johnson. Augmented reality: An overview and five directions for ar in education. *Journal of Educational Technology Development and Exchange (JETDE)*, 4(1):11, 2011.

[131] Mark Billinghurst and Andreas Duenser. Augmented reality in the classroom. *Computer*, 45(7):56–63, 2012.

[132] Mehmet Kesim and Yasin Ozarslan. Augmented reality in education: current technologies and the potential for education. *Procedia-social and behavioral sciences*, 47:297–302, 2012.

[133] Hsin-Yi Chang, Hsin-Kai Wu, and Ying-Shao Hsu. Integrating a mobile augmented reality activity to contextualize student learning of a socioscientific issue. *British Journal of Educational Technology*, 44(3), 2013.

[134] Jun Rekimoto. Navicam: A magnifying glass approach to augmented reality. *Presence: Teleoperators & Virtual Environments*, 6(4):399–412, 1997.

[135] Wolfgang Narzt, Gustav Pomberger, Alois Ferscha, Dieter Kolb, Reiner Müller, Jan Wieghardt, Horst Hörtner, and Christopher Lindinger. Augmented reality navigation systems. *Universal Access in the Information Society*, 4:177–187, 2006.

[136] Chee Oh Chung, Yilun He, and Hoe Kyung Jung. Augmented reality navigation system on android. *International Journal of Electrical and Computer Engineering*, 6 (1):406, 2016.

[137] Vassilios Vlahakis, M Ioannidis, John Karigiannis, Manolis Tsotros, Michael Gounaris, Didier Stricker, Tim Gleue, Patrick Daehne, and Luis Almeida. Archeoguide: an augmented reality guide for archaeological sites. *IEEE Computer Graphics and Applications*, 22(5):52–60, 2002.

[138] David Ingram. Trust-based filtering for augmented reality. In *Trust Management: First International Conference, iTrust 2003 Heraklion, Crete, Greece, May 28–30, 2003 Proceedings 1*, pages 108–122. Springer, 2003.

[139] Margarita Martínez and Guadalupe Muñoz. Designing augmented interfaces for guided tours using multimedia sketches. In *MIXER*. Citeseer, 2004.

[140] Fabian Fritz, Ana Susperregui, and Maria Teresa Linaza. Enhancing cultural tourism experiences with augmented reality technologies. 6th International Symposium on Virtual Reality, Archaeology and Cultural . . . , 2005.

[141] THJ Collett and Bruce A MacDonald. Developer oriented visualisation of a robot program. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*, pages 49–56, 2006.

[142] Enkelejda Tafaj, Gjergji Kasneci, Wolfgang Rosenstiel, and Martin Bogdan. Bayesian online clustering of eye movement data. In *Proceedings of the symposium on eye tracking research and applications*, pages 285–288, 2012.

[143] Thiago Santini, Wolfgang Fuhl, Thomas Kübler, and Enkelejda Kasneci. Bayesian identification of fixations, saccades, and smooth pursuits. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 163–170, 2016.

[144] Shahram Eivazi, Ahmad Hafez, Wolfgang Fuhl, Hoorieh Afkari, Enkelejda Kasneci, Martin Lehecka, and Roman Bednarik. Optimal eye movement strategies: a comparison of neurosurgeons gaze patterns when using a surgical microscope. *Acta neurochirurgica*, 159:959–966, 2017.

[145] Thomas C Kübler, Enkelejda Kasneci, and Wolfgang Rosenstiel. Subsmatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 319–322, 2014.

[146] Maria Laura Mele and Stefano Federici. Gaze and eye-tracking solutions for psychological research. *Cognitive processing*, 13:261–265, 2012.

[147] Jacob L Orquin and Kenneth Holmqvist. Threats to the validity of eye-movement research in psychology. *Behavior research methods*, 50:1645–1656, 2018.

[148] Rima-Maria Rahal and Susann Fiedler. Understanding cognitive and affective mechanisms in social psychology through eye-tracking. *Journal of Experimental Social Psychology*, 85:103842, 2019.

[149] Renê de Oliveira Joaquim dos Santos, Jorge Henrique Caldeira de Oliveira, Jéssica Bonaretto Rocha, and Janaina de Moura Engracia Giraldi. Eye tracking in neuromarketing: a research agenda for marketing studies. *International journal of psychological studies*, 7(1):32, 2015.

[150] H Zamani, A Abas, and MKM Amin. Eye tracking application on emotion analysis for marketing strategy. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 8(11):87–91, 2016.

[151] Michel Wedel and Rik Pieters. A review of eye-tracking research in marketing. *Review of marketing research*, pages 123–147, 2017.

[152] Wen-Bing Horng, Chih-Yuan Chen, Yi Chang, and Chun-Hai Fan. Driver fatigue detection based on eye tracking and dynamic template matching. In *IEEE International Conference on Networking, Sensing and Control, 2004*, volume 1, pages 7–12. IEEE, 2004.

[153] Mandalapu Sarada Devi and Preeti R Bajaj. Driver fatigue detection based on eye tracking. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 649–652. IEEE, 2008.

