# Recognition of Grasping Patterns for Human–Robot Collaboration

Subjects: Computer Science, Artificial Intelligence | Robotics

Contributor: Pedro Amaral , Filipe Silva , Vítor Santos

Recent advances in the field of collaborative robotics aim to endow industrial robots with prediction and anticipation abilities. In many shared tasks, the robot's ability to accurately perceive and recognize the objects being manipulated by the human operator is crucial to make predictions about the operator's intentions.

collaborative robotics | object recognition | hand–object interaction | grasping posture

## 1. Introduction

Human–robot collaboration (HRC) is a research topic becoming increasingly important in modern industry, driven by the need to enhance productivity, efficiency and safety in work environments [1][2][3][4][5][6]. The combination of human skills and robotic capabilities provides significant potential to improve the execution of complex and repetitive tasks. However, the effective synchronization of actions and seamless communication between partners are open challenges that need to be further addressed [7][8][9]. In recent years, there has been a remarkable trend toward endowing collaborative robots with cognitive abilities, transforming them from simple automated machines into intelligent and adaptable collaborators. This shift is driven by the increasing demand for robots that can work alongside humans, understand their intentions and actively contribute to complex tasks in dynamic environments. Collaborative cognition encompasses a range of essential abilities in order to enable robots to learn, predict and anticipate human actions [6][10][11].

In collaborative scenarios, assistive robots are designed to work alongside humans in assembly processes or maintenance operations, providing timely support in order to enhance the overall efficiency of the task. Robots can assist the human worker by delivering a component, tool or part, by holding a part while the operator works on it or by autonomously performing a specific sub-task. In any case, the ability of an assistive robot to anticipate the upcoming needs of a human operator plays a pivotal role in supporting efficient teamwork. By anticipating human intentions, actions and needs, robots can proactively assist or complement human tasks, providing timely support and improving overall efficiency [12][13][14][15].

## 2. Object Sensing

Approaches within the "object sensing" category leverage visual information extracted from images or videos of the objects the user is interacting with, by using techniques from computer vision and machine learning to discern object identities based on their visual attributes. A common approach involves the extraction of visual features that

can encompass color histograms, texture descriptors, contour shapes and local keypoints. Early works in this domain [16][17][18] applied traditional image processing techniques to extract features such as shape moments and color histograms, leading to initial success in recognizing simple objects. The surge of progress seen in recent years is largely due to the latest developments in deep learning [19], particularly convolutional neural networks (CNNs) and geometric reasoning [20].

Deep learning has had an enormous impact in perception tasks with the design of effective architectures for real-time object recognition, providing significant advancements in accuracy and robustness. CNNs have demonstrated remarkable performance in extracting hierarchical features from images [21]. Transfer learning, where pre-trained models are fine-tuned for specific tasks, has enabled efficient object recognition even with limited training data [22]. A relevant vision-based approach is the one in which the process of recognizing the human-grasped object, across consecutive frames, comprises two sub-processes: hand tracking and object recognition. The hand detection and tracking system is commonly used for defining a bounding box around the grasped object that describes its spatial location. This initial step can, in turn, simplify the object recognition algorithm as it can focus attention solely on the region where the object is likely to be present. This reduces the search space and the required computational resources. Object detection frameworks like YOLO (You Only Look Once) and Faster R-CNN fall under this category. They divide the RGB image into a grid and predict bounding boxes and class probabilities directly from the grid.

In parallel to deep learning, the recent availability of inexpensive RGB-D sensors has enabled significant improvements in scene modeling and human pose estimation. Some studies explore the fusion of multiple modalities to enhance object recognition. These approaches combine visual information with other sensory data, such as depth information from 3D sensors [23][24]. This integration of modalities has shown promise in improving recognition accuracy, especially in scenarios with varying lighting conditions or occlusions. Researchers have also studied how to leverage information from multiple viewpoints (i.e., multi-view 3D object recognition) to enhance recognition accuracy [25]. This approach is particularly relevant for 3D objects, where recognizing an object's 3D structure from different viewpoints can aid in robust recognition. Techniques like using 3D point clouds, multi-view CNNs or methods that combine RGB images and depth information fall under this category.

