Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review

NICOLE ROBINSON, Queensland University of Technology and Monash University
BRENDAN TIDD, Queensland University of Technology
DYLAN CAMPBELL, University of Oxford
DANA KULIĆ, Monash University
PETER CORKE, Queensland University of Technology

Robotic vision, otherwise known as computer vision for robots, is a critical process for robots to collect and interpret detailed information related to human actions, goals, and preferences, enabling robots to provide more useful services to people. This survey and systematic review presents a comprehensive analysis on robotic vision in human-robot interaction and collaboration (HRI/C) over the past 10 years. From a detailed search of 3,850 articles, systematic extraction and evaluation was used to identify and explore 310 papers in depth. These papers described robots with some level of autonomy using robotic vision for locomotion, manipulation, and/or visual communication to collaborate or interact with people. This article provides an in-depth analysis of current trends, common domains, methods and procedures, technical processes, datasets and models, experimental testing, sample populations, performance metrics, and future challenges. Robotic vision was often used in action and gesture recognition, robot movement in human spaces, object handover and collaborative actions, social communication, and learning from demonstration. Few high-impact and novel techniques from the computer vision field had been translated into HRI/C. Overall, notable advancements have been made on how to develop and deploy robots to assist people.

CCS Concepts: • **General and reference** → **Surveys and overviews**; • **Human-centered computing** → *HCI theory, concepts and models*; • **Computing methodologies** → **Vision for robotics;**

Additional Key Words and Phrases: Robotic vision, computer vision, human-robot interaction, gesture recognition, robot movement in human spaces, object handover, collaborative actions, learning from demonstration, social communication

This research was supported by the Australian Research Council (project no. CE140100016). D. Campbell received funding from Continental AG.

Authors' addresses: N. Robinson, Australian Research Council Centre of Excellence for Robotic Vision, School of Electrical Engineering & Robotics, QUT Centre for Robotics, Queensland University of Technology, and Faculty of Engineering, Turner Institute for Brain and Mental Health, Monash University, 18 Alliance Lane, Clayton, Victoria, Australia, 3800; email: nicole.robinson@monash.edu; B. Tidd and P. Corke, Australian Research Council Centre of Excellence for Robotic Vision, School of Electrical Engineering & Robotics, QUT Centre for Robotics, Queensland University of Technology, 2 George Street, Brisbane, Queensland, Australia, 4000; emails: brendan.tidd@hdr.qut.edu.au, peter.corke@qut.edu.au; D. Campbell, Visual Geometry Group, Department of Engineering Science, University of Oxford, 17 Parks Road, Oxford, Oxfordshire, UK, OX1 3PJ; email: dylan@robots.ox.ac.uk; D. Kulić, Australian Research Council Centre of Excellence for Robotic Vision, Faculty of Engineering, Monash University, 18 Alliance Lane, Clayton, Victoria, Australia, 3800; email: dana.kulic@monash.edu.

2573-9522/2023/02-ART12 $15.00
https://doi.org/10.1145/3570731

ACM Reference format:
Nicole Robinson, Brendan Tidd, Dylan Campbell, Dana Kulić, and Peter Corke. 2023. Robotic Vision for Human-Robot Interaction and Collaboration: A Survey and Systematic Review. *ACM Trans. Hum.-Robot Interact.* 12, 1, Article 12 (February 2023), 66 pages.
https://doi.org/10.1145/3570731

1 INTRODUCTION

This article presents a comprehensive survey and review of robotic vision methods for Human-Robot Interaction and Collaboration (HRI/C) based on a review of 3,850 articles to create a collection of 310 eligible articles for in-depth analysis. The selected 310 published papers examine how robotic vision is used to facilitate human-robot interaction tasks such as robot navigation in human spaces, social interaction with people to exchange information, and human-robot handovers of everyday objects. Such a combination of a systematic review to calculate trends and prevalence alongside a comprehensive survey for each section will help to explore emerging patterns, statistical trends, and recommendations on how robotic vision can help to improve HRI/C.

1.1 Purpose

The purpose of this systematic review and survey was to provide detailed insight into underlying emergent research themes pursued by the community, and to explore the trajectory and impact that robotic vision will have on enabling robots to better interact and collaborate with humans. For the purpose of this work, robotic vision will be defined as computer vision that is used to inform or direct a robot on what actions to perform that will contribute to achieving the chosen goal. In practice, robotic vision can enable robots to sense, perceive, and respond to people through capturing and responding to a rich continuous information stream. Visual information provided by humans can help robots to better understand the scenario and to plan their actions, such as interpreting hand gesture movements as communication signals and human body movements as an indication of future intent to perform an action. Robots with robotic vision can therefore help to create and facilitate an important information exchange between the human and the robot, opening up new communication channels using a method that is natural and intuitive to people, improving the effectiveness of collaborative tasks.

1.2 Scope

This survey and review explored published papers from the past 10 years using a systematic search, screen, and evaluation protocol to extract a general overview of current research trends, common applications and domains, methods and procedures, technical processes, relevant datasets and models, experimental testing setups, sample populations, vision algorithm metrics, and performance evaluations. To create the systematic search strategy, several key parameters needed to be defined before commencing the extraction and evaluation of papers. First, given the extensive scope of reviewing all relevant papers in the broad field of robotic vision for HRI/C, this review focused on the past 10 years (i.e., 2010–2020). This time frame helped to showcase the more contemporary use of robotic vision based on newly emergent techniques, and was chosen to coincide with the introduction of critical camera hardware that boosted the applied use of robotic vision to enable robots to be more reactive and suitable for human interaction, such as the release date of the Kinect camera [173].

1.3 Related Surveys and Systematic Reviews

To the best of our knowledge, no systematic review or comprehensive survey on the development and use of robotic vision for HRI/C had been conducted. Current surveys or reviews have focused

on other areas, such as specific domains including robotics in industry [183, 436], agriculture [429], public areas [140], healthcare [365], and education [43]. All of these works described different robots or methods relevant to the domain of interest, including describing robot types and use cases that did not use robotic vision. Other reviews or surveys focused on components related to the process of HRI/C, such as the use of physical touch and tactile sensing techniques [22], safety bounds for vision-based safety systems [171, 480], trust modeling and trust-related factors [174, 218], distance between humans and robots [243], and the use of non-verbal communication [377]. There have also been other published works on specific methods used in human-robot interaction that did not have a direct focus on vision, such as tests with psycho-physiological measures [47], exploring general robot perception methods [462], investigating a single robot platform [413], or a specific form of robot behavior [295].

The computer vision field has contributed to providing detailed surveys and reviews that show the technical process for computer vision related to humans, such as gesture-based human-machine interaction [399] and multi-modal machine collaboration with a focus on body, gesture, gaze, and affective interaction [200]. Others have explored more detailed and specific use cases such as action recognition [486], hand gesture recognition [256, 357, 456], and human motion capture [302]. There have also been detailed surveys and reviews that explored vision-based techniques in robots—for instance, reviews or surveys that included recent developments in robotic vision techniques [82], learning for robotic vision [381], and the use of computer vision for a specific type of robot, such as aerial robots [261]. Other surveys and reviews instead had more general overviews of vision for robots such as object recognition and modeling, site reconstruction and inspection, robotic manipulation, localization, path following, map construction, autonomous navigation, and exploration [53, 83]. Others also included a brief mention of applications to people but did not provide a detailed analysis on how this could better facilitate HRI/C across different technique types. In collection, the identified surveys and reviews provided an excellent commentary on their respective fields and target areas, but there were limited works that presented a detailed investigation into robotic vision techniques, hardware integration, and evaluation of its use in real-world scenarios for HRI/C.

1.4 Contribution

The contribution of this work is the systematic extraction, discovery, and detailed synthesis of the literature to showcase the current use of robotic vision for robots that can interact and collaborate with people. This survey and systematic review contributes new knowledge on how robots can be improved by integrating and refining functionality related to robotic vision, showcases real-world use of robots with vision capabilities to improve collaborative outcomes, and provides a critical discussion to help push the field forward.

2 BACKGROUND

2.1 A Brief History of the Field of Computer Vision

Computer vision is important to help machines better understand and interact with the real world, making relevant actions and decisions based on visual information [98]. Common sensors in computer vision include RGB (red, green, and blue wavelength) cameras that provide detailed information by capturing light in RGB to create a color representation of the world. The use of visual information to understand the world can help emulate how humans perceive the world, creating a common language and understanding between humans and robots when sharing details, objects, and task-related information. Computer vision involves techniques such as object detection to localize where an object is in the scene, image classification to determine what it is in the image, and

pixel-level classification to classify what part of the image belongs to an area of interest [98, 138]. In relation to computer vision for humans, computer vision can address the detection and analysis of humans in visual scenes, including methods such as face detection [438], pose estimation [64], and human motion tracking [182]. This type of visual information can then further assist in creating shared knowledge and understanding between humans and machines.

The field of computer vision has evolved rapidly in the past decade from 2010 to 2020. Deep learning has played a dominant role since its success at the 2012 ImageNet competition [227]. Learning complex parameterized functions from data has also served to make computer vision algorithms more robust and effective in real-world situations, making it ideal for the field of HRI/C. However, this comes at the expense of increased hardware requirements and longer development time, such as the need for data collection, labeling, and network training. Another significant change at the start of this period was the advent of more readily available RGB-D sensors with the Microsoft Kinect camera released in 2010 [173]. This allowed researchers to reason about color and 2.5D geometry jointly, facilitating new breakthroughs such as real-time 3D reconstruction [197].

In the decades before this period, computer vision had several major successes relevant to HRI/C. The first was the codification of the principles of multiple-view geometry [175] and their successful application in large-scale reconstruction tasks [7] using the techniques of structure from motion. The period was also marked by increasingly sophisticated handcrafted features such as SIFT [266] and HOG (histogram of oriented gradients) [103] features and the use of increasingly sophisticated learning algorithms, such as kernel Support Vector Machine (SVM) [54] and AdaBoost [143], the latter used to great effect in the Viola–Jones face detector [437]. The topics of image classification, object detection, image segmentation, and optical flow received significant research attention, among many others. Some highlights include deformable part models [133, 134] that demonstrated unprecedented performance on object detection benchmarks before deep learning, conditional random fields for image segmentation [156, 226, 419], graph cuts for tasks such as stereo depth estimation [57], and variational methods for optical flow estimation [186, 268]. These approaches continue to be used in robotics and embodied vision settings due to their efficiency and low hardware requirements.

2.2 A Brief History of the Field of Robotic Vision

Robotic vision, by contrast, exists at the intersection of robotics and computer vision, enabling robots to sense, perceive, and respond to people by providing rich, continuous information about human states, actions, intentions, and communication. Robotic vision involves a vision sensor (RGB, RGB-D) and supporting algorithms that translate raw images to a control signal for a robot. In other words, any computer vision techniques used to guide a robot on what action to perform can be considered robotic vision. Robotic vision benefited from the advancements in the computer vision research community, such as large datasets, computing power, complex algorithms, and scientific methods. Robotic vision has started to become a key perception channel for the robot to interact with and provide assistance to people. Robotic vision has important advantages for enabling robots to smartly interact with the environment, such as better camera control, physical movement around the space, and the capacity to adapt its viewpoint to gather further information [98, 402]. There have also been notable advances to effectively handle multi-modal data in robotic sensing, including visual processing for intelligent robot decisions and actions [344, 402]. Robotic vision can also create new opportunities for humans to interact with robots in a way that does not inhibit natural actions, such as removing the need to use a computer terminal or to wear a physical apparatus. Improvements to robots through visual perception can therefore help to contribute to creating more general-purpose robots, extending the potential for a wide range of tasks that a robot can complete for a person [381].

2.3 Human-Robot Interaction and Collaboration

Human-robot interaction focuses on the interactivity between humans and robots, and often involves creating a robotic system that can identify and respond to the complexities of human behavior. For the robot to behave in socially acceptable ways, the robot should be able to sense, perceive, and respond to human states, actions, intentions, and emotions. Human-robot interaction related topics include improving robot social acuity using visual perception of the person [415]. Human-robot collaboration instead focuses on how humans and robots work together to achieve shared goals with a common purpose and directed outcome. In HRC, robots work to complement or add value to the intended goal of the human [35]. Collaboration with a robot can help to improve task speed and work productivity, reduce the number of errors, and improve human safety to minimize repetition fatigue and injuries [166, 429].

2.3.1 Robots with Computer Vision to Improve Collaborative Outcomes. Robotic vision techniques have been used to create new interaction methods and improve the current process of human-robot interaction, such as using vision to create the ability for people to communicate with the robot, such as to signal information or commands. These contribute to the ability for the robot to provide a more functional service to the human. Visual information captured by the robot through the camera system can then be used to help enable the robot to make more informed decisions about its next set of actions. Examples could include to detect target objects in its field of view when humans request a specific object, to understand events and scenarios that are occurring in the scene for social group dynamics, and to classify and better understand human actions to offer predictive assistance [66, 132, 151, 173, 179, 258, 368]. For instance, visual information from people can help robots make informed decisions on how to interact or assist the person, such as to help the robot to decide how to approach a person [363], how to follow a person [196], or when to offer to hand over an item to a person [326]. Robotic vision can help to identify, classify, or predict human movements through action or activity recognition [39]. Activity recognition to perceive human movements can give robots the ability to better predict or recognize what a human is doing in the environment so that the robot can better provide useful information, advice, or assistance to the person in settings such as in industrial settings [367] or in different contexts such as recognition for multiple people in a robot's field of view [154]. Gesture recognition has often been tested as a communication and control method through the translation of human pose into a command signal, an action to trigger a state change, or to signify the start of an information exchange between the human and the robot [154, 202, 213, 388, 421]. Gestures can also be used to signal to the robot which object the robot should use [388, 389] and where a robot should move [202]. Visual information can also support collaborative robot actions with the person such as human-robot object handover through perception and interpretation of humans and objects in the scene, including human reach ability, motion, and collision range [237, 326]. There is significant opportunity to draw from principles and concepts of computer vision to improve robot capacity to perceive and act upon visual information to improve human-robot collaboration. This includes robotic vision with the intention to improve robot functionality, user experience, interface design, control methods, and robot utility for certain actions or tasks.

3 REVIEW PROTOCOL

This survey and systematic review will provide insight into the underlying emergent research themes pursued by the community, and explore the broader use of robotic vision to enhance human-robot interactivity and collaborative outcomes. The purpose of this systematic review is to inform readers about the current state of robotic vision applied to interpreting and responding to human actions, activities, tasks, states, and emotions. For the purpose of this survey and review, a

robot was defined as a system that can perform (semi-)autonomously through an algorithm/s, action through actuator/s in the world in response to perception through sensor/s, with the potential inclusion of an externally provided goal directive.

This systematic review protocol followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology for systematic research, which specifies the search, screening, and evaluation steps. This method involves a comprehensive and reproducible search strategy to present and critically appraise the research findings related to the topic of interest [306]. PRISMA guidelines [306] are considered to be a gold-standard reporting method with more than 80,000 citations in the past 15 years. The PRISMA method also allows for clear conclusions to be drawn across a expansive pool of studies with minimal selection bias alongside balanced reporting of research findings [306]. All of the studies were assessed for inclusion/exclusion criteria to provide a defined view of the topic, as well as to assess the included studies for quality assurance. The search strategy of the systematic review was designed to capture different aspects related to robotic vision for HRI/C, including robot behaviors, collaborative tasks, and communicative behaviors. This review was designed to answer the following Research Questions (RQ):

RQ1. What is the general trend of robotic vision work in human-robot collaboration and interaction in the past 10 years?
RQ2. What are the most common application areas and domains for robotic vision in human-robot collaboration and interaction?
RQ3. What is the human-robot interaction taxonomy for robots with robotic vision in human-robot collaboration and interaction?
RQ4. What are the vision techniques and tools used in human-robot collaboration and interaction?
RQ5. What are the datasets and models that have been used for robotic vision in human-robot collaboration and interaction?
RQ6. What has been the main participant sample, and how is robotic vision in human-robot collaboration and interaction evaluated?
RQ7. What is the state of the art in vision algorithm performance for robotic vision in human-robot collaboration and interaction?
RQ8. What are the upcoming challenges for robotic vision in human-robot collaboration and interaction?

Preliminary searches were undertaken in field-relevant journals and conferences to help inform search criteria keywords. The search terms went through extensive iteration and the final terms were chosen to be broad enough to capture works across multiple disciplines and topic keywords, but scoped as best as possible to systematically extract papers on the intended topic of interest: robots, vision, and humans: ((((“Abstract”: Robot* OR “Abstract”: UAV OR “Abstract”: AUV OR “Abstract”: UUV OR “Abstract”: Drone OR “Abstract”: Humanoid OR “Abstract”: Manipulator) AND (“Abstract”: Vision OR Abstract: Image OR “Abstract”: Camera OR “Abstract”: RGB* OR “Abstract”: Primesense OR “Abstract”: Realsense OR “Abstract”: Kinect) AND (“Abstract”: Human OR “Abstract”: Humans OR “Abstract”: Person OR “Abstract”: People OR “Abstract”: User OR “Abstract”: Users) AND (“Abstract”: HRI OR “Abstract”: HRC OR “Abstract”: Collaborat* OR “Abstract”: Interact* OR “Abstract”: “Human-in-the-Loop” OR “Abstract”: Team* OR “Abstract”: “Human-to-Robot” OR “Abstract”: “Robot-to-Human”)))). To create the search method, the following databases were chosen for systematic search and data extraction, representing multi-disciplinary avenues for published works: IEEE Xplore, ACM Library, and Scopus. Inclusion and exclusion criteria were generated, reviewed, and approved by subject matter experts across robotics, human-robot interaction, and behavioral science to confirm keyword relevance to identify suitable papers for the topic of

interest, and to reduce the chance of extracting unrelated works. The final inclusion criteria markers were used in a sequential order when categorizing extracted papers to determine its inclusion into this systematic review:

(C1) The research must include at least one physically embodied robot that can perceive through a vision system.
(C2) The robot(s) must be capable of at least one closed-loop interaction or information exchange between the human and the robot(s), where the robot(s) vision system is utilized in the exchange, and a human is the focus of the vision system.
(C3) The robot(s) must be able to make a decision and/or action based on visual input that is real time or at least fast enough for an interactive channel to occur between the human and the robot (i.e., 60 seconds).

The purpose of C1 was to ensure that only physical robot systems with a vision system were analyzed with digital avatars and software systems running on computers removed from analysis. The purpose of C2 was to ensure that robots were able to perceive visual information relevant to creating a robot signal, task, or action based on the vision system, and that the robot could in fact perceive some or part of the person during the interaction or collaborative exchange. The purpose of C3 was to ensure that robots could perform a decision and/or action based on interpretation of the visual information, and the interaction exchange occurs without extended wait times. Taken together, these chosen criteria would ensure that the robot was acting on the visual information, that the human was classified and/or involved in the process, and the information and/or exchange was occurring in a functional amount of time for an interaction.

To maintain the proposed review theme of humans, robotic vision, and interaction or collaboration, several exclusions were created and used in this review. Papers were excluded if they contained non-embodied agents that did not operate as a robotic system, such as having no actuation system and/or capacity to make or execute decisions (i.e., cameras, computers, smartphones, tele-operated devices, avatars). Papers were also excluded if there was no physical or verbal robot action involved in the process as a result of processing visual information, as well as instances in which the robot could have been substituted with a camera on its own, a computer screen on its own, or another simple input signal such as using the robot as a speaker only. Given the clear focus on robotic vision, this review excluded papers that did not meet the criteria of a vision sensor (RGB, RGB-D) paired with an algorithm/s that could translate raw images to a control signal for a robot. Papers were also excluded if vision was not central to the system's operation or lacked control, such as vision being a function of the robot, but not used to inform or update the decision-making process or resulting action of the robot.

Papers that did not have any human-relevant information, use case application, or research experiment with people were also excluded because it did not meet inclusion criteria for the intention to explore robotic vision in relation to human-robot interaction or collaboration. Examples of these papers include early-stage design work on proposed concepts of robot systems that had not yet been built, and robot competition papers where the robot was intended for a human environment, but its proposed performance or relationship with people was not reported at all. For this review, only papers in which there was a clear interaction or collaboration between the human and the robot were included. Therefore, robots that only used an open-loop interaction were excluded from analysis for not meeting the criteria for an interaction, especially if the visual signal input was independent of the robot output and did not influence the action or decision making of the robot. Simple devices such as children's toys were also excluded given the limited interaction set often involved in these devices, and the intention to focus on robots that could provide benefit to support a person's work or lifestyle.

All papers must have been published and available for access from the publication venue between January 1, 2010 and December 31, 2020 in a peer-reviewed journal or conference. Papers that were not formally published between these dates were not extracted. If authors published multiple versions of the work, the most complete version was included. E-Print services (e.g., arXiv.org) were not included for three reasons: (1) the abundance of early stage work that did not yet include humans into the proposed system, (2) there was limited quality control without a peer-review process to ensure that only high-quality papers were identified in an unbiased way, and (3) reporting on early-stage work that has not yet undergone peer review could have created a skewed commentary on the current prevalence and impact of the field. We do acknowledge the importance of robotic vision use cases that occurred in works that would have fallen into the excluded criteria category for this systematic review. As such, we will present a section in the following to acknowledge and investigate the use cases that did not meet the search strategy criteria, including any key papers that were not captured as part of the systematic search. Examples include tele-operated robots, robots with an open-loop system, simple devices, and early-stage work that did not have tests with people.

3.1 Review Information and Categorization

Each eligible article underwent systematic data extraction informed by robot classification and human-robot interaction taxonomies (e.g., [40, 464]). For each, manuscript information was extracted into categories such as task type, task criticality (low, medium, high), robot morphology (anthropomorphic, zoomorphic, functional), ratio of people to robots (i.e., a non-reduced fraction with number of humans over number of robots), composition of robot teams (formation), level of shared interaction among teams, interaction roles, type of human-robot physical proximity, time/space taxonomy [464], level of autonomy [40], task evaluation, sensor fusion (i.e., vision and speech), camera system and type, vision techniques and algorithm, training method, and datasets. User study information was extracted if a user study was reported, including participant details and experimental outcomes. A custom metric for overall task evaluation was computed using 3-point scaling (low, medium, high) for task complexity, risk, importance, and robot complexity.

Application areas were clustered and labeled using the following criteria. Gestures were defined as a hand, arm, head, or body movement intended to indicate, convey a message, or send information. Action recognition was defined as the recognition of human actions or activities that were not related to explicit gestures. Robot movement in human spaces was classified if the robot had physical movement in a human environment and robot movement did not require the human to perform a set pose to signal movement commands to the robot, including if the robot was classified as a (semi-)autonomous vehicle. Object handover and collaborative action papers included a robot capable of manipulating objects while the interaction did not require the person to perform a set pose (i.e., the person was detected without performing a gesture or action). Categorization for social communication captured papers in which the robot needed to perform a social behavior, or be capable of socially interacting with a person. Last, learning from demonstration must have used some form of demonstration learning.

4 RESULTS

4.1 Selected Articles

The initial search across three databases found 6,771 papers, 2,034 of which were identified as duplicate records (Figure 1). The remaining 4,737 papers were screened for titles and abstracts to assess initial eligibility and 887 papers excluded based on format: textbook chapters that did not include original research work (n = 63, 7%), reviews or surveys (n = 92, 10%), no English version

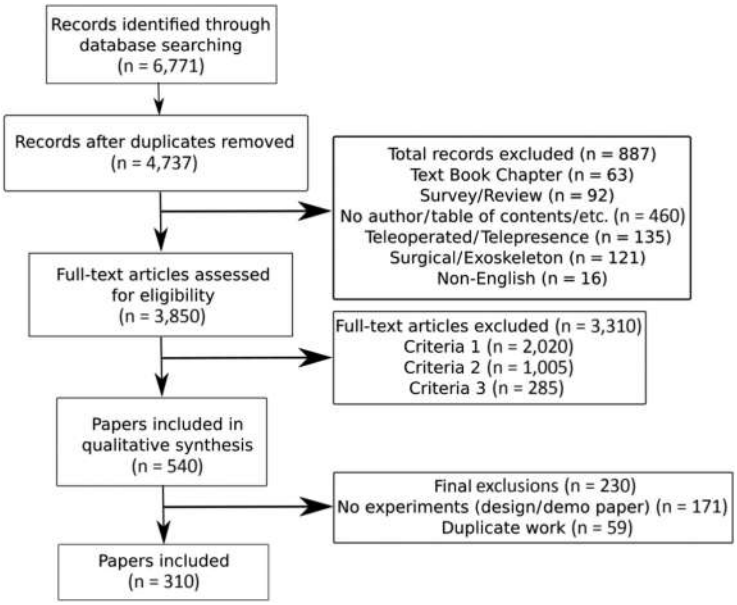


Fig. 1. CONSORT chart for systematic review to determine inclusion.

of the work (n = 16, 2%), and other non-related works (n = 460, 52%) such as front pages, table of contents, plenary talks, copyright notices, and keynotes. A total of 255 papers were excluded based on title alone: robotic surgical tools only (n = 121, 14%) and tele-operation only (n = 135, 15%). The remaining 3,850 papers were assessed by detailed review of the text and 3,310 papers omitted for not meeting the C1–3 inclusion criteria: 2020 on C1 (61%), 1,005 on C2 (30%), and 285 on C3 (9%). The large volume of papers omitted on C1 showed that most works had no physical robot (e.g., [165, 288, 422, 495]), involved simulation testing such as using a virtual robot (e.g., [44, 119, 329, 461]), involved cameras on their own that were not connected to robotic systems (e.g., [307]), or the robot/s did not use a vision system as part of the interaction with the person (e.g., [294, 346]). Papers omitted on C2 often had a clear focus on other components such as speech (e.g., [257]) or visual servoing (e.g., [205]). C2 papers often had no humans at all (e.g., [110]), or no human vision involved in the vision process of the interaction or collaboration (e.g., [93, 167, 217, 355, 460]). Papers omitted on C3 often did not have a robot perform an action based on visual input (e.g., [153, 284, 289, 417, 485]), robots that were tele-operated (e.g., [360]), or no near real-time information exchange (e.g., [77]). A total of 540 papers that met the C1–3 inclusion criteria were then subject to another round of investigation. Papers were then excluded if it was the same study published across multiple venues (n = 59, 26%), or there was no clear experiment or demonstration of human-robot interaction or collaboration despite reporting on a system designed for HRI/C (n = 171, 74%). A final total of 310 papers (8% of full articles assessed for eligibility) met final inclusion criteria, which provides a significant pool of research works for detailed analysis on the chosen topic. The CONSORT chart of inclusion and exclusion steps can be seen in Figure 1. Two independent raters went through 10% of 310 eligible papers and achieved a 100% consensus on inclusion and exclusion criteria. Some notable works that would have fallen into the excluded criteria category for this systematic review were reported in a separate section as part of presenting a comprehensive survey on the topic, but these papers were not included in the final systematic review.

5 RQ1. WHAT IS THE GENERAL OVERALL TREND OF ROBOTIC VISION IN HUMAN-ROBOT COLLABORATION AND INTERACTION IN THE PAST 10 YEARS?

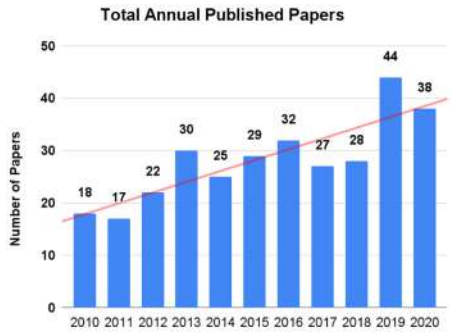
This section presents the general trend of robotic vision in human-robot collaboration and interaction in the past 10 years. Robotic vision for HRI/C had a moderate but steady increase, which might be attributed to several components, such as limited accessibility to robot platforms, integration challenges, the interdisciplinary nature of HRI/C, technical capacity for robots to operate consistently for robust use cases with people, limited engineering knowledge of human-robot testing, limited capacity to test robots in human spaces, and human-centered robotics representing a much smaller research field compared to its robotics and computer vision counterparts. Figure 2(a) depicts a modest increase in publications over the period but a small decline from 2020/2021 that was predicted to be attributed to the COVID-19 pandemic. Figure 2(b) depicts the publication themes of robotic vision work, including interaction (human-robot interaction, human-machine systems), robotics (robotics, automation, mechatronics), sensors (sensors, vision, signal processing), engineering (engineering, systems, industry, control, science), and computers and Artificial Intelligence (AI). Figure 2(c) depicts the most relevant papers that were published in conferences, journals, and then book series.

Figure 2(d) depicts the most common application areas clustered into groups (N = 335): action recognition (13%), gesture recognition (35%), robot movement in human spaces (22%), object handover and collaborative actions (17%), learning from demonstration (3%), and social communication (10%). If a paper had more than one application, each area was included in the final total. Individual application breakdowns will be seen in the next section. Figure 2(e) depicts that common domain areas involved field, industrial (i.e., manufacturing and warehouses), domestic (i.e., home use), and urban settings (i.e., shopping centers, schools, restaurants, and hotels). Robots that work with and around humans were often proposed for domestic and urban environments. Figure 2(f) depicts common robot types being mobile robots, followed by fixed manipulators, social robots, mobile manipulators, and aerial robots. If multiple robots were tested, only the first or most detailed test was reported in the total. Most works had a single focus on a specific vision application for a target purpose, and the intended outcome was often for robots to better integrate into human-populated environments in a direct (i.e., controlled via gesture) or non-direct way (i.e., following a person).

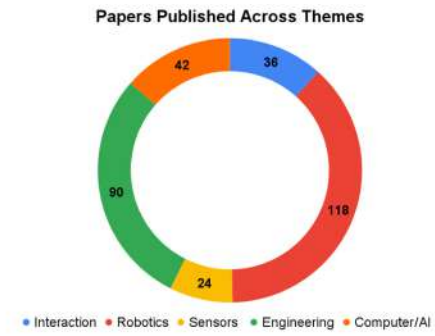
Figure 3 depicts that for camera type, RGB-D cameras such as the Kinect were the most frequently used, followed by monocular, stereo, and omni-directional cameras. Figure 3 also shows an increased uptake of RGB-D cameras. RGB-D cameras were extensively used across all use cases, environments, and robot types, showing the value of this sensor capacity to provide critical visual information to improve robotic vision for robot tasks. Figure 4 depicts global trends in domain and types. Figure 4(a) depicts that the highest volume of work was conducted in gesture recognition or robot movement in human spaces, with the exception of Europe with a higher focus on object handover and collaborative actions. Figure 4(b) depicts that the most common robot types was mobile robots and fixed manipulators across all continents.

6 RQ2. WHAT ARE THE MOST COMMON APPLICATION AREAS AND DOMAINS FOR ROBOTIC VISION IN HUMAN-ROBOT COLLABORATION AND INTERACTION?

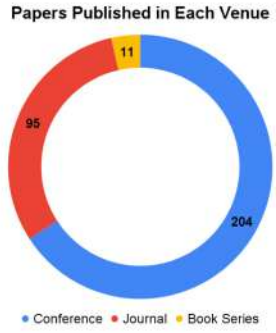
This section provides a detailed breakdown of the following application areas: gesture and action recognition, robot movement in human spaces, object handover and collaborative actions, learning from demonstration, and social communication. If a paper had more than one application, each area was included in the final total. Papers often reported tasks and actions that were simplified or well contained in their relevant context or domain. In addition, papers often focused on using the human to improve the robot's performance, such as human gestures for more control over the robot



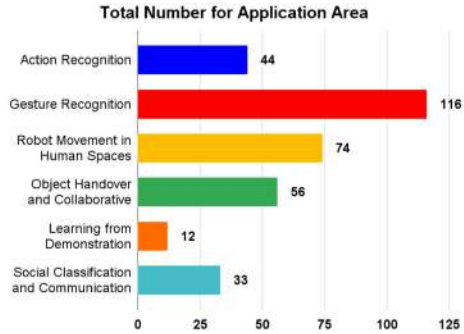
(a) Total annual published papers



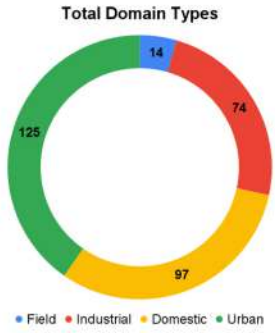
(b) Number of papers published acroos themes



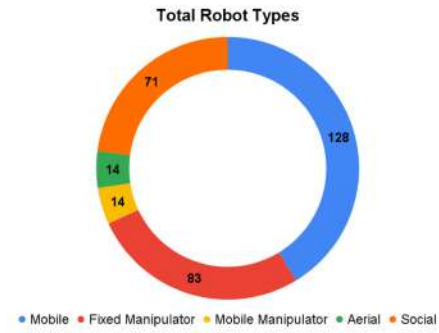
(c) Number of published papers in each venue



(d) Application area total



(e) Domain area total



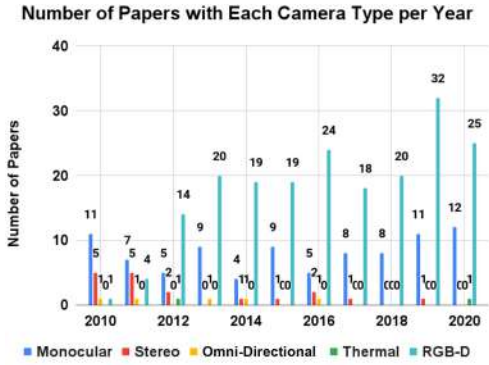
(f) Robot type total

Fig. 2. Statistical summaries of total paper counts and trends.