[154] Hardeep Singh, Jagjit Singh Bhatia, and Jasbir Kaur. Eye tracking based driver fatigue monitoring and warning system. In *India International Conference on Power Electronics 2010 (IICPE2010)*, pages 1–6. IEEE, 2011.

[155] Enkelejda Tafaj, Thomas C Kübler, Gjergji Kasneci, Wolfgang Rosenstiel, and Martin Bogdan. Online classification of eye tracking data for automated analysis of traffic hazard perception. In *Artificial Neural Networks and Machine Learning–ICANN 2013: 23rd International Conference on Artificial Neural Networks Sofia, Bulgaria, September 10-13, 2013. Proceedings 23*, pages 442–450. Springer, 2013.

[156] Christian Braunagel, Enkelejda Kasneci, Wolfgang Stolzmann, and Wolfgang Rosenstiel. Driver-activity recognition in the context of conditionally autonomous driving. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 1652–1657. IEEE, 2015.

[157] Christian Braunagel, Wolfgang Rosenstiel, and Enkelejda Kasneci. Ready for take-over? a new driver assistance system for an automated classification of driver take-over readiness. *IEEE Intelligent Transportation Systems Magazine*, 9(4):10–22, 2017.

[158] Muhammad Qasim Khan and Sukhan Lee. Gaze and eye tracking: Techniques and applications in adas. *Sensors*, 19(24):5540, 2019.

[159] Efe Bozkir, David Geisler, and Enkelejda Kasneci. Assessment of driver attention during a safety critical situation in vr to generate vr-based training. In *ACM Symposium on Applied Perception 2019*, pages 1–5, 2019.

[160] David B Carr and Prateek Grover. The role of eye tracking technology in assessing older driver safety. *Geriatrics*, 5(2):36, 2020.

[161] André Roca, Paul R Ford, Allistair P McRobert, and A Mark Williams. Perceptual-cognitive skills and their interaction as a function of task constraints in soccer. *Journal of Sport and Exercise Psychology*, 35(2):144–155, 2013.

[162] Derek Panchuk, Samuel Vine, and Joan N Vickers. Eye tracking methods in sport expertise. In *Routledge handbook of sport expertise*, pages 176–187. Routledge, 2015.

[163] Stephanie Hüttermann, Benjamin Noël, and Daniel Memmert. Eye tracking in high-performance sports: Evaluation of its application in expert athletes. *International Journal of Computer Science in Sport*, 17(2):182–203, 2018.

[164] Benedikt Hosp, Florian Schultz, Enkelejda Kasneci, and Oliver Höner. Expertise classification of soccer goalkeepers in highly dynamic decision tasks: a deep learning approach for temporal and spatial feature recognition of fixation image patch sequences. *Frontiers in Sports and Active Living*, 3:692526, 2021.

[165] Teresa Busjahn, Carsten Schulte, Bonita Sharif, Andrew Begel, Michael Hansen, Roman Bednarik, Paul Orlov, Petri Ihantola, Galina Shchekotova, and Maria Antropova. Eye tracking in computing education. In *Proceedings of the tenth annual conference on International computing education research*, pages 3–10, 2014.

[166] Hajra Ashraf, Mikael H Sodergren, Nabeel Merali, George Mylonas, Harsimrat Singh, and Ara Darzi. Eye-tracking technology in medical education: A systematic review. *Medical teacher*, 40(1):62–69, 2018.

[167] Omer Sumer, Patricia Goldberg, Kathleen Sturmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Teachers' perception in the classroom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2315–2324, 2018.

[168] Anselm R Strohmaier, Kelsey J MacKay, Andreas Obersteiner, and Kristina M Reiss. Eye-tracking methodology in mathematics education research: A systematic literature review. *Educational Studies in Mathematics*, 104:147–200, 2020.

[169] Patricia Goldberg, Ömer Sümer, Kathleen Stürmer, Wolfgang Wagner, Richard Göllner, Peter Gerjets, Enkelejda Kasneci, and Ulrich Trautwein. Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, 33:27–49, 2021.

[170] Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 2021.

[171] L Hahn and Pascal Klein. Eye tracking in physics education research: A systematic literature review. *Physical Review Physics Education Research*, 18(1):013102, 2022.

[172] Hong Gao, Lisa Hasenbein, Efe Bozkir, Richard Göllner, and Enkelejda Kasneci. Evaluating the effects of virtual human animation on students in an immersive vr classroom using eye movements. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*, pages 1–11, 2022.

[173] Sean Anthony Byrne, Nora Castner, Ard Kastrati, Martyna Plomecka, William Schaefer, Enkelejda Kasneci, and Zoya Bylinskii. Leveraging eye tracking in digital classrooms: A step towards multimodal model for learning assistance. In *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, pages 1–6, 2023.

[174] Mark R Wilson, Samuel J Vine, Elizabeth Bright, Rich SW Masters, David Defriend, and John S McGrath. Gaze training enhances laparoscopic technical skill acquisition and multi-tasking performance: a randomized, controlled study. *Surgical endoscopy*, 25:3731–3739, 2011.