Despite their successes, methods within the "Object Sensing" category are often constrained by the variability in object appearances, limited viewpoint coverage and sensitivity to illumination changes. As a result, the focus on object characteristics alone may not provide a complete solution, particularly in situations where the human hand's interaction with the object plays a crucial role.

## 3. Hand Sensing

Recognizing objects based on the interactions of the human hand is a complex problem due to the intricate nature of hand–object interactions (HOIs) and the variability in grasp patterns and gestures [26][27][28][29]. Achieving accurate and real-time recognition involves understanding the relationships and dynamics between a human hand and the objects it interacts with (e.g., the interaction context, the person's actions and the patterns that emerge

over time), as well as the the tactile and kinesthetic feedback generated during manipulation. Additionally, variations in grasp styles, object sizes and orientation further worsen the complexity of the task. Several works propose interaction reasoning networks for modeling spatio-temporal relationships between hands and objects in egocentric video during activities of the daily life, such as playing an instrument, kicking a ball, opening a drawer (one-handed interaction), opening a bottle (two-handed interaction) or cutting a vegetable with a knife. Main advances are due to the development of several human-centric datasets (e.g., V-COCO [30], HICO-DET [27] and HCVRD [31]) that annotate the bounding boxes of each human actor, the object with which he/she is interacting and the corresponding interaction. However, the creation of large-scale, diverse and annotated datasets remains an ongoing effort.

Some works consider the hand–object interaction (HOI) as a manifestation of human intention or purpose of action [32][33][34][35][36]. Despite the growing need for detection and inference of HOIs in practical applications, such as collaborative robotics, the problem of recognizing objects based on hand–object interactions is inherently complex. Instead of addressing the full complexity of HOI recognition, several works have adopted targeted approaches that address specific aspects of the problem without necessarily delving into the entire spectrum of interactions. A recent work investigated the influence of physical properties of objects such as shape, size and weight on forearm electromyiography (EMG) signals and the opportunities that this sensing technology brings in hand–object interaction recognition and/or for object-based activity tracking [37]. Despite the relevance of the work, it is difficult to be applied in collaborative assembly scenarios given the complexity of the required setup that requires sensor attachment, calibration and training. Some other limitations may include user-dependent variability, muscle fatigue and discomfort and/or interference from other electrical devices.

Another line of research focuses on tracking the positions of hand and finger landmarks during interactions. By monitoring the spatial relationships of these landmarks, these methods aim to deduce the object's identity based on the specific manipulations applied. This approach captures critical information about the hand's interaction without necessarily modeling the full complexity of interactions. A glove-based interaction approach has been proposed by Paulson et al. [38] in the HCI domain to investigate a grasp-based selection of objects in office settings. The authors showed that hand posture information alone can be used to recognize various activities in an office, such dialing a number, holding a mug, typing at the keyboard or handling the mouse. The classification of hand posture is performed using the nearest-neighbor algorithm. In a similar work based on a data glove, Vatavu et al. [39] proposed the automatic recognition of the size and shape of objects using the posture of the hand during prehension. The objects used in the experiments consisted of six basic shapes (cube, parallelepiped, cylinder, sphere, pyramid and a thin plate) and, for each shape, three different sizes (small, medium and large). Twelve right-handed participants took part in the experiments using a 5DT Data Glove Ultra equipped with 14 optical sensors. These sensors were distributed as follows: 10 sensors measure finger flexion (two sensors per finger) and four sensors measure abduction between fingers.

The study compared several classifiers derived from the nearest-neighbour approach with a multi-layer perceptron (MLP) and a multi-class support vector machine (SVM). The best results were achieved with the *K*-nearest-neighbor classification approach when combining the results of individual postures across an entire time window of

half a second. The experiments carried out included the capture of hand postures when grasping and maintaining a stable grip for a reliable translation of the objects. The results show that object size and shape can be recognized with up to 98% accuracy when using user-specific metrics. The authors also pointed out the lower accuracy for user-independent training and the variability in the individual grasping postures during object exploration. Although in general the proposed approach recognizes the physical properties of the grasped objects with high accuracy, wearing a glove directly on the hand is intrusive and troublesome, interfering with the natural movement of the fingers.