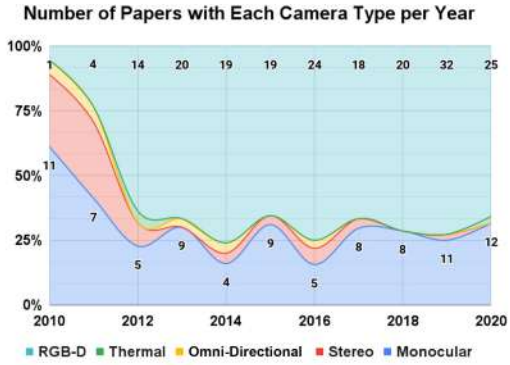
actions, humans to improve robot handover accuracy, and humans contributing to better mobile robot safety on pathways. A summary of the identified papers will be reviewed in Section 6, and detailed exploration into the technical content of the papers will be discussed in Section 10.

6.1 Gesture Recognition

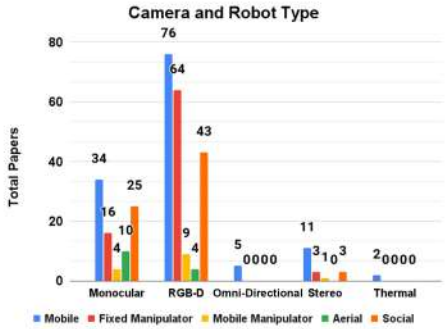
6.1.1 Overview. Gestures were defined as a hand, arm, head, or body movement intended to indicate, convey a message, or send information. A total of 116 papers (37% of the eligible total) were



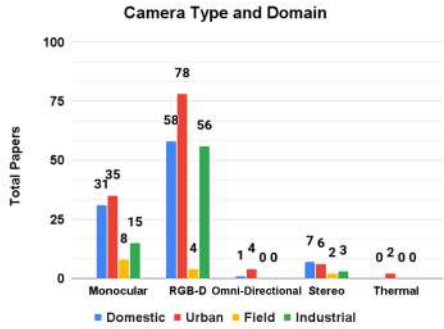
(a) Total camera types–bar chart totals over time



(b) Total camera type–100% stacked area

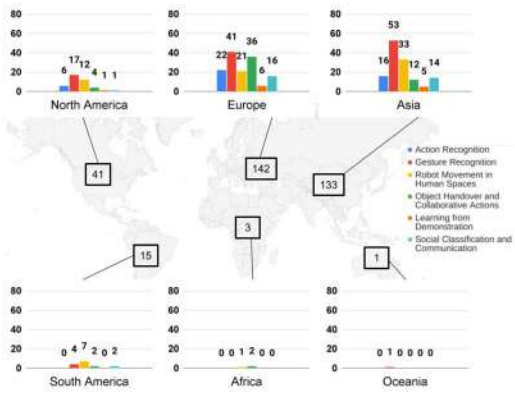


(c) Camera type and robot type

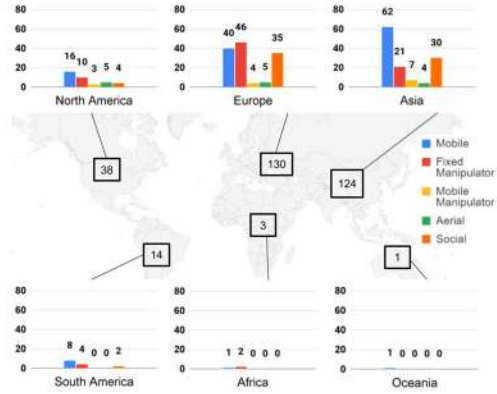


(d) Camera type and domain

Fig. 3. Statistical summary of application, domain, and robot types.



(a) Application area



(b) Robot type

Fig. 4. Statistical summaries of total paper counts and trends per continents.

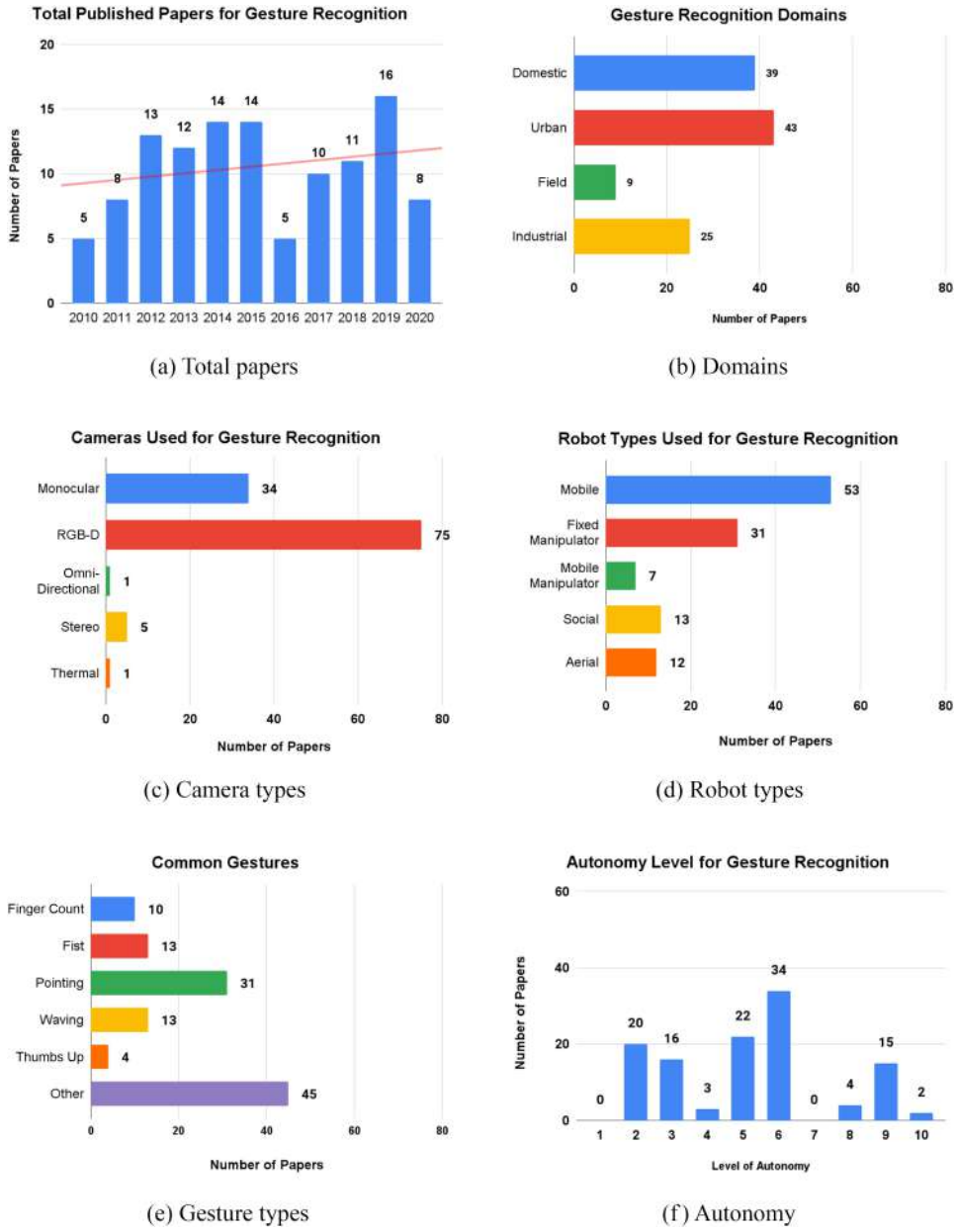


Fig. 5. Gesture recognition totals and summaries.

found to have at least one form of gesture recognition that used vision to identify and respond to the person. Figure 5 depicts the number of gesture-related works, common domains, camera types, robot types, gesture types, and level of autonomy. In 75 papers that used RGB-D cameras, 66 were the Kinect (88%). From the total 116 papers, 79 (68%) involved gestures from body pose, and 37 (32%) involved a hand gesture. Most papers used static gestures (i.e., stationary human pose, $N = 91$, 78%) that did not require visual detection of movement. A smaller portion used

dynamic methods (N = 25, 22%), requiring multiple frames to classify the gesture. Some papers had a blended approach—for example, a static gesture to signify the start of a dynamic gesture (e.g., [458]). Gesture recognition to control robots was used for different robot types: industrial robot arms (e.g., [129, 254, 282, 393]), mobile ground robots (e.g., [79, 296, 491]), and mobile manipulators (e.g., [63, 117, 303, 397]), and less commonly for social robots (e.g., [209]), and aerial robots (e.g., [253]). In robot type, robots had little consistency across different categories with a wide variety of models used for gesture recognition. Continuous control was often used, such as to interact with mobile robots using hand position [333] and small mobile robots with head positions [230]. Last, gestures were also used with teams of robots [13, 63, 253, 297, 323] and by multiple humans in the same scene [270].

6.1.2 Use Case Examples: Mobile Robot Control. In mobile ground robots, hand gestures were often used to control the robot to move forward, left, right, or stop (e.g., [79, 84, 94, 129, 131, 150, 229, 236, 248, 263, 277, 296, 338, 447, 459, 491]). Pointing gestures were often used to direct mobile robots to a specified location (e.g., [6, 81, 202, 330, 348, 416, 426, 472]). Human body movements were used as the control signal, such as shoulder angle to control robot direction based on discrete angles [445]. More dynamic motions were also used to control a robot to move forward, back, left, and right by movement of an arm up and down, left to right, or in circles (e.g., [127]), or hand waving to signify a follow-me or goodbye command (e.g., [145]). Body gestures were used to control an otherwise autonomous mobile robot to turn and stop by moving arms up or down [150], or human shoulder position to control robot velocity [279]. Last, a spherical vision system was used with a mobile robot with three omni-directional wheels to detect pointing gestures from a person wearing a red coat with a blue glove [472].

6.1.3 Use Case Examples: Manipulator Robot Control. In manipulators, robots were controlled using hand gestures such as an open palm [129, 254, 282]. Hand gestures were used to command robot actions, such as to lift or lower the arm [109, 129, 272, 393], rotate the arm [91, 393], open or close the gripper [129, 254, 282], place an object into an open palm [21], return to position when the palm is closed [21], and to set positions for lifting and lowering [109, 129, 272]. Pointing was commonly used for selecting an object for grasping [303, 353, 397, 425], including having the robot arm confirm object selection with the robot arm pointing at the object [353]. Hand gestures were also paired with other body movements for controlling manipulators [254, 282]. Other works included more collaborative actions such as the robot helping to cook by dropping confirmed toppings over a pizza base [353]. A robot equipped with two arms, stereo vision, and tactile sensors could also pick up an object (sponge cube, wooden cube, ping-pong ball) that was selected by a hand pointing gesture from a human, and could release the object onto the palm of the person [192].

6.1.4 Use Case Examples: Mobile Manipulators and Aerial Robots. In mobile manipulators, pointing gestures were similarly used to select desired objects for the robot to pick up (e.g., [63, 117, 303, 351, 397]). In one instance, a mobile manipulator responded to gestures (left and right hand) and user speech to identify, fetch, and handover objects such as a water bottle [62]. Last, a mobile base with arms could wave back to a person waving at the robot and perform a behavior as commanded by a dynamic gesture [248]. Aerial examples include the use of body pose to control an aerial robot, such as right arm up to take off and right arm out to turn right [375] and pointing gestures to select an aerial robot and confirm the selection by touching the right arm to the left hand [253].

6.1.5 Use Case Examples: State Changes. Gestures were also used to signal the robot to commence state changes (e.g., [79, 122, 229, 339]). Some examples include to initiate person guiding or following [339] or to indicate a path direction change for an otherwise autonomous robot [122, 229].

Hand and body gestures were used to start/stop a walk action for a small humanoid [356, 387], body gestures to start/stop person following in an indoor environment [263], or left/right arm raised to change between robot following or parking behavior [296]. In one example, an autonomous mobile navigation robot explored a laboratory and asked humans for directions when a person was detected, translating pointing gestures to a goal in the robot's map [426]. Gestures were also used in learning from demonstration to determine when a demonstration has commenced or concluded (hand [287] and body [398]), or to update a robot's behavior online [118, 340]. Further works on learning from demonstration will be discussed in Section 6.6.

6.1.6 Use Case Examples: Team-Based Scenarios. Gestures were also used in team-based scenarios, such as four mobile robots responding to gestures from a human operator [297]. In this example, the human selected a group of robots by drawing a circle around robots, and directing the robots to go to a chosen location [297]. Other team-based examples include the use of gesture-based interaction to signal to aerial and ground robot teams [318]. Gestures were used to command a small swarm of mobile robots to move into a set configuration using body poses (i.e., arms out front, or above the person's head) [13]. Pointing gestures were often used to select a specific robot from a team of aerial robots [253], and to command a selected group of mobile robots [297]. This included pointing to direct robot attention to other human targets [270]. Last, one example showed a multi-person interaction, including a mobile robot that identified a person by localizing from an audio source, then determining which person to track when they waved at the robot [322].

6.1.7 Use Case Examples: Implicit (Non-Verbal) Communication and Social Interactivity. There were fewer papers around gestures being performed by anthropomorphic robots to mimic human gestures for the purpose of social interaction. An example includes hand waving from a humanoid robot in response to a human human wave [63], helping to facilitate non-verbal communication. In another example, gesture recognition was used for humanoid robots (Pepper and NAO) to perform finger spelling gestures to communicate with hearing-impaired individuals at a public service center [209]. Social interactivity with robots also involved gesture-based games, such as paper-scissors-rock, which required the robot to classify human pose to determine the result [213, 476]. Last, a game with the iCub robot required the robot to recognize each gesture performed by the person to participate in the game [155].

6.1.8 Included Papers. Papers related to gesture recognition are listed here: [6, 13, 15, 21, 28, 30, 32, 58, 60, 62, 63, 71, 74, 76, 78–80, 84, 87, 91, 94, 99, 101, 107, 109, 117, 118, 122, 127, 129–131, 145, 146, 150, 155, 159, 168, 172, 176, 187, 192, 202, 207, 209, 214, 223, 229, 231, 236, 241, 248, 252–254, 263, 270, 272, 276–282, 286, 287, 291, 296, 297, 303, 304, 310, 311, 318, 322, 323, 330, 338–341, 348, 351–353, 356, 372, 375, 384, 387, 393, 397, 398, 408, 416, 423–426, 439, 442, 445, 447, 457–459, 469, 472, 474–477, 483, 484, 491].

6.2 Action Recognition

6.2.1 Overview. Action recognition was defined as the recognition of human actions or activities that were not related to explicit gestures. A total of 44 papers (14% of eligible total) involved some form of action or activity recognition. Figure 6 shows the number of action recognition-related works, common domains, camera types, robot types, action types, and level of autonomy. Of the 36 that used RGB-D cameras, 33 were the Kinect (92%). Action recognition often involved recognizing the person's activities such as action recognition and response (N = 23, 52%), activities of daily living (N = 7, 16%), exercise pose (N = 6, 14%), and recognition of their walking motion (N = 3, 7%). Humanoid robots often used action recognition: NAO [27, 121, 158, 452], Pepper [160, 233, 394], other humanoids [27, 347], and mobile robots [239, 430, 493].

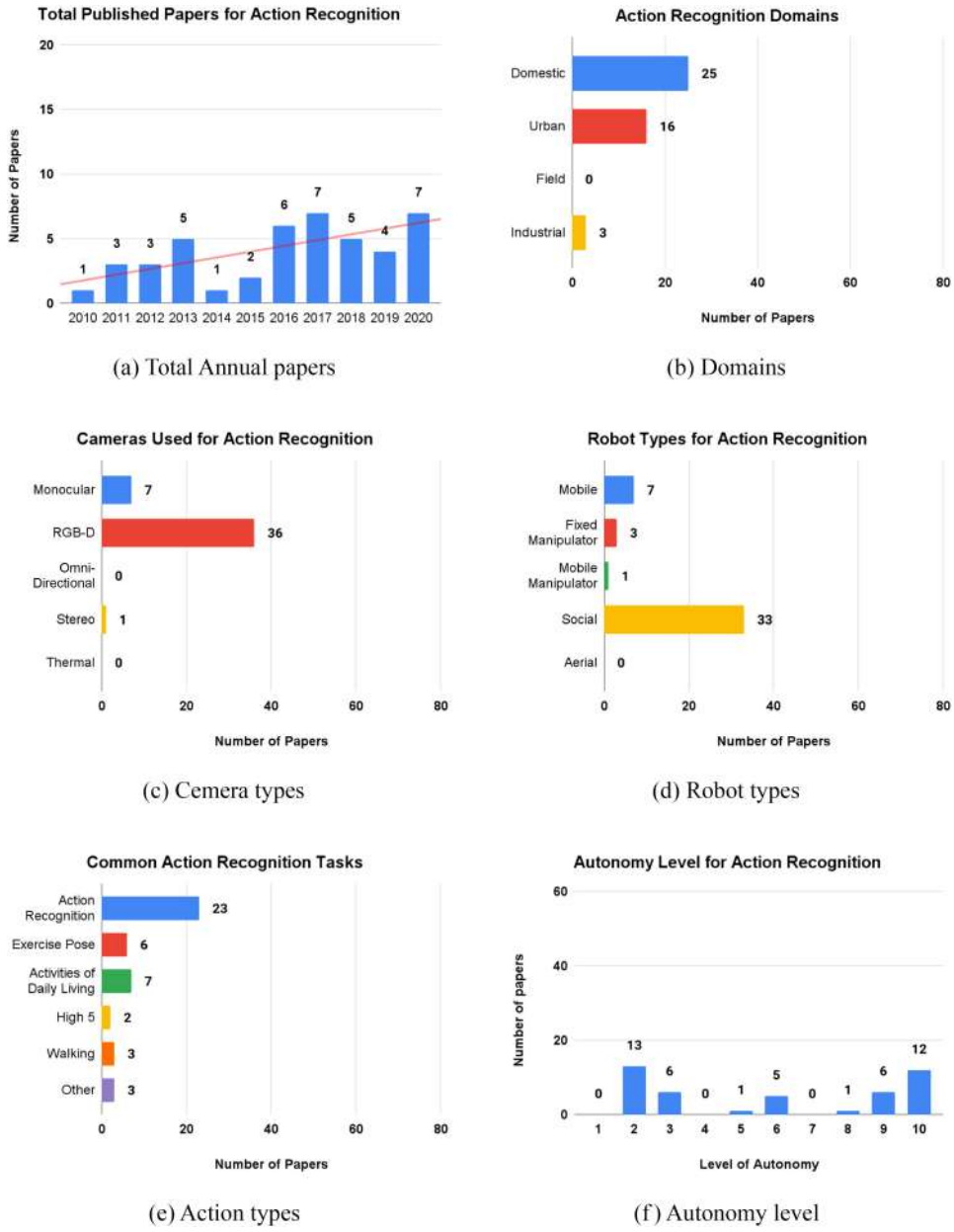


Fig. 6. Action recognition totals and summaries.

6.2.2 Use Case Examples. Action recognition often involved the identification of states. For action recognition, a robot could make a decision on when to offer a person a footrest to rest their feet [493], if the person had fallen down to ask them if they could call an ambulance [394], or when to respond to a human who was handing a bottle to the robot [347]. This included to help robots to recognize multiple actions such as eating, brushing teeth, and making a phone call (e.g., [239]). Action recognition was used to help the robot predict human motion such as walking, eating, and smoking, and to infer the remaining motion sequence after the camera was occluded [160], as well

as to predict human actions in a shared workspace, such as to avoid collision during tool use by recognizing the use of a hammer or reaching for a cup [446]. Action recognition was also used to allow a robot to detect other body actions (i.e., shake head, wave hand) and then perform a corresponding behavior [233]. Other examples include to understand and copy human motions, such as joint positions [26, 102, 191, 193, 267, 405, 418, 467, 499], head positions [69], facial expressions [283, 292, 395, 405], or following continuous position of the person's hand [333] or head [230]. This also involved more rigorous body motions such as humans performing physical activity, and the robot could give feedback to the person on a chosen exercise about pose quality [27, 158, 452]. Other physical activity examples include recording pose count and signal to change exercises if the person waved their hand [27], and the robot learning exercises through action recognition and response to a human demonstrator [158]. In a more applied environment, action recognition was used to detect walking ability for a service robot to be able to guide people of different walking capacities (i.e., wheelchairs, crutches, or walkers) to a suitable entrance [430]. In particular, motion analysis helped the robot to track the person's position to adapt the motion of a robotic walking aid for different mobility levels [72]. For service assistance, action recognition was used to determine when to provide domestic chore assistance, such as filling a glass of water, opening a fridge door [224], clearing a table, or pushing a trivet to the person [240]. Last, some forms of action recognition were used in playful contexts, such as a child and humanoid robot taking turns to perform and recognize a pantomime action such as swimming, painting a wall, or digging a hole [121].

6.2.3 Included Papers. Papers related to action recognition are listed here: [5, 8, 26, 27, 69, 72, 95, 102, 111, 121, 158, 160, 191, 193, 224, 230, 233, 239, 240, 259, 267, 283, 292, 305, 324, 333, 347, 371, 376, 390, 394–396, 405, 418, 430, 435, 446, 452, 463, 467, 493, 498, 499].

6.3 Robot Movement in Human Spaces

6.3.1 Overview. Robot movement in human spaces was classified if the robot had physical movement in a human-based environment and robot movement did not require the human to perform a set pose to signal movement commands to the robot, including if the robot was classified as a (semi-)autonomous vehicle. A total of 74 papers (24% of the eligible total) used robot movement in human spaces. Figure 7 shows the number of action recognition related works, common domains, camera types, robot types, common tasks, and level of autonomy. In the 37 papers that used RGB-D cameras, 29 were the Kinect (78%). Common robot tasks were following the person (n = 55, 74%), avoiding a person (n = 9, 12%), and approaching one or more people (n = 7, 9%). In total, body pose detection was the most common method for identifying a person in an image (n = 57, 77%), followed by face detection (n = 14, 19%). Other methods involved tracking clothing or detection of clothing [298, 453]. Mobile robots often had laser range sensors for person detection [14, 188, 219, 271, 339, 450], obstacle avoidance [10, 465, 479], and for navigation (i.e., Simultaneous Localization and Mapping (SLAM)) [16, 113, 296, 475]. Depth images were also used for obstacle avoidance [37] and SLAM [406]. Ultrasonic sensors were also used for person following [178], and for navigation [89], as well as audio to localize a person not in view [37, 271, 322]. Re-identification when a person who had become occluded when following the person was addressed in several papers [97, 271, 450, 453, 492]. Some papers used multi-modal detectors such as laser or ultrasonic range sensors to identify a person (i.e., detecting legs (n = 9) or shoulders (n = 1)), and audio localization to determine if a person was out of view (n = 4). Others required minimal intervention from the person through gesture commands (n = 9). Proxemics was often considered for appropriate social distance to approach [290] and avoid [41, 406, 428, 465] people, including velocity when a person is detected [428]. Person following environments included both urban [113] and indoor settings [219, 263, 453–455]. Commonly used mobile platforms included

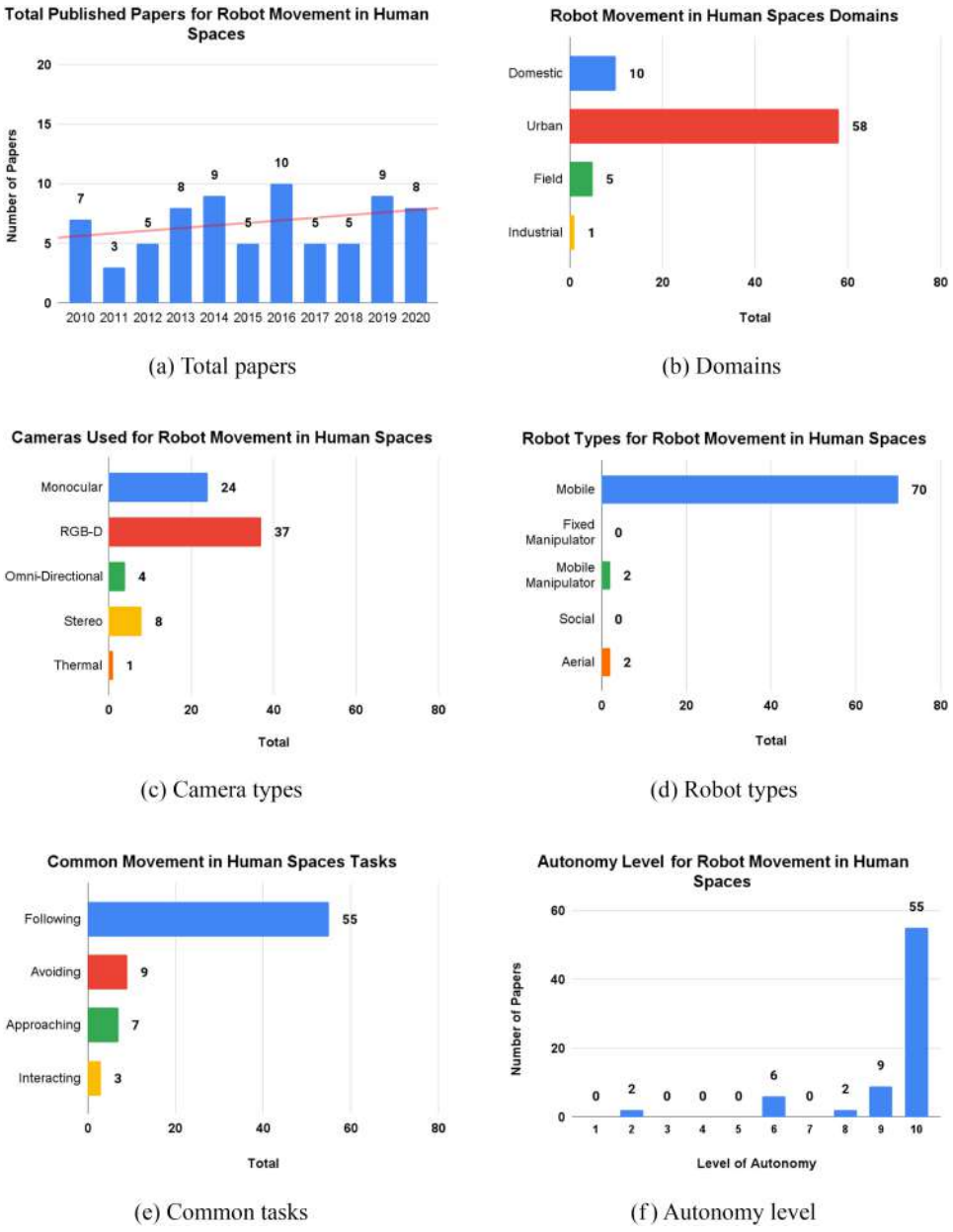


Fig. 7. Robot movement in human spaces totals and summaries.

Pioneer mobile robots ($n = 16$), SCITOS G5 [448, 450], iRobot create [113], a wheelchair robot [453], Turtlebot [97], and even a robotic blimp [470].

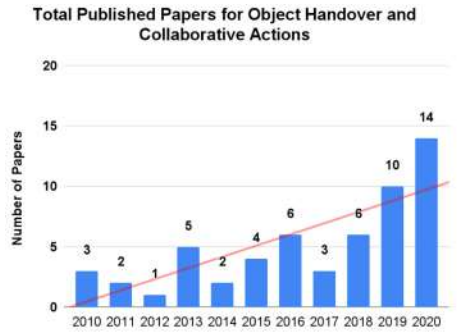
6.3.2 Use Case Examples: Social and Functional Navigation. There were several key examples for robots that followed people for both social and functional reasons [18, 97, 190, 263, 273, 454, 455, 492]. Many works required the robot to approach the person to commence navigation [61, 81, 86, 137, 290], and with human detection and tracking to know where the person was

to avoid colliding with them [41, 150, 406, 428, 465]. In addition, other works included integrating humans into a map for a mobile robot [16]. Other works had the robot approach and then interact with the person by initiating a dialogue [137], or by creating a new goal through the use of human gestures to change the robot's state, such as to commence a variant of a person following behavior [79, 145, 263, 296, 339, 348, 475]. Other state changes included to guide the people avoiding the robot toward a particular path [150]. In works when the robot approached the person, a line following robot approached a person when their face was detected [61] and a PR2 robot approached a person using a real-time proxemic controller [290], helping to bridge functional navigation methods toward social interaction points. In one example, a robot had an omni-directional camera for person detection, and a laser that is used to navigate through the environment using SLAM [113]. This included a social force model that allowed for appropriate social distance when passing people and to avoid having the robot cross into a person's personal or intimate space when passing a person from behind [41, 465]. Robot movement through human spaces was also tested in busy environments, such as an iRobot that could navigate through urban environments, detect human faces, and report back to a supervisory person [113]. Multiple people were also detected and used as landmarks for integration into a SLAM solution for a Pioneer robot following a path [16]. This included robust person following, even when multiple people were present in a scene and when the target became occluded [316]. The iRobot was also used to identify the face of a specific person and keep the target person's face in view [490]. In robot navigation for human environments, some robots also conducted multiple tasks together as part of its use case. For example, a mobile R2D2 robot with arms carried a smaller mobile robot with a gripper. The robot could wave when it detected a person's face, and delivered a drink based on the distance to the target face. The main robot could also deploy a smaller robot with a gripper to pick up small objects in its path [86]. Others had different methods of person following, such as following a person through airspace. Two examples involved using a two-camera vision system with a small aerial robot to detect and hover above the hand of a person wearing a glove with the intention to pass the robot between people [298], and a monocular camera attached to an autonomous blimp robot to detect and follow a person's face [470].

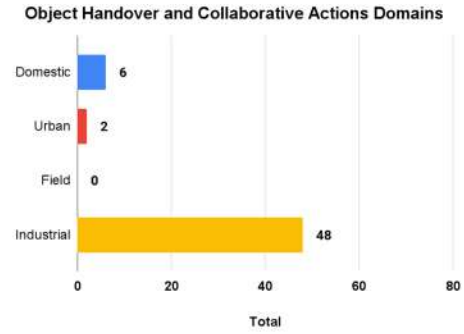
6.3.3 Included Papers. Papers related to robot movement in human spaces are listed here: [10–12, 14, 16–18, 34, 37, 41, 59, 61, 79, 81, 85, 86, 90, 97, 108, 113, 125, 137, 145, 148–150, 157, 163, 178, 181, 188, 190, 204, 208, 215, 219, 242, 263, 265, 271, 273, 285, 290, 296, 298–300, 316, 322, 328, 337, 339, 348, 349, 361, 379, 406, 428, 440, 448, 450, 451, 453–455, 465, 470, 475, 478, 479, 488, 490, 492, 494].

6.4 Object Handover and Collaborative Actions

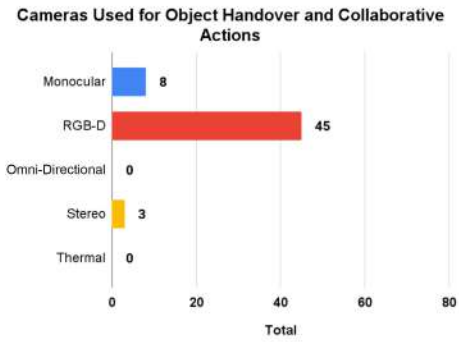
6.4.1 Overview. Object handover and collaborative action papers included a robot capable of manipulating objects while the interaction did not require the person to perform a set pose (i.e., the person was detected without performing a gesture or action). A total of 56 papers (18% of the eligible total) involved an object handover or collaborative action. Figure 8 shows the number of works, common domains, camera types, robot types, common tasks, and level of autonomy. In the 45 papers that used RGB-D cameras, 34 were the Kinect (75.5%). The most common use case involved a human-aware work space where the robot had to operate safely in a shared space (N = 26, 46%), followed by direct control of a robot arm (N = 15, 27%), object handover (N = 7, 13%), and collaborative manipulation (N = 3, 5%). For shared space work, most actions were to improve safety outcomes for the human. For instance, if a human was detected in the shared area, the robot would came to a halt [20, 312, 319, 385, 407], slow down [51, 482], or change its trajectory to avoid contact [73, 112, 136, 232, 262, 313, 314, 331, 409]. In some shared work instances, the person was required to be in a safe standing pose before robot commands were accepted [169]. In object



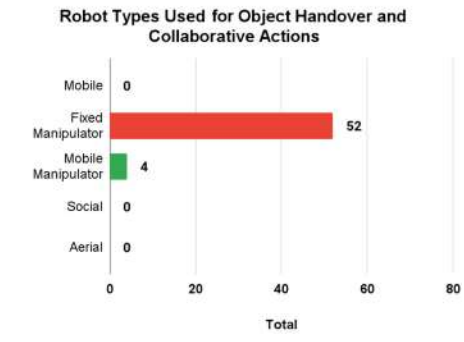
(a) Total papers



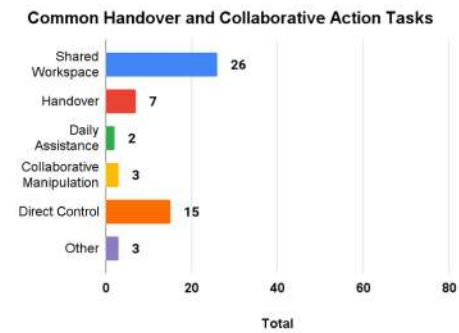
(b) Domains



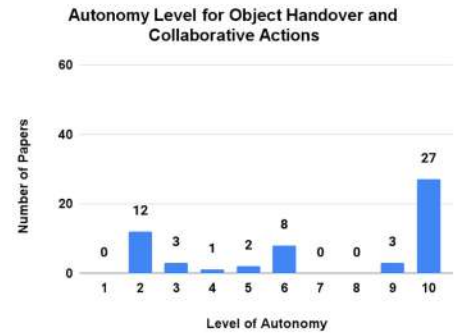
(c) Camera types



(d) Robot types



(e) Common tasks



(f) Autonomy level

Fig. 8. Object handover and collaborative actions totals and summaries.

handover cases, the handover process often involved both passing objects from robot to human [25, 38, 192, 392] and from human to robot [382]. These handover actions often used force [38] or tactile sensors [192] to determine when to release the object. Another use case was robotic vision to control robot arms by matching tool center point with human hand positions [25, 52, 92, 169, 185]. Last, other related works also used both robotic vision and an IMU [169]. Commonly used platforms were the Kuka ($N = 9$), Universal Robots (UR) series ($N = 8$), ABB Industrial ($N = 4$),

Franka Emika Panda ($N = 3$), Stäubli ($N = 2$), Baxter ($N = 3$), Lynxmotion ($N = 2$), Sawyer ($N = 1$), WAM robot ($N = 1$), and other types of arms ($N = 19$).