[175] Cristina Almansa, Muhammad W Shahid, Michael G Heckman, Susan Preissler, and Michael B Wallace. Association between visual gaze patterns and adenoma detection rate during colonoscopy: a preliminary investigation. *Official journal of the American College of Gastroenterology| ACG*, 106(6):1070–1074, 2011.

[176] Sophie Voisin, Frank Pinto, Songhua Xu, Garnetta Morin-Ducote, Kathy Hudson, and Georgia D Tourassi. Investigating the association of eye gaze pattern and diagnostic error in mammography. In *Medical Imaging 2013: Image Perception, Observer Performance, and Technology Assessment*, volume 8673, page 867302. SPIE, 2013.

[177] Nora Castner, Thomas C Kuebler, Katharina Scheiter, Juliane Richter, Thérése Eder, Fabian Hüttig, Constanze Keutel, and Enkelejda Kasneci. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In *ACM symposium on eye tracking research and applications*, pages 1–10, 2020.

[178] Benedikt Hosp, Myat Su Yin, Peter Haddawy, Ratthapoom Watcharopas, Paphon Sa-Ngasoongsong, and Enkelejda Kasneci. Differentiating surgeons' expertise solely by eye movement features. In *Companion Publication of the 2021 International Conference on Multimodal Interaction*, pages 371–375, 2021.

[179] Benedikt Hosp, Myat Su Yin, Peter Haddawy, Ratthaphum Watcharopas, Paphon Sa-Ngasoongsong, and Enkelejda Kasneci. States of confusion: Eye and head tracking reveal surgeons' confusion during arthroscopic surgery. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 753–757, 2021.

[180] Lubaina T Arsiwala-Scheppach, Nora Castner, Csaba Rohrer, Sarah Mertens, Enkelejda Kasneci, Jose Eduardo Cejudo Grano de Oro, Joachim Krois, and Falk Schwendicke. Gaze patterns of dentists while evaluating bitewing radiographs. *Journal of Dentistry*, page 104585, 2023.

[181] Chien-Ming Huang and Bilge Mutlu. Anticipatory robot control for efficient human-robot collaboration. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 83–90. IEEE, 2016.

[182] Yann-Seing Law-Kam Cio, Maxime Raison, Cedric Leblond Menard, and Sofiane Achiche. Proof of concept of an assistive robotic arm control using artificial stereovision and eye-tracking. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(12):2344–2352, 2019.

[183] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5. Kobe, Japan, 2009.

[184] Martin Bischoff. ROS#, June 2019. URL https://github.com/siemens/ros-sharp.

[185] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[186] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015.

[187] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[188] Glenn Jocher et al. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, February 2022. URL https://doi.org/10.5281/zenodo.6222936.

[189] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. *Advances in neural information processing systems*, 19, 2006.

[190] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[191] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully convolutional one-stage object detection. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[192] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[193] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014.

[194] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[195] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.

[196] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1742–1749. IEEE, 2019.

[197] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[198] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[199] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[200] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

[201] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021.

[202] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022.

[203] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[204] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[205] Jonathan Harel, C Koch, and P Perona. A saliency implementation in matlab. *URL: http://www.klab.caltech.edu/~harel/share/gbvs.php*, 2006.

[206] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[207] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023.

[208] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[209] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[210] Christian Braunagel, David Geisler, Wolfgang Stolzmann, Wolfgang Rosenstiel, and Enkelejda Kasneci. On the necessity of adaptive eye movement classification in conditionally automated driving scenarios. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 19–26, 2016.

[211] Katarzyna Harezlak and Pawel Kasprowski. Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, 65:176–190, 2018.

[212] Benedikt W Hosp, Florian Schultz, Oliver Höner, and Enkelejda Kasneci. Soccer goalkeeper expertise identification based on eye movements. *PloS one*, 16(5): e0251070, 2021.

[213] Karthikeyan Shanmuga Vadivel, Thuyen Ngo, Miguel Eckstein, and BS Manjunath. Eye tracking assisted extraction of attentionally important objects from videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3241–3250, 2015.

[214] Jeffrey Mahler, Matthew Matl, Vishal Satish, Michael Danielczuk, Bill DeRose, Stephen McKinley, and Ken Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26), 2019.

[215] Fen Xiao, Liangchan Peng, Lei Fu, and Xieping Gao. Salient object detection based on eye tracking data. *Signal Processing*, 144:392–397, 2018.

[216] Ran Shi, Ngi King Ngan, and Hongliang Li. Gaze-based object segmentation. *IEEE Signal Processing Letters*, 24(10):1493–1497, 2017.

[217] Xiaoxue Luo, Junjie Shen, Hong Zeng, Aiguo Song, Baoguo Xu, Huijun Li, Pengcheng Wen, and Cong Hu. Interested object detection based on gaze using low-cost remote eye tracker. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 1101–1104. IEEE, 2019.

[218] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020.

[219] Jiatong Bao, Yunyi Jia, Yu Cheng, and Ning Xi. Saliency-guided detection of unknown objects in rgb-d indoor scenes. *Sensors*, 15(9):21054–21074, 2015.

[220] Wolfgang Fuhl, Nikolai Sanamrad, and Enkelejda Kasneci. The gaze and mouse signal as additional source for user fingerprints in browser applications. *arXiv preprint arXiv:2101.03793*, 2021.