When attempting to model human grasping, researchers have focused their attention on defining a comprehensive taxonomy of human grasp types [40] and the multifaceted factors that influence the choice of grasping, including user intentions [41], object properties [42] and environmental constraints [43]. Mackenzie and Iberall [41] theorize the existence of a cognitive model that converts the object's geometry properties and user's intent into a motor program driving the hand and finger motions. From this seminal work, several studies on human reach-to-grasp actions have consistently shown that the natural kinematics of prehension allows for predicting the object he/she is going to grasp, as well as the subsequent actions that will be carried out with that object. Feix et al. [42] provided an analysis of human grasping behaviors showing the correlation between the properties of the objects and the grasp choice. More recently, the works of Betti et al. [44] and Egmose and Koppe [45] focus on the reach-to-grasp phase. Their finding shows that grasp formation is highly correlated with the size and shape of the object to be grasped, as well as strongly related to the intended action. These insights promise improved interaction by exploring the ability with in which the robot can predict the object the user intends to grasp or to recognize the one he/she is already holding, provided that the hand kinematics information is extracted and processed in real time.

In line with this, Valkov et al. [46] investigated the feasibility and accuracy of recognizing objects based on hand kinematics and long short-term memory (LSTM) networks. The data are extracted from a Polhemus Viper16 electromagnetic tracking system with 12 sensors attached to the hand and fingers. On the one hand, the study focuses on the size discrimination of nine synthetic objects: three regular solids (sphere, box and cylinder) in three different sizes (small—2 cm, medium—4 cm and large—6 cm). On the other hand, a different set of seven objects (pen, glue, bottle, Rubik's cube, volcano-egg, toy and scissor) was used for object discrimination. The data recorded during the experiments include a phase in which participants were asked to reach and grasp the object starting from a fixed initial position. The results demonstrated that LSTM networks can predict the time point at which the user grasps an object with 23 ms precision and the current distance to it with a precision better than 1 cm. Furthermore, the size and the object discrimination during the reach-to-grasp actions were achieved successfully with an accuracy above 90% using *K*-fold cross-validation. Although the results are still preliminary, the leave-one-out cross-validation showed a significant degradation in the performance of the models compared to the *K*-fold validation. While the tracking system offers many advantages, there are also practical limitations such as sensor attachment and comfort, line-of-sight requirements, interference and noise as well as calibration and drift.