6.4.2 Use Case Examples. Collaborative actions often involved the robot providing assistance during a specific task. In relation to specific tasks, robotic vision was used to track a person's hand, and a force sensor to sense contact in an screwing task [88] and to match the end-effector position with hand position to pick up and pass an item [25]. Others included moving the tool center point of a robot arm to follow a human hand to execute a grasp action when the human placed both hands out in front [52] or a robot performing cooperative sheet folding with a hand detected against the fabric corner with the robot arm picking up the opposite corner to perform the fold [225]. Some collaborative actions took a mixed sensor approach such as joint manipulation of a wooden stick with a Franka Emika arm with the goal to keep the stick horizontal using data from a wearable IMU device attached to the wrists, elbows, and the hip fused with extracted skeleton information to provide the robot with an accurate human pose [487], a camera mounted to the end effector of a Stäubli industrial arm to follow a human hand using visual servoing, and a force sensor to know when the robot should release the object [38]. Collaborative actions also had specific person-centered applications, such as robots to assist persons with disabilities through assistive dressing with a jacket [425], soft robots to assist in bathing [114], and for a surgeon to instruct a robot to fetch an item during an operation [221]. Other applications were orientated around a specific outcome, such as to improve safety. Safety outcomes included robotic vision to identify a potential collision with a person and stop the robot in collaborative assembly task [312], to inform the robot to deactivate if a person was too close [319], to assist in consideration of proxemics when handing over a water bottle [392], and image and torque sensing to help reduce robot arm speed based on human proximity and contact [51]. Other safety-related functions included the robot using the shoulder and hand position to determine if the person was directed toward the task and, if not, to pause until the person was again engaged or no person was present [55], and robotic vision to help address ergonomics, such as to adjust the working height of the end effector based on human pose (i.e., height and arm position) [427].

6.4.3 Included Papers. Papers related to object handover and collaborative actions are listed here: [20, 25, 29, 30, 38, 45, 51, 52, 55, 56, 73, 88, 92, 100, 107, 108, 112, 114, 126, 136, 144, 147, 162, 169, 185, 192, 221, 225, 232, 262, 301, 312–314, 319–321, 331, 374, 378, 380, 382, 385, 392, 401, 404, 407, 409, 412, 425, 427, 444, 482, 484, 487, 496].

6.5 Social Communication

6.5.1 Overview. Categorization for social communication required that the robot needed to perform a social behavior, or be capable of socially interacting with a person. A total of 33 papers (33% of the eligible total) involved a social interaction between a person and a robot. Figure 9 shows the number of social interaction works, common domains, camera types, robot types, common social tasks, and level of autonomy. In the 16 papers that used RGB-D cameras, 12 were the Kinect (75%). Common tasks required a robot to converse with a person ($N = 10$, 30%), detect social engagement ($N = 6$, 18%), or to approach people in a social way ($N = 3$, 9%).

6.5.2 Use Case Examples. Social actions included having the robot face toward the person who was talking [70, 102, 249, 343], to detect the active speaker in a group of people using facial recognition and audio [70], to commence a conversation when a person is detected [343], to identify when the person had finished talking [49, 201], to perform face detection and gestures during a conversation with a person [102], to wave when a waving gesture was detected [152], and to recognize engagement levels through facial expression and gaze detection [68, 373, 405] from head

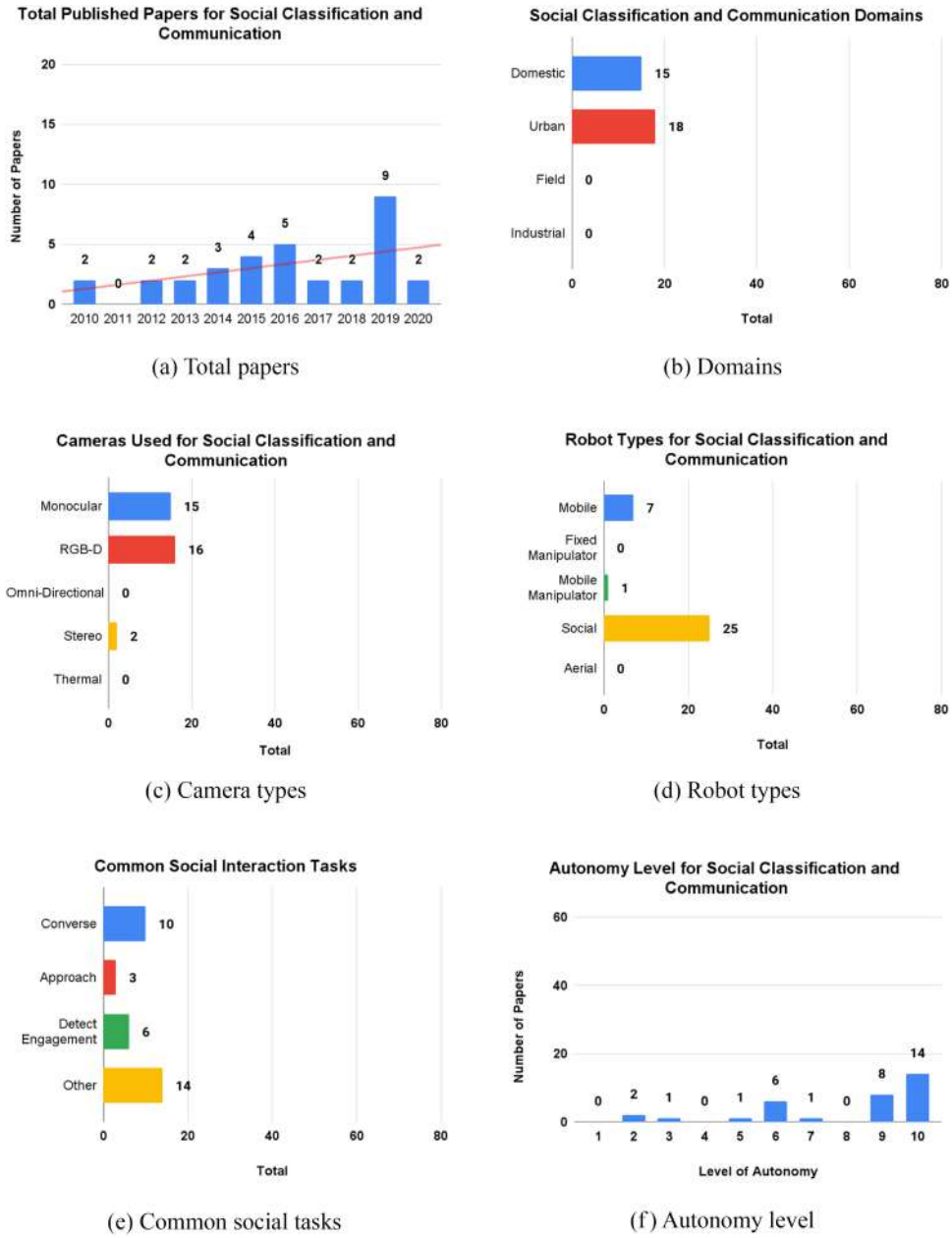


Fig. 9. Social communication totals and summaries.

movements such as nodding or shaking [373]. Other actions related to social interaction included classifying facial expressions [216, 250], gender and person identification [89], and age and gender estimation using facial cues [70]. Social communication was also used for mobile robot situations, such as to detect if the person wanted to interact with it [247, 309, 432], and to determine social group configurations to enact appropriate social distance conventions [420]. Multi-person applications were also explored, such as speech and vision sensing with an iCat robot head with two

arms to greet people, take orders, and serve drinks to multiple people [139], and the Pepper robot in a restaurant where the robot was required to point to seating locations, repeat bar orders, relocate a person, and deliver an item to a person even if they had moved from their original position [238]. This included other tests for if there was more than one human in the robot's field of view (e.g., [247, 420]).

6.5.3 Included Papers. Papers related to social communication are listed here: [19, 33, 42, 49, 68, 70, 89, 102, 106, 139, 152, 177, 201, 206, 216, 234, 238, 246, 247, 249, 250, 309, 327, 343, 354, 373, 391, 400, 405, 420, 432, 473, 489].

6.6 Learning from Demonstration

6.6.1 Overview. A total of 12 papers (4% of the eligible total) involved some form of learning from demonstration. Figure 10 shows the number of learning from demonstration works, common domains, camera types, robot types, common tasks, and level of autonomy. In the 10 papers that used RGB-D cameras, 9 were the Kinect (90%). Learning from demonstration tasks included manufacturing assistance ($N = 6$, 25%), human interaction ($N = 2$, 16%), scene understanding ($N = 2$, 16%), and behavior learning ($N = 2$, 16%). Robots could learn by watching a person perform a task, or through collecting data from an interaction between two people. Gestures were often used to enter a demonstration mode, including when a human would move the robot ($N = 3$, 25%), or to provide instructions ($N = 2$, 16%). Robots were often humanoid robots ($N = 7$), which included the iCub ($N = 2$) [370, 481], Pepper ($N = 1$) [473], SARCOS humanoid ($N = 1$) [340], imNeu ($N = 1$) [251], a small humanoid ($N = 1$) [347], and an anthropomorphic robot head ($N = 1$) [471]. There were five industrial robot arms: Kuka ($N = 3$) [118, 287, 463], WAM robot ($N = 1$) [424], and FANUC ($N = 1$) [398]).

6.6.2 Use Case Examples. Gestures were often used in learning from demonstration tasks, such as hand and body gestures to signal the beginning and end of a demonstration [287, 398], or teach the robot online [118, 340]. Other examples involved pointing and speech to show an industrial arm where to work [118], teaching the robot to perform a peg grasping task [251], and to follow actions in a watering task from visual gaze [471]. Learning methods included programming the robot to become compliant once a hand gesture command had been received to move the end effector for a chosen action, with a second gesture to signal the completion of programming, and for the robot to begin executing a new task [287]. Other methods included fine-tuning actions when the robot end effector was close to the desired position [118], to change the periodic motion of a humanoid end effector [340], and change the motion of its hand in response to a human coaching gesture [340]. Learning from demonstration also included humans teaching a small humanoid [347], or from mirrored human examples such as doing a task with a bottle by observing related actions (hold, place, and take) and learning to replicate it [370].

6.6.3 Included Papers. Papers related to learning from demonstrations are listed here: [118, 251, 287, 340, 347, 370, 398, 424, 463, 471, 473, 481].

7 RQ3. WHAT IS THE HUMAN-ROBOT INTERACTION TAXONOMY FOR ROBOTIC VISION IN HUMAN ROBOT COLLABORATION AND INTERACTION?

7.1 Summary

This section will explore the human-robot interaction taxonomy data (Figures 11 and 12) as informed by human-robot interaction and robot classification taxonomies (e.g., [40, 464]). Detailed explanation of taxonomy hierarchy and their relevant classification labels can be found elsewhere [40, 464], and a brief summary of labels has been listed in the review information and

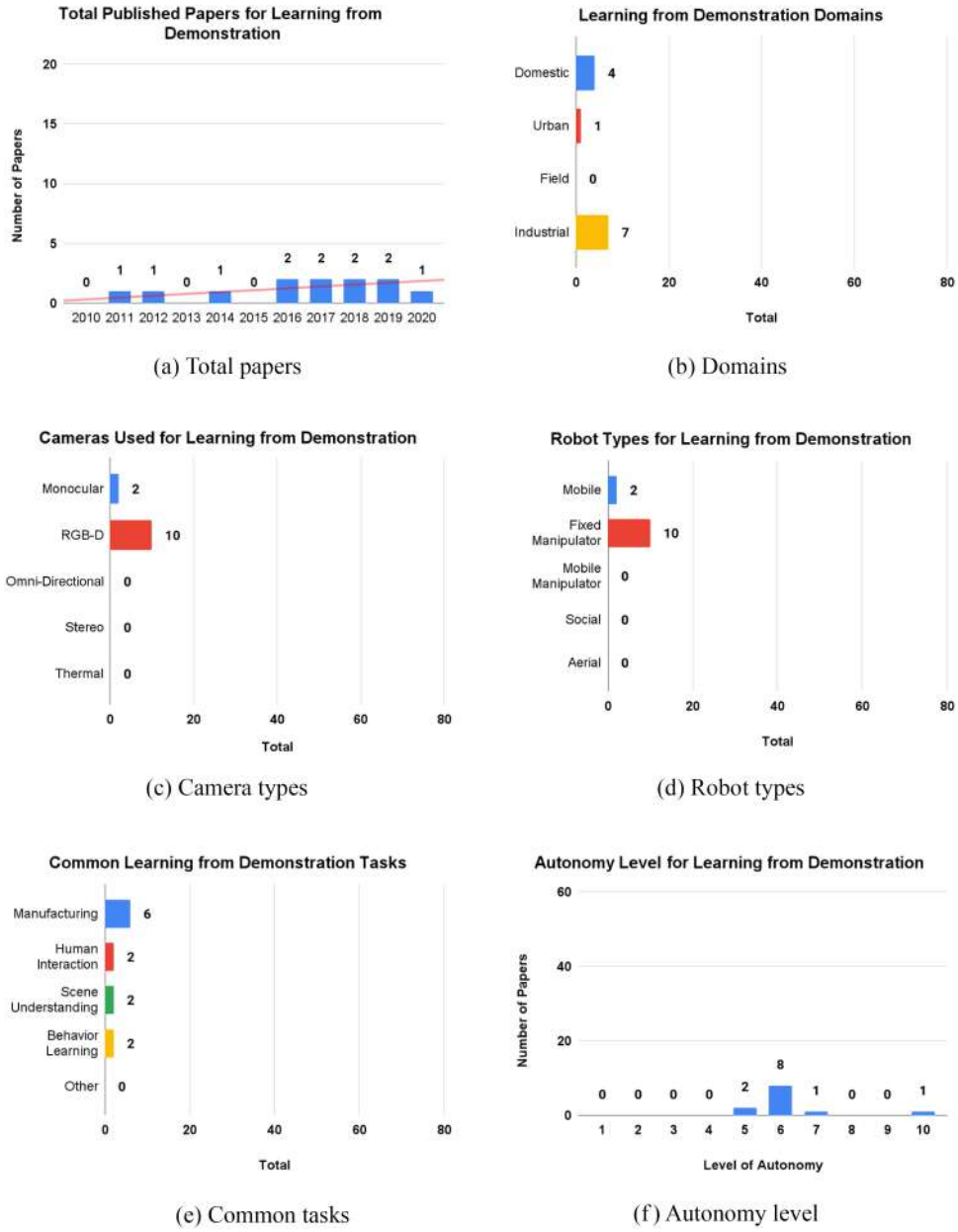


Fig. 10. Learning from demonstration.

categorization section (Section 3.1). Task type was relatively broad with limited consistency between studies and has been reported in individual sections listed prior to this section (see Section 6). Task criticality for most robot use cases was identified as low criticality (N = 271, 88%) compared to medium (N = 32, 10%) or high (N = 7, 2%) classification, which may further support the emergent nature of robot roles in easier use cases as a first application. There was also a link between task difficulty and frequency, where less difficult tasks are more commonly investigated and harder tasks are less represented. These classification patterns were similar to our own custom metric on

(a) Task criticality

(b) Robot morphology

(c) Task criticality and robot domain

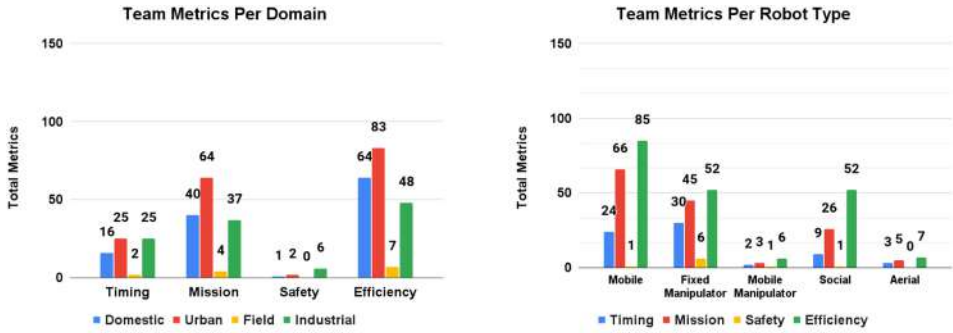
(d) Robot interaction formation

(e) Human-robot proximity

(f) Type of interaction

Fig. 11. Taxonomy data summary.

a single score for overall task evaluation using task complexity, risk, importance, and robot complexity: low (N = 249, 80%), medium (N = 51, 17%), and high (N = 10, 3%). Robot morphology was categorized as anthropomorphic (human-like), zoomorphic (animal-like), and functional (neither human-like nor animal-like, but related to function). Most systems were functional (N = 203, 65%) compared to anthropomorphic (N = 104, 34%) or zoomorphic (N = 3, 1%). Functional robots were more likely to be used in medium and high task criticality studies compared to anthropomorphic or zoomorphic robots. A high volume of works had a 1:1 human to robot ratio (N = 281, 91%) with

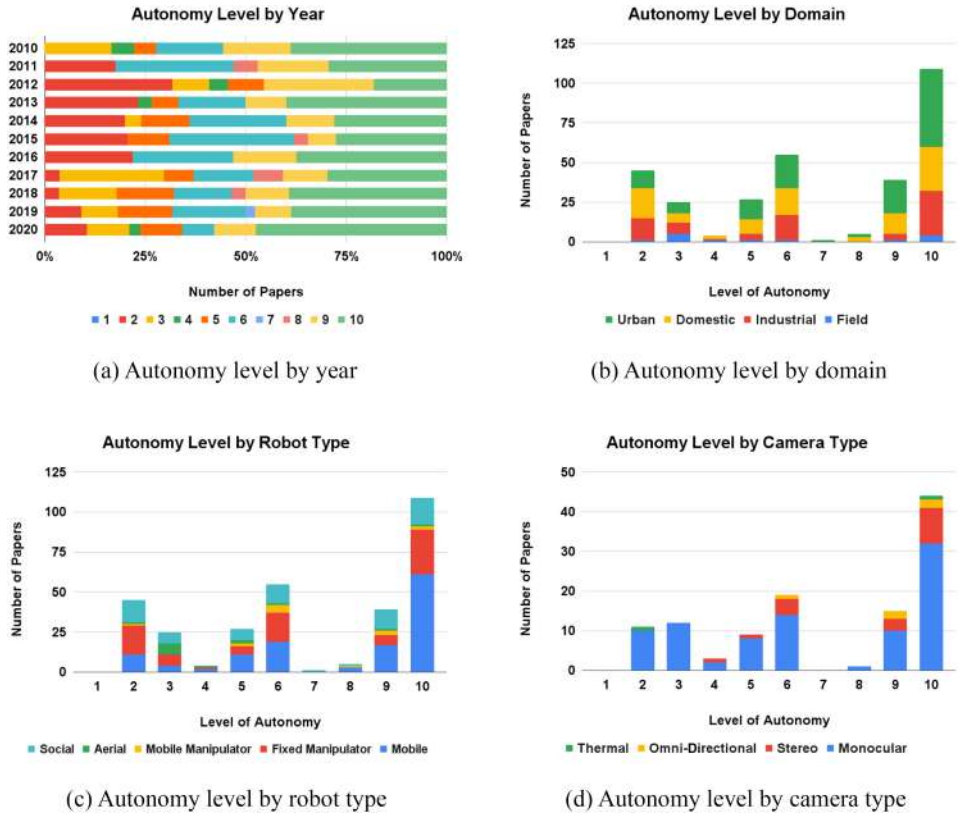


(a) Metrics and domain (b) Metrics and robot type

Fig. 12. Taxonomy data summary: team metrics.

an overall mean of 1.08, 9 with more than one robot, 20 with more than one human, a maximum reported ratio as 5 [246], and minimum reported ratio between 0.1 [13] and 0.07 [318]. Human-robot teams often had 1:1 human-robot team compositions, showing that the methods and team setups focused on a single human to potentially assist in robustness and utility of robotic vision in the collaborative scenario. Homogeneous teams were used for 307 (99%) of the reported studies, with only 3 (1%) of studies using heterogeneous robot team compositions (e.g., [86, 238, 253]). Level of shared interaction among teams was high in the 'A' formation (N = 280, 90%) as predicated on the earlier reported ratio of people to robots (N = 281, 90%). There were 8 papers with a 'B' formation (one human with multiple robots using a single interaction), 5 with a 'C' formation (one human with multiple robots using a separate interaction for each robot), 6 with a 'D' formation (multiple humans with one robot, where the robot interacts with the humans through a single interaction), and 11 with an 'E' formation (multiple humans with one robot, where each human interacts with the robot separately). In terms of the type of human-robot physical proximity, interacting (N = 186, 60%) was the highest followed by following (N = 64, 21%) and then avoiding (N = 24, 8%). No studies that used tele-operation were eligible in this review, but for the remainder of eligible studies, nearly all (N = 309, 99%) had the robot as synchronous (same time) and collocated (same place) with the exception of one study that was non-collocated (e.g., [229]). Autonomy level scoring by Beer et al. [40] was used, but no scores were classified on level 1 due to exclusion criteria that the robot must not be manually operated by the human. For the remainder, robots were often high on autonomy, which may have been skewed by initial entry criteria that required the robot to use robotic vision to perform an action or response. Figure 13(a) depicts that papers over the past 10 years often had level 2 (tele-operation: robot prompted to assist but sensing and planning left to the human), level 6 (shared control with human initiative: robot senses the environment, develops plans/goals, and implements actions while the human monitors the robot's progress), or level 10 (full autonomy: robot performs all task aspects autonomously without human intervention). Figure 13(b) depicts that there was a relatively even spread of autonomy level across the four domains, Figure 13(c) depicts mobile and fixed manipulators were most often used with level 10 autonomy, with similar trends seen across the autonomy levels, and Figure 13(d) depicts camera types per robot autonomy level.

7.1.1 Included Papers. Papers where a mobile robot was used are listed here: [6, 10–18, 28, 34, 37, 41, 59, 61, 63, 72, 79–81, 84, 85, 89, 90, 94, 97, 101, 113, 117, 122, 125, 127, 131, 137, 145, 148–150, 157–159, 163, 177, 178, 181, 187, 188, 190, 202, 204, 207, 208, 215, 219, 223, 229, 230, 236, 239, 241, 242, 247, 248, 263, 265, 271, 273, 277–281, 285, 290, 296, 297, 299, 300, 309, 316, 322, 323, 328, 330, 333, 337–339, 343, 348, 349, 361, 372, 379, 406, 408, 416, 420, 423, 426, 428, 430, 432, 440, 445, 447, 448, 450,



(a) Autonomy level by year

(b) Autonomy level by domain

(c) Autonomy level by robot type

(d) Autonomy level by camera type

Fig. 13. Autonomy level trends and summaries.

451, 453–455, 458, 459, 465, 472, 475, 477–479, 483, 488, 490–494]. Papers where a fixed manipulator was used are listed here: [20, 21, 29, 30, 38, 45, 51, 52, 55, 56, 60, 71, 73, 74, 78, 88, 91, 92, 100, 107, 109, 112, 114, 118, 126, 130, 136, 144, 147, 169, 176, 185, 192, 214, 221, 225, 231, 232, 240, 251, 252, 254, 262, 282, 286, 287, 301, 312–314, 319–321, 331, 340, 352, 353, 370, 374, 378, 380, 382, 385, 393, 398, 401, 404, 407, 409, 412, 424, 425, 427, 442, 444, 446, 463, 469, 474, 482, 484, 487, 496]. Papers that used a mobile manipulator are listed here: [25, 58, 62, 86, 108, 129, 162, 224, 238, 272, 303, 351, 392, 397]. Papers that used an aerial robot are listed here: [76, 99, 253, 276, 291, 298, 304, 310, 311, 318, 341, 375, 384, 470]. Papers that used a social robot are listed here: [5, 8, 19, 26, 27, 32, 33, 42, 49, 68–70, 87, 95, 102, 106, 111, 121, 139, 146, 152, 155, 160, 168, 172, 191, 193, 201, 206, 209, 216, 233, 234, 246, 249, 250, 259, 267, 270, 283, 292, 305, 324, 327, 347, 354, 356, 371, 373, 376, 387, 390, 391, 394–396, 400, 405, 418, 435, 439, 452, 457, 467, 471, 473, 476, 481, 489, 498, 499].

8 RQ4. WHAT ARE THE VISION TECHNIQUES AND TOOLS USED IN HUMAN-ROBOT COLLABORATION AND INTERACTION?

This section provides a detailed review of robotic vision techniques used in selected papers, including methods, algorithms, datasets, cameras, and methods to allow robots to provide information or take action. Common techniques are discussed in detail to provide clear trends on how methods and techniques have been adapted from computer vision to robotic vision problems. However, many emergent techniques from computer vision are yet to be seen in HRI/C works. The use of these techniques may create many new opportunities for robots to help people in domains that

Table 1. Camera Types Represented in the Corpus of Papers and Their Properties

Camera	RGB Resolution	Depth Resolution	Depth Range	Fieldof View (Degrees)	Frame Rate	Examples
Microsoft Kinect	1920 × 1080	512 × 424	0.5–4.5 m	70 × 60	30 fps	[262, 267, 487]
Intel RealSense	1920 × 1080	1280 × 720	0.3–3 m	69 × 42	30 fps	[28, 327, 390]
PrimeSense	1280 × 960	640 × 480	0.35–3 m	54 × 45	30 fps	[406, 406, 484]
Asus Xtion Pro Live	1920 × 1080	640 × 480	0.5–6 m	57 × 44	30 fps	[8, 314, 409]
Logitech c9xx	1920 × 1080	–	–	78	30 fps	[32, 68, 427]

have not yet been explored due to technical challenges, and to the speed or accuracy needed to be useful to the person, as discussed in Section 12. Yet this level of development and testing will likely experience a translation delay from the advances seen in computer vision, given the need for human study approvals, extensive testing on hardware components that are subject to error and malfunction, and robust results that can meet peer-review standards for publication.

This RQ is important to the field of HRI/C because understanding the visual world is fundamental for interacting with the environment and its actors. It is important for the HRI/C researcher to understand the computer vision techniques and tools that have been used so far, since these are often robust, well understood, and well suited to HRI/C applications. More significantly, these patterns of usage allow us to identify areas for improvement, where an over-reliance on traditional or proprietary techniques might be inhibiting progress.

To begin, robot vision requires the robot to perceive a human and/or their actions to be able to provide a function, task, or service to the person during human-robot collaboration and interaction. Robotic vision is often performed in a two-step process. First, localization detects where the human is located in the robot's field of view, often to the granularity of the position of specific parts of the body, such as the hands. Second, classification determines what gesture, action, or expression is being shown by the person. This can include tracking across image frames to help resolve actions that are ambiguous or reliant on motion cues, and to facilitate continuous interaction between the person and the robot. This visual information can then provide the robot with important information that can help the robot to determine its next action or movement, making the vision process central to the function and utility of the robot to the person. Multiple sensors are also commonly used to provide different types of relevant information, especially the combination of color and depth measurements. The next section will discuss implemented solutions for human detection and pose estimation, gesture classification, action classification, tracking, and multi-sensor fusion. A summary of camera types, including relevant information and examples, is listed in Table 1.

8.1 Human Detection and Pose Estimation

Detection of the location of the person or people in a stream of visual data is often the first step for reasoning about them. Pose estimation techniques go beyond coarse detection (e.g., of bounding boxes) to find the location of body parts and their connections (i.e., skeleton estimation).

8.1.1 Commercial Software. The most common approach from the corpus of selected papers was to apply skeleton extraction and tracking software to depth images from RGB-D cameras (N = 107), which combines detection and pose estimation. This software was primarily sourced from the Microsoft SDK for the Kinect camera or OpenNI for PrimeSense cameras. This approach has the advantage of using commercial-grade software that is easily available, real time, and robust. Example papers include the following: [18, 20, 26, 41, 55, 58, 73, 79, 88, 92, 94, 102, 118, 127, 136, 150, 158, 185, 191, 193, 225, 232, 233, 251, 253, 262, 263, 267, 292, 296, 309, 312, 313, 340, 343, 347, 382, 394, 398, 405, 418, 420, 425, 444, 454, 465, 467, 475, 487, 493, 499].

8.1.2 Face Detection. The next most common method (especially for earlier papers) was to detect a person by first detecting their face using the Viola–Jones method [438] ($N = 34$). This approach uses Haar filters to extract features from an image, followed by AdaBoost [143] to make predictions using a cascade of classifiers. This approach has the advantage of being extremely fast and performant, without requiring depth information. Depth can be used if available to disambiguate false positives based on the realistic size of a face [490]. Example papers that use this method include the following: [10, 25, 37, 61, 97, 101, 102, 113, 117, 137, 160, 216, 219, 230, 250, 271, 281, 283, 339, 373, 426, 453, 470, 471]. Once the face location is identified, this was sometimes used to help detect other body parts, such as hands [117, 270, 297].

8.1.3 Segmentation. Color segmentation is a simple and fast method to distinguish regions containing skin from a video stream. However, color segmentation is less robust than many other methods to different skin tones, scene colors, and occlusions such as clothing and hair. Segmentation is typically performed in HSV [231, 277, 397] or YCrCb [91, 270, 459] color spaces. This includes segmentation using skin threshold values [84, 91, 139, 270], or by using histogram analysis from a skin sample [231, 277, 459]. For example, Xu et al. [459] used color segmentation thresholds tuned for human skin tones and incorporated depth to remove segmentation errors. Accurate color segmentation facilitated by using colored items was also used, such as clothing [80, 131, 178, 472], or colored tape around the person's body [236]. For example, Miyoshi et al. [298] had an aerial robot follow a glove of a known color that can be easily segmented. Given a segmented image, morphology and edge detection operations can be used to approximately extract the person from the image [385]. Depth segmentation is an alternative that relies on some assumptions about the scene and setup [117, 287, 333, 459]. For example, Paulo et al. [333] segmented hand gestures by setting minimum and maximum distances to the sensor, and Mazhar et al. [287] expanded a hand keypoint to a hand segmentation using depth.

8.1.4 Region of Interest. Regions of interest, such as a bounding box around a body, face, or hands, can also be acquired using deep learning techniques. The most prevalent models include the Region-Based Convolutional Neural Network (R-CNN) family of architectures [151], and the single shot detectors [258, 359]. These approaches are fast and robust, and are able to detect multiple people in a single image. The R-CNN family includes two-stage networks that generate many object proposals, before filtering and classifying them, whereas single-stage networks predict the final bounding boxes in one go. This saves computation time at the expense of accuracy. From the corpus of selected papers, Vasquez et al. [430] use Fast R-CNN to extract a human bounding box from an image of a person using a walking aid. The SSD network is used in other papers to locate bodies [97, 190] or faces [69, 247, 448] in an image. For example, Weber et al. [448] used an SSD network for face detection, fit faces to a deformable template model, and tracked the detections using LSD-SLAM (large-scale direct monocular SLAM) [123]. While there are many other methods for detecting humans that have been deployed on robots, such as detection from 2D range data [24, 48, 203], these approaches were not present in the corpus of papers.

8.1.5 Pose Estimation. Deep learning is also commonly used for human pose estimation from RGB images. Image-based pose estimation usually refers to extracting keypoints and skeletons from an image of one or more people, without any additional depth information. Out of the HRI/C corpus surveyed, Convolutional Pose Machines (CPM) [449] and OpenPose [64] were used most often ($N = 15$) for robotic vision. The OpenPose approach trains a network to predict the location of all joints in the image, then assembles skeletons using learned part affinity fields in a post-processing step. The main advantage of this bottom-up approach is that it runs in real time, and runtime is independent of the number of people in the image, making it particularly suitable

for human-robot interaction. An additional explanation for its prevalence among HRI/C papers is that it has a well-known, well-maintained, and high-quality codebase that is user-friendly and publicly available. Example papers can be seen here: [72, 160, 239, 282, 287, 296, 427, 482].