[221] Wolfgang Fuhl, Daniel Weber, and Enkelejda Kasneci. Pistol: Pupil invisible supportive tool to extract pupil, iris, eye opening, eye movements, pupil and iris gaze vector, and 2d as well as 3d gaze. *arXiv preprint arXiv:2201.06799*, 01 2022.

[222] DarkLabel. https://github.com/darkpgmr/DarkLabel, 2021. Accessed: 2021-12-07.

[223] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[224] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. SSD: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[225] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for

mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019.

[226] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[227] Daan R van Renswoude, Ingmar Visser, Maartje EJ Raijmakers, Tawny Tsang, and Scott P Johnson. Real-world scene perception in infants: What factors guide attention allocation? *Infancy*, 24(5):693–717, 2019.

[228] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[229] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000.

[230] Iuliia Kotseruba and John K Tsotsos. 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, pages 1–78, 2018.

[231] Qi Zhao and Christof Koch. Learning saliency-based visual attention: A review. *Signal Processing*, 93(6):1401–1407, 2013.

[232] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. CalibMe: Fast and unsupervised eye tracker calibration for gaze-based pervasive human-computer interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 05 2017.

[233] Thiago Santini, Wolfgang Fuhl, David Geisler, and Enkelejda Kasneci. EyeRecToo: Open-source software for real-time pervasive head-mounted eye-tracking. In *12th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, 02 2017.

[234] Thiago Santini, Diederick C Niehorster, and Enkelejda Kasneci. Get a grip: slippage-robust and glint-free gaze estimation for real-time pervasive head-mounted eye tracking. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, page 17. ACM, 2019.

[235] B. Hosp, S. Evazi, M. Maurer, W. Fuhl, and E. Kasneci. Remoteeye: An open source remote eye tracker. *Behavior Research Methods, BRM*, dec 2019.

[236] Tanja Blascheck, Michael Burch, Michael Raschke, and Daniel Weiskopf. Challenges and perspectives in big eye-movement data visual analytics. In *2015 Big Data Visual Analytics (BDVA)*, pages 1–8. IEEE, 2015.

[237] Zoya Bylinskii, Michelle A Borkin, Nam Wook Kim, Hanspeter Pfister, and Aude Oliva. Eye fixation metrics for large scale evaluation and comparison of information visualizations. In *Workshop on Eye Tracking and Visualization*, pages 235–255. Springer, 2015.

[238] Andrew T Duchowski, Margaux M Price, Miriah Meyer, and Pilar Orero. Aggregate gaze visualization with real-time heatmaps. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 13–20, 2012.

[239] Thanh-Chung Dao, Roman Bednarik, and Hana Vrzakova. Heatmap rendering from large-scale distributed datasets using cloud computing. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 215–218, 2014.

[240] Thomas Kübler, Wolfgang Fuhl, Raphael Rosenberg, Wolfgang Rosenstiel, and Enkelejda Kasneci. Novel methods for analysis and visualization of saccade trajectories. In *European Conference on Computer Vision*, pages 783–797. Springer, 2016.

[241] Michael Burch, Hansjörg Schmauder, Michael Raschke, and Daniel Weiskopf. Saccade plots. In ACM, editor, *Proceedings of Symposium on Eye Tracking Research and Applications*, 2014. URL http://dx.doi.org/10.1145/2578153.2578205.

[242] Michael Raschke, Dominik Herr, Tanja Blascheck, Michael Burch, Michael Schrauf, Sven Willmann, and Thomas Ertl. A visual approach for scan path comparison. In *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '14*. ACM, 2014. URL http://dx.doi.org/10.1145/2578153.2578173.

[243] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Stefan Strohmaier, Daniel Weiskopf, and Thomas Ertl. Aoi hierarchies for visual exploration of fixation sequences. In *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '16*, 2016.

[244] Kuno Kurzhals, Marcel Hlawatsch, Christof Seeger, and Daniel Weiskopf. Visual analytics for mobile eye tracking. *IEEE transactions on visualization and computer graphics*, 23(1):301–310, 2016.

[245] Kuno Kurzhals, Marcel Hlawatsch, Florian Heimerl, Michael Burch, Thomas Ertl, and Daniel Weiskopf. Gaze stripes: Image-based visualization of eye tracking data. *IEEE transactions on visualization and computer graphics*, 22(1):1005–1014, 2015.

[246] Kuno Kurzhals, Marcel Hlawatsch, Michael Burch, and Daniel Weiskopf. Fixation-image charts. In *Proceedings of the Ninth Biennial ACM Symposium on Eye Tracking Research & Applications*, pages 11–18, 2016.

[247] Michael Raschke, Bernhard Schmitz, Michael Wörner, Thomas Ertl, and Frederik Diederichs. Application design for an eye tracking analysis based on visual analytics. In *Proceedings of the Symposium on Eye Tracking Research and Applications, ETRA '16*, 2016. Rezensiertes Poster.