## References

1. Robla-Gómez, S.; Becerra, V.M.; Llata, J.R.; González-Sarabia, E.; Torre-Ferrero, C.; Pérez-Oria, J. Working Together: A Review on Safe Human-Robot Collaboration in Industrial Environments. IEEE Access 2017, 5, 26754–26773.
2. Villani, V.; Pini, F.; Leali, F.; Secchi, C. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. Mechatronics 2018, 55, 248–266.
3. Ajoudani, A.; Zanchettin, A.M.; Ivaldi, S.; Albu-Schäffer, A.; Kosuge, K.; Khatib, O. Progress and Prospects of the Human-Robot Collaboration. Auton. Robot. 2018, 42, 957–975.
4. Matheson, E.; Minto, R.; Zampieri, E.G.G.; Faccio, M.; Rosati, G. Human-Robot Collaboration in Manufacturing Applications: A Review. Robotics 2019, 8, 100.
5. Kumar, S.; Savur, C.; Sahin, F. Survey of Human-Robot Collaboration in Industrial Settings: Awareness, Intelligence, and Compliance. IEEE Trans. Syst. Man Cybern. Syst. 2021, 51, 280–297.
6. Castro, A.; Silva, F.; Santos, V. Trends of human-robot collaboration in industry contexts: Handover, learning, and metrics. Sensors 2021, 21, 4113.
7. Michalos, G.; Kousi, N.; Karagiannis, P.; Gkournelos, C.; Dimoulas, K.; Koukas, S.; Mparis, K.; Papavasileiou, A.; Makris, S. Seamless human robot collaborative assembly—An automotive case study. Mechatronics 2018, 55, 194–211.
8. Papanastasiou, S.; Kousi, N.; Karagiannis, P.; Gkournelos, C.; Papavasileiou, A.; Dimoulas, K.; Baris, K.; Koukas, S.; Michalos, G.; Makris, S. Towards seamless human robot collaboration: Integrating multimodal interaction. Int. J. Adv. Manuf. Technol. 2019, 105, 3881–3897.
9. Hoffman, G. Evaluating Fluency in Human–Robot Collaboration. IEEE Trans. Hum.-Mach. Syst. 2019, 49, 209–218.
10. Rozo, L.; Ben Amor, H.; Calinon, S.; Dragan, A.; Lee, D. Special issue on learning for human–robot collaboration. Auton. Robot. 2018, 42, 953–956.
11. Jiao, J.; Zhou, F.; Gebraeel, N.Z.; Duffy, V. Towards augmenting cyber-physical-human collaborative cognition for human-automation interaction in complex manufacturing and operational environments. Int. J. Prod. Res. 2020, 58, 5089–5111.
12. Hoffman, G.; Breazeal, C. Effects of anticipatory action on human-robot teamwork efficiency, fluency, and perception of team. In Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction, Arlington, VA, USA, 10–12 March 2007; pp. 1–8.
13. Williams, A.M. Perceiving the intentions of others: How do skilled performers make anticipation judgments? Prog. Brain Res. 2009, 174, 73–83.
14. Huang, C.M.; Mutlu, B. Anticipatory robot control for efficient human-robot collaboration. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction

(HRI), Christchurch, New Zealand, 7–10 March 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 83–90.

15. Duarte, N.F.; Raković, M.; Tasevski, J.; Coco, M.I.; Billard, A.; Santos-Victor, J. Action anticipation: Reading the intentions of humans and robots. IEEE Robot. Autom. Lett. 2018, 3, 4132–4139.
16. Taubin, G.; Cooper, D.B. Object Recognition Based on Moment (or Algebraic) Invariants. In Geometric Invariance in Computer Vision; MIT Press: Cambridge, MA, USA, 1992; pp. 375–397.
17. Mindru, F.; Moons, T.; Van Gool, L. Color-Based Moment Invariants for Viewpoint and Illumination Independent Recognition of Planar Color Patterns. In Proceedings of the International Conference on Advances in Pattern Recognition, Plymouth, UK, 23–25 November 1998; Singh, S., Ed.; Springer: London, UK, 1999; pp. 113–122.
18. Sarfraz, M. Object Recognition Using Moments: Some Experiments and Observations. In Proceedings of the Geometric Modeling and Imaging–New Trends (GMAI'06), London, UK, 5–7 July 2006; pp. 189–194.
19. Wu, X.; Sahoo, D.; Hoi, S.C. Recent advances in deep learning for object detection. Neurocomputing 2020, 396, 39–64.
20. Barabanau, I.; Artemov, A.; Burnaev, E.; Murashkin, V. Monocular 3D Object Detection via Geometric Reasoning on Keypoints. In Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020)—Volume 5: VISAPP. INSTICC, Valletta, Malta, 27–29 February 2020; SciTePress: Setúbal, Portugal, 2020; pp. 652–659.
21. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 1137–1149.
22. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. Proc. IEEE 2021, 109, 43–76.
23. Zimmermann, C.; Welschehold, T.; Dornhege, C.; Burgard, W.; Brox, T. 3D Human Pose Estimation in RGBD Images for Robotic Task Learning. In Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, Australia, 21–25 May 2018; IEEE Press: Piscataway, NJ, USA, 2018; pp. 1986–1992.
24. Rato, D.; Oliveira, M.; Santos, V.; Gomes, M.; Sappa, A. A sensor-to-pattern calibration framework for multi-modal industrial collaborative cells. J. Manuf. Syst. 2022, 64, 497–507.
25. Qi, S.; Ning, X.; Yang, G.; Zhang, L.; Long, P.; Cai, W.; Li, W. Review of multi-view 3D object recognition methods based on deep learning. Displays 2021, 69, 102053.
26. Gkioxari, G.; Girshick, R.; Dollár, P.; He, K. Detecting and recognizing human-object interactions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake

City, UT, USA, 18–22 June 2018; pp. 8359–8367.