8.2 Gesture Classification

For human-robot interaction or collaboration, the human is often required to perform a specific hand or body configuration to interact with the robot. Once the region of interest or skeleton is extracted, hand and body gestures are classified through a variety of methods, depending on the dynamic or static nature of the gestures being performed. For static hand gestures, classification was often performed from segmented images by determining the contours (boundary pixels of a region) and the convex hull (smallest convex polygon to contain the region) as a method of counting the number of fingers being held up [91, 109, 231, 397, 459]. The distance of each contour point to the centroid of the segmented image was also used for finger counting [270, 277]. From skeleton information, gestures were often classified from the 3D joint positions of the human skeleton using the angular configuration of the joints [27, 87, 129, 145, 158, 172, 233, 253, 296, 303, 323, 444, 452, 469]. For example, Ghandour et al. [150] trained a neural network classifier to recognize body pose gestures using joint positions from a single image as input. In contrast, dynamic gestures, which consist of a sequence of poses, require multiple image frames for classification. Common algorithms used to track and classify dynamic gestures include hidden Markov models [62, 117, 145, 330, 408, 458], particle filters [62, 118, 330, 348], and dynamic time warping [63, 338, 351, 439]. For example, Tao and Liu [408] used a hidden Markov model to classify hand waves in various directions, Cicirelli et al. [94] used a neural network to recognize gestures from an input of Fourier-transformed joint positions, and Li et al. [248] reason about temporal information using an LSTM to determine the intention of the person. Although traditional machine learning techniques have been used for gesture and pose classification, such as k-nearest neighbors [146, 339], SVM [80, 84, 122, 155], and multi-layered perceptrons [281, 408], deep neural networks have become increasingly prevalent. Convolutional neural networks have been used to classify gestures directly from color images [21, 287, 459] or depth images [209, 254]. Other architectures have been used to classify gestures from intermediate representations such as skeletal information [94, 150, 248] and other sensory inputs [281, 408]. Practitioners can fine-tune pre-trained models on task-specific data [287], or train from scratch with a custom dataset [447, 459]. As previously indicated, neural networks have been used to classify both static [150] and dynamic [94, 250] body poses.

8.3 Non-Gestural Action Classification

Human action classification involves the identification of specific types of human motions from video streams. For this section, gestural actions are considered to involve gestures where the person is explicitly trying to communicate with the robot, and non-gestural actions involve actions such as walking, eating, running, and sitting down. Non-gestural actions can be important for robots to recognize and facilitate contextual understanding, but the robot may not require an immediate response from the person, unlike a gesture action. Several action classification methods operated on pre-detected human keypoints [72, 158, 160, 233, 239, 394, 493]. For example, Görer et al. [158] compared the keypoint position of motion identifier joints on an elderly user performing an exercise pose with those of a human demonstrator to identify any disparities. In another example, Vasquez et al. [430] classified the category and estimated the 3D position and velocity of a person with a walking aid using a ResNet-50 [180] network and a hidden Markov model. Efthymiou et al. [121] showed that dense trajectories [443] provide better features that result in

more accurate action classifications than convolutional neural networks, when there is a mismatch between training and testing data (e.g., identifying the actions of children from models trained on large action recognition datasets where adults are more prevalent). Last, Gui et al [160] trained a generative adversarial network from a sequence of skeleton keypoints over time, extracted using OpenPose [64], to generate plausible motion predictions.

8.4 Social Classification

This section explores the classification of social factors that were not directly associated with a specific gesture or action, including facial expression, level of engagement, and intent to interact with the robot. For example, Saleh and Berns [373] classify head movements, such as nodding and shaking, with an SVM, using the direction and magnitude patterns of depth pixels as features. Facial expressions are classified by fitting an active appearance model [283], by using Gabor filters with principal component analysis and an SVM [216], or using a neural network classifier on facial keypoints [292]. Gender classification was performed from principal component analysis of face regions [216], or using an SVM with local binary patterns and histogram-of-gradient features [89]. Intention to interact with a robot was identified using random forest regression on facial expression features [247], or from a combination of the user's line of sight to the robot, shoulder orientation, and speech activity [309]. Last, level of engagement was estimated using gaze analysis techniques to determine if a person was averting their gaze during conversation with a social robot [49].

8.5 Human Motion Tracking

To interact with a human, robots are often required to track the person through multiple frames, which includes detection, tracking their motion, and re-acquiring the intended person if they have been occluded or fall out-of-frame. Human motion tracking for the purpose of robotic vision has been performed using particle filters [10, 25, 125, 219, 309, 312, 319, 348], optical flow [454], or SLAM [16, 41, 448]. In specific examples, Fahn and Lin [125] used a particle filter to track a face from the center of the image using its color properties, whereas Nair et al. [319] tracked multiple people with a particle filter by first segmenting out the static background and then tracking bounding boxes in the foreground. Image keypoints and features, such as SURF features [36], have be used to find correspondences across frames [453], where the track can be initialized by providing a known pattern worn by the target [85] or automatically identifying features on a detected person's clothing [453]. Kernalized correlation filters [182] have been used for efficient tracking [190, 273], and Kalman filters have also been frequently used to reason about and reduce tracking errors from noisy sensors and odometry [139, 163, 263, 273, 312, 406, 430]. For example, Foster et al. [139] propagated a set of pixel hypotheses for segmented skin-colored blobs with a Kalman filter. Last, human motion prediction [79, 163, 188, 232, 312] and robot kinematic models [444, 492] have been used to perform better tracking of the person. To demonstrate, Landi et al. [232] used a neural network to predict hand positions to avoid collisions and inverse kinematics to plan a trajectory that could avoid the human [313, 482]. Although not expressly used in the corpus of HRI/C papers, there is a significant body of work on techniques for detecting and tracking groups of people, which is likely to be used in future work [199, 235, 255, 317, 410, 411, 433, 434]. For example, Lau et al. [235] and Linder and Arras [255] cast group detection and tracking as a multi-hypothesis model selection problem, a probabilistic model that allows for the splitting and merging of clusters. RGB-D sensors are frequently used by robots for multi-person tracking [199, 317, 410]. The latter [410]predicts social groups from egocentric RGBD by reasoning about joint motion and proximity estimates. These approaches have significant potential for use in HRI/C applications, since reasoning about group behavior is likely to be critical for robots in social settings.

Table 2. Datasets Used in Robotic Vision for Human-Robot Interaction/Collaboration

Name of Dataset	Type of Data	Volume of Data	Usage	Used by
NTU RGB-D dataset [383]	RGB-D images and 3D skeletal data	60 action classes, 56,880 video samples	Action recognition	[239, 394]
Manipulation action dataset (MAD) [135]	Video of object/action samples (e.g., cup | drink, pound, shake, move)	625 recordings	Action recognition	[446]
H3.6M dataset [194]	RGB-D images with bounding regions, 3D pose data	3.6 million 3D human poses	Human motion prediction	[160]
Market-1501 dataset [497]	RGB images	32,000 annotated bounding boxes	Person detection	[238]
INRIA dataset [104]	RGB images	1,805 images of humans	Person detection	[14, 316]
HollywoodHeads dataset [441]	RGB video frames	Annotated head regions for 224,740 video frames	Head detection	[448]
The extended Cohn-Kanade (CK+) dataset [210, 269]	RGB images	593 image sequences	Facial expression recognition	[216, 250, 395]
The AffectNet Database [308]	RGB images	1,000,000 facial images	Facial expression recognition	[250]
The WIDER FACE dataset [468]	RGB images	32,203 images, with 393,703 bounding regions	Facial expression recognition	[247]
The Aberdeen Facial Database [2]	RGB images	687 faces	Face detection	[89]
Glasgow Unfamiliar Face Database (GUFD) [3]	RGB images	6,000 images of faces	Face detection	[89]
Utrecht ECVP Facial Database [4]	RGB images	131 images	Face detection	[89]
ChaLearn Looking at People Challenge [124]	RGB-D images	14,000 images of hand gestures	Hand gestures	[60]
Kinect Tracking Precision (KTP) dataset [316]	RGB-D images	8,475 frames, 14,766 instances of people	Person detection	[316]
Annotated hospital dataset [430]	RGB-D images	17,000 annotated images	Mobility detection	[430]
OpenSign [287]	RGB-D images	20,950 images	Hand gestures	[287]
Dataset by Lima et al. [254]	RGB images	160,000 images of open and closed hands	Gesture recognition	[254]

8.6 Multiple Sensors

In the selected papers, robotic vision was often paired with other sensors to enhance the robot's capacity to perceive and respond to the human, see Table 2. Other sensors often involved microphones for speech recognition [62, 118, 145, 286, 442, 475], laser range sensors [14, 188, 219, 271, 339, 450], and ultrasonic sensors [178] to help determine distance, audio sensors for locating the active speaker [70, 322]) and to relocate people who were out-of-view [37, 271, 322], and Leap Motion sensors [209] and inertial measurement units to help track movement and orientation [118, 169, 487]. Force sensors were used with vision sensors during applied tasks such as object handover [38], collaborative manipulation [88, 225], and to determine contact in a safe workspace scenario [51]. Tactile sensors were also used in object handover to assist physical interaction with the environment [192]. Humans who operated the robot (N = 138) were often provided additional information from other sensors or information sources. Operators made informed decisions by information provided from sources such as different LED colors [13, 101, 146, 253, 297], feedback on a display screen [52, 79, 145, 236], video feedback [229, 397], augmented reality [231], haptic feedback [379], spoken response from the robot [145, 146, 287, 490], or robot movement [303, 490] to signal or confirm that the command had been received by the robot. In most examples, the position or configuration of the robot was sufficient (N = 96, e.g., [21, 26]).

9 RQ6. WHAT HAS BEEN THE MAIN PARTICIPANT SAMPLE, AND HOW IS ROBOTIC VISION IN HUMAN-ROBOT COLLABORATION AND INTERACTION EVALUATED?

9.1 Main Participant Sample

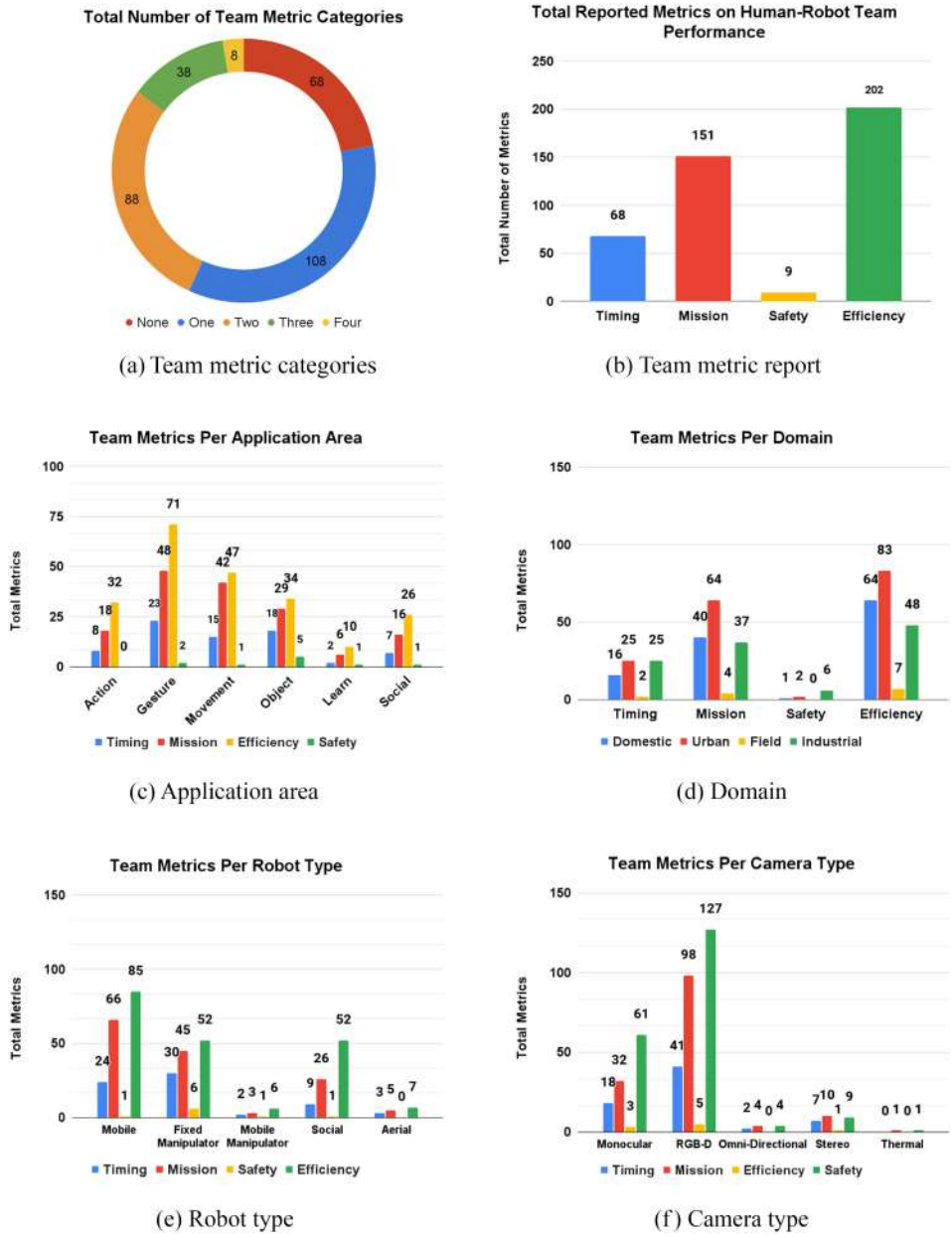
Many published works reported little human-relevant information. In the 310 papers, only 66 (21%) reported details around a human experiment or testing with people. In the studies that did report

participant numbers, there was a calculated total of 1228 participants across all papers ($M = 20$, range = 1–150, $SD = 22$; Christiernin and Augustsson [92] did not report numbers). In studies that reported participant age (22%), participants were on average 32 years old (range = 1–88, $SD = 24$). For papers that reported gender (35%), there was an average percentage split of 70% male and 30% female participants. No studies reported participant country of origin. Gesture recognition had the highest number of participants at 302 (20%, 23 out of 116) papers. Few experiments had direct evaluation for robotic vision performance. Instead, experiments often had a clear focus on robot evaluation as part of its overall intended task or role in the interaction. Evaluation metrics were more likely to involve objective metrics ($N = 384$) compared to subjective metrics ($N = 40$), with the nature of the metrics around robot components such as overall perception, robot task performance, or robot preference when compared to other modalities or system setups. Common quantitative questionnaires included the NASA-TLX (e.g., [202, 221, 425]) and GODSPEED (e.g., [139, 452, 493]). Others included the PARADISE framework [139], Positive or Negative Affect Scale [452], Robot Acceptance Scale [452], or a custom-made scale, such as a 9-point evaluation [348], 7-point comfort rating [379], or 5-point robot performance rating [240]. Figure 14 depicts the number of team metrics and team metric categories, team metrics for each application area, domain, robot type and camera type. A selection of exemplar use cases will be provided in the next section (Section 10) to describe state-of-the-art robotic vision in HRI/C.

9.2 Evaluation Metrics in Each Application Area

9.2.1 Gesture Recognition. Robot evaluation scores were often on preference ratings, how enjoyable people found the interaction, and how favorable people found the robot. For instance, the robot was found to be engaging during a social interaction [102], the interaction was enjoyable in a navigation task [229], and the robot was the preferred choice when compared to other control methods such as joystick or gamepad [472], or when compared to other modalities such as a screen in a public service center [209]. People reported high rating or preference for a robotic vision component involved in the interaction, such as for gestures were natural and easy to use (79%, $n = 24$ [63]), for specific gesture styles (62% preference for elbow to finger, and 38% for eye to finger [6]). This was not systematic, with other modalities not related to robotic vision also rated more favorably, such as physical interaction rated as the least demanding and most accurate method to guide a mobile robot to different waypoints ($n = 24$ [202]), or handheld devices being easier to use than gestures ($n = 23$ [229]). Other robot evaluations included learning speed for use of the robot in a gesture controlled grasping task ($n = 10$ [254]), number of errors in terms of distance from the robot when signaling to pick up items ($n = 16$ [117]), or accuracy as seen during a gesture game with a robot (15% lost due to out of distribution gestures and 10% due to classification error ($n = 30$ [155]).

9.2.2 Action Recognition. Robot evaluations often involved preference scores, willingness to use, and satisfaction levels with the robot. People rated their impression of the robot's behavior compared to several baselines on a 1 to 5 scale ($n = 12$ [240]). Trust was assessed in a home service robot for when it could detect the persons' actions, with 75% of people ($n = 16$) reporting that the feature was important [394], and more than 40 out of 50 people (exact number not reported) reporting a satisfaction level of at least 4 out of 5 [233]. In preference scores, 77% ($n = 30$) preferred a robot for exercise compared to training videos [452], and 92% ($n = 32$) reported willingness to continue the robot program [27]. Other evaluations included 65% (13 out of 20) who were unable to tell if the robot was autonomously operated or tele-operated from its behavior [493]. Last, a robot that used activity recognition for physical activity had people report an increase in exercise success rates with 12 elderly users over a 3-week period [158].



(a) Team metric categories (b) Team metric report

(c) Application area (d) Domain

(e) Robot type (f) Camera type

Fig. 14. Trends in robotic vision evaluation in HRI/C.

9.2.3 Robot Movement in Human Spaces. Evaluations were often in performance, preference, and acceptability. In one example, a mobile robot was successfully controlled in 77.4% of the interactions, with 8.9% of unsuccessful interactions attributed to fast rotational robot movement or large distance between the camera and person [122]. Considering preference, a human following task revealed that people in general found the robot's behavior appropriate, but most reported being uneasy with it ($n = 13$ [348]), and 46% (6 out of 13) were willing to adapt their living environment to accommodate a mobile robot, but 10 did not want to adjust their walking speed [348].

9.2.4 Object Handover and Collaborative Actions. Robot evaluations involved both performance and preference rubrics. For performance, 12 people used an UR10 for a user-controlled pick and place task and after three trials, and those with no experience achieved greater than 75% success with a interaction time of 69 seconds on average [52]. In addition, higher mental load was found for seven novice and experienced operators in a shared work space when the robot operated closer to the person with a higher velocity [407]. For preference, 4 people (100% of the sample) reported that a robot arm did not always match their expectations [92], and collaborative actions from robots in surgical assistance could help them perform their role more efficiently (n = 16 [221]) and reduced workload for a shoe fitting task when the robot was personalized with preferences, including shorter time and fewer commands [425]. Last, some found no differences in time or error rates between humans for passing instruments to a person [221].

9.2.5 Social Communication. There were performance and preference scores for social communication. For performance, 16 people had high detection accuracy for interest to engage (99%) as well as if the person was not interested (92% [373]). A robot served drinks successfully with high accuracy (100% for single-person scenario) with a good response (658 ms) and expected interaction time (49.4 seconds [139]). Gesture recognition and speed in a human-robot gesture game had good recognition accuracy (92%, n = 5 [476]). Other performance results were for a PR2 robot that achieved a 100% success rate to answer user requests (72.7% first attempt, 18.2% second, remainder on the third [309]). Some behaviors were also improved with robotic vision—for example, a humanoid receptionist robot conversed with 26 people to compare engagement-aware behavior, which included small improvements in eye gaze toward the robot (78.8% with and 73.7% without [249]). People perceived the robot as more intelligent and were more satisfied with the interaction, although no effect was noticed on task performance [249]. A robot bartender was rated as likable, intelligent, and safe by 31 people [139]. Others found larger increases with robotic vision—for example, gaze and pause detection to determine when a person had finished speaking resulted in a two times increase in talking time compared with filled pause detection (n = 28 [50]). Some also had reported engagement with the robot but no statistics [405].

10 RQ7. WHAT IS THE STATE OF THE ART IN VISION ALGORITHM PERFORMANCE FOR ROBOTIC VISION IN HUMAN-ROBOT COLLABORATION AND INTERACTION?

Considering robot evaluation and vision algorithm performance, state-of-the-art performance is paramount to the functional benefit of the robot, including what tasks or services the robot could provide to the human. However, Section 9 demonstrated that few papers reported standardized metrics that directly evaluated robotic vision performance. This makes it challenging to fairly compare the performance of the vision algorithms used in these works. It is nonetheless important to highlight works that make superior use of robotic vision in HRI/C systems. Therefore, we present selected works that well represent the use of robotic vision in human-robot collaboration and interaction with respect to the criteria of novelty, impact, and/or robustness. Exemplar studies are identified that showcased creative and/or robust use cases of robotic vision, given that systematic differences could not be calculated across studies from metric and result reporting. These examples help to direct to future pathways in the field to increase experimental rigor with more experimentation and more systematically evaluate the feasibility for the capacity, speed, and accuracy for robotic vision to be used with people.

10.1 Gesture Recognition

Mazhar et al. [287] demonstrated control of a KUKA arm via hand gestures by fine-tuning an Inception V3 convolutional neural network on a custom dataset (OpenSign). This resulted in a

system that was able to detect 5 gestures in a row at 40 Hz (250 ms per detection). OpenPose was used to localize hands in the dataset images, and the Kinect V2 depth map was used to segment the hands from the background, allowing background substitution for data augmentation. Inception V3 fine-tuning resulted in a validation accuracy of 99.1% and a test accuracy of 98.9%. The dataset had RGB-D images with 10 gestures performed by 10 people, including 8,646 original images and 12,304 synthetic images from background substitution. Waskito et al. [447] tested the robustness of their hand gesture classifier as the hand was rotated or when lighting conditions were varied to find an average total accuracy of 96.7% with each gesture having a 0.141-second average response time. Last, Pentiuc and Vultur [338] used skeleton data from the Kinect and the dynamic time warping algorithm to detect 5 gestures with an accuracy of more than 86%. Although these methods cannot be compared directly, since different gestures and settings were being evaluated, the overall trend was to use higher-capacity models trained with more data to increase the accuracy and robustness of gesture recognition.

10.2 Human Detection and Tracking

To detect and track a specific person, Hwang et al. [190] integrated a Single-Shot Detector, FaceNet, and a Kernelized Correlation Filter. With this system, Hwang et al. [190] were able to detect humans up to 8 m away and recognize specific faces, achieving a maximum position error of 4 cm and 5-degree orientation. For tracking a person from a mobile robot, Weber et al. [448] achieved a 59.7% mean average precision with a Single-Shot Detector and tracking-as-repeated-detection strategy. Zhang et al. [492] compared their detection and tracking method using target contour bands to several others that used videos from the object tracking benchmark dataset. This work showed the presented method was more accurate (94%) with the fastest processing time (34 fps). Once deployed on a mobile robot, the robot could follow an identified target for around 648 m. Fang et al. [127] found that dynamic body poses could be recognized with high accuracy (>96% classification accuracy on 300 tests), but limited details were provided on the method [127]. To detect and distinguish people based on walking aids, Vasquez et al. [430] found that combining a Kalman filter, a hidden Markov model, and a Fast R-CNN region extractor improved system performance by a factor of 7 compared to a dense sliding window method.

10.3 Non-Gestural Action Recognition

In an action recognition task, Lee and Ahn [239] achieved an accuracy of 71% from an RGB camera on the NTU RGB-D dataset (75% from Kinect) at 15 fps. For recognizing actions from a child, Efthymiou et al. [121] used dense trajectories from multi-view fusion as the input to their action recognition system and evaluated on a test set of 25 children performing 12 actions with comparisons to a test set of 14 adults [121]. Finally, to adapt the motion of a robotic assistant rollator to the patients, Chalvatzaki et al. [72] found that their model-based reinforcement learning method that uses predicted human actions obtained a smaller tracking error than several other control methods [72].

11 RQ8. WHAT ARE THE UPCOMING CHALLENGES FOR ROBOTIC VISION IN HUMAN-ROBOT COLLABORATION AND INTERACTION?

This section will discuss potential and known challenges for robotic vision in human-robot collaboration and interaction, as well as a brief discussion of general robotic vision challenges. Overall summaries on future human-related challenges and general robotic vision challenges demonstrate target areas for consideration in the design, deployment, and future use of robotic vision in human spaces. Challenges specific to vision during human-robot collaboration include the ethical use of

human data, human model selection and optimization, experimental design and validation, and appropriate trust.

11.1 General Challenges of Robotic Vision

In addition to challenges specific to human-robot interaction, there are also more general challenges related to robotic and computer vision. Although the performance of robotic vision systems are often bounded by what can be achieved by state-of-the-art computer vision algorithms, there are many reasons why state-of-the-art computer vision techniques have not been transferred to robotic platforms.

An algorithm that uses visual data may not be sufficiently robust to perform in real-world conditions and edge cases relevant for HRI/C [98, 402]. Although visual data may contribute to better multi-modal understandings of human states, actions, and involvement with the robot [39], vision still presents with challenges. For the robot to understand its given task or action, robots may require large training data for new tasks and/or need access to processing capabilities that are unavailable on the robotic platform (e.g., multiple GPUs). Hardware performance can be limited by robot on-board processing compared to cloud processing and the availability, performance, and cost of hardware solutions, such as the RGB-D camera as an inexpensive source of depth information compared with expensive alternatives like laser range sensors [403]. Computationally intensive algorithms can lead to the requirement to have a GPU in the robot, which can be heavy and noisy, and require a lot of power [336].

Sim-to-real transfer can be a particular challenge for many learning-based robotics applications, and progress is often relinquished to large companies that can implement large-scale data collection and testing [244]. In addition, deep neural networks often fail to generalize, with a reduction in accuracy when tested outside of benchmark datasets [358]; a method that achieves state-of-the-art performance on a benchmark dataset may not generalize immediately to a real-world setting on a robotic platform. Transferring cutting-edge computer vision techniques may also be too recent to have been adopted in physically embodied robots, let alone applied to scenarios related to HRI/C. In the scenarios when techniques have been transferred into HRI/C applications, robots can encounter failures due to computational delays or challenges around the complexity of the vision-based activities related to the task, such as to perceive and understand the diversity of hand-based gestures from multiple people across different countries [399]. Real-world vision-based challenges can also occur with robots operating around people, such as important information being occluded when perceiving the person, and critical information not being identified or perceived during the interaction. However, robots can better overcome some of these challenges, such as by controlling camera positioning, adjusting to capture missing information, and orienting visual capture to help fill in the missing gaps [402]. Additional challenges include software challenges such as the availability of open source libraries, software development kits, and the lack of training data for specific use cases.

Robots must also be able to act upon the visual information in a relevant and suitable way. For instance, there continue to be challenges around translating signals from computer vision into actionable and useful robot functions for robots, such as movement and manipulation actions that can improve the robot's utility for a given task [275]. This could be impacted by the limited use of participatory design to select suitable applications for robots to assist people [315], or the inability of current vision systems to address tasks and actions that people want robots to assist them on [462]. Either of these could have contributed to slowing down the deployment of robotic vision use cases in human domains. However, human detection and tracking is clearly a key capability for many HRI/C systems, and therefore it is likely that the current state of the art in computer vision will be rapidly transferred to these systems in the near future.

11.2 Fair and Ethical Use of Human Data

First, there are important challenges around the fair and ethical use of human data for the purpose of HRI/C when interacting and collaborating with robots [96, 293]. For example, the General Data Protection Regulation (GDPR) describes regulations on the processing of personal data in Europe, including the consent of individuals for use of their personal data [1]. Fair and ethical use will therefore need to use data processing and management methods that comply with national and/or international data protection regulation and laws. For the next decade, robots are likely to continue to enter human spaces, and there may be limited public knowledge and awareness on how interacting with a robot may use their personal information to help facilitate the interaction. Examples include the robot using its camera system to perceive and classify the person's facial features, body pose, and actions, as well as using visual information to make inferences on ways to engage the person, such as by classifying the person's age range, intent to interact with the robot, and future actions. Similar to other data-intensive fields, there are important implications on the use of human data in HRI/C, such as the capacity for people to give informed consent and for the appropriate collection, management, and storage of data in the context of human-robot interaction. Consent to use data should be obtained with clear explanation or capacity to access detailed information on what data is being collected, how it will be used, and how long it will be stored if data will be used for personal, private, or third-party use. For instance, previously recorded data (images and videos) could be captured and approved to be used to improve future robotic interactions as commonly used in computer vision by fine-tuning pre-trained neural networks (e.g., [50]). Other future challenges around fair and ethical use will include data storage and/or ownership of any images and video collected from robots in human spaces, including the right for people to access, edit, or request deletion of any or all images or video streams collected from them. Robot interactions with a physical hardware system do not always include screens or terminals, and these interactions can be located in high-volume areas with frequent turnover of people, such as in a public space. These robot interactions also often do not facilitate similar user agreements or consent notices as other digital methods such as website or smartphone application use [274]. Future challenges therefore should involve methods to address clear and transparent notices of intent to use human data when images or video captured for the purpose of core vision-based features, such as to follow the correct person or identification of a specific customer to complete an order transaction. This could include consent as provided by active or passive consent, and/or accessible information about the robot deployed in the public space with the potential for people to avoid or remove themselves from the robots field of view. This could also include detailed consideration for processes to obtain appropriate informed consent for certain groups that can have a second person involved in the consent process, such as guardians of young children or those who are unable to consent for themselves.

Other challenges also involve the concept of privacy and helping to mediate negative effects around invasion of privacy from robot use [274]. For instance, weighing up potential benefits with risks for each deployment ensures that visual information is collected only when required for functionality, and if so, it is handled and stored with proper care. Furthermore, there have been advances made in the domain of privacy-preserving computer vision, such as to anonymize faces during action detection [362], as well as privacy-preserving visual SLAM [386], given that point clouds can retain a sufficient level of information to re-create the surrounding environment, potentially compromising privacy if people were intentionally or unintentionally involved in the scene [345]. Continued deployment of robots in public spaces or in private or sensitive contexts, such as at home, may require the consistent use of privacy-preserving vision techniques to ensure that human data is handled appropriately, and that humans do not have unresolved concerns about how robots with robotic vision capacity will operate safely in their own space.

11.3 Human Models

Second, there are important challenges on model selection and optimization for human behaviors, including the reliability and validity of behavioral phenomena to be captured and responded to through robotic vision. This systematic review presents several important use cases, including gesture and action recognition, human walking trajectories, object handover, social communication, and learning from human demonstration. These range from both simple and complex behaviors that require the robot to understand the person. However, some of these behaviors are not as simple to interpret through the use of model selection and optimization. One key example of this is the use of emotion recognition. Robotic vision that is dependent on state classification of human intent or emotion into static categories (i.e., happy or sad) can result in inaccurate identification and/or irrelevant responses provided from the robot without consideration of a more complex human emotional spectrum [31]. For instance, behavioral science research has drawn attention to the unsuitable use of current computer vision methods to detect a person's emotional state from their facial movements, instead calling for research that explores how people actually move their face to express emotion or other signals in different contexts [31]. This research demonstrates that care should be taken around selection for what robotic vision should and should not be used during human interaction with robots—for example, inferring other human characteristics from visual data that could cause more harm than benefit to the interaction, such as classifying sexuality, race, or serious underlying medical conditions not known to the person. Incorrect model selection and optimization could also cause notable long-term problems in future robot deployments, if development and testing continues to optimize for behavior that is not accurate or representative of the person, further contributing to bias [293]. This also raises the question of whether simulated humans need to be involved in the simulation process, and the level of realism needed for this to be meaningful to the learning process. In addition, current methods that require large datasets may use datasets originally collected for other purposes, and therefore may not easily translate to a new context, such as a dataset of a busy crowd repurposed to help robots learn about social norms in small groups. Last, people may eventually develop long-term hesitancy and rejection to use robotic systems due to perceptions that their capacities are non-functional after repeated errors, given the importance of robot performance on trust in the system [174].

11.4 Experimental Design and Evaluation

Third, there has been limited human experimentation and evaluation of robotic vision with humans. As reported in the preceding sections, few studies report direct testing with humans, and for those that did report a form of testing with people, there were limited participant numbers across all studies. This is a challenge for future deployments, because exploration of participant characteristics found that many did not involve a large range of people who were representative of the general population, and instead involved a narrow sample with limited diversity [293]. Therefore, robotic vision for HRI/C could lead toward design and optimization for a very restricted sample, further contributing to bias [293]. For instance, people who provide feedback in experimental testing become the leading designers in future iterations of robot behavior and function. This can create barriers for wider-scale adoption when robots are deployed in the general community and inevitably encounter different kinds of people who have not been taken into consideration during design and refinement stages. Greater inclusion of different people has been the recent focus of co-design methodology for human-robot interaction and collaboration to ensure that robots demonstrate a more inclusive behavior for a wider range of people who are likely to use them (e.g., [366]). In addition, reported experiments often used single or simple evaluation metrics to measure robot perceptions and human-robot team performance, which may skew robotic vision

evaluation in collaborative scenarios, without taking into consideration the human, the robot, and the team dynamic [105, 184]. Such a simplified approach to testing and evaluation could further contribute to skewed development around how robotic vision should work to help people, if testing on human participants continues to remain low and only on restricted samples.