[248] T. C. Kübler, K. Sippel, W. Fuhl, G. Schievelbein, J. Aufreiter, R. Rosenberg, W. Rosenstiel, and E. Kasneci. *Analysis of eye movements with Eyetrace*, volume 574. Biomedical Engineering Systems and Technologies. Communications in Computer and Information Science (CCIS). Springer International Publishing, 2015.

[249] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. State-of-the-art of visualization for eye tracking data. In *EuroVis STAR*, 2014. URL http://dx.doi.org/10.2312/eurovisstar.20141173.

[250] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. Visualization of eye tracking data: A taxonomy and survey. In *Computer Graphics Forum*, volume 36, pages 260–284. Wiley Online Library, 2017.

[251] David Geisler, Wolfgang Fuhl, Thiago Santini, and Enkelejda Kasneci. Saliency sandbox-bottom-up saliency framework. In *VISIGRAPP (4: VISAPP)*, pages 657–664, 2017.

[252] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[253] Peng Bian and Liming Zhang. Biological plausibility of spectral domain approach for spatiotemporal visual saliency. In *International conference on neural information processing*, pages 251–258. Springer, 2008.

[254] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013.

[255] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.

[256] Vijay John, Keisuke Yoneda, B Qi, Zheng Liu, and Seiichi Mita. Traffic light recognition in varying illumination using deep learning and saliency map. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on*, pages 2286–2291. IEEE, 2014.

[257] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015.

[258] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–668, 2016.

[259] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

[260] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014.

[261] Xiaowu Chen, Anlin Zheng, Jia Li, and Feng Lu. Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1050–1058, 2017.

[262] Rachit Dubey, Joshua Peterson, Aditya Khosla, Ming-Hsuan Yang, and Bernard Ghanem. What makes an object memorable? In *Proceedings of the ieee international conference on computer vision*, pages 1089–1097, 2015.

[263] Wenbin Zou and Nikos Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 406–414, 2015.

[264] Raphael Yuster and Uri Zwick. Fast sparse matrix multiplication. *ACM Transactions On Algorithms (TALG)*, 1(1):2–13, 2005.

[265] Ton Roosendaal. Big buck bunny. In *ACM SIGGRAPH ASIA 2008 Computer Animation Festival*, SIGGRAPH Asia '08, page 62, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605585307. doi:10.1145/1504271.1504321. URL https://doi.org/10.1145/1504271.1504321.

[266] Anneli Olsen. The tobii i-vt fixation filter. *Tobii Technology*, pages 1–21, 2012.

[267] Andrew YC Nee, SK Ong, George Chryssolouris, and Dimitris Mourtzis. Augmented reality applications in design and manufacturing. *CIRP annals*, 61(2):657–679, 2012.

[268] Loukas Rentzos, Stergios Papanastasiou, Nikolaos Papakostas, and George Chryssolouris. Augmented reality for human-based assembly: using product and process semantics. *IFAC Proceedings Volumes*, 46(15):98–101, 2013.

[269] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. Does object recognition work for everyone? In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 52–59. Computer Vision Foundation / IEEE, 2019.

[270] Mesay Belete Bejiga, Abdallah Zeggada, Abdelhamid Nouffidj, and Farid Melgani. A convolutional neural network approach for assisting avalanche search and rescue operations with UAV imagery. *Remote Sensing*, 9(2):100, 2017.

[271] Junghwan Cho, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *arXiv preprint arXiv:1511.06348*, 2015.

[272] Dim P. Papadopoulos, Jasper R. R. Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[273] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[274] Vsevolod Peysakhovich, Frédéric Dehais, and Andrew Duchowski. ArUco/gaze tracking in real environments. In *Eye Tracking for Spatial Research, Proceedings of the 3rd International Workshop*. ETH Zurich, 2018.

[275] Mahmoud Kalash, Karishma Singh, Rasit Eskicioglu, and Neil DB Bruce. Gaze-contingent interactive visualization of high-dynamic-range imagery. In *2016 IEEE*

Second Workshop on Eye Tracking and Visualization (ETVIS), pages 16–20. IEEE, 2016.

[276] Jeff J MacInnes, Shariq Iqbal, John Pearson, and Elizabeth N Johnson. Wearable eye-tracking for research: Automated dynamic gaze mapping and accuracy/precision comparisons across devices. *bioRxiv*, page 299925, 2018.

[277] Reuben M Aronson and Henny Admoni. Semantic gaze labeling for human-robot shared manipulation. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2019.

[278] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.

[279] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick. Mask R-CNN. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[280] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 1202–1211, 2017.

[281] Yu-Hsin Chen, Tushar Krishna, Joel S Emer, and Vivienne Sze. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits*, 52(1):127–138, 2016.

[282] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *arXiv preprint arXiv:1906.02243*, 2019.

[283] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

[284] Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the 2012 ACM Symposium on Eye Tracking Research & Applications*, pages 91–98. ACM, 2012.

[285] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014.

[286] Rainer Stiefelhagen, Christian Fügen, Petra Gieselmann, Hartwig Holzapfel, Kai Nickel, and Alex Waibel. Natural human-robot interaction using speech, head pose and gestures. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, Sendai, Japan, September 28 - October 2, 2004*, pages 2422–2427. IEEE, 2004. doi:10.1109/IROS.2004.1389771. URL https://doi.org/10.1109/IROS.2004.1389771.