27. Chao, Y.W.; Liu, Y.; Liu, X.; Zeng, H.; Deng, J. Learning to detect human-object interactions. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 381–389.
28. Cao, Z.; Radosavovic, I.; Kanazawa, A.; Malik, J. Reconstructing hand-object interactions in the wild. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 12417–12426.
29. Liu, S.; Jiang, H.; Xu, J.; Liu, S.; Wang, X. Semi-supervised 3d hand-object poses estimation with interactions in time. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14687–14697.
30. Gupta, S.; Malik, J. Visual semantic role labeling. arXiv 2015, arXiv:1505.04474.
31. Zhuang, B.; Wu, Q.; Shen, C.; Reid, I.; van den Hengel, A. HCVRD: A benchmark for large-scale human-centered visual relationship detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
32. Koppula, H.S.; Saxena, A. Anticipating human activities using object affordances for reactive robotic response. IEEE Trans. Pattern Anal. Mach. Intell. 2015, 38, 14–29.
33. Hayes, B.; Shah, J.A. Interpretable models for fast activity recognition and anomaly explanation during collaborative robotics tasks. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 6586–6593.
34. Furnari, A.; Farinella, G.M. What would you expect? Anticipating egocentric actions with rolling-unrolling lstms and modality attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6252–6261.
35. Xu, B.; Li, J.; Wong, Y.; Zhao, Q.; Kankanhalli, M.S. Interact as you intend: Intention-driven human-object interaction detection. IEEE Trans. Multimed. 2019, 22, 1423–1432.
36. Roy, D.; Fernando, B. Action anticipation using pairwise human-object interactions and transformers. IEEE Trans. Image Process. 2021, 30, 8116–8129.
37. Fan, J.; Fan, X.; Tian, F.; Li, Y.; Liu, Z.; Sun, W.; Wang, H. What is that in your hand? Recognizing grasped objects via forearm electromyography sensing. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies; Association for Computing Machinery: New York, NY, USA, 2018; Volume 2, pp. 1–24.
38. Paulson, B.; Cummings, D.; Hammond, T. Object interaction detection using hand posture cues in an office setting. Int. J. Hum.-Comput. Stud. 2011, 69, 19–29.

39. Vatavu, R.D.; Zaiţi, I.A. Automatic recognition of object size and shape via user-dependent measurements of the grasping hand. Int. J. Hum.-Comput. Stud. 2013, 71, 590–607.
40. Feix, T.; Romero, J.; Schmiedmayer, H.B.; Dollar, A.M.; Kragic, D. The grasp taxonomy of human grasp types. IEEE Trans. Hum.-Mach. Syst. 2015, 46, 66–77.
41. MacKenzie, C.L.; Iberall, T. The Grasping Hand; Elsevier: Amsterdam, The Netherlands, 1994.
42. Feix, T.; Bullock, I.M.; Dollar, A.M. Analysis of human grasping behavior: Object characteristics and grasp type. IEEE Trans. Haptics 2014, 7, 311–323.
43. Puhlmann, S.; Heinemann, F.; Brock, O.; Maertens, M. A compact representation of human single-object grasping. In Proceedings of the 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Republic of Korea, 9–14 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1954–1959.
44. Betti, S.; Zani, G.; Guerra, S.; Castiello, U.; Sartori, L. Reach-to-grasp movements: A multimodal techniques study. Front. Psychol. 2018, 9, 990.
45. Egmose, I.; Køppe, S. Shaping of reach-to-grasp kinematics by intentions: A meta-analysis. J. Mot. Behav. 2018, 50, 155–165.
46. Valkov, D.; Kockwelp, P.; Daiber, F.; Krüger, A. Reach Prediction using Finger Motion Dynamics. In Proceedings of the Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–8.