11.5 Appropriate Trust

Last, there is the high potential that humans perceive robots that can interpret visual information to have a high sense of intelligence and general capacity to function with the person and within the environment [228, 414]. For instance, it may not be clear to the person to what extent the robot can identify only a limited visual field or target areas of interest, instead assuming the robot can view all of its surroundings and the activities that occur within it. The use of visual information to interact with the person may also contribute to an increased sense of anthropomorphic interpretation of the robot, leading people to perceive the robot as having more emotional expression or intelligence [120]. This can result in people who inflate their confidence, trust, or perception of the system, meaning that people will relinquish greater autonomy or responsibility to the robot beyond what the robot is capable to perform on its own [364]. For instance, people may falsely assume that the robot has human-like perception and cognitive abilities [120, 228, 414], which can lead people to assume that the robot can perform better than it can. Misunderstanding around the capability of the robot could have notable consequences for safe and effective HRI/C—for instance, the human assuming that the robot will detect visual hazards for the person, or recognize its own errors or mistakes in collaborative work. Therefore, visual information in human-robot collaboration and interaction should be used for the functional purpose of the robot, as well as be explained to the people who use the robot, which includes both its potential strengths and limitations within the intended context to help regulate expectations and define the intended role of the robot [228, 414].

12 PROMISING AREAS OF FUTURE RESEARCH

There were several relevant computer vision methods that were not represented in the corpus of selected HRI/C papers. Four prominent examples are video convolutional networks, 3D human pose estimators, human–object interaction classifiers, and sign language recognition. Each of these could potentially have a significant impact on HRI/C research in the future.

To begin, recent state-of-the-art methods for action classification process video data using 3D convolutional networks [66, 132], unlike the predominantly frame-based classification approaches used in the HRI/C literature (see Section 8). New techniques include the inflated 3D convolutional networks [66] and the two-stream SlowFast network [132]. Action classification from video is critical to many HRI/C systems, but 3D convolutional networks tend to require significant training data for fine-tuning and significant GPU resources for inference, making it challenging to transfer to many HRI/C systems. However, the expansion of models pre-trained on diverse datasets, coupled with developments in transfer learning, help mitigate this difficulty. Therefore, it is expected that pre-trained video action recognition networks could become a commonly used tool in HRI/C research. Although 2D human pose estimation was well represented, 3D human pose estimation was mostly absent, despite the fact that providing spatial and shape information could be very useful for HRI/C to enable the robot to perform more accurate and functional actions with the person. Monocular 3D human pose estimation from a single image or video is a very popular topic in computer vision. Model-free approaches [334, 335] include VideoPose3D [335],which estimates 3D joint locations using temporal convolutions over a sequence of 2D joint detections. Model-based approaches [211, 212, 220, 222] predict the parameters (e.g., joint angles, shape, and transformation) of a body model, such as the SMPL mesh model [264]. Adversarial learning can

be used [211, 212, 220] to generate realistic body poses and motions, which tends to generalize better to unseen datasets, and therefore may be more appropriate for HRI/C tasks. Modeling humans in 3D allows physics to be taken into account, allowing the robot to plan and respond more appropriately and preventing it from making non-physical predictions.

Another set of techniques of relevance to HRI/C are those developed for human-object interaction classification [75, 164, 350]. This task aims to extend action recognition to interaction recognition: localizing and describing pairs of interacting humans and objects in the scene. A robot that collaborates with people to perform a task would strongly benefit from knowing which object the person is interacting with at that point in time, and what type of interaction is taking place. For example, an airport assistance robot may need to detect instances of "person carrying suitcase" to determine where best to provide support to the person. Methods for this task almost always detect human and object bounding boxes first, before combining information from different modalities (appearance, relative geometry, human pose) using multi-stream networks [75, 164] or graph neural networks [142, 350]. There is a clear case for widespread use of these techniques in HRI/C, to facilitate higher-order reasoning about what the people proximal to the robot are doing, and with what objects.

In general, substantial progress has been made in computer vision and machine learning since 2020. Although beyond the scope of this work, there are significant opportunities for HRI/C arising from these developments. In particular, the Transformer architecture [431], originally proposed for natural language processing, has begun to supplant or supplement convolutional neural networks for vision tasks, with large performance increases across many tasks. These include image recognition [115], object detection [65], video understanding [23, 46, 332], and human-object interaction [141]. This represents a significant opportunity for the HRI/C community, because it is a general-purpose architecture that facilitates multi-modal sensor processing [198], allowing a robot to reason about its video, audio, and other inputs jointly. This is likely to expand and robustify robot capabilities while interacting or collaborating with people.

There are also additional areas in which computer vision research has potential impact to adapt or improve HRI/C across different settings. One notable area is sign language recognition [9, 245, 342],which was not present in the corpus of papers. This represents an opportunity for further developing methods for non-verbal communication in HRI/C. The techniques developed, involving fine-grained gesture recognition and multi-modal learning, are relevant for HRI/C, since these techniques can provide benefit to human-robot communication, as well as general situational awareness. Other area is autonomous and assisted driving, in which robotic vision for HRI/C could have a notable impact to increase the uptake, efficiency, and safety of autonomous vehicles [116, 170, 369]. For example, autonomous driving requires that the car can detect and predict the trajectories of others around and on the road, such as drivers, cyclists, and pedestrians. This process can involve close monitoring and coordination to ensure that humans can safely move around autonomous vehicles while vehicles can also get to their intended destination. There is currently a growing body of work around the human component in autonomous driving, but most of these works have so far been tested in simulated environments and without vision, creating notable opportunities to explore new areas of robotic vision for HRI/C-style tasks in the near future [116, 128, 170, 369, 466]. Other areas also include to further explore the capacity to anticipate human actions in advance, resulting in robot behavior that can be more reactive than passive to respond to dynamic interaction patterns over time [161, 189, 325].

13 CONCLUSION

This survey and systematic review provided a comprehensive overview on robot vision for HRI/C with a detailed review of papers published in the past 10 years for robots that can perceive and take

action to facilitate a high-level task. Robotic vision had the capacity to improve HRI/C, including to create new ways for humans and robots to work together. This survey and systematic review provided an extensive analysis on the use of robotic vision in human-robot collaboration and interaction into common domains, areas, and performance metrics for robotic vision. This includes exploring how computer vision has been adapted and translated through robotic vision to improve aspects of HRI/C. This survey and systematic review also contributed to identifying application areas that had not yet been attempted, and how techniques from the computer vision research could help to inform human-focused vision research in robotics. It was found that robotic vision for improving the capacity of robots to collaborate with people is still an emerging domain. Most works involved a one-on-one interaction, and focused on using robotic vision to enhance a specific feature or function related to the interaction. It was also found that only some high-impact and novel techniques from the computer vision field had been translated for HRI/C, highlighting an important opportunity to improve the capacity of robots to engage and assist people. More novel and emerging areas in the HRI/C field such as multi-human, multi-robot teams were less represented in the corpus of papers [67, 195, 260]. Furthermore, robotic vision was often tested in a simple or single application field for each specific use case, showing limited depth in its current form.

Future pathways for HRI/C involve the creation and development of robotic platforms using vision-related information to create more competent robots that can operate in dynamic environments with people—for instance, improving robots to better handle multiple visual inputs at once to open up new domains or collaborative tasks, such as multi-human multi-robot teams. Robotic vision could therefore help to break down some barriers present in long-term human-robot teamwork, such as better adaptation to dynamic environments and different kinds of people over a long period of time.

REFERENCES

[1] European Commission. n.d. 2018 Reform of EU Data Protection Rules. Retrieved December 11, 2022 from https://ec.europa.eu/commission/sites/beta-political/files/data-protection-factsheet-changes_en.pdf.

[2] University of Stirling. n.d. Aberdeen Facial Database. Retrieved December 11, 2022 from http://pics.psych.stir.ac.uk/zips/Aberdeen.zip.

[3] University of York. n.d. Glasgow Unfamiliar Face Database (GUFD). Retrieved December 11, 2022 from http://www.facevar.com/glasgow-unfamiliar-face-database.

[4] University of Stirling. n.d. Utrecht ECVP Facial Database. Retrieved December 11, 2022 from http://pics.psych.stir.ac.uk/zips/utrecht.zip.

[5] W. Z. B. W. Z. Abiddin, R. Jailani, and A. R. Omar. 2018. Development of robot-human imitation program for telerehabilitation system. In *Proceedings of the 2018 11th International Conference on Developments in eSystems Engineering (DeSE'18)*. 198–201. https://doi.org/10.1109/DeSE.2018.00045

[6] S. Abidi, M. Williams, and B. Johnston. 2013. Human pointing as a robot directive. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 67–68. https://doi.org/10.1109/HRI.2013.6483504

[7] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. 2009. Building Rome in a day. In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV'09)*. IEEE, Los Alamitos, CA, 72–79.

[8] R. Agrigoroaie, F. Ferland, and A. Tapus. 2016. The ENRICHME project: Lessons learnt from a first interaction with the elderly. In *Social Robotics*. Lecture Notes in Computer Science, Vol. 9979. Springer, 735–745. https://doi.org/10.1007/978-3-319-47437-3_72

[9] Samuel Albanie, Gül Varol, Liliane Momeni, Triantafyllos Afouras, Joon Son Chung, Neil Fox, and Andrew Zisserman. 2020. BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues. In *Proceedings of the European Conference on Computer Vision*. 35–53.

[10] B. Ali, Y. Ayaz, M. Jamil, S. O. Gilani, and N. Muhammad. 2015. Improved method for stereo vision-based human detection for a mobile robot following a target person. *South African Journal of Industrial Engineering* 26, 1 (2015), 102–119. https://doi.org/10.7166/26-1-891

[11] S. Ali, A. Lam, H. Fukuda, Y. Kobayashi, and Y. Kuno. 2019. Smart wheelchair maneuvering among people. In *Intelligent Computing Methodologies*. Lecture Notes in Computer Science, Vol. 11645. Springer, 32–42. https://doi.org/10.1007/978-3-030-26766-7_4

[12] D. Almonfrey, A. P. do Carmo, F. M. de Queiroz, R. Picoreti, R. F. Vassallo, and E. O. T. Salles. 2018. A flexible human detection service suitable for Intelligent Spaces based on a multi-camera network. *International Journal of Distributed Sensor Networks* 14, 3 (2018), 1–22. https://doi.org/10.1177/1550147718763550

[13] J. Alonso-Mora, R. Siegwart, and P. Beardsley. 2014. Human - Robot swarm interaction for entertainment: From animation display to gesture based control. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 98. https://doi.org/10.1145/2559636.2559645

[14] V. Alvarez-Santos, X. M. Pardo, R. Iglesias, A. Canedo-Rodriguez, and C. V. Regueiro. 2012. Feature analysis for human recognition and discrimination: Application to a person-following behaviour in a mobile robot. *Robotics and Autonomous Systems* 60, 8 (2012), 1021–1036. https://doi.org/10.1016/j.robot.2012.05.014

[15] A. Angani, J.-W. Lee, T. Talluri, J.-Y. Lee, and K. J. Shin. 2020. Human and robotic fish interaction controlled using hand gesture image processing. *Sensors and Materials* 32, 10 (2020), 3479–3490. https://doi.org/10.18494/SAM.2020.2925

[16] A. T. Angonese and P. F. Ferreira Rosa. 2017. Multiple people detection and identification system integrated with a dynamic simultaneous localization and mapping system for an autonomous mobile robotic platform. In *Proceedings of the 6th International Conference on Military Technologies (ICMT'17)*. 779–786. https://doi.org/10.1109/MILTECHS.2017.7988861

[17] Alberto Torres Angonese and Paulo Fernando Ferreira Rosa. 2016. Integration of people detection and simultaneous localization and mapping systems for an autonomous robotic platform. In *Proceedings of the 2016 XIII Latin American Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR'16)*. 251–256. https://doi.org/10.1109/LARS-SBR.2016.49

[18] U. A. D. N. Anuradha, K. W. S. N. Kumari, and K. W. S. Chathuranga. 2020. Human detection and following robot. *International Journal of Scientific and Technology Research* 9, 3 (2020), 6359–6363. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85082858300&partnerID=40&md5=76978175cda39acb033058f4d669c257.

[19] S. M. Anzalone, S. Ivaldi, O. Sigaud, and M. Chetouani. 2013. Multimodal people engagement with iCub. *Advances in Intelligent Systems and Computing* 196 AISC (2013), 59–64. https://doi.org/10.1007/978-3-642-34274-5_16

[20] D. Araiza-Lllan and A. De San Bernabe Clemente. 2018. Dynamic regions to enhance safety in human-robot interactions. In *Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA'18)*. 693–698. https://doi.org/10.1109/ETFA.2018.8502453

[21] J. O. P. Arenas, R. D. H. Beleno, and R. J. Moreno. 2017. Deep convolutional neural network for hand gesture recognition used for human-robot interaction. *Journal of Engineering and Applied Sciences* 12, 11 (2017), 9278–9285. https://doi.org/10.3923/jeasci.2017.9278.9285

[22] Brenna D. Argall and Aude G. Billard. 2010. A survey of tactile human–robot interactions. *Robotics and Autonomous Systems* 58, 10 (2010), 1159–1176.

[23] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. ViVit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6836–6846.

[24] Kai O. Arras, Oscar Martinez Mozos, and Wolfram Burgard. 2007. Using boosted features for the detection of people in 2D range data. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA, 3402–3407.

[25] A. K. Arumbakkam, T. Yoshikawa, B. Dariush, and K. Fujimura. 2010. A multi-modal architecture for human robot communication. In *Proceedings of the 2010 10th IEEE-RAS International Conference on Humanoid Robots (Humanoids'10)*. 639–646. https://doi.org/10.1109/ICHR.2010.5686337

[26] A. Augello, A. Ciulla, A. Cuzzocrea, S. Gaglio, G. Pilato, and F. Vella. 2020. Towards an intelligent system for supporting gesture acquisition and reproduction in humanoid robots. In *Proceedings of the 26th International DMS Conference on Visualization and Visual Languages (DMSVIVA'20)*. 82–86. https://doi.org/10.18293/DMSVIVA2020-017

[27] O. Avioz-Sarig, S. Olatunji, V. Sarne-Fleischmann, and Y. Edan. 2020. Robotic system for physical training of older adults. *International Journal of Social Robotics* 13 (2020), 1109–1124. https://doi.org/10.1007/s12369-020-00697-y

[28] Bita Azari, Angelica Lim, and Richard Vaughan. 2019. Commodifying pointing in HRI: Simple and fast pointing gesture detection from RGB-D images. In *Proceedings of the 2019 16th Conference on Computer and Robot Vision (CRV'19)*. 174–180. https://doi.org/10.1109/CRV.2019.00031

[29] X. Bai, C. Li, K. Chen, Y. Feng, Z. Yu, and M. Xu. 2018. Kinect-based hand tracking for first-person-perspective robotic arm teleoperation. In *Proceedings of the 2018 IEEE International Conference on Information and Automation (ICIA'18)*. 684–691. https://doi.org/10.1109/ICInfA.2018.8812561

[30] G. Baron, P. Czekalski, D. Malicki, and K. Tokarz. 2013. Remote control of the artificial arm model using 3D hand tracking. In *Proceedings of the 2013 International Symposium on Electrodynamic and Mechatronic Systems (SELM'13)*. 9–10. https://doi.org/10.1109/SELM.2013.6562954

[31] Lisa Feldman Barrett, Ralph Adolphs, Stacy Marsella, Aleix M. Martinez, and Seth D. Pollak. 2019. Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest* 20, 1 (2019), 1–68. https://doi.org/10.1177/1529100619832930

[32] Pablo Barros, German I. Parisi, Doreen Jirak, and Stefan Wermter. 2014. Real-time gesture recognition using a humanoid robot with a deep neural architecture. In *Proceedings of the 2014 IEEE-RAS International Conference on Humanoid Robots*. 646–651. https://doi.org/10.1109/HUMANOIDS.2014.7041431

[33] Michael Barz, Peter Poller, and Daniel Sonntag. 2017. Evaluating remote and head-worn eye trackers in multi-modal speech-based HRI. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. ACM, New York, NY, 79–80. https://doi.org/10.1145/3029798.3038367

[34] N. C. Batista and G. A. S. Pereira. 2015. A probabilistic approach for fusing people detectors. *Journal of Control, Automation and Electrical Systems* 26, 6 (2015), 616–629. https://doi.org/10.1007/s40313-015-0202-6

[35] Andrea Bauer, Dirk Wollherr, and Martin Buss. 2008. Human–robot collaboration: A survey. *International Journal of Humanoid Robotics* 5, 01 (2008), 47–66.

[36] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up robust features (SURF). *Computer Vision and Image Understanding* 110, 3 (June 2008), 346–359.

[37] B. Bayram and G. Ince. 2016. Audio-visual multi-person tracking for active robot perception. In *Proceedings of the 2015 IEEE/SICE International Symposium on System Integration (SII'15)*. 575–580. https://doi.org/10.1109/SII.2015.7405043

[38] M. Bdiwi, J. Suchý, and A. Winkler. 2013. Handing-over model-free objects to human hand with the help of vision/force robot control. In *Proceedings of the 10th International MultiConference on Systems, Signals, and Devices (SSD'13)*. 1–6. https://doi.org/10.1109/SSD.2013.6564138

[39] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. 2020. Vision-based human activity recognition: A survey. *Multimedia Tools and Applications* 79, 41-42 (Nov. 2020), 30509–30555. https://doi.org/10.1007/s11042-020-09004-3

[40] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction* 3, 2 (2014), 74.

[41] A. Bellarbi, S. Kahlouche, N. Achour, and N. Ouadah. 2017. A social planning and navigation for tour-guide robot in human environment. In *Proceedings of the 2016 8th International Conference on Modelling, Identification, and Control (ICMIC'16)*. 622–627. https://doi.org/10.1109/ICMIC.2016.7804186

[42] José Pedro Ribeiro Belo, Felipe Padula Sanches, and Roseli Aparecida Francelin Romero. 2019. Facial recognition experiments on a robotic system using one-shot learning. In *Proceedings of the 2019 Latin American Robotics Symposium (LARS'19), the 2019 Brazilian Symposium on Robotics (SBR'19), and the 2019 Workshop on Robotics in Education (WRE'19)*. 67–73. https://doi.org/10.1109/LARS-SBR-WRE48964.2019.00020

[43] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. 2018. Social robots for education: A review. *Science Robotics* 3, 21 (2018), eaat5954.

[44] Ismail Ben Abdallah, Yassine Bouteraa, and Chokri Rekik. 2016. Kinect-based sliding mode control for Lynxmotion robotic arm. *Advances in Human-Computer Interaction* 2016, 7 (2016), 1–10.

[45] Ismail Benabdallah, Yassine Bouteraa, Rahma Boucetta, and Chokri Rekik. 2015. Kinect-based computed Torque control for Lynxmotion robotic arm. In *Proceedings of the 2015 7th International Conference on Modelling, Identification, and Control (ICMIC'15)*. 1–6. https://doi.org/10.1109/ICMIC.2015.7409416

[46] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*. 813–824.

[47] Cindy L. Bethel, Kristen Salomon, Robin R. Murphy, and Jennifer L. Burke. 2007. Survey of psychophysiology measurements applied to human-robot interaction. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'07)*. 732–737. https://doi.org/10.1109/ROMAN.2007.4415182

[48] Lucas Beyer, Alexander Hermans, Timm Linder, Kai O. Arras, and Bastian Leibe. 2018. Deep person detection in two-dimensional range data. *IEEE Robotics and Automation Letters* 3, 3 (2018), 2726–2733.

[49] M. Bilac, M. Chamoux, and A. Lim. 2017. Gaze and filled pause detection for smooth human-robot conversations. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. 297–304. https://doi.org/10.1109/HUMANOIDS.2017.8246889

[50] Hakan Bilen, Basura Fernando, Efstratios Gavves, Andrea Vedaldi, and Stephen Gould. 2016. Dynamic image networks for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3034–3042.

[51] M. C. Bingol and O. Aydogmus. 2020. Practical application of a safe human-robot interaction software. *Industrial Robot* 47, 3 (2020), 359–368. https://doi.org/10.1108/IR-09-2019-0180

[52] G. Bolano, A. Tanev, L. Steffen, A. Roennau, and R. Dillmann. 2018. Towards a vision-based concept for gesture control of a robot providing visual feedback. In *Proceedings of the 2018 IEEE International Conference on Robotics and Biometrics (ROBIO'18)*. 386–392. https://doi.org/10.1109/ROBIO.2018.8665314

[53] Francisco Bonin-Font, Alberto Ortiz, and Gabriel Oliver. 2008. Visual navigation for mobile robots: A survey. *Journal of Intelligent and Robotic Systems* 53, 3 (2008), 263–296.

[54] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*. 144–152.
[55] K. Bothe, A. Winkler, and L. Goldhahn. 2018. Effective use of lightweight robots in human-robot workstations with monitoring via RGBD-camera. In *Proceedings of the 2018 23rd International Conference on Methods and Models in Automation and Robotics (MMAR'18)*. 698–702. https://doi.org/10.1109/MMAR.2018.8486036
[56] Y. Bouteraa and I. B. Abdallah. 2017. A gesture-based telemanipulation control for a robotic arm with biofeedback-based grasp. *Industrial Robot* 44, 5 (2017), 575–587. https://doi.org/10.1108/IR-12-2016-0356
[57] Yuri Boykov, Olga Veksler, and Ramin Zabih. 2001. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 11 (2001), 1222–1239.
[58] G. Broccia, M. Livesu, and R. Scateni. 2011. Gestural interaction for robot motion control. In *Proceedings of the 2011 Eurographics Italian Chapter Conference*. 61–66. https://doi.org/10.2312/LocalChapterEvents/ItalChap/ItalianChapConf2011/061-066
[59] J. Brookshire. 2010. Person following using histograms of oriented gradients. *International Journal of Social Robotics* 2, 2 (2010), 137–146. https://doi.org/10.1007/s12369-010-0046-y
[60] A. Brás, M. Simão, and P. Neto. 2018. Gesture recognition from skeleton data for intuitive human-machine interaction. In *Transdisciplinary Engineering Methods for Social Innovation of Industry 4.0*, Vol. 7. Advances in Transdisciplinary Engineering. IOS Press, 271–280. https://doi.org/10.3233/978-1-61499-898-3-271
[61] W. Budiharto, A. Jazidie, and D. Purwanto. 2010. Indoor navigation using adaptive neuro fuzzy controller for servant robot. In *Proceedings of the 2010 2nd International Conference on Computer Engineering and Applications (ICCEA'10)*, Vol. 1. 582–586. https://doi.org/10.1109/ICCEA.2010.119
[62] B. Burger, I. Ferrané, F. Lerasle, and G. Infantes. 2012. Two-handed gesture recognition and fusion with speech to command a robot. *Autonomous Robots* 32, 2 (2012), 129–147. https://doi.org/10.1007/s10514-011-9263-y
[63] G. Canal, C. Angulo, and S. Escalera. 2015. Gesture based human multi-robot interaction. In *Proceedings of the International Joint Conference on Neural Networks*. https://doi.org/10.1109/IJCNN.2015.7280540
[64] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*.
[65] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision*. 213–229.
[66] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'17)*. 6299–6308.
[67] J. Casper and R. R. Murphy. 2003. Human-robot interactions during the robot-assisted urban search and rescue response at the World Trade Center. *IEEE Transactions on Systems, Man, and Cybernetics: Part B (Cybernetics)* 33, 3 (2003), 367–385. https://doi.org/10.1109/TSMCB.2003.811794
[68] G. Castellano, I. Leite, A. Pereira, C. Martinho, A. Paiva, and P. W. McOwan. 2014. Context-sensitive affect recognition for a robotic game companion. *ACM Transactions on Interactive Intelligent Systems* 4, 2 (2014), Article 10, 25 pages. https://doi.org/10.1145/2622615
[69] D. Cazzato, C. Cimarelli, J. L. Sanchez-Lopez, M. A. Olivares-Mendez, and H. Voos. 2019. Real-time human head imitation for humanoid robots. In *Proceedings of the 2019 3rd International Conference on Artificial Intelligence and Virtual Reality (AIVR'19)*. 65–69. https://doi.org/10.1145/3348488.3348501
[70] J. Cech, R. Mittal, A. Deleforge, J. Sanchez-Riera, X. Alameda-Pineda, and R. Horaud. 2015. Active-speaker detection and localization with microphones and cameras embedded into a robotic head. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. 203–210. https://doi.org/10.1109/HUMANOIDS.2013.7029977
[71] Ibrahim Baran Celik and Mehmet Kuntalp. 2012. Development of a robotic-arm controller by using hand gesture recognition. In *Proceedings of the 2012 International Symposium on Innovations in Intelligent Systems and Applications*. 1–5. https://doi.org/10.1109/INISTA.2012.6246985
[72] G. Chalvatzaki, X. S. Papageorgiou, P. Maragos, and C. S. Tzafestas. 2019. Learn to adapt to human walking: A model-based reinforcement learning approach for a robotic assistant rollator. *IEEE Robotics and Automation Letters* 4, 4 (Oct. 2019), 3774–3781. https://doi.org/10.1109/LRA.2019.2929996
[73] C.-C. Chan and C.-C. Tsai. 2020. Collision-free path planning based on new navigation function for an industrial robotic manipulator in human-robot coexistence environments. *Journal of the Chinese Institute of Engineers* 43, 6 (2020), 508–518. https://doi.org/10.1080/02533839.2020.1771210
[74] F. Chao, F. Chen, Y. Shen, W. He, Y. Sun, Z. Wang, C. Zhou, and M. Jiang. 2014. Robotic free writing of Chinese characters via human-robot interactions. *International Journal of Humanoid Robotics* 11, 1 (2014), 1450007. https://doi.org/10.1142/S0219843614500078
[75] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. 2018. Learning to detect human-object interactions. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*.

[76] B. Chen, C. Hua, B. Dai, Y. He, and J. Han. 2019. Online control programming algorithm for human-robot interaction system with a novel real-time human gesture recognition method. *International Journal of Advanced Robotic Systems* 16, 4 (2019), 1–18. https://doi.org/10.1177/1729881419861764

[77] Dingping Chen, Jilin He, Guanyu Chen, Xiaopeng Yu, Miaolei He, Youwen Yang, Junsong Li, and Xuanyi Zhou. 2020. Human-robot skill transfer systems for mobile robot based on multi sensor fusion. In *Proceedings of the 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN'20)*. IEEE, Los Alamitos, CA, 1354–1359.

[78] H. Chen, M. C. Leu, W. Tao, and Z. Yin. 2020. Design of a real-time human-robot collaboration system using dynamic gestures. In *Proceedings of the ASME International Mechanical Engineering Congress and Exposition (IMECE'20)*, Vol. 2B. https://doi.org/10.1115/IMECE2020-23650

[79] J. Chen and W.-J. Kim. 2019. A human-following mobile robot providing natural and universal interfaces for control with wireless electronic devices. *IEEE/ASME Transactions on Mechatronics* 24, 5 (2019), 2377–2385. https://doi.org/10.1109/TMECH.2019.2936395

[80] K.-Y. Chen, C.-C. Chien, W.-L. Chang, and J.-T. Teng. 2010. An integrated color and hand gesture recognition approach for an autonomous mobile robot. In *Proceedings of the 2010 3rd International Congress on Image and Signal Processing (CISP'10)*, Vol. 5. 2496–2500. https://doi.org/10.1109/CISP.2010.5647930

[81] L. Chen, Z. Dong, S. Gao, B. Yuan, and M. Pei. 2014. Stereovision-only based interactive mobile robot for human-robot face-to-face interaction. In *Proceedings of the International Conference on Pattern Recognition*. 1840–1845. https://doi.org/10.1109/ICPR.2014.322

[82] S. Y. Chen. 2011. Kalman filter for robot vision: A survey. *IEEE Transactions on Industrial Electronics* 59, 11 (2011), 4409–4420.

[83] Shengyong Chen, Youfu Li, and Ngai Ming Kwok. 2011. Active vision in robotic systems: A survey of recent developments. *International Journal of Robotics Research* 30, 11 (2011), 1343–1377.

[84] T.-D. Chen. 2010. Approaches to robotic vision control using image pointing recognition techniques. In *Advances in Neural Network Research and Applications*. Lecture Notes in Electrical Engineering, Vol. 67. Springer, 321–328. https://doi.org/10.1007/978-3-642-12990-2_36

[85] W. Chen and S. Guo. 2012. Person following of a mobile robot using Kinect through features detection based on SURF. *Advanced Materials Research* 542-543 (2012), 779–784. https://doi.org/10.4028/www.scientific.net/AMR.542-543.779

[86] C.-Y. Cheng, Y.-Y. Zhuo, and G.-H. Kuo. 2013. A multiple-robot system for home service. In *Proceedings of the 2013 CACS International Automatic Control Conference (CACS'13)*. 79–84. https://doi.org/10.1109/CACS.2013.6734111

[87] L. Cheng, Q. Sun, H. Su, Y. Cong, and S. Zhao. 2012. Design and implementation of human-robot interactive demonstration system based on Kinect. In *Proceedings of the 2012 24th Chinese Control and Decision Conference (CCDC'12)*. 971–975. https://doi.org/10.1109/CCDC.2012.6242992

[88] A. Cherubini, R. Passama, P. Fraisse, and A. Crosnier. 2015. A unified multimodal control framework for human-robot interaction. *Robotics and Autonomous Systems* 70 (2015), 106–115. https://doi.org/10.1016/j.robot.2015.03.002

[89] J.-C. Chien, Z.-Y. Dang, and J.-D. Lee. 2019. Navigating a service robot for indoor complex environments. *Applied Sciences (Switzerland)* 9, 3 (2019), 491. https://doi.org/10.3390/app9030491

[90] Hui-Min Chou, Yu-Cheng Chou, and Hsin-Hung Chen. 2020. Development of a monocular vision deep learning-based AUV diver-following control system. In *Proceedings of Global Oceans 2020: Singapore—U.S. Gulf Coast*. 1–4. https://doi.org/10.1109/IEEECONF38699.2020.9389477

[91] G.B. Choudhary and R.B.V. Chethan. 2015. Real time robotic arm control using hand gestures. In *Proceedings of the 2014 International Conference on High Performance Computing and Applications, ICHPCA 2014*. https://doi.org/10.1109/ICHPCA.2014.7045349

[92] L. G. Christiernin and S. Augustsson. 2016. Interacting with industrial robots—A motion-based interface. In *Proceedings of the Workshop on Advanced Visual Interfaces (AVI'16)*, Vol. 07. 310–311. https://doi.org/10.1145/2909132.2926073

[93] Marcelo Cicconet, Mason Bretan, and Gil Weinberg. 2013. Human-robot percussion ensemble: Anticipation on the basis of visual cues. *IEEE Robotics & Automation Magazine* 20, 4 (2013), 105–110.

[94] G. Cicirelli, C. Attolico, C. Guaragnella, and T. D'Orazio. 2015. A Kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems* 12 (2015), 22. https://doi.org/10.5772/59974

[95] Felipe Cid, José Augusto Prado, Pablo Bustos, and Pedro Núñez. 2013. A real time and robust facial expression recognition and imitation approach for affective human-robot interaction using Gabor filtering. In *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2188–2193. https://doi.org/10.1109/IROS.2013.6696662

[96] Karin Clark, Matt Duckham, Marilys Guillemin, Assunta Hunter, Jodie McVernon, Christine O'Keefe, Cathy Pitkin, et al. 2019. Advancing the ethical use of digital data in human research: Challenges and strategies to promote ethical practice. *Ethics and Information Technology* 21, 1 (2019), 59–73.