[287] Wolfgang Fuhl, Thiago Santini, Carsten Reichert, Daniel Claus, Alois Herkommer, Hamed Bahmani, Katharina Rifai, Siegfried Wahl, and Enkelejda Kasneci. Non-intrusive practitioner pupil detection for unmodified microscope oculars. *Comp. in Bio. and Med.*, 79:36–44, 2016. doi:10.1016/j.compbiomed.2016.10.005. URL https://doi.org/10.1016/j.compbiomed.2016.10.005.

[288] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, page 12. ACM, 2018.

[289] Oskar Palinko, Francesco Rea, Giulio Sandini, and Alessandra Sciutti. 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)*, pages 5048–5054. IEEE, 2016.

[290] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[291] Joao Carreira and Cristian Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3241–3248. IEEE, 2010.

[292] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE transactions on pattern analysis and machine intelligence*, 34 (11):2189–2202, 2012.

[293] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *The IEEE 12th International Conference on Computer Vision (ICCV)*, pages 606–613. IEEE, 2009.

[294] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

[295] Pupil Labs. https://pupil-labs.com/, 2019. Accessed: 2020-02-09.

[296] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. PuRe: Robust pupil detection for real-time pervasive eye tracking. *Computer Vision and Image Understanding*, 170:40 – 50, 2018. ISSN 1077-3142. doi:https://doi.org/10.1016/j.cviu.2018.02.002. URL http://www.sciencedirect.com/science/article/pii/S1077314218300146.

[297] Thiago Santini, Wolfgang Fuhl, and Enkelejda Kasneci. PuReST: Robust pupil tracking for real-time pervasive eye tracking. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, ETRA '18, pages 61:1–61:5, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5706-7. doi:10.1145/3204493.3204578. URL http://doi.acm.org/10.1145/3204493.3204578.

[298] MetraLabs. https://www.metralabs.com/mobiler-roboter-scitos-g5/, 2023. Accessed: 2023-02-18.

[299] Lingzhu Xiang, Florian Echtler, Christian Kerl, et al. libfreenect2: Release 0.2, apr 2016. URL https://doi.org/10.5281/zenodo.50641.

[300] Thiemo Wiedemeyer. IAI Kinect2. https://github.com/code-iai/iai_kinect2, 2014 – 2015. Accessed: 2019-11-21.

[301] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[302] Diederick C Niehorster, Thiago Santini, Roy S Hessels, Ignace TC Hooge, Enkelejda Kasneci, and Marcus Nyström. The impact of slippage on the data quality of head-worn eye trackers. *Behavior Research Methods*, pages 1–21, 2020.

[303] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*, 37(3):583–596, 2014.

[304] Alan Lukezic, Tomas Vojir, Luka ˇCehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6309–6318, 2017.

[305] Sebastian Thrun, Maren Bennewitz, Wolfram Burgard, Armin B Cremers, Frank Dellaert, Dieter Fox, Dirk Hahnel, Charles Rosenberg, Nicholas Roy, Jamieson Schulte, et al. Minerva: A second-generation museum tour-guide robot. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No. 99CH36288C)*, volume 3. IEEE, 1999.

[306] H-M Gross, H-J Boehme, Christof Schröter, Steffen Müller, Alexander König, Ch Martin, Matthias Merten, and Andreas Bley. Shopbot: Progress in developing an interactive mobile shopping assistant for everyday use. In *2008 IEEE International Conference on Systems, Man and Cybernetics*, pages 3471–3478. IEEE, 2008.

[307] Sebastian Blankemeyer, Rolf Wiemann, Lukas Posniak, Christoph Pregizer, and Annika Raatz. Intuitive robot programming using augmented reality. *Procedia CIRP*, 76:155–160, 2018.

[308] AY Elatta, Li Pei Gen, Fan Liang Zhi, Yu Daoyuan, and Luo Fei. An overview of robot calibration. *Information Technology Journal*, 3(1):74–78, 2004.

[309] Dražen Brščić, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Takayuki Kanda. Escaping from children's abuse of social robots. In *Proceedings of the tenth annual acm/ieee international conference on human-robot interaction*, pages 59–66, 2015.

[310] Balasubramaniyan Chandrasekaran and James M Conrad. Human-robot collaboration: A survey. In *SoutheastCon 2015*, pages 1–8. IEEE, 2015.

[311] Christopher Reardon, Kevin Lee, and Jonathan Fink. Come see this! augmented reality to enable human-robot cooperative search. In *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pages 1–7. IEEE, 2018.

[312] Brian Gleeson, Karon MacLean, Amir Haddadi, Elizabeth Croft, and Javier Alcazar. Gestures for industry intuitive human-robot communication from human observation. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 349–356. IEEE, 2013.

[313] André Driemeyer Wilbert, Barbara Behrens, Olaf Dambon, and Fritz Klocke. Robot assisted manufacturing system for high gloss finishing of steel molds. In *International Conference on Intelligent Robotics and Applications*, pages 673–685. Springer, 2012.

[314] Rainer Müller, Matthias Vette, and Aaron Geenen. Skill-based dynamic task allocation in human-robot-cooperation with the example of welding application. *Procedia Manufacturing*, 11:13–21, 2017.

[315] Iroju Olaronke, Ojerinde Oluwaseun, and Ikono Rhoda. State of the art: a study of human-robot interaction in healthcare. *International Journal of Information Engineering and Electronic Business*, 9(3):43, 2017.

[316] Ankur Kapoor, Ming Li, and Russell H Taylor. Constrained control for surgical assistant robots. In *ICRA*, pages 231–236, 2006.

[317] Joost Broekens, Marcel Heerink, Henk Rosendal, et al. Assistive social robots in elderly care: a review. *Gerontechnology*, 8(2):94–103, 2009.

[318] Hayley Robinson, Bruce A MacDonald, Ngaire Kerse, and Elizabeth Broadbent. Suitability of healthcare robots for a dementia unit and suggested improvements. *Journal of the American Medical Directors Association*, 14(1):34–40, 2013.

[319] Dennis Krupke, Frank Steinicke, Paul Lubos, Yannick Jonetzko, Michael Görner, and Jianwei Zhang. Comparison of multimodal heading and pointing gestures for co-located mixed reality human-robot interaction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–9. IEEE, 2018.

[320] Eric Rosen, David Whitney, Michael Fishman, Daniel Ullman, and Stefanie Tellex. Mixed reality as a bidirectional communication interface for human-robot interaction. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11431–11438. IEEE, 2020.

[321] Linh Kästner and Jens Lambrecht. Augmented-reality-based visualization of navigation data of mobile robots on the microsoft hololens-possibilities and limitations. In *2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM)*, pages 344–349. IEEE, 2019.

[322] Linh Kästner, Vlad Catalin Frasineanu, and Jens Lambrecht. A 3d-deep-learning-based augmented reality calibration method for robotic environments using depth sensor data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1135–1141. IEEE, 2020.

[323] Fu-Jen Chu, Ruinian Xu, Zhenxuan Zhang, Patricio A Vela, and Maysam Ghovanloo. The helping hand: An assistive manipulation framework using augmented reality and tongue-drive interfaces. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2158–2161. IEEE, 2018.

[324] Qilong Zhang and Robert Pless. Extrinsic calibration of a camera and laser range finder (improves camera calibration). In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 2301–2306. IEEE, 2004.

[325] Andreas Geiger, Frank Moosmann, Ömer Car, and Bernhard Schuster. Automatic camera and range sensor calibration using a single shot. In *2012 IEEE International Conference on Robotics and Automation*, pages 3936–3943. IEEE, 2012.

[326] Guangda Chen, Guowei Cui, Zhongxiao Jin, Feng Wu, and Xiaoping Chen. Accurate intrinsic and extrinsic calibration of rgb-d cameras with gp-based depth correction. *IEEE Sensors Journal*, 19(7):2685–2694, 2018.

[327] Vicon. https://www.vicon.com/. Accessed: 2021-02-24.

[328] OptiTrack. https://optitrack.com/. Accessed: 2021-02-17.

[329] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

[330] Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. Unseen object instance segmentation for robotic environments. *arXiv preprint arXiv:2007.08073*, 2020.

[331] Kenneth Holmqvist, Saga Lee Örbom, Ignace TC Hooge, Diederick C Niehorster, Robert G Alexander, Richard Andersson, Jeroen S Benjamins, Pieter Blignaut, Anne-Marie Brouwer, Lewis L Chuang, et al. Eye tracking: empirical foundations for a minimal reporting guideline. *Behavior Research Methods*, 2021.

[332] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press, 2011.

[333] Christian Müller-Tomfelde. Dwell-based pointing in applications of human computer interaction. In *IFIP Conference on Human-Computer Interaction*, pages 560–573. Springer, 2007.

[334] Michael Montemerlo, Sebastian Thrun, Daphne Koller, Ben Wegbreit, et al. Fastslam: A factored solution to the simultaneous localization and mapping problem. *Aaai/iaai*, 593598, 2002.

[335] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: A versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5): 1147–1163, 2015.

[336] Shusheng Bi, Dongsheng Yang, and Yueri Cai. Automatic calibration of odometry and robot extrinsic parameters using multi-composite-targets for a differential-drive robot with a camera. *Sensors*, 18(9):3097, 2018.

[337] Li Xiao and Vikas Kumar. Robotics for customer service: a useful complement or an ultimate substitute? *Journal of Service Research*, 24(1):9–29, 2021.

[338] Robert Bogue. Strong prospects for robots in retail. *Industrial Robot: the international journal of robotics research and application*, 2019.

[339] Yi Li and Chongli Wang. Effect of customer's perception on service robot acceptance. *International Journal of Consumer Studies*, 46(4):1241–1261, 2022.

[340] Anna-Maria Velentza, Dietmar Heinke, and Jeremy Wyatt. Human interaction and improving knowledge through collaborative tour guide robots. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–7. IEEE, 2019.