[97] I. Condés and J. M. Cañas. 2019. Person following robot behavior using deep learning. *Advances in Intelligent Systems and Computing* 855 (2019), 147–161. https://doi.org/10.1007/978-3-319-99885-5_11
[98] Peter I. Corke. 2011. *Robotics, Vision and Control: Fundamental Algorithms in MATLAB*, Vol. 73. Springer.
[99] Gabriele Costante, Enrico Bellocchio, Paolo Valigi, and Elisa Ricci. 2014. Personalizing vision-based gestural interfaces for HRI with UAVs: A transfer learning approach. In *Proceedings of the 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 3319–3326. https://doi.org/10.1109/IROS.2014.6943024
[100] Marco Costanzo, Giuseppe De Maria, Gaetano Lettera, Ciro Natale, and Dario Perrone. 2019. A multimodal perception system for detection of human operators in robotic work cells. In *Proceedings of the 2019 IEEE International Conference on Systems, Man, and Cybernetics (SMC'19)*. 692–699. https://doi.org/10.1109/SMC.2019.8914519
[101] A. Couture-Beil, R. T. Vaughan, and G. Mori. 2010. Selecting and commanding individual robots in a vision-based multi-robot system. In *Proceedings of the 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*. 355–356. https://doi.org/10.1109/HRI.2010.5453167
[102] A. Csapó, E. Gilmartin, J. Grizou, J. Han, R. Meena, D. Anastasiou, K. Jokinen, and G. Wilcock. 2012. Multimodal conversational interaction with a humanoid robot. In *Proceedings of the 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom'12)*. 667–672. https://doi.org/10.1109/CogInfoCom.2012.6421935
[103] Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1. IEEE, Los Alamitos, CA, 886–893.
[104] N. Dalal and B. Triggs. 2005. Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'05)*.
[105] Praveen Damacharla, Ahmad Y. Javaid, Jennie J. Gallimore, and Vijay K. Devabhaktuni. 2018. Common metrics to benchmark human-machine teams (HMT): A review. *IEEE Access* 6 (2018), 38637–38655.
[106] Dipankar Das, Yoshinori Kobayashi, and Yoshinori Kuno. 2013. Attracting attention and establishing a communication channel based on the level of visual focus of attention. In *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2194–2201. https://doi.org/10.1109/IROS.2013.6696663
[107] Alessandro De Luca and Fabrizio Flacco. 2012. Integrated control for pHRI: Collision avoidance, detection, reaction and collaboration. In *Proceedings of the 2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob'12)*. 288–295. https://doi.org/10.1109/BioRob.2012.6290917
[108] D. De Schepper, B. Moyaers, G. Schouterden, K. Kellens, and E. Demeester. 2020. Towards robust human-robot mobile co-manipulation for tasks involving the handling of non-rigid materials using sensor-fused force-torque, and skeleton tracking data. *Procedia CIRP* 97 (2020), 325–330. https://doi.org/10.1016/j.procir.2020.05.245
[109] M. Deepan Raj, I. Gogul, M. Thangaraja, and V. S. Kumar. 2017. Static gesture recognition based precise positioning of 5-DOF robotic arm using FPGA. In *Proceedings of the 9th International Conference on Trends in Industrial Measurement and Automation (TIMA'17)*. https://doi.org/10.1109/TIMA.2017.8064804
[110] Angel P. del Pobil, Mario Prats, and Pedro J. Sanz. 2011. Interaction in robotics with a combination of vision, tactile and force sensing. In *Proceedings of the 2011 5th International Conference on Sensing Technology*. IEEE, Los Alamitos, CA, 21–26.
[111] Maxime Devanne, Sao Mai Nguyen, Olivier Remy-Neris, Beatrice Le Gals-Garnett, Gilles Kermarrec, and Andre Thepaut. 2018. A co-design approach for a rehabilitation robot coach for physical rehabilitation based on the error classification of motion errors. In *Proceedings of the 2018 2nd IEEE International Conference on Robotic Computing (IRC'18)*. 352–357. https://doi.org/10.1109/IRC.2018.00074
[112] H. Ding, K. Wijaya, G. Reißig, and O. Stursberg. 2011. Optimizing motion of robotic manipulators in interaction with human operators. In *Intelligent Robotics and Applications*. Lecture Notes in Computer Science, Vol. 7101. Springer, 520–531. https://doi.org/10.1007/978-3-642-25486-4_52
[113] H. M. Do, C. Mouser, M. Liu, and W. Sheng. 2014. Human-robot collaboration in a Mobile Visual Sensor Network. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 2203–2208. https://doi.org/10.1109/ICRA.2014.6907163
[114] A. C. Dometios, X. S. Papageorgiou, A. Arvanitakis, C. S. Tzafestas, and P. Maragos. 2017. Real-time end-effector motion behavior planning approach using on-line point-cloud data towards a user adaptive assistive bath robot. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 5031–5036. https://doi.org/10.1109/IROS.2017.8206387
[115] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*.
[116] Katherine Driggs-Campbell, Vijay Govindarajan, and Ruzena Bajcsy. 2017. Integrating intuitive driver models in autonomous planning for interactive maneuvers. *IEEE Transactions on Intelligent Transportation Systems* 18, 12 (2017), 3461–3472.

[117] D. Droeschel, J. Stückler, D. Holz, and S. Behnke. 2011. Towards joint attention for a domestic service robot—Person awareness and gesture recognition using Time-of-Flight cameras. In *Proceedings of the 2011 IEEE International Conference on Robotics and Automation*. 1205–1210. https://doi.org/10.1109/ICRA.2011.5980067

[118] G. Du, M. Chen, C. Liu, B. Zhang, and P. Zhang. 2018. Online robot teaching with natural human-robot interaction. *IEEE Transactions on Industrial Electronics* 65, 12 (2018), 9571–9581. https://doi.org/10.1109/TIE.2018.2823667

[119] Guanglong Du and Ping Zhang. 2014. Markerless human–robot interface for dual robot manipulators using Kinect sensor. *Robotics and Computer-Integrated Manufacturing* 30, 2 (2014), 150–159.

[120] Brian R. Duffy. 2003. Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42, 3-4 (2003), 177–190.

[121] N. Efthymiou, P. Koutras, P. P. Filntisis, G. Potamianos, and P. Maragos. 2018. Multi-view fusion for action recognition in child-robot interaction. In *Proceedings of the International Conference on Image Processing (ICIP'18)*. 455–459. https://doi.org/10.1109/ICIP.2018.8451146

[122] K. Ehlers and K. Brama. 2016. A human-robot interaction interface for mobile and stationary robots based on real-time 3D human body and hand-finger pose estimation. In *Proceedings of the IEEE International Conference on Emerging Technologies and Factory Automation (ETFA'16)*. https://doi.org/10.1109/ETFA.2016.7733719

[123] Jakob Engel, Thomas Schöps, and Daniel Cremers. 2014. LSD-SLAM: Large-scale direct monocular SLAM. In *Proceedings of the European Conference on Computer Vision*.

[124] Sergio Escalera, Xavier Baró, Jordi Gonzàlez, Miguel A. Bautista, Meysam Madadi, Miguel Reyes, Víctor Ponce-López, Hugo J. Escalante, Jamie Shotton, and Isabelle Guyon. 2015. ChaLearn Looking at People Challenge 2014: Dataset and results. In *Computer Vision—ECCV 2014 Workshops*. Lecture Notes in Computer Science, Vol. 8925. Springer, 459–473. https://doi.org/10.1007/978-3-319-16178-5_32

[125] C.-S. Fahn and Y.-T. Lin. 2010. Real-time face tracking techniques used for the interaction between humans and robots. In *Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications (ICIEA'10)*. 12–17. https://doi.org/10.1109/ICIEA.2010.5514736

[126] Navid Fallahinia and Stephen A. Mascaro. 2020. Comparison of constrained and unconstrained human grasp forces using fingernail imaging and visual servoing. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA'20)*. 2668–2674. https://doi.org/10.1109/ICRA40945.2020.9196963

[127] J. Fang, M. Qiao, and Y. Pei. 2019. Vehicle-mounted with tracked robotic system based on the Kinect. In *Proceedings of the 2019 2nd World Conference on Mechanical Engineering and Intelligent Manufacturing (WCMEIM'19)*. 521–524. https://doi.org/10.1109/WCMEIM48965.2019.00110

[128] Zhijie Fang and Antonio M. López. 2020. Intention recognition of pedestrians and cyclists by 2D pose estimation. *IEEE Transactions on Intelligent Transportation Systems* 21, 11 (2020), 4773–4783. https://doi.org/10.1109/TITS.2019.2946642

[129] M. M. F. M. Fareed, Q. I. Akram, S. B. A. Anees, and A. H. Fakih. 2015. Gesture based wireless single-armed robot in Cartesian 3D space using Kinect. In *Proceedings of the 2015 5th International Conference on Communication Systems and Network Technologies (CSNT'15)*. 1210–1215. https://doi.org/10.1109/CSNT.2015.86

[130] G. A. Farulla, L. O. Russo, C. Pintor, D. Pianu, G. Micotti, A. R. Salgarella, D. Camboni, et al. 2014. Real-time single camera hand gesture recognition system for remote deaf-blind communication. In *Augmented and Virtual Reality*. Lecture Notes in Computer Science, Vol. 8853. Springer, 35–52. https://doi.org/10.1007/978-3-319-13969-2_3

[131] A. A. M. Faudzi, M. H. K. Ali, M. A. Azman, and Z. H. Ismail. 2012. Real-time hand gestures system for mobile robots control. *Procedia Engineering* 41 (2012), 798–804. https://doi.org/10.1016/j.proeng.2012.07.246

[132] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6202–6211.

[133] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. 2009. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 9 (2009), 1627–1645.

[134] Pedro F. Felzenszwalb and Daniel P. Huttenlocher. 2005. Pictorial structures for object recognition. *International Journal of Computer Vision* 61, 1 (2005), 55–79.

[135] Cornelia Fermüller, Fang Wang, Yezhou Yang, Konstantinos Zampogiannis, Yi Zhang, Francisco Barranco, and Michael Pfeiffer. 2018. Prediction of manipulation actions. *International Journal of Computer Vision* 126, 2 (April 2018), 358–374.

[136] F. Ferraguti, C. Talignani Landi, S. Costi, M. Bonfè, S. Farsoni, C. Secchi, and C. Fantuzzi. 2020. Safety barrier functions and multi-camera tracking for human-robot shared environment. *Robotics and Autonomous Systems* 124 (2020), 103388. https://doi.org/10.1016/j.robot.2019.103388

[137] G. Ferrer, A. Garrell, M. Villamizar, I. Huerta, and A. Sanfeliu. 2013. Robot interactive learning through human assistance. *Intelligent Systems Reference Library* 48 (2013), 185–203. https://doi.org/10.1007/978-3-642-35932-3_11

[138] David Forsyth. 2012. *Computer Vision: A Modern Approach* (2nd ed.). Pearson, Boston, MA.

[139] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. P. A. Petrick. 2012. Two people walk into a bar: Dynamic multi-party social interaction with a robot agent. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI'12)*. 3–10. https://doi.org/10.1145/2388676.2388680

[140] Mary Ellen Foster, Rachid Alami, Olli Gestranius, Oliver Lemon, Marketta Niemelä, Jean-Marc Odobez, and Amit Kumar Pandey. 2016. The MuMMER project: Engaging human-robot interaction in real-world public spaces. In *Social Robotics*, Arvin Agah, John-John Cabibihan, Ayanna M. Howard, Miguel A. Salichs, and Hongsheng He (Eds.). Springer International Publishing, Cham, Switzerland, 753–763.

[141] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. 2021. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. *arXiv preprint arXiv:2112.01838* (2021).

[142] Frederic Z. Zhang, Dylan Campbell, and Stephen Gould. 2021. Spatially conditioned graphs for detecting human-object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV'21)*. 13319–13327.

[143] Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* 55, 1 (1997), 119–139.

[144] Muhammad Fuad. 2015. Skeleton based gesture to control manipulator. In *Proceedings of the 2015 International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA'15)*. 96–101. https://doi.org/10.1109/ICAMIMIA.2015.7508010

[145] T. Fujii, J. H. Lee, and S. Okamoto. 2014. Gesture recognition system for human-robot interaction and its application to robotic service task. In *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS'14)*. 63–68. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84938237053&partnerID=40&md5=65a01757df8b0aa92518a19dc3e25b06.

[146] X. Gao, M. Zheng, and M. Q.-H. Meng. 2015. Humanoid robot locomotion control by posture recognition for human-robot interaction. In *Proceedings of the 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO'15)*. 1572–1577. https://doi.org/10.1109/ROBIO.2015.7418995

[147] Y. Gao, H. J. Chang, and Y. Demiris. 2020. User modelling using multimodal information for personalised dressing assistance. *IEEE Access* 8 (2020), 45700–45714. https://doi.org/10.1109/ACCESS.2020.2978207

[148] A. Gardel, F. Espinosa, R. Nieto, J. L. Lázaro, and I. Bravo. 2016. Wireless camera nodes on a cyber-physical system. In *Proceedings of the 10th International Conference on Distributed Smart Camera (ICDSC'16)*. ACM, New York, NY, 31–36. https://doi.org/10.1145/2967413.2967423

[149] J. Gemerek, S. Ferrari, B. H. Wang, and M. E. Campbell. 2019. Video-guided camera control for target tracking and following. *IFAC-PapersOnLine* 51, 34 (2019), 176–183. https://doi.org/10.1016/j.ifacol.2019.01.062

[150] M. Ghandour, H. Liu, N. Stoll, and K. Thurow. 2017. Human robot interaction for hybrid collision avoidance system for indoor mobile robots. *Advances in Science, Technology and Engineering Systems* 2, 3 (2017), 650–657. https://doi.org/10.25046/aj020383

[151] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*. IEEE, Los Alamitos, CA, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

[152] J. Gong, H. Wang, Z. Lu, N. Feng, and F. Hu. 2018. Research on human-robot interaction security strategy of movement authorization for service robot based on people's attention monitoring. In *Proceedings of the 2018 International Conference on Intelligence and Safety for Robotics (ISR'18)*. 521–526. https://doi.org/10.1109/IISR.2018.8535908

[153] Jonas Gonzalez-Billandon, Alessandra Sciutti, Matthew Tata, Giulio Sandini, and Francesco Rea. 2020. Audiovisual cognitive architecture for autonomous learning of face localisation by a Humanoid Robot. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA'20)*. IEEE, Los Alamitos, CA, 5979–5985.

[154] I. Gori, J. K. Aggarwal, L. Matthies, and M. S. Ryoo. 2016. Multitype activity recognition in robot-centric scenarios. *IEEE Robotics and Automation Letters* 1, 1 (Jan. 2016), 593–600. https://doi.org/10.1109/LRA.2016.2525002

[155] I. Gori, S. R. Fanello, G. Metta, and F. Odone. 2012. All gestures you can: A memory game against a humanoid robot. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. 330–336. https://doi.org/10.1109/HUMANOIDS.2012.6651540

[156] Stephen Gould, Jim Rodgers, David Cohen, Gal Elidan, and Daphne Koller. 2008. Multi-class segmentation with relative location prior. *International Journal of Computer Vision* 80, 3 (2008), 300–316.

[157] Consuelo Granata, Joseph Salini, Ragou Ady, and Philippe Bidaud. 2013. Human whole body motion characterization from embedded Kinect. In *Proceedings of the 2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom'13)*. 133–138. https://doi.org/10.1109/CogInfoCom.2013.6719228

[158] B. Görer, A. A. Salah, and H. L. Akin. 2017. An autonomous robotic exercise tutor for elderly people. *Autonomous Robots* 41, 3 (2017), 657–678. https://doi.org/10.1007/s10514-016-9598-5

[159] Ye Gu, Ha Do, Yongsheng Ou, and Weihua Sheng. 2012. Human gesture recognition through a Kinect sensor. In *Proceedings of the 2012 IEEE International Conference on Robotics and Biomimetics (ROBIO'12)*. 1379–1384. https://doi.org/10.1109/ROBIO.2012.6491161

[160] L.-Y. Gui, K. Zhang, Y.-X. Wang, X. Liang, J. M. F. Moura, and M. Veloso. 2018. Teaching robots to predict human motion. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 562–567. https://doi.org/10.1109/IROS.2018.8594452

[161] Liang-Yan Gui, Kevin Zhang, Yu-Xiong Wang, Xiaodan Liang, José M. F. Moura, and Manuela Veloso. 2018. Teaching robots to predict human motion. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*. 562–567. https://doi.org/10.1109/IROS.2018.8594452

[162] Liang Guo, Chenxi Liu, Xiaoyan Wen, Haohua Chen, and Jianghui Zhang. 2016. A control system of human-computer interaction based on Kinect somatosensory equipment. In *Proceedings of the 2016 Chinese Control and Decision Conference (CCDC'16)*. 5170–5175. https://doi.org/10.1109/CCDC.2016.7531921

[163] M. Gupta, L. Behera, V. K. Subramanian, and M. M. Jamshidi. 2015. A robust visual human detection approach with UKF-based motion tracking for a mobile robot. *IEEE Systems Journal* 9, 4 (2015), 1363–1375. https://doi.org/10.1109/JSYST.2014.2317777

[164] Tanmay Gupta, Alexander Schwing, and Derek Hoiem. 2019. No-frills human-object interaction detection: Factorization, layout encodings, and training techniques. *Proceedings of the IEEE International Conference on Computer Vision*.

[165] Akif Hacinecipoglu, Erhan Konukseven, and Ahmet Koku. 2020. Pose invariant people detection in point clouds for mobile robots. *International Journal of Mechanical Engineering and Robotics Research* 9, 5 (2020), 709–715.

[166] Sami Haddadin, Alin Albu-Schäffer, and Gerd Hirzinger. 2009. Requirements for safe robots: Measurements, analysis and new insights. *International Journal of Robotics Research* 28, 11–12 (2009), 1507–1527.

[167] Sami Haddadin, Michael Suppa, Stefan Fuchs, Tim Bodenmüller, Alin Albu-Schäffer, and Gerd Hirzinger. 2011. Towards the robotic co-worker. In *Robotics Research*. Springer, 261–282.

[168] Saad Hafiane, Yasir Salih, and Aamir S. Malik. 2013. 3D hand recognition for telerobotics. In *Proceedings of the 2013 IEEE Symposium on Computers Informatics (ISCI'13)*. 132–137. https://doi.org/10.1109/ISCI.2013.6612390

[169] A. Haghighi, M. Bdiwi, and M. Putz. 2019. Integration of camera and inertial measurement unit for entire human robot interaction using machine learning algorithm. In *Proceedings of the 16th International MultiConference on Systems, Signals, and Devices (SSD'19)*. 741–746. https://doi.org/10.1109/SSD.2019.8893167

[170] Anaïs Halin, Jacques G. Verly, and Marc Van Droogenbroeck. 2021. Survey and synthesis of state of the art in driver monitoring. *Sensors* 21, 16 (2021), 5558.

[171] Roni-Jussi Halme, Minna Lanz, Joni Kämäräinen, Roel Pieters, Jyrki Latokartano, and Antti Hietanen. 2018. Review of vision-based safety systems for human-robot collaboration. *Procedia CIRP* 72 (2018), 111–116.

[172] J. Han, W. Jang, D. Jung, and E. C. Lee. 2017. Human robot interaction method by using hand gesture recognition. In *Advanced Multimedia and Ubiquitous Engineering*. Lecture Notes in Electrical Engineering, Vol. 448. Springer, 97–102. https://doi.org/10.1007/978-981-10-5041-1_17

[173] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. 2013. Enhanced computer vision with Microsoft Kinect sensor: A review. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1318–1334.

[174] Peter A. Hancock, Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors* 53, 5 (2011), 517–527. https://doi.org/10.1177/0018720811417254

[175] Richard Hartley and Andrew Zisserman. 2003. *Multiple View Geometry in Computer Vision*. Cambridge University Press.

[176] F. Hartmann and A. Schlaefer. 2013. Feasibility of touch-less control of operating room lights. *International Journal of Computer Assisted Radiology and Surgery* 8, 2 (2013), 259–268. https://doi.org/10.1007/s11548-012-0778-2

[177] Md. Hasanuzzaman and Tetsunari Inamura. 2010. Adaptation to new user interactively using dynamically calculated principal components for user-specific human-robot interaction. In *Proceedings of the 2010 IEEE/SICE International Symposium on System Integration*. 164–169. https://doi.org/10.1109/SII.2010.5708319

[178] M. S. Hassan, A. F. Khan, M. W. Khan, M. Uzair, and K. Khurshid. 2016. A computationally low cost vision based tracking algorithm for human following robot. In *Proceedings of the 2016 2nd International Conference on Control, Automation, and Robotics (ICCAR'16)*. 62–65. https://doi.org/10.1109/ICCAR.2016.7486699

[179] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision*. 2961–2969.

[180] K. He, X. Zhang, S. Ren, and J. Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[181] F. Hegger, N. Hochgeschwender, G. K. Kraetzschmar, and P. G. Ploeger. 2013. People detection in 3D point clouds using local surface normals. In *RoboCup 2012: Robot Soccer World Cup XVI*. Lecture Notes in Computer Science, Vol. 7500. Springer, 154–165. https://doi.org/10.1007/978-3-642-39250-4_15

[182] João F. Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 3 (March 2015), 583–596.

[183] Abdelfetah Hentout, Mustapha Aouache, Abderraouf Maoudj, and Isma Akli. 2019. Human–robot interaction in industrial collaborative robotics: A literature review of the decade 2008–2017. *Advanced Robotics* 33, 15–16 (2019), 764–799.

[184] Guy Hoffman. 2019. Evaluating fluency in human–robot collaboration. *IEEE Transactions on Human-Machine Systems* 49, 3 (2019), 209–218.

[185] C. Hong, Z. Chen, J. Zhu, and X. Zhang. 2018. Interactive humanoid robot arm imitation system using human upper limb motion tracking. In *Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO'17)*. 2746–2751. https://doi.org/10.1109/ROBIO.2017.8324706

[186] Berthold K. P. Horn and Brian G. Schunck. 1981. Determining optical flow. *Artificial Intelligence* 17, 1–3 (1981), 185–203.

[187] Roy Chaoming Hsu, Po-Cheng Su, Jia-Le Hsu, and Chi-Yong Wang. 2020. Real-time interaction system of human-robot with hand gestures. In *Proceedings of the 2020 IEEE Eurasia Conference on IOT, Communication, and Engineering (ECICE'20)*. 396–398. https://doi.org/10.1109/ECICE50847.2020.9301957

[188] J.-S. Hu, J.-J. Wang, and D. M. Ho. 2014. Design of sensing system and anticipative behavior for human following of mobile robots. *IEEE Transactions on Industrial Electronics* 61, 4 (2014), 1916–1927. https://doi.org/10.1109/TIE.2013.2262758

[189] Chien-Ming Huang and Bilge Mutlu. 2016. Anticipatory robot control for efficient human-robot collaboration. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 83–90. https://doi.org/10.1109/HRI.2016.7451737

[190] C.-L. Hwang, D.-S. Wang, F.-C. Weng, and S.-L. Lai. 2020. Interactions between specific human and omnidirectional mobile robot using deep learning approach: SSD-FN-KCF. *IEEE Access* 8 (2020), 41186–41200. https://doi.org/10.1109/ACCESS.2020.2976712

[191] R. R. Igorevich, E. P. Ismoilovich, and D. Min. 2011. Behavioral synchronization of human and humanoid robot. In *Proceedings of the 2011 8th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI'11)*. 655–660. https://doi.org/10.1109/URAI.2011.6145902

[192] T. Ikai, S. Kamiya, and M. Ohka. 2016. Robot control using natural instructions via visual and tactile sensations. *Journal of Computer Science* 12, 5 (2016), 246–254. https://doi.org/10.3844/jcssp.2016.246.254

[193] W. Indrajit and A. Muis. 2013. Development of whole body motion imitation in humanoid robot. In *Proceedings of the 2013 International Conference on Quality in Research (QiR'13) in Conjunction with ICCS 2013: The 2nd International Conference on Civic Space*. 138–141. https://doi.org/10.1109/QiR.2013.6632552

[194] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 7 (July 2014), 1325–1339. https://doi.org/10.1109/TPAMI.2013.248

[195] Tariq Iqbal and Laurel D. Riek. 2017. Coordination dynamics in multihuman multirobot teams. *IEEE Robotics and Automation Letters* 2, 3 (2017), 1712–1717.

[196] Md. Jahidul Islam, Jungseok Hong, and Junaed Sattar. 2019. Person-following by autonomous robots: A categorical overview. *International Journal of Robotics Research* 38, 14 (2019), 1581–1618. https://doi.org/10.1177/0278364919881683

[197] Shahram Izadi, David Kim, Otmar Hilliges, David Molyneaux, Richard Newcombe, Pushmeet Kohli, Jamie Shotton, et al. 2011. KinectFusion: Real-time 3D reconstruction and interaction using a moving depth camera. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*. 559–568.

[198] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, et al. 2021. Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795* (2021).

[199] Omid Hosseini Jafari, Dennis Mitzel, and Bastian Leibe. 2014. Real-time RGB-D based people detection and tracking for mobile robots and head-worn cameras. In *Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA'14)*. IEEE, Los Alamitos, CA, 5636–5643.

[200] Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer Vision and Image Understanding* 108, 1-2 (2007), 116–134.

[201] M. Jarosz, P. Nawrocki, L. Placzkiewicz, B. Sniezynski, M. Zielinski, and B. Indurkhya. 2019. Detecting gaze direction using robot-mounted and mobile-device cameras. *Computer Science* 20, 4 (2019), 455–476. https://doi.org/10.7494/csci.2019.20.4.3435

[202] A. Jevtić, G. Doisy, Y. Parmet, and Y. Edan. 2015. Comparison of interaction modalities for mobile indoor robot guidance: Direct physical interaction, person following, and pointing control. *IEEE Transactions on Human-Machine Systems* 45, 6 (2015), 653–663. https://doi.org/10.1109/THMS.2015.2461683

[203] Dan Jia, Alexander Hermans, and Bastian Leibe. 2020. DR-SPAAM: A spatial-attention and auto-regressive model for person detection in 2D range data. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'20)*. IEEE, Los Alamitos, CA, 10270–10277.

[204] S. Jia, L. Zhao, X. Li, W. Cui, and J. Sheng. 2011. Autonomous robot human detecting and tracking based on stereo vision. In *Proceedings of the 2011 IEEE International Conference on Mechatronics and Automation (ICMA'11)*. 640–645. https://doi.org/10.1109/ICMA.2011.5985736

[205] Lihua Jiang, Weitian Wang, Yi Chen, and Yunyi Jia. 2018. Personalize vision-based human following for mobile robots by learning from human-driven demonstrations. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. IEEE, Los Alamitos, CA, 726–731.

[206] Mitsuru Jindai and Tomio Watanabe. 2010. A small-size handshake robot system based on a handshake approaching motion model with a voice greeting. In *Proceedings of the 2010 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*. 521–526. https://doi.org/10.1109/AIM.2010.5695738

[207] Z. Ju, X. Ji, J. Li, and H. Liu. 2017. An integrative framework of human hand gesture segmentation for human-robot interaction. *IEEE Systems Journal* 11, 3 (2017), 1326–1336. https://doi.org/10.1109/JSYST.2015.2468231

[208] H. M. Kahily and A. P. Sudheer. 2016. Real-time human detection and tracking from a mobile armed robot using RGB-D sensor. In *Proceedings of the 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (WCTFTR'16)*. https://doi.org/10.1109/STARTUP.2016.7583953

[209] N. Kalidolda and A. Sandygulova. 2018. Towards interpreting robotic system for fingerspelling recognition in real time. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 141–142. https://doi.org/10.1145/3173386.3177085

[210] T. Kanade, J. F. Cohn, and Yingli Tian. 2000. Comprehensive database for facial expression analysis. In *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*. 3124.

[211] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. 2018. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 7122–7131.

[212] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. 2019. Learning 3D human dynamics from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 5614–5623.

[213] Yugo Katsuki, Yuji Yamakawa, and Masatoshi Ishikawa. 2015. High-speed human/robot hand interaction system. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts (HRI'15 Extended Abstracts)*. ACM, New York, NY, 117–118. https://doi.org/10.1145/2701973.2701984

[214] Y. Katsuki, Y. Yamakawa, and M. Ishikawa. 2015. High-speed human/robot hand interaction system. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 117–118. https://doi.org/10.1145/2701973.2701984

[215] Y. Kawasaki, A. Yorozu, M. Takahashi, and E. Pagello. 2020. A multimodal path planning approach to human robot interaction based on integrating action modeling. *Journal of Intelligent and Robotic Systems: Theory and Applications* 100, 3–4 (2020), 955–972. https://doi.org/10.1007/s10846-020-01244-7

[216] X. Ke, Y. Zhu, Y. Yang, J. Xing, and Z. Luo. 2016. Vision system of facial robot SHFR-III for human-robot interaction. In *Proceedings of the 13th International Conference on Informatics in Control, Automation, and Robotics (ICINCO'16)*, Vol. 2. 472–478. https://doi.org/10.5220/0005994804720478

[217] Maram Khatib, Khaled Al Khudir, and Alessandro De Luca. 2017. Visual coordination task for human-robot collaboration. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'17)*. IEEE, Los Alamitos, CA, 3762–3768.

[218] Zahra Rezaei Khavas, S. Reza Ahmadzadeh, and Paul Robinette. 2020. Modeling trust in human-robot interaction: A survey. In *Proceedings of the International Conference on Social Robotics*. 529–541.

[219] Y. Kobayashi and Y. Kuno. 2010. People tracking using integrated sensors for human robot interaction. In *Proceedings of the IEEE International Conference on Industrial Technology*. 1617–1622. https://doi.org/10.1109/ICIT.2010.5472444

[220] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. 2020. VIBE: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 5253–5263.

[221] A. Kogkas, A. Ezzat, R. Thakkar, A. Darzi, and G. Mylonas. 2019. Free-View, 3D gaze-guided robotic scrub nurse. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*. Lecture Notes in Computer Science, Vol. 11768. Springer, 164–172. https://doi.org/10.1007/978-3-030-32254-0_19

[222] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. 2019. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 4501–4510.

[223] Kishore Reddy Konda, Achim Königs, Hannes Schulz, and Dirk Schulz. 2012. Real time interaction with mobile robots using hand gestures. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*. ACM, New York, NY, 177–178. https://doi.org/10.1145/2157689.2157743

[224] H. S. Koppula and A. Saxena. 2016. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 1 (2016), 14–29. https://doi.org/10.1109/TPAMI.2015.2430335

[225] P. N. Koustoumpardis, K. I. Chatzilygeroudis, A. I. Synodinos, and N. A. Aspragathos. 2016. Human robot collaboration for folding fabrics based on force/RGB-D feedback. *Advances in Intelligent Systems and Computing* 371 (2016), 235–243. https://doi.org/10.1007/978-3-319-21290-6_24

[226] Philipp Krähenbühl and Vladlen Koltun. 2011. Efficient inference in fully connected CRFs with Gaussian edge potentials. In *Advances in Neural Information Processing Systems 24*.

[227] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25*.

[228] Minae Kwon, Malte F. Jung, and Ross A. Knepper. 2016. Human expectations of social robots. In *Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 463–464. https://doi.org/10.1109/HRI.2016.7451807

[229] A. Lalejini, D. Duckworth, R. Sween, C. L. Bethel, and D. Carruth. 2015. Evaluation of supervisory control interfaces for mobile robot integration with tactical teams. In *Proceedings of IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO'15)*. 1–6. https://doi.org/10.1109/ARSO.2014.7020971

[230] M. C. Lam, A. S. Prabuwono, H. Arshad, and C. S. Chan. 2011. A real-time vision-based framework for human-robot interaction. In *Visual Informatics: Sustaining Research and Innovations*. Lecture Notes in Computer Science, Vol. 7066. Springer, 257–267. https://doi.org/10.1007/978-3-642-25191-7_25

[231] J. Lambrecht and J. Kruger. 2012. Spatial programming for industrial robots based on gestures and Augmented Reality. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 466–472. https://doi.org/10.1109/IROS.2012.6385900

[232] C. T. Landi, Y. Cheng, F. Ferraguti, M. Bonfe, C. Secchi, and M. Tomizuka. 2019. Prediction of human arm target for robot reaching movements. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 5950–5957. https://doi.org/10.1109/IROS40897.2019.8968559

[233] X. Lang, Z. Feng, and X. Yang. 2020. Research on human-robot natural interaction algorithm based on body potential perception. In *Proceedings of the 2020 ACM 6th International Conference on Computing and Data Engineering (ICCDE'20)*. 260–264. https://doi.org/10.1145/3379247.3379256

[234] Stéphane Lathuilière, Benoit Massé, Pablo Mesejo, and Radu Horaud. 2018. Deep reinforcement learning for audio-visual gaze control. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*. 1555–1562. https://doi.org/10.1109/IROS.2018.8594327

[235] Boris Lau, Kai O. Arras, and Wolfram Burgard. 2010. Multi-model hypothesis group tracking and group size estimation. *International Journal of Social Robotics* 2, 1 (2010), 19–30.

[236] K. N. Lavanya, D. R. Shree, B. R. Nischitha, T. Asha, and C. Gururaj. 2018. Gesture controlled robot. In *Proceedings of the International Conference on Electrical, Electronics, Communication Computer Technologies, and Optimization Techniques (ICEECCOT'17)*. 465–469. https://doi.org/10.1109/ICEECCOT.2017.8284549

[237] D. Leal and Y. Yihun. 2019. Progress in human-robot collaboration for object handover. In *Proceedings of the 2019 IEEE International Symposium on Measurement and Control in Robotics (ISMCR'19)*. C3-2-1–C3-2-6. https://doi.org/10.1109/ISMCR47492.2019.8955665

[238] C.-Y. Lee, H. Lee, I. Hwang, and B.-T. Zhang. 2020. Visual perception framework for an intelligent mobile robot. In *Proceedings of the 2020 17th International Conference on Ubiquitous Robots (UR'20)*. 612–616. https://doi.org/10.1109/UR49135.2020.9144932

[239] J. Lee and B. Ahn. 2020. Real-time human action recognition with a low-cost RGB camera and mobile robot platform. *Sensors (Switzerland)* 20, 10 (2020), 2886. https://doi.org/10.3390/s20102886

[240] J. Lee and M. S. Ryoo. 2017. Learning robot activities from first-person human videos using convolutional future regression. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 1497–1504. https://doi.org/10.1109/IROS.2017.8205953

[241] J.-E. Lee, J. Park, G.-S. Kim, J.-H. Lee, and M.-H. Kim. 2012. Interactive multi-resolution display using a projector mounted mobile robot in intelligent space. *International Journal of Advanced Robotic Systems* 9, 5 (2012), 196. https://doi.org/10.5772/54232

[242] S. J. Lee, G. Shah, A. A. Bhattacharya, and Y. Motai. 2012. Human tracking with an infrared camera using curve matching framework. *Eurasip Journal on Advances in Signal Processing* 2012, 1 (2012), 99. https://doi.org/10.1186/1687-6180-2012-99

[243] Benedikt Leichtmann and Verena Nitsch. 2020. How much distance do humans keep toward robots? Literature review, meta-analysis, and theoretical considerations on personal space in human-robot interaction. *Journal of Environmental Psychology* 68 (2020), 101386.

[244] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. 2018. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *International Journal of Robotics Research* 37, 4–5 (2018), 421–436. https://doi.org/10.1177/0278364917710318

[245] Dongxu Li, Xin Yu, Chenchen Xu, Lars Petersson, and Hongdong Li. 2020. Transferring cross-domain knowledge for video sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 6205–6214.

[246] Hanchuan Li, Peijin Zhang, Samer Al Moubayed, Shwetak N. Patel, and Alanson P. Sample. 2016. ID-Match: A hybrid computer vision and RFID system for recognizing individuals in groups. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA'16)*. ACM, New York, NY, 7. https://doi.org/10.1145/2851581.2889430

[247] K. Li, S. Sun, X. Zhao, J. Wu, and M. Tan. 2019. Inferring user intent to interact with a public service robot using bimodal information analysis. *Advanced Robotics* 33, 7–8 (2019), 369–387. https://doi.org/10.1080/01691864.2019.1599727

[248] K. Li, J. Wu, X. Zhao, and M. Tan. 2019. Real-time human-robot interaction for a service robot based on 3D human activity recognition and human-mimicking decision mechanism. In *Proceedings of the 8th Annual IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER'18)*. 498–503. https://doi.org/10.1109/CYBER.2018.8688272

[249] L. Li, Q. Xu, G. S. Wang, X. Yu, Y. K. Tan, and H. Li. 2015. Visual perception based engagement awareness for multiparty human-robot interaction. *International Journal of Humanoid Robotics* 12, 4 (2015), 1550019. https://doi.org/10.1142/S021984361550019X

[250] T.-H. S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan. 2019. CNN and LSTM based facial expression analysis model for a humanoid robot. *IEEE Access* 7 (2019), 93998–94011. https://doi.org/10.1109/ACCESS.2019.2928364

[251] X. Li, H. Cheng, G. Ji, and J. Chen. 2018. Learning complex assembly skills from Kinect based human robot interaction. In *Proceedings of the 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO'17)*. 2646–2651. https://doi.org/10.1109/ROBIO.2017.8324818

[252] Y.-T. Li, M. Jacob, G. Akingba, and J. P. Wachs. 2013. A cyber-physical management system for delivering and monitoring surgical instruments in the OR. *Surgical Innovation* 20, 4 (2013), 377–384. https://doi.org/10.1177/1553350612459109

[253] M. Lichtenstern, M. Frassl, B. Perun, and M. Angermann. 2012. A prototyping environment for interaction between a human and a robotic multi-agent system. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*. 185–186. https://doi.org/10.1145/2157689.2157747

[254] B. Lima, G. L. N. Júnior, L. Amaral, T. Vieira, B. Ferreira, and T. Vieira. 2019. Real-time hand pose tracking and classification for natural human-robot control. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging, and Computer Graphics Theory and Applications (VISIGRAPP'19)*, Vol. 5. 832–839. https://doi.org/10.5220/0007384608320839

[255] Timm Linder and Kai O. Arras. 2014. Multi-model hypothesis tracking of groups of people in RGB-D data. In *Proceedings of the 17th International Conference on Information Fusion (FUSION'14)*. IEEE, Los Alamitos, CA, 1–7.

[256] Hongyi Liu and Lihui Wang. 2018. Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics* 68 (2018), 355–367.

[257] Phoebe Liu, Dylan F. Glas, Takayuki Kanda, and Hiroshi Ishiguro. 2018. Learning proactive behavior for interactive social robots. *Autonomous Robots* 42, 5 (2018), 1067–1085.

[258] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. 2016. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision*. 13811.

[259] Xiaofeng Liu, Xu Zhou, Ce Liu, Jianmin Wang, Xiaoqin Zhou, Ning Xu, and Aimin Jiang. 2016. An interactive training system of motor learning by imitation and speech instructions for children with autism. In *Proceedings of the 2016 9th International Conference on Human System Interactions (HSI'16)*. 56–61. https://doi.org/10.1109/HSI.2016.7529609

[260] Yugang Liu and Goldie Nejat. 2016. Multirobot cooperative learning for semiautonomous control in urban search and rescue applications. *Journal of Field Robotics* 33, 4 (2016), 512–536.

[261] Yu-Chi Liu and Qiong-Hai Dai. 2010. A survey of computer vision applied in Aerial robotic Vehicles. In *Proceedings of the 2010 International Conference on Optics, Photonics, and Energy Engineering (OPEE'10)*, Vol. 1. 277–280. https://doi.org/10.1109/OPEE.2010.5508131

[262] Z. Liu, X. Wang, Y. Cai, W. Xu, Q. Liu, Z. Zhou, and D. T. Pham. 2020. Dynamic risk assessment and active response strategy for industrial human-robot collaboration. *Computers and Industrial Engineering* 141 (2020), 106302. https://doi.org/10.1016/j.cie.2020.106302

[263] Y. Long, Y. Xu, Z. Xiao, and Z. Shen. 2018. Kinect-based human body tracking system control of medical care service robot. In *Proceedings of the 2018 WRC Symposium on Advanced Robotics and Automation (WRC SARA'18)*. 65–69. https://doi.org/10.1109/WRC-SARA.2018.8584246

[264] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics* 34, 6 (2015), 1–16.

[265] Percy W. Lovon-Ramos, Yessica Rosas-Cuevas, Claudia Cervantes-Jilaja, Maria Tejada-Begazo, Raquel E. Patiño-Escarcina, and Dennis Barrios-Aranibar. 2016. People detection and localization in real time during navigation of autonomous robots. In *Proceedings of the 2016 XIII Latin American Robotics Symposium and IV Brazilian Robotics Symposium (LARS/SBR'16)*. 239–244. https://doi.org/10.1109/LARS-SBR.2016.47

[266] David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the 7th IEEE International Conference on Computer Vision*, Vol. 2. IEEE, Los Alamitos, CA, 1150–1157.

[267] G. Lu, W. Tang, J. Zheng, T. Chen, and X. Zou. 2020. Research and implementation of real-time motion control of robot based on Kinect. *Smart Innovation, Systems and Technologies* 166 (2020), 779–792. https://doi.org/10.1007/978-3-030-57745-2_65

[268] Bruce D. Lucas and Takeo Kanade. 1981. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81)*.

[269] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. 2010. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 3019.

[270] R. C. Luo, S.-R. Chang, and Y.-P. Yang. 2011. Tracking with pointing gesture recognition for human-robot interaction. In *Proceedings of the 2011 IEEE/SICE International Symposium on System Integration (SII'11)*. 1220–1225. https://doi.org/10.1109/SII.2011.6147623

[271] R. C. Luo, C. H. Huang, and T. T. Lin. 2010. Human tracking and following using sound source localization for multisensor based mobile assistive companion robot. In *Proceedings of the 36th Annual Conference on IEEE Industrial Electronics Society (IECON'10)*. 1552–1557. https://doi.org/10.1109/IECON.2010.5675451

[272] X. Luo, A. Amighetti, and D. Zhang. 2019. A human-robot interaction for a mecanum wheeled mobile robot with real-time 3D two-hand gesture recognition. *Journal of Physics: Conference Series* 1267 (2019), 012056. https://doi.org/10.1088/1742-6596/1267/1/012056

[273] X. Luo, D. Zhang, and X. Jin. 2019. A real-time moving target following mobile robot system with depth camera. *IOP Conference Series: Materials Science and Engineering* 491 (2019), 012004. https://doi.org/10.1088/1757-899X/491/1/012004

[274] Christoph Lutz, Maren Schöttler, and Christian Pieter Hoffmann. 2019. The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication* 7, 3 (2019), 412–434. https://doi.org/10.1177/2050157919843961

[275] Ryan A. MacDonald and Stephen L. Smith. 2019. Active sensing for motion planning in uncertain environments via mutual information policies. *International Journal of Robotics Research* 38, 2–3 (2019), 146–161.

[276] A. Maher, C. Li, H. Hu, and B. Zhang. 2017. Realtime human-UAV interaction using deep learning. In *Biometric Recognition*. Lecture Notes in Computer Science, Vol. 10568. Springer, 511–519. https://doi.org/10.1007/978-3-319-69923-3_55

[277] M. Manigandan and I. M. Jackin. 2010. Wireless vision based mobile robot control using hand gesture recognition through perceptual color space. In *Proceedings of the 2010 International Conference on Advances in Computer Engineering (ACE'10)*. 95–99. https://doi.org/10.1109/ACE.2010.69

[278] Sotiris Manitsaris, Apostolos Tsagaris, Alina Glushkova, Fabien Moutarde, and Frédéric Bevilacqua. 2016. Fingers gestures early-recognition with a unified framework for RGB or depth camera. In *Proceedings of the 3rd International Symposium on Movement and Computing (MOCO'16)*. ACM, New York, NY. https://doi.org/10.1145/2948910.2948947

[279] L. Mao and P. Zhu. 2018. The medical service robot interaction based on Kinect. In *Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization, and Signal Processing (INCOS'17)*. 1–7. https://doi.org/10.1109/ITCOSP.2017.8303077

[280] Dardan Maraj, Arianit Maraj, and Adhurim Hajzeraj. 2016. Application interface for gesture recognition with Kinect sensor. In *Proceedings of the 2016 IEEE International Conference on Knowledge Engineering and Applications (ICKEA'16)*. 98–102. https://doi.org/10.1109/ICKEA.2016.7803000

[281] C. Martin, F.-F. Steege, and H.-M. Gross. 2010. Estimation of pointing poses for visually instructing mobile robots under real world conditions. *Robotics and Autonomous Systems* 58, 2 (2010), 174–185. https://doi.org/10.1016/j.robot.2009.09.013

[282] J. B. Martin and F. Moutarde. 2019. Real-time gestural control of robot manipulator through deep learning human-pose inference. In *Computer Vision Systems*. Lecture Notes in Computer Science, Vol. 11754. Springer, 565–572. https://doi.org/10.1007/978-3-030-34995-0_51

[283] R. Masmoudi, M. Bouchouicha, and P. Gorce. 2011. Expressive robot to support elderly. *Assistive Technology Research Series* 29 (2011), 557–564. https://doi.org/10.3233/978-1-60750-814-4-557

[284] Jean Massardi, Mathieu Gravel, and Éric Beaudry. 2020. PARC: A plan and activity recognition component for assistive robots. In *Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA'20)*. IEEE, Los Alamitos, CA, 3025–3031.

[285] A. Mateus, P. Miraldo, P. U. Lima, and J. Sequeira. 2016. Human-aware navigation using external omnidirectional cameras. *Advances in Intelligent Systems and Computing* 417 (2016), 283–295. https://doi.org/10.1007/978-3-319-27146-0_22

[286] I. Maurtua, I. Fernández, A. Tellaeche, J. Kildal, L. Susperregi, A. Ibarguren, and B. Sierra. 2017. Natural multimodal communication for human-robot collaboration. *International Journal of Advanced Robotic Systems* 14, 4 (2017), 1–12. https://doi.org/10.1177/1729881417716043

[287] O. Mazhar, B. Navarro, S. Ramdani, R. Passama, and A. Cherubini. 2019. A real-time human-robot interaction framework with robust background invariant hand gesture detection. *Robotics and Computer-Integrated Manufacturing* 60 (2019), 34–48. https://doi.org/10.1016/j.rcim.2019.05.008

[288] Grace McFassel, Sheng-Jen Hsieh, and Bo Peng. 2018. Prototyping and evaluation of interactive and customized interface and control algorithms for robotic assistive devices using Kinect and infrared sensor. *International Journal of Advanced Robotic Systems* 15, 2 (2018), 1729881418769521.

[289] Stephen McKeague, Jindong Liu, and Guang-Zhong Yang. 2013. Hand and body association in crowded environments for human-robot interaction. In *Proceedings of the 2013 IEEE International Conference on Robotics and Automation*. IEEE, Los Alamitos, CA, 2161–2168.

[290] R. Mead and M. J. Mataric. 2012. A probabilistic framework for autonomous proxemic control in situated and mobile human-robot interaction. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*. 193–194. https://doi.org/10.1145/2157689.2157751

[291] A. C. S. Medeiros, P. Ratsamee, Y. Uranishi, T. Mashita, and H. Takemura. 2020. Human-drone interaction: Using pointing gesture to define a target object. In *Human-Computer Interaction. Multimodal and Natural Interaction*. Lecture Notes in Computer Science, Vol. 12182. Springer, 688–705. https://doi.org/10.1007/978-3-030-49062-1_48

[292] A. Meghdari, S. B. Shouraki, A. Siamy, and A. Shariati. 2017. The real-time facial imitation by a social humanoid robot. In *Proceedings of the 4th RSI International Conference on Robotics and Mechatronics (ICRoM'16)*. 524–529. https://doi.org/10.1109/ICRoM.2016.7886797

[293] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys* 54, 6 (July 2021), Article 115, 35 pages. https://doi.org/10.1145/3457607

[294] Nuno Mendes, João Ferrer, João Vitorino, Mohammad Safeea, and Pedro Neto. 2017. Human behavior and hand gesture classification for smart human-robot interaction. *Procedia Manufacturing* 11 (2017), 91–98.

[295] Zhen-Qiang Mi and Yang Yang. 2013. Human-robot interaction in UVs swarming: A survey. *International Journal of Computer Science Issues* 10, 2 Pt. 1 (2013), 273.

[296] J. Miller, S. Hong, and J. Lu. 2019. Self-driving mobile robots using human-robot interactions. In *Proceedings of the 2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC'18)*. 1251–1256. https://doi.org/10.1109/SMC.2018.00219

[297] B. Milligan, G. Mori, and R. Vaughan. 2011. Selecting and commanding groups in a multi-robot vision based system. In *Proceedings of the 2011 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*. 415–415. https://doi.org/10.1145/1957656.1957809

[298] K. Miyoshi, R. Konomura, and K. Hori. 2014. Above your hand: Direct and natural interaction with aerial robot. In *Proceedings of ACM SIGGRAPH 2014 Emerging Technologics (SIGGRAPH'14)*. https://doi.org/10.1145/2614066.2614086

[299] S. Müller, T. Wengefeld, T. Q. Trinh, D. Aganian, M. Eisenbach, and H.-M. Gross. 2020. A multi-modal person perception framework for socially interactive mobile service robots. *Sensors (Switzerland)* 20, 3 (2020), 722. https://doi.org/10.3390/s20030722

[300] J. A. Méndez-Polanco, A. Muñoz-Meléndez, and E. F. Morales-Manzanares. 2010. Detection of multiple people by a mobile robot in dynamic indoor environments. In *Advances in Artificial Intelligence—IBERAMIA 2010*. Lecture Notes in Computer Science, Vol. 6433. Springer, 522–531. https://doi.org/10.1007/978-3-642-16952-6_53

[301] Signe Moe and Ingrid Schjølberg. 2013. Real-time hand guiding of industrial manipulator in 5 DOF using Microsoft Kinect and accelerometer. In *Proceedings of the 2013 IEEE International Workshop on Robot and Human Communication (RO-MAN'13)*. 644–649. https://doi.org/10.1109/ROMAN.2013.6628421

[302] Thomas B. Moeslund and Erik Granum. 2001. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding* 81, 3 (2001), 231–268.

[303] J.J. Moh, T. Kijima, B. Zhang, and H.-O. Lim. 2019. Gesture recognition and effective interaction based dining table cleaning robot. In *Proceedings of the 2019 7th International Conference on Robot Intelligence Technology and Applications (RiTA'19)*. 72–77. https://doi.org/10.1109/RITAPP.2019.8932802

[304] Sepehr MohaimenianPour and Richard Vaughan. 2018. Hands and faces, fast: Mono-camera user detection robust enough to directly control a UAV in flight. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'18)*. 5224–5231. https://doi.org/10.1109/IROS.2018.8593709

[305] F. Mohammad, K. R. Sudini, V. Puligilla, and P. R. Kapula. 2013. Tele-operation of robot using gestures. In *Proceedings of the 2013 7th Asia Modelling Symposium (AMS'13)*. 67–71. https://doi.org/10.1109/AMS.2013.15

[306] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and Prisma Group Collaborators. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine* 6, 7 (2009), e1000097.

[307] M. W. C. N. Moladande and B. G. D. A. Madhusanka. 2019. Implicit intention and activity recognition of a human using neural networks for a service robot eye. In *Proceedings of the 2019 International Research Conference on Smart Computing and Systems Engineering (SCSE'19)*. IEEE, Los Alamitos, CA, 38–43.

[308] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing* 10, 1 (Jan. 2019), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

[309] C. Mollaret, A. A. Mekonnen, J. Pinquier, F. Lerasle, and I. Ferrane. 2016. A multi-modal perception based architecture for a non-intrusive domestic assistant robot. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 481–482. https://doi.org/10.1109/HRI.2016.7451816

[310] Mani Monajjemi, Jake Bruce, Seyed Abbas Sadat, Jens Wawerla, and Richard Vaughan. 2015. UAV, do you see me? Establishing mutual attention between an uninstrumented human and an outdoor UAV in flight. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*. 3614–3620. https://doi.org/10.1109/IROS.2015.7353882

[311] V. M. Monajjemi, J. Wawerla, R. Vaughan, and G. Mori. 2013. HRI in the sky: Creating and commanding teams of UAVs with a vision-mediated gestural interface. In *Proceedings of the 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 617–623. https://doi.org/10.1109/IROS.2013.6696415

[312] C. Morato, K. N. Kaipa, B. Zhao, and S. K. Gupta. 2014. Toward safe human robot collaboration by using multiple Kinects based real-time human tracking. *Journal of Computing and Information Science in Engineering* 14, 1 (2014), 011006. https://doi.org/10.1115/1.4025810

[313] R. J. Moreno, M. Mauledoux, and O. F. Avilés. 2016. Path optimization planning for human-robot interaction. *International Journal of Applied Engineering Research* 11, 22 (2016), 10822–10827. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85002736320&partnerID=40&md5=f881e78233bc762de3474a62d985513c.

[314] D. Mronga, T. Knobloch, J. de Gea Fernández, and F. Kirchner. 2020. A constraint-based approach for human-robot collision avoidance. *Advanced Robotics* 34, 5 (2020), 265–281. https://doi.org/10.1080/01691864.2020.1721322

[315] Michael J. Muller. 2007. *Participatory Design: The Third Space in HCI*. CRC Press, Boca Raton, FL.

[316] M. Munaro and E. Menegatti. 2014. Fast RGB-D people tracking for service robots. *Autonomous Robots* 37, 3 (2014), 227–242. https://doi.org/10.1007/s10514-014-9385-0

[317] Matteo Munaro and Emanuele Menegatti. 2014. Fast RGB-D people tracking for service robots. *Autonomous Robots* 37, 3 (2014), 227–242.

[318] J. Nagi, H. Ngo, L. M. Gambardella, and Gianni A. Di Caro. 2015. Wisdom of the swarm for cooperative decision-making in human-swarm interaction. In *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA'15)*. 1802–1808. https://doi.org/10.1109/ICRA.2015.7139432

[319] S. Nair, E. Dean-Leon, and A. Knoll. 2011. 3D position based multiple human servoing by low-level-control of 6 DOF industrial robot. In *Proceedings of the 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO'11)*. 2816–2823. https://doi.org/10.1109/ROBIO.2011.6181732

[320] Hugo Nascimento, Martin Mujica, and Mourad Benoussaad. 2020. Collision avoidance in human-robot interaction using Kinect vision system combined with robot's model and data. In *Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'20)*. 10293–10298. https://doi.org/10.1109/IROS45743.2020.9341248

[321] Samira Nazari, Mostafa Charmi, Maryam Hassani, and Ghazale Ahmadi. 2015. A simplified method in human to robot motion mapping schemes. In *Proceedings of the 2015 3rd RSI International Conference on Robotics and Mechatronics (ICROM'15)*. 545–550. https://doi.org/10.1109/ICRoM.2015.7367842

[322] Q. Nguyen, S.-S. Yun, and J. Choi. 2014. Audio-visual integration for human-robot interaction in multi-person scenarios. In *Proceedings of the 19th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA'14)*. https://doi.org/10.1109/ETFA.2014.7005303

[323] T. Nishiyama, M. Takimoto, and Y. Kambayashi. 2013. Human intervention for searching targets using mobile agents in a multi-robot environment. *Frontiers in Artificial Intelligence and Applications* 254 (2013), 154–163. https://doi.org/10.3233/978-1-61499-262-2-154

[324] Takenori Obo, Chu Kiong Loo, and Naoyuki Kubota. 2015. Robot posture generation based on genetic algorithm for imitation. In *Proceedings of the 2015 IEEE Congress on Evolutionary Computation (CEC'15)*. 552–557. https://doi.org/10.1109/CEC.2015.7256938

[325] Dimitri Ognibene, Eris Chinellato, Miguel Sarabia, and Yiannis Demiris. 2013. Contextual action recognition and target localization with an active allocation of attention on a humanoid robot. *Bioinspiration & Biomimetics* 8, 3 (2013), 035002.

[326] Valerio Ortenzi, Akansel Cosgun, Tommaso Pardi, Wesley P. Chan, Elizabeth Croft, and Dana Kulić. 2021. Object handovers: A review for robotics. *IEEE Transactions on Robotics* 37, 6 (2021), 1855–1873.

[327] Maike Paetzel and Ginevra Castellano. 2019. Let me get to know you better: Can interactions help to overcome uncanny feelings? In *Proceedings of the 7th International Conference on Human-Agent Interaction (HAI'19)*. ACM, New York, NY, 59–67. https://doi.org/10.1145/3349537.3351894

[328] L. Pang, Y. Zhang, S. Coleman, and H. Cao. 2020. Efficient hybrid-supervised deep reinforcement learning for person following robot. *Journal of Intelligent and Robotic Systems: Theory and Applications* 97, 2 (2020), 299–312. https://doi.org/10.1007/s10846-019-01030-0

[329] Christos Papadopoulos, Ioannis Mariolis, Angeliki Topalidou-Kyniazopoulou, Grigorios Piperagkas, Dimosthenis Ioannidis, and Dimitrios Tzovaras. 2019. An advanced human-robot interaction interface for collaborative robotic assembly tasks. In *Rapid Automation: Concepts, Methodologies, Tools, and Applications*. IGI Global, 794–812.

[330] C.-B. Park and S.-W. Lee. 2011. Real-time 3D pointing gesture recognition for mobile robots with cascade HMM and particle filter. *Image and Vision Computing* 29, 1 (2011), 51–63. https://doi.org/10.1016/j.imavis.2010.08.006

[331] S. Pasinetti, C. Nuzzi, M. Lancini, G. Sansoni, F. Docchio, and A. Fornaser. 2018. Development and characterization of a safety system for robotic cells based on multiple time of flight (TOF) cameras and point cloud analysis. In *Proceedings of the 2018 Workshop on Metrology for Industry 4.0 and IoT (MetroInd 4.0 and IoT'18)*. 34–39. https://doi.org/10.1109/METROI4.2018.8439037

[332] Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. 2021. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems 34*.

[333] T. Paulo, R. Fernando, and L. Gil. 2012. Vision-based hand segmentation techniques for human-robot interaction for real-time applications. In *Proceedings of the 3rd ECCOMAS Thematic Conference on Computational Vision and Medical Image Processing*. 31–35. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84856694609&partnerID=40&md5=a6a3bbed4537f2d964f66ea6f3d5bf9c.

[334] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. 2018. Ordinal depth supervision for 3D human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 7307–7316.

[335] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 2019. 3D human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 7753–7762.

[336] Dexmont Pena, Andrew Forembski, Xiaofan Xu, and David Moloney. 2017. Benchmarking of CNNs for low-cost, low-power robotics applications. In *Proceedings of the RSS 2017 Workshop: New Frontier for Deep Learning in Robotics*. 1–5.

[337] A. Pennisi, F. Previtali, C. Gennari, D. D. Bloisi, L. Iocchi, F. Ficarola, A. Vitaletti, and D. Nardi. 2015. Multi-robot surveillance through a distributed sensor network. *Studies in Computational Intelligence* 604 (2015), 77–98. https://doi.org/10.1007/978-3-319-18299-5_4

[338] S.-G. Pentiuc and O.-M. Vultur. 2018. "Drive me": A interaction system between human and robot. In *Proceedings of the 2018 14th International Conference on Development and Application Systems (DAS'18)*. 144–149. https://doi.org/10.1109/DAAS.2018.8396087

[339] F. G. Pereira, R. F. Vassallo, and E. O. T. Salles. 2013. Human-robot interaction and cooperation through people detection and gesture recognition. *Journal of Control, Automation and Electrical Systems* 24, 3 (2013), 187–198. https://doi.org/10.1007/s40313-013-0040-3

[340] T. Petric, A. Gams, L. Zlajpah, A. Ude, and J. Morimoto. 2014. Online approach for altering robot behaviors based on human in the loop coaching gestures. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 4770–4776. https://doi.org/10.1109/ICRA.2014.6907557

[341] K. P. Pfeil, S. L. Koh, and J. J. LaViola Jr. 2013. Exploring 3D gesture metaphors for interaction with unmanned aerial vehicles. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI'13)*. 257–266. https://doi.org/10.1145/2449396.2449429

[342] Tomas Pfister, James Charles, and Andrew Zisserman. 2013. Large-scale learning of sign language by watching TV (using co-occurrences). In *Proceedings of the British Machine Vision Conference*.

[343] N. T. T. Phong, L. H. T. Nam, and N. T. Thinh. 2020. Vietnamese service robot based on artificial intelligence. *International Journal of Mechanical Engineering and Robotics Research* 9, 5 (2020), 701–708. https://doi.org/10.18178/ijmerr.9.5.701-708

[344] Harry A. Pierson and Michael S. Gashler. 2017. Deep learning in robotics: A review of recent research. *Advanced Robotics* 31, 16 (2017), 821–835. https://doi.org/10.1080/01691864.2017.1365009

[345] Francesco Pittaluga, Sanjeev J. Koppal, Sing Bing Kang, and Sudipta N. Sinha. 2019. Revealing scenes by inverting structure from motion reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*. 145–154.

[346] Alexandru Pop and Ovidiu Stan. 2019. Control a 6DOF anthropomorphic robotic structure with computer vision as MEMS input. In *Proceedings of the 2019 22nd International Conference on Control Systems and Computer Science (CSCS'19)*. IEEE, Los Alamitos, CA, 700–706.

[347] S. Potdar, A. Sawarkar, and F. Kazi. 2016. Learning by demonstration from multiple agents in humanoid robots. In *Proceedings of the 2016 IEEE Students' Conference on Electrical, Electronics, and Computer Science (SCEECS'16)*. https://doi.org/10.1109/SCEECS.2016.7509324

[348] M. Prediger, A. Braun, A. Marinc, and A. Kuijper. 2014. Robot-supported pointing interaction for intelligent environments. In *Distributed, Ambient, and Pervasive Interations*. Lecture Notes in Computer Science, Vol. 8530. Springer, 172–183. https://doi.org/10.1007/978-3-319-07788-8_17

[349] Alexandru Ionut Pustianu, Adriana Serbencu, and Daniela Cristina Cernega. 2011. Mobile robot control using face recognition algorithms. In *Proceedings of the 15th International Conference on System Theory, Control, and Computing*. 1–6.

[350] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. 2018. Learning human-object interactions by graph parsing neural networks. In *Proceedings of the European Conference on Computer Vision*.

[351] K. Qian and C. Hu. 2013. Visually gesture recognition for an interactive robot grasping application. *International Journal of Multimedia and Ubiquitous Engineering* 8, 3 (2013), 189–196. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84878477530&partnerID=40&md5=983edd9a03f6a4690e16308da763716e.

[352] C. P. Quintero, R. T. Fomena, A. Shademan, O. Ramirez, and M. Jagersand. 2014. Interactive teleoperation interface for semi-autonomous control of robot arms. In *Proceedings of the Conference on Computer and Robot Vision (CRV'14)*. 357–363. https://doi.org/10.1109/CRV.2014.55

[353] C. P. Quintero, R. Tatsambon, M. Gridseth, and M. Jagersand. 2015. Visual pointing gestures for bi-directional human robot interaction in a pick-and-place task. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*. 349–354. https://doi.org/10.1109/ROMAN.2015.7333604

[354] A. A. Ramírez-Duque, A. Frizera-Neto, and T. F. Bastos. 2019. Robot-assisted autism spectrum disorder diagnostic based on artificial reasoning. *Journal of Intelligent and Robotic Systems: Theory and Applications* 96, 2 (2019), 267–281. https://doi.org/10.1007/s10846-018-00975-y

[355] Gabriele Randelli, Taigo Maria Bonanni, Luca Iocchi, and Daniele Nardi. 2013. Knowledge acquisition through human–robot multimodal interaction. *Intelligent Service Robotics* 6, 1 (2013), 19–31.

[356] A. U. Ratul, M. T. Ali, and R. Ahasan. 2016. Gesture based wireless shadow robot. In *Proceedings of the 2016 5th International Conference on Informatics, Electronics, and Vision (ICIEV'16)*. 351–355. https://doi.org/10.1109/ICIEV.2016.7760024

[357] Siddharth S. Rautaray and Anupam Agrawal. 2015. Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review* 43, 1 (2015), 1–54.

[358] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do ImageNet classifiers generalize to ImageNet? *arXiv:1902.10811* [cs.CV] (2019).

[359] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. 779–788.

[360] Matthias Rehm and Anders Krogsager. 2013. Negative affect in human robot interaction—Impoliteness in unexpected encounters with robots. In *Proceedings of the 2013 IEEE International Workshop on Robot and Human Communication (RO-MAN'13)*. IEEE, Los Alamitos, CA, 45–50.

[361] S. Rehman, T. Ashraf, M. Umair, U. Zubair, Y. Ayaz, and H. Khan. 2017. Target detection and tracking using intelligent wheelchair. *International Journal of Simulation: Systems, Science and Technology* 18, 1 (2017), 4.1–4.8. https://doi.org/10.5013/IJSSST.a.18.01.04 0.

[362] Zhongzheng Ren, Yong Jae Lee, and Michael S. Ryoo. 2018. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the European Conference on Computer Vision (ECCV'18)*.

[363] Jorge Rios-Martinez, Anne Spalanzani, and Christian Laugier. 2015. From proxemics theory to socially-aware navigation: A survey. *International Journal of Social Robotics* 7, 2 (2015), 137–153.

[364] Paul Robinette, Ayanna Howard, and Alan R. Wagner. 2017. *Conceptualizing Overtrust in Robots: Why Do People Trust a Robot That Previously Failed?* Springer International Publishing, Cham, Switzerland, 129–155. https://doi.org/10.1007/978-3-319-59719-5_6

[365] Nicole Lee Robinson, Timothy Vaughan Cottier, and David John Kavanagh. 2019. Psychosocial health interventions by social robots: Systematic review of randomized controlled trials. *Journal of Medical Internet Research* 21, 5 (May 2019), e13203. https://doi.org/10.2196/13203

[366] Wendy A. Rogers, Travis Kadylak, and Megan A. Bayles. 2022. Maximizing the benefits of participatory design for human-robot interaction research with older adults. *Human Factors* 64, 3 (2022), 441–450. https://doi.org/10.1177/00187208211037465

[367] Alina Roitberg, Alexander Perzylo, Nikhil Somani, Manuel Giuliani, Markus Rickert, and Alois Knoll. 2014. Human activity recognition in the context of industrial human-robot interaction. In *Proceedings of the 2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA'14)*. 1–10. https://doi.org/10.1109/APSIPA.2014.7041588

[368] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*. 234–241.

[369] Dorsa Sadigh, Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. 2016. Planning for autonomous cars that leverage effects on human actions. In *Proceedings of Robotics: Science and Systems*, Vol. 2. 1–9.

[370] R. Saegusa, L. Natale, G. Metta, and G. Sandini. 2011. Cognitive robotics-active perception of the self and others. In *Proceedings of the 4th International Conference on Human System Interaction (HSI'11)*. 419–426. https://doi.org/10.1109/HSI.2011.5937403

[371] Mohammad Taghi Saffar, Mircea Nicolescu, Monica Nicolescu, and Banafsheh Rekabdar. 2015. Context-based intent understanding using an Activation Spreading architecture. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*. 3002–3009. https://doi.org/10.1109/IROS.2015.7353791

[372] N. SaiChinmayi, Ch. Hasitha, B. Sravya, and V. K. Mittal. 2015. Gesture signals processing for a silent spybot. In *Proceedings of the 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN'15)*. 756–761. https://doi.org/10.1109/SPIN.2015.7095406

[373] S. Saleh and K. Berns. 2015. Nonverbal communication with a humanoid robot via head gestures. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments (PETRA'15)*. https://doi.org/10.1145/2769493.2769543

[374] Ricardo Sanchez-Matilla, Konstantinos Chatzilygeroudis, Apostolos Modas, Nuno Ferreira Duarte, Alessio Xompero, Pascal Frossard, Aude Billard, and Andrea Cavallaro. 2020. Benchmark for human-to-robot handovers of unseen containers with unknown filling. *IEEE Robotics and Automation Letters* 5, 2 (April 2020), 1642–1649. https://doi.org/10.1109/LRA.2020.2969200

[375] A. Sanna, F. Lamberti, G. Paravati, E. A. Henao Ramirez, and F. Manuri. 2012. A Kinect-based natural interface for quadrotor control. In Intelligent Technologies for Interactive Entertainment. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 78. Springer, 48–56. https://doi.org/10.1007/978-3-642-30214-5_6

[376] L. Santos, A. Geminiani, I. Olivieri, J. Santos-Victor, and A. Pedrocchi. 2020. CopyRobot: Interactive mirroring robotics game for ASD children. *IFMBE Proceedings* 76 (2020), 2014–2027. https://doi.org/10.1007/978-3-030-31635-8_239

[377] Shane Saunderson and Goldie Nejat. 2019. How robots influence humans: A survey of nonverbal communication in social human–robot interaction. *International Journal of Social Robotics* 11, 4 (2019), 575–608.