[341] Florian Alexander Schmidt. Crowdsourced production of ai training data: How human workers teach self-driving cars how to see. Technical report, Working Paper Forschungsförderung, 2019.

[342] Zhanat Makhataeva and Huseyin Atakan Varol. Augmented reality for robotics: A review. *Robotics*, 9(2):21, 2020.

[343] Long Qian, Jie Ying Wu, Simon P DiMaio, Nassir Navab, and Peter Kazanzides. A review of augmented reality in robotic-assisted surgery. *IEEE Transactions on Medical Robotics and Bionics*, 2(1):1–16, 2019.

[344] Gianluca Vadalà, Sergio De Salvatore, Luca Ambrosio, Fabrizio Russo, Rocco Papalia, and Vincenzo Denaro. Robotic spine surgery and augmented reality systems: a state of the art. *Neurospine*, 17(1):88, 2020.

[345] Camilo Perez Quintero, Sarah Li, Matthew KXJ Pan, Wesley P Chan, HF Machiel Van der Loos, and Elizabeth Croft. Robot programming through augmented trajectories in augmented reality. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1838–1844. IEEE, 2018.

[346] Sebastian von Mammen, Heiko Hamann, and Michael Heider. Robot gardens: an augmented reality prototype for plant-robot biohybrid systems. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pages 139–142, 2016.

[347] Andreagiovanni Reina, Alex J Cope, Eleftherios Nikolaidis, James AR Marshall, and Chelsea Sabo. Ark: Augmented reality for kilobots. *IEEE Robotics and Automation letters*, 2(3):1755–1761, 2017.

[348] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016.

[349] Thies Pfeiffer and Patrick Renner. Eyesee3d: a low-cost approach for analyzing mobile 3d eye tracking data using computer vision and augmented reality technology. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 195–202, 2014.

[350] Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. Active segmentation with fixation. In *2009 IEEE 12th international conference on computer vision*, pages 468–475. IEEE, 2009.

[351] Raihan Kabir, Yutaka Watanobe, Md Rashedul Islam, Keitaro Naruse, and Md Mostafizer Rahman. Unknown object detection using a one-class support vector machine for a cloud–robot system. *Sensors*, 22(4):1352, 2022.

[352] Yimeng Li and Jana Košecká. Uncertainty aware proposal segmentation for unknown object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 241–250, 2022.

[353] Weiwei Wan, Feng Lu, Zepei Wu, and Kensuke Harada. Teaching robots to do object assembly using multi-modal 3d vision. *Neurocomputing*, 259:85–93, 2017.

[354] Guglielmo Gemignani, Emanuele Bastianelli, and Daniele Nardi. Teaching robots parametrized executable plans through spoken interaction. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 851–859. International Foundation for Autonomous Agents and Multiagent Systems, 2015.

[355] Hakan Karaoguz and Patric Jensfelt. Object detection approach for robot grasp detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4953–4959. IEEE, 2019.

[356] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2013.

[357] Evan Krause, Michael Zillich, Thomas Williams, and Matthias Scheutz. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[358] Lotfi El Hafi, Hitoshi Nakamura, Akira Taniguchi, Yoshinobu Hagiwara, and Tadahiro Taniguchi. Teaching system for multimodal object categorization by human-robot interaction in mixed reality. In *2021 IEEE/SICE International Symposium on System Integration (SII)*, pages 320–324. IEEE, 2021.

[359] Sepehr Valipour, Camilo Perez, and Martin Jagersand. Incremental learning for robot perception through hri. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2772–2777. IEEE, 2017.

[360] Masood Dehghan, Zichen Zhang, Mennatullah Siam, Jun Jin, Laura Petrich, and Martin Jagersand. Online object and task learning via human robot interaction. In *2019 international conference on robotics and automation (ICRA)*, pages 2132–2138. IEEE, 2019.

[361] Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *2011 IEEE international conference on robotics and automation*, pages 1–4. IEEE, 2011.

[362] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[363] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.

[364] Kinova. https://www.kinovarobotics.com/product/gen2-robots, 2023. Accessed: 2023-02-18.

[365] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[366] B. E. Moore and J. J. Corso. Fiftyone. https://github.com/voxel51/fiftyone, 2020.

[367] Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2022.

[368] Michele Mazzamuto, Francesco Ragusa, Antonino Furnari, Giovanni Signorello, and Giovanni Maria Farinella. Weakly supervised attended object detection using gaze data as annotations. In *Image Analysis and Processing–ICIAP 2022: 21st International Conference, Lecce, Italy, May 23–27, 2022, Proceedings, Part II*, pages 263–274. Springer, 2022.

[369] Padraig Higgins, Ryan Barron, and Cynthia Matuszek. Head pose for object deixis in vr-based human-robot interaction. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 610–617. IEEE, 2022.

[370] Yang Liu, Feng Yang, Cheng Zhong, Ying Tao, Bing Dai, and Mengxiao Yin. Visual tracking via salient feature extraction and sparse collaborative model. *AEU-International Journal of Electronics and Communications*, 87:134–143, 2018.

[371] Intel. https://www.intelrealsense.com/depth-camera-d435/, 2023. Accessed: 2023-02-18.