[378] Matteo Saveriano and Dongheui Lee. 2014. Safe motion generation and online reshaping using dynamical systems. In *Proceedings of the 2014 11th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI'14)*. 45–45. https://doi.org/10.1109/URAI.2014.7057407

[379] S. Scheggi, F. Morbidi, and D. Prattichizzo. 2014. Human-robot formation control via visual and vibrotactile haptic feedback. *IEEE Transactions on Haptics* 7, 4 (2014), 499–511. https://doi.org/10.1109/TOH.2014.2332173

[380] B. Schmidt and L. Wang. 2013. Contact-less and programming-less human-robot collaboration. *Procedia CIRP* 7 (2013), 545–550. https://doi.org/10.1016/j.procir.2013.06.030

[381] Tanner Schmidt and Dieter Fox. 2020. Self-directed lifelong learning for robot vision. In *Robotics Research*, Nancy M. Amato, Greg Hager, Shawna Thomas, and Miguel Torres-Torriti (Eds.). Springer International Publishing, Cham, Switzerland, 109–114.

[382] L. S. Scimmi, M. Melchiorre, S. Mauro, and S. Pastorelli. 2019. Experimental real-time setup for vision driven hand-over with a collaborative robot. In *Proceedings of the 2019 International Conference on Control, Automation, and Diagnosis (ICCAD'19)*. https://doi.org/10.1109/ICCAD46983.2019.9037961

[383] A. Shahroudy, J. Liu, T. Ng, and G. Wang. 2016. NTU RGB+D: A large scale dataset for 3D human activity analysis. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

[384] N. G. Shakev, S. A. Ahmed, A. V. Topalov, V. L. Popov, and K. B. Shiev. 2018. Autonomous flight control and precise gestural positioning of a small quadrotor. *Studies in Computational Intelligence* 756 (2018), 179–197. https://doi.org/10.1007/978-3-319-75181-8_9

[385] M. Shariatee, H. Khosravi, and E. Fazl-Ersi. 2017. Safe collaboration of humans and SCARA robots. In *Proceedings of the 4th RSI International Conference on Robotics and Mechatronics (ICRoM'16)*. 589–594. https://doi.org/10.1109/ICRoM.2016.7886809

[386] Mikiya Shibuya, Shinya Sumikura, and Ken Sakurada. 2020. Privacy preserving visual SLAM. In *Computer Vision—ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, Switzerland, 102–118.

[387] M.-Y. Shieh, C.-Y. Hsieh, and T.-M. Hsieh. 2014. Fuzzy visual detection for human-robot interaction. *Engineering Computations (Swansea, Wales)* 31, 8 (2014), 1709–1719. https://doi.org/10.1108/EC-11-2012-0292

[388] D. Shukla, O. Erkent, and J. Piater. 2015. Probabilistic detection of pointing directions for human-robot interaction. In *Proceedings of the 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA'15)*. 1–8. https://doi.org/10.1109/DICTA.2015.7371296

[389] D. Shukla, O. Erkent, and J. Piater. 2017. Proactive, incremental learning of gesture-action associations for human-robot collaboration. In *Proceedings of the 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'17)*. 346–353. https://doi.org/10.1109/ROMAN.2017.8172325

[390] Vinicius Silva, Filomena Soares, and João Sena Esteves. 2016. Mirroring emotion system—On-line synthesizing facial expressions on a robot face. In *Proceedings of the 2016 8th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT'16)*. 213–218. https://doi.org/10.1109/ICUMT.2016.7765359

[391] Nishikanto Sarkar Simul, Nusrat Mubin Ara, and Md. Saiful Islam. 2016. A support vector machine approach for real time vision based human robot interaction. In *Proceedings of the 2016 19th International Conference on Computer and Information Technology (ICCIT'16)*. 496–500. https://doi.org/10.1109/ICCITECHN.2016.7860248

[392] E. A. Sisbot, L. F. Marin-Urias, X. Broquère, D. Sidobre, and R. Alami. 2010. Synthesizing robot motions adapted to human presence: A planning and control framework for safe and socially acceptable robot motions. *International Journal of Social Robotics* 2, 3 (2010), 329–343. https://doi.org/10.1007/s12369-010-0059-6

[393] H. Song, W. Feng, N. Guan, X. Huang, and Z. Luo. 2017. Towards robust ego-centric hand gesture analysis for robot control. In *Proceedings of the 2016 IEEE International Conference on Signal and Image Processing (ICSIP'16)*. 661–666. https://doi.org/10.1109/SIPROCESS.2016.7888345

[394] M. Sorostinean and A. Tapus. 2018. Activity recognition based on RGB-D and thermal sensors for socially assistive robots. In *Proceedings of the 2018 15th International Conference on Control, Automation, Robotics, and Vision (ICARCV'18)*. 1298–1304. https://doi.org/10.1109/ICARCV.2018.8581349

[395] S. Sosnowski, C. Mayer, K. Kühnlenz, and B. Radig. 2010. Mirror my emotions! Combining facial expression analysis and synthesis on a robot. In *Proceedings of the 2nd International Symposium on New Frontiers in Human-Robot Interaction—A Symposium at the AISB 2010 Convention*. 108–112. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84863926598&partnerID=40&md5=109248e6985b1ff75829cc333a0d272e.

[396] A. Sripada, H. Asokan, A. Warrier, A. Kapoor, H. Gaur, R. Patel, and R. Sridhar. 2019. Teleoperation of a humanoid robot with motion imitation and legged locomotion. In *Proceedings of the 2018 3rd International Conference on Advanced Robotics and Mechatronics (ICARM'18)*. 375–379. https://doi.org/10.1109/ICARM.2018.8610719

[397] K. N. V. Sriram and S. Palaniswamy. 2019. Mobile robot assistance for disabled and senior citizens using hand gestures. In *Proceedings of the 1st International Conference on Power Electronics Applications and Technology in Present Energy Scenario (PETPES'19)*. https://doi.org/10.1109/PETPES47060.2019.9003821

[398] T. Stipancic, B. Jerbic, A. Bucevic, and P. Curkovic. 2012. Programming an industrial robot by demonstration. In *Proceedings of the 2012 23rd DAAAM International Symposium on Intelligent Manufacturing and Automation*, Vol. 1. 15–18. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84896948058&partnerID=40&md5=6e496dc214a79d3349c1c4326736d866.

[399] V. Suma. 2019. Computer vision for human-machine interaction-review. *Journal of Trends in Computer Science and Smart Technology* 1, 02 (2019), 131–139.

[400] Xiaowen Sun, Ran Zhao, Abdul Mateen Khattak, Kaite Shi, Yanzhao Ren, Wanlin Gao, and Minjuan Wang. 2019. Intelligent interactive robot system for agricultural knowledge popularity and achievements display. In *Proceedings of the 2019 IEEE 4th Advanced Information Technology, Electronic, and Automation Control Conference (IAEAC'19)*, Vol. 1. 511–518. https://doi.org/10.1109/IAEAC47372.2019.8997911

[401] Yongdian Sun, Xiangpeng Liang, Hua Fan, Muhammad Imran, and Hadi Heidari. 2019. Visual hand tracking on depth image using 2-D matched filter. In *Proceedings of 2019 UK/China Emerging Technologies (UCET'19)*. 1–4. https://doi.org/10.1109/UCET.2019.8881866

[402] Niko Sünderhauf, Oliver Brock, Walter Scheirer, Raia Hadsell, Dieter Fox, Jürgen Leitner, Ben Upcroft, et al. 2018. The limits and potentials of deep learning for robotics. *International Journal of Robotics Research* 37, 4–5 (2018), 405–420.

[403] Loreto Susperregi, Jose Maria Martínez-Otzeta, Ander Ansuategui, Aitor Ibarguren, and Basilio Sierra. 2013. RGB-D, laser and thermal sensor fusion for people following in a mobile robot. *International Journal of Advanced Robotic Systems* 10, 6 (2013), 271.

[404] Petr Svarny, Michael Tesar, Jan Kristof Behrens, and Matej Hoffmann. 2019. Safe physical HRI: Toward a unified treatment of speed and separation monitoring together with power and force limiting. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'19)*. 7580–7587. https://doi.org/10.1109/IROS40897.2019.8968463

[405] A. R. Taheri, M. Alemi, A. Meghdari, H. R. Pouretemad, and N. M. Basiri. 2014. Social robots as assistants for autism therapy in Iran: Research in progress. In *Proceedings of the 2014 2nd RSI/ISM International Conference on Robotics and Mechatronics (ICRoM'14)*. 760–766. https://doi.org/10.1109/ICRoM.2014.6990995

[406] Z. Talebpour, I. Navarro, and A. Martinoli. 2016. On-board human-aware navigation for indoor resource-constrained robots: A case-study with the ranger. In *Proceedings of the 2015 IEEE/SICE International Symposium on System Integration (SII'15)*. 63–68. https://doi.org/10.1109/SII.2015.7404955

[407] J. T. C. Tan, F. Duan, R. Kato, and T. Arai. 2010. Safety strategy for human-robot collaboration: Design and development in cellular manufacturing. *Advanced Robotics* 24, 5–6 (2010), 839–860. https://doi.org/10.1163/016918610X493633

[408] C. Tao and G. Liu. 2013. A multilayer hidden Markov models-based method for human-robot interaction. *Mathematical Problems in Engineering* 2013 (2013), 384865. https://doi.org/10.1155/2013/384865

[409] S. Tarbouriech and W. Suleiman. 2020. Bi-objective motion planning approach for safe motions: Application to a collaborative robot. *Journal of Intelligent and Robotic Systems: Theory and Applications* 99, 1 (2020), 45–63. https://doi.org/10.1007/s10846-019-01110-1

[410] Angelique Taylor, Darren M. Chan, and Laurel D. Riek. 2020. Robot-centric perception of human groups. *ACM Transactions on Human-Robot Interaction* 9, 3 (2020), 1–21.

[411] Angelique Taylor and Laurel D. Riek. 2016. Robot perception of human groups in the real world: State of the art. In *Proceedings of the 2016 AAAI Fall Symposium Series*.

[412] M. Terreran, E. Lamon, S. Michieletto, and E. Pagello. 2020. Low-cost scalable people tracking system for human-robot collaboration in industrial environment. *Procedia Manufacturing* 51 (2020), 116–124. https://doi.org/10.1016/j.promfg.2020.10.018

[413] Dante Tezza and Marvin Andujar. 2019. The state-of-the-art of human–drone interaction: A survey. *IEEE Access* 7 (2019), 167438–167454.

[414] Sam Thellman, Annika Silvervarg, and Tom Ziemke. 2020. Anthropocentric attribution bias in human prediction of robot behavior. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction (HRI'20)*. ACM, New York, NY, 476-478. https://doi.org/10.1145/3371382.3378347

[415] Leimin Tian and Sharon Oviatt. 2021. A taxonomy of social errors in human-robot interaction. *ACM Transactions on Human-Robot Interaction* 10, 2 (Feb. 2021), Article 13, 32 pages. https://doi.org/10.1145/3439720

[416] M. Tölgyessy, M. Dekan, F. Duchoň, J. Rodina, P. Hubinský, and L. Chovanec. 2017. Foundations of visual linear human-robot interaction via pointing gesture navigation. *International Journal of Social Robotics* 9, 4 (2017), 509–523. https://doi.org/10.1007/s12369-017-0408-9

[417] Michael Tornow, Ayoub Al-Hamadi, and Vinzenz Borrmann. 2013. A multi-agent mobile robot system with environment perception and HMI capabilities. In *Proceedings of the 2013 IEEE International Conference on Signal and Image Processing Applications*. IEEE, Los Alamitos, CA, 252–257.

[418] N. A. Torres, N. Clark, I. Ranatunga, and D. Popa. 2012. Implementation of interactive arm playback behaviors of social robot Zeno for autism spectrum disorder therapy. In *Proceedings of the 5th International Conference on Pervasive Technologies Related to Assistive Environments (PETRA'12)*. https://doi.org/10.1145/2413097.2413124

[419] Bill Triggs and Jakob Verbeek. 2007. Scene segmentation with CRFs learned from partially labeled images. In *Advances in Neural Information Processing Systems 20*.

[420] S.-H. Tseng, Y.-H. Hsu, Y.-S. Chiang, T.-Y. Wu, and L.-C. Fu. 2014. Multi-human spatial social pattern understanding for a multi-modal robot through nonverbal social signals. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'14)*. 531–536. https://doi.org/10.1109/ROMAN.2014.6926307

[421] A. Tsiami, P. Koutras, N. Efthymiou, P. P. Filntisis, G. Potamianos, and P. Maragos. 2018. Multi3: Multi-sensory perception system for multi-modal child interaction with multiple robots. In *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA'18)*. 4585–4592. https://doi.org/10.1109/ICRA.2018.8461210

[422] Satoshi Ueno, Sei Naito, and Tsuhan Chen. 2014. An efficient method for human pointing estimation for robot interaction. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP'14)*. IEEE, Los Alamitos, CA, 1545–1549.

[423] Alvaro Uribe, Silas Alves, João M. Rosário, Humberto Ferasoli Filho, and Byron Pérez-Gutiérrez. 2011. Mobile robotic teleoperation using gesture-based human interfaces. In *Proceedings of the 2011 IEEE IX Latin American Robotics Symposium and IEEE Colombian Conference on Automatic Control*. 1–6. https://doi.org/10.1109/LARC.2011.6086812

[424] Sepehr Valipour, Camilo Perez, and Martin Jagersand. 2017. Incremental learning for robot perception through HRI. In *Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'17)*. 2772–2777. https://doi.org/10.1109/IROS.2017.8206106

[425] A.F. Valle, G. Alenyà, G. Chance, P. Caleb-Solly, S. Dogramadzi, and C. Torras. 2019. Personalized robot assistant for support in dressing. *IEEE Transactions on Cognitive and Developmental Systems* 11, 3 (2019), 363–374. https://doi.org/10.1109/TCDS.2018.2817283

[426] M. Van Den Bergh, D. Carton, R. De Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, and M. Buss. 2011. Real-time 3D hand gesture interaction with a robot for understanding directions from humans. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*. 357–362. https://doi.org/10.1109/ROMAN.2011.6005195

[427] M. K. Van Den Broek and T. B. Moeslund. 2020. Ergonomic adaptation of robotic movements in human-robot collaboration. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*. 499–501. https://doi.org/10.1145/3371382.3378304

[428] P. A. A. Vasconcelos, H. N. S. Pereira, D. G. Macharet, and E. R. Nascimento. 2016. Socially acceptable robot navigation in the presence of humans. In *Proceedings of the 12th Latin American Robotics Symposium and the 3rd SBR Brazilian Robotics Symposium (LARS-SBR'15)*. 222–227. https://doi.org/10.1109/LARS-SBR.2015.14

[429] Juan P. Vasconez, George A. Kantor, and Fernando A. Auat Cheein. 2019. Human-robot interaction in agriculture: A survey and current challenges. *Biosystems Engineering* 179 (2019), 35–48. https://doi.org/10.1016/j.biosystemseng.2018.12.005

[430] A. Vasquez, M. Kollmitz, A. Eitel, and W. Burgard. 2017. Deep detection of people and their mobility aids for a hospital robot. In *Proceedings of the 2017 European Conference on Mobile Robots (ECMR'17)*. https://doi.org/10.1109/ECMR.2017.8098665

[431] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*.

[432] D. Vaufreydaz, W. Johal, and C. Combe. 2016. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems* 75 (2016), 4–16. https://doi.org/10.1016/j.robot.2015.01.004

[433] Marynel Vázquez, Aaron Steinfeld, and Scott E. Hudson. 2015. Parallel detection of conversational groups of free-standing people and tracking of their lower-body orientation. In *Proceedings of the 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'15)*. IEEE, Los Alamitos, CA, 3010–3017.

[434] Marynel Vázquez, Aaron Steinfeld, and Scott E. Hudson. 2016. Maintaining awareness of the focus of attention of a conversation: A robot-centric reinforcement learning approach. In *Proceedings of the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'16)*. IEEE, Los Alamitos, CA, 36–43.

[435] A. Vignolo, A. Sciutti, F. Rea, N. Noceti, F. Odone, and G. Sandini. 2017. Computational vision for social intelligence. In *Proceedings of the 2017 AAAI Spring Symposium Series*, Vols. SS-17-01 and SS-17-08. 647–651. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85028705748&partnerID=40&md5=9b703a5d5e1eac69dd73d2883ecaf39f.

[436] Valeria Villani, Fabio Pini, Francesco Leali, and Cristian Secchi. 2018. Survey on human-robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* 55 (2018), 248–266. https://doi.org/10.1016/j.mechatronics.2018.02.009

[437] Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'01)*, Vol. 1. IEEE, Los Alamitos, CA.

[438] Paul Viola and Michael J. Jones. 2004. Robust real-time face detection. *International Journal of Computer Vision* 57 (2004), 137–154.

[439] M. Virčíkova and P. Sinčák. 2015. Teach your robot how you want it to express emotions: On the personalized affective human-humanoid interaction. *Advances in Intelligent Systems and Computing* 316 (2015), 81–92. https://doi.org/10.1007/978-3-319-10783-7_9

[440] Emil-Ioan Voisan, Bogdan Paulis, Radu-Emil Precup, and Florin Dragan. 2015. ROS-based robot navigation and human interaction in indoor environment. In *Proceedings of the 2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*. 31–36. https://doi.org/10.1109/SACI.2015.7208244

[441] Tuan-Hung Vu, Anton Osokin, and Ivan Laptev. 2015. Context-aware CNNs for person head detection. In *Proceedings of the International Conference on Computer Vision (ICCV'15)*.

[442] A. Vysocký, R. Pastor, and P. Novák. 2019. Interaction with collaborative robot using 2D and TOF camera. In *Modelling and Simulation for Autonomous Systems*. Lecture Notes in Computer Science, Vol. 11472. Springer, 477–489. https://doi.org/10.1007/978-3-030-14984-0_35

[443] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. 2011. Action recognition by dense trajectories. In *Proceedings of the 2011 Conference on Computer Vision and Pattern Recognition (CVPR'11)*.

[444] L. Wang, B. Schmidt, and A. Y. C. Nee. 2013. Vision-guided active collision avoidance for human-robot collaborations. *Manufacturing Letters* 1, 1 (2013), 5–8. https://doi.org/10.1016/j.mfglet.2013.08.001

[445] Y. Wang, G. Song, G. Qiao, Y. Zhang, J. Zhang, and W. Wang. 2013. Wheeled robot control based on gesture recognition using the Kinect sensor. In *Proceedings of the 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO'13)*. 378–383. https://doi.org/10.1109/ROBIO.2013.6739488

[446] Y. Wang, X. Ye, Y. Yang, and W. Zhang. 2017. Collision-free trajectory planning in human-robot interaction through hand movement prediction from vision. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. 305–310. https://doi.org/10.1109/HUMANOIDS.2017.8246890

[447] T. B. Waskito, S. Sumaryo, and C. Setianingsih. 2020. Wheeled robot control with hand gesture based on image processing. In *Proceedings of the 2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT'20)*. 48–54. https://doi.org/10.1109/IAICT50021.2020.9172032

[448] T. Weber, S. Triputen, M. Danner, S. Braun, K. Schreve, and M. Rätsch. 2018. Follow me: Real-time in the wild person tracking application for autonomous robotics. In *RoboCup 2017: Robot World Cup XXI*. Lecture Notes in Computer Science, Vol. 11175. Springer, 156–167. https://doi.org/10.1007/978-3-030-00308-1_13

[449] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional pose machines. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. IEEE, Los Alamitos, CA, 4724–4732. https://doi.org/10.1109/CVPR.2016.511

[450] C. Weinrich, M. Volkhardt, and H.-M. Gross. 2013. Appearance-based 3D upper-body pose estimation and person re-identification on mobile robots. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC'13)*. 4384–4390. https://doi.org/10.1109/SMC.2013.748

[451] Astrid Weiss, Judith Igelsböck, Manfred Tscheligi, Andrea Bauer, Kolja Kühnlenz, Dirk Wollherr, and Martin Buss. 2010. Robots asking for directions: The willingness of passers-by to support robots. In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI'10)*. IEEE, Los Alamitos, CA, 23–30.

[452] F. Werner, D. Krainer, J. Oberzaucher, and K. Werner. 2013. Evaluation of the acceptance of a social assistive robot for physical training support together with older users and domain experts. *Assistive Technology Research Series* 33 (2013), 137–142. https://doi.org/10.3233/978-1-61499-304-9-137

[453] B.-F. Wu, C.-L. Jen, T.-Y. Tsou, W.-F. Li, and P.-Y. Tseng. 2012. Accompanist detection and following for wheelchair robots with fuzzy controller. In *Proceedings of the 2012 International Conference on Advanced Mechatronic Systems (ICAMechS'12)*. 638–643. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84869774498&partnerID=40&md5=cf2b7d1a68212e98c613c1440ef11df6.

[454] Y. Wu, Q. Yang, and X. Zhou. 2019. An improved method of optical flow using human body-following wheeled robot. *International Journal of Modeling, Simulation, and Scientific Computing* 10, 2 (2019), 1950003. https://doi.org/10.1142/S179396231950003X

[455] Z. Wu and S. Payandeh. 2020. Toward design of a drip-stand patient follower robot. *Journal of Robotics* 2020 (2020), 9080642. https://doi.org/10.1155/2020/9080642

[456] Z. Xia, Q. Lei, Y. Yang, H. Zhang, Y. He, W. Wang, and M. Huang. 2019. Vision-based hand gesture recognition for human-robot collaboration: A survey. In *Proceedings of the 2019 5th International Conference on Control, Automation, and Robotics (ICCAR'19)*. 198–205. https://doi.org/10.1109/ICCAR.2019.8813509

[457] Y. Xiao, Z. Zhang, A. Beck, J. Yuan, and D. Thalmann. 2014. Human-robot interaction by understanding upper body gestures. *Presence: Teleoperators and Virtual Environments* 23, 2 (2014), 133–154. https://doi.org/10.1162/PRES_a_00176

[458] D. Xu, X. Wu, Y.-L. Chen, and Y. Xu. 2014. Online dynamic gesture recognition for human robot interaction. *Journal of Intelligent and Robotic Systems: Theory and Applications* 77, 3-4 (2014), 583–596. https://doi.org/10.1007/s10846-014-0039-4

[459] J. Xu, J. Li, S. Zhang, C. Xie, and J. Dong. 2020. Skeleton guided conflict-free hand gesture recognition for robot control. In *Proceedings of the 2020 11th International Conference on Awareness Science and Technology (iCAST'20)*. 1–6. https://doi.org/10.1109/iCAST51195.2020.9319483

[460] Yuji Yamakawa, Yutaro Matsui, and Masatoshi Ishikawa. 2018. Human–robot collaborative manipulation using a high-speed robot hand and a high-speed camera. In *Proceedings of the 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS'18)*. IEEE, Los Alamitos, CA, 426–429.

[461] Takafumi Yamamoto, Yoji Yamada, Masaki Onishi, and Yoshihiro Nakabo. 2011. A 2D safety vision system for human-robot collaborative work environments based upon the safety preservation design policy. In *Proceedings of the 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO'11)*. IEEE, Los Alamitos, CA, 2049–2054.

[462] Haibin Yan, Marcelo H. Ang, and Aun Neow Poo. 2014. A survey on perception methods for human–robot interaction in social robots. *International Journal of Social Robotics* 6, 1 (2014), 85–119.

[463] Jihong Yan, Chao Chen, Zipeng Wang, Lizhong Zhao, and Dianguo Li. 2020. An optimization method for human-robot collaboration in a production unit in the context of intelligent manufacturing. In *Proceedings of the 2020 8th International Conference on Information Technology: IoT and Smart City (ICIT'20)*. ACM, New York, NY, 244–250. https://doi.org/10.1145/3446999.3447640

[464] Holly A. Yanco and Jill Drury. 2004. Classifying human-robot interaction: An updated taxonomy. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, Vol. 3. IEEE, Los Alamitos, CA, 2841–2846.

[465] C.-T. Yang, T. Zhang, L.-P. Chen, and L.-C. Fu. 2019. Socially-aware navigation of omnidirectional mobile robot with extended social force model in multi-human environment. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. 1963–1968. https://doi.org/10.1109/SMC.2019.8913844

[466] Dongfang Yang, Haolin Zhang, Ekim Yurtsever, Keith A. Redmill, and Ümit Özgüner. 2022. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention. *IEEE Transactions on Intelligent Vehicles* 7, 2 (2022), 221–230. https://doi.org/10.1109/TIV.2022.3162719

[467] N. Yang, F. Duan, Y. Wei, C. Liu, J. T. C. Tan, B. Xu, and J. Zhang. 2013. A study of the human-robot synchronous control system based on skeletal tracking technology. In *Proceedings of the 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO'13)*. 2191–2196. https://doi.org/10.1109/ROBIO.2013.6739794

[468] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*.

[469] Y. Yang, H. Yan, M. Dehghan, and M. H. Ang. 2015. Real-time human-robot interaction in complex environment using Kinect v2 image recognition. In *Proceedings of the 2015 7th IEEE International Conference on Cybernetics and Intelligent Systems (CIS'15) and Robotics, Automation, and Mechatronics (RAM'15)*. 112–117. https://doi.org/10.1109/ICCIS.2015.7274606

[470] N. Yao, E. Anaya, Q. Tao, S. Cho, H. Zheng, and F. Zhang. 2017. Monocular vision-based human following on miniature robotic blimp. In *Proceedings of the IEEE International Conference on Robotics and Automation*. 3244–3249. https://doi.org/10.1109/ICRA.2017.7989369

[471] B.-S. Yoo and J.-H. Kim. 2017. Gaze control of humanoid robot for learning from demonstration. *Advances in Intelligent Systems and Computing* 447 (2017), 263–270. https://doi.org/10.1007/978-3-319-31293-4_21

[472] K. Yoshida, F. Hibino, Y. Takahashi, and Y. Maeda. 2011. Evaluation of pointing navigation interface for mobile robot with spherical vision system. In *Proceedings of the IEEE International Conference on Fuzzy Systems*. 721–726. https://doi.org/10.1109/FUZZY.2011.6007673

[473] C. Yu and A. Tapus. 2019. Interactive robot learning for multimodal emotion recognition. In *Social Robotics*. Lecture Notes in Computer Science, Vol. 11876. Springer, 633–642. https://doi.org/10.1007/978-3-030-35888-4_59

[474] J. Yu and W. Paik. 2019. Efficiency and learnability comparison of the gesture-based and the mouse-based telerobotic systems. *Studies in Informatics and Control* 28, 2 (2019), 213–220. https://doi.org/10.24846/v28i2y201909

[475] W. Yuan and Z. Li. 2018. Development of a human-friendly robot for socially aware human-robot interaction. In *Proceedings of the 2017 2nd International Conference on Advanced Robotics and Mechatronics (ICARM'17)*. 76–81. https://doi.org/10.1109/ICARM.2017.8273138

[476] X. Yuan, S. Dai, and Y. Fang. 2020. A natural immersive closed-loop interaction method for human-robot "Rock-Paper-Scissors" game. In *Recent Trends in Intelligent Computing, Communication and Devices*. Advances in Intelligent Systems and Computing, Vol. 1006. Springer, 103–111. https://doi.org/10.1007/978-981-13-9406-5_14

[477] Yufeng Yue, Xiangyu Liu, Yuanzhe Wang, Jun Zhang, and Danwei Wang. 2020. Human-robot teaming and coordination in day and night environments. In *Proceedings of the 2020 16th International Conference on Control, Automation, Robotics, and Vision (ICARCV'20)*. 375–380. https://doi.org/10.1109/ICARCV50220.2020.9305408

[478] S. Yun, C. G. Kim, M. Kim, and M.-T. Choi. 2010. Robust robot's attention for human based on the multi-modal sensor and robot behavior. In *Proceedings of the IEEE Workshop on Advanced Robotics and Its Social Impacts (ARSO'10)*. 117–122. https://doi.org/10.1109/ARSO.2010.5680037

[479] W.-H. Yun, Y.-J. Cho, D. Kim, J. Lee, H. Yoon, and J. Kim. 2013. Robotic person-tracking with modified multiple instance learning. In *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*. 198–203. https://doi.org/10.1109/ROMAN.2013.6628445

[480] Angeliki Zacharaki, Ioannis Kostavelis, Antonios Gasteratos, and Ioannis Dokas. 2020. Safety bounds in human robot interaction: A survey. *Safety Science* 127 (2020), 104667. https://doi.org/10.1016/j.ssci.2020.104667

[481] Martina Zambelli and Yiannis Demiris. 2016. Multimodal imitation using self-learned sensorimotor representations. In *Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'16)*. 3953–3958. https://doi.org/10.1109/IROS.2016.7759582

[482] D. Zardykhan, P. Svarny, M. Hoffmann, E. Shahriari, and S. Haddadin. 2019. Collision preventing phase-progress control for velocity adaptation in human-robot collaboration. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*. 266–273. https://doi.org/10.1109/Humanoids43949.2019.9035065

[483] Ayberk Özgur, Stéphane Bonardi, Massimo Vespignani, Rico Möckel, and Auke J. Ijspeert. 2014. Natural user interface for Roombots. In *Proceedings of the 23rd IEEE International Symposium on Robot and Human Interactive Communication*. 12–17. https://doi.org/10.1109/ROMAN.2014.6926223

[484] Bo Zhang, Guanglong Du, Wenming Shen, and Fang Li. 2019. Gesture-based human-robot interface for dual-robot with hybrid sensors. *Industrial Robot* 46, 6 (Oct. 2019), 800–811. https://doi.org/10.1108/IR-11-2018-0245

[485] Hao Zhang, Christopher Reardon, and Lynne E. Parker. 2013. Real-time multiple human perception with color-depth cameras on a mobile robot. *IEEE Transactions on Cybernetics* 43, 5 (2013), 1429–1441.

[486] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. 2019. A comprehensive survey of vision-based human action recognition methods. *Sensors* 19, 5 (2019), 1005.

[487] J. Zhang, P. Li, T. Zhu, W.-A. Zhang, and S. Liu. 2020. Human motion capture based on Kinect and IMUs and its application to human-robot collaboration. In *Proceedings of the 2020 5th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM'20)*. 392–397. https://doi.org/10.1109/ICARM49381.2020.9195342

[488] K. Zhang and L. Zhang. 2018. Indoor omni-directional mobile robot that track independently. *Journal of Computers (Taiwan)* 29, 2 (2018), 118–135. https://doi.org/10.3966/199115992018042902013

[489] L. Zhang, K. Mistry, M. Jiang, S. Chin Neoh, and M. A. Hossain. 2015. Adaptive facial point detection and emotion recognition for a humanoid robot. *Computer Vision and Image Understanding* 140 (2015), 93–114. https://doi.org/10.1016/j.cviu.2015.07.007

[490] L. Zhang and R. Vaughan. 2016. Optimal robot selection by gaze direction in multi-human multi-robot interaction. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*. 5077–5083. https://doi.org/10.1109/IROS.2016.7759745

[491] L. Zhang and K. Zhang. 2018. An interactive control system for mobile robot based on cloud services. *Concurrency Computation* 30, 24 (2018). e4983. https://doi.org/10.1002/cpe.4983

[492] M. Zhang, X. Liu, D. Xu, Z. Cao, and J. Yu. 2019. Vision-based target-following guider for mobile robot. *IEEE Transactions on Industrial Electronics* 66, 12 (2019), 9360–9371. https://doi.org/10.1109/TIE.2019.2893829

[493] Z. Zhang, Z. Chen, and W. Li. 2018. Automating robotic furniture with a collaborative vision-based sensing scheme. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. 719–725. https://doi.org/10.1109/ROMAN.2018.8525783

[494] C. Zhao, W. Pan, and H. Hu. 2013. Interactive indoor environment mapping through visual tracking of human skeleton. *International Journal of Modelling, Identification and Control* 20, 4 (2013), 319–328. https://doi.org/10.1504/IJMIC.2013.057565

[495] Lijun Zhao, Xiaoyu Li, Peidong Liang, Chenguang Yang, and Ruifeng Li. 2016. Intuitive robot teaching by hand guided demonstration. In *Proceedings of the 2016 IEEE International Conference on Mechatronics and Automation*. IEEE, Los Alamitos, CA, 1578–1583.

[496] L. Zhao, Y. Liu, K. Wang, P. Liang, and R. Li. 2016. An intuitive human robot interface for tele-operation. In *Proceedings of the 2016 IEEE International Conference on Real-Time Computing and Robotics (RCAR'16)*. 454–459. https://doi.org/10.1109/RCAR.2016.7784072

[497] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV'15)*. IEEE, Los Alamitos, CA, 1116–1124. https://doi.org/10.1109/ICCV.2015.133

[498] M.-D. Zhu, L.-X. Xia, and J.-B. Su. 2016. Real-time imitation framework for humanoid robots based on posture classification. In *Proceedings of the International Conference on Machine Learning and Cybernetics*, Vol. 2. 489–494. https://doi.org/10.1109/ICMLC.2016.7872936

[499] T. Zhu, Q. Zhao, W. Wan, and Z. Xia. 2017. Robust regression-based motion perception for online imitation on humanoid robot. *International Journal of Social Robotics* 9, 5 (2017), 705–725. https://doi.org/10.1007/s12369-017-0416-9

Received 28 September 2021; revised 18 May 2022; accepted 22 September 2022