Localization Based Grasping for Robotic Grippers

Shangquan Li *

Ronald Reagan Secondary School, Irvine, 92620, USA

* Corresponding Author Email: lisqu@ldy.edu.rs

Abstract. To meet the demands of grasping in complex environments, robotic grippers are moving from traditional open-loop control to perception-based closed-loop localization for grasping. Open-loop control lacks real-time feedback and struggles with the uncertainty of unstructured settings. In contrast, closed-loop grasping with multimodal sensing such as vision and touch adapts the strategy online and increases success rates. This paper reviews three categories of grippers and their development, namely industrial, medical, and soft. Industrial grippers provide high stiffness, high payload capacity, and high precision. They meet heavy-load and high-accuracy requirements in production and use vision and control algorithms to raise grasp success. Medical grippers emphasize compliance and micro-manipulation so that surgery meets strict requirements on fine control and safety. They combine force sensing and vision to ensure safe and stable operation. Soft grippers use compliant structures and under-actuated designs to perform enveloping grasps on irregular objects that improves tolerance to error and adaption to shape. Finally, the paper summarizes the value and challenges of fusing vision, tactile sensing, and force sensing with localization strategies in closed-loop grasp control.

Keywords: Robotic grippers, grasping methods, sensors, localization.

1. Introduction

In recent years, with the continued advancement of Industry 4.0 and intelligent manufacturing as well as the rapid development of service robotics [1], the robotic gripper is a crucial end-effector that enables physical interaction between robots and the working environment, and its importance has been steadily growing. Initially, robotic arms were developed as a means of supplementing or replacing humans doing laborious, dirty, or dangerous tasks. However, simple grasping is no longer adequate for applications in complex, unstructured environments, such as the precision components assembly, automated fruits and vegetables harvesting, and tissue manipulation in laparoscopic surgery [2]. These tasks require the gripper to perceive the environment and adapt its grasping strategy accordingly. Traditional industrial grasping typically uses rigid parallel-jaw grippers or suction, along with open-loop [3] force and form-closure planning, which is analytically designed for known geometries within fixtures and structured cells. These strategies are fast and repeatable but depend on accurate models and tight calibration. As a result, they degrade under pose or model errors and struggle with transparent, reflective, deformable, or cluttered objects. Additionally, these approaches often require frequent re-teaching, further limiting their flexibility and robustness.

Currently, with the integration of multiple sensors such as vision, force, touch, and proximity, robotic grippers are transitioning from blind grasping to perception-driven adaptive grasping [4]. Grippers for different domains (industrial, medical, and soft) integrate different sensor combinations and grasping strategies according to application needs, leading to distinct design approaches. For example, warehouses use "depth vision and vacuum pressure sensing" with a retry policy, favoring multi-chamber suction cups [5], and medical information systems use "endoscopic vision and fiber-optic force sensing" for safety control, favoring miniaturized, compliant tips [6]. However, a systematic framework to organize and compare these technologies is still lacking. This paper aims to provide a systematic review of grasping tasks for robotic manipulators, focusing on industrial, medical, and soft robotics, as well as the novel sensing and positioning methods used in applications. We discuss the different structures and characteristics of robotic grippers, categorized according to their application areas. Then, we explore the methods for localization and grasping control, with a focus on perception-driven strategies across diverse domains. Furthermore, we outline directions for

future research, followed by a discussion of the challenges and evaluation criteria. We also identify open problems and trends to guide future research and development in the field.

2. Types of Robotics Grippers and Typical Grasping Modes

With the rapid development of industrial automation, robotic grippers are widely used in precision manufacturing and assembly-line operations. They serve as the end-effector of industrial manipulators. To meet the demands for efficient and precise operation, localization for grasping is essential to robotic grippers. The core challenge is to integrate high-precision localization systems with efficient grasp control algorithms in complex and dynamic environments. Accurate target grasping requires a robotic gripper that provides high stiffness and precise repeatability. It also requires a strong load capacity to adapt to varied materials and operating conditions. This chapter classifies robotic grippers by application and capability into three types: industrial, medical, and soft, and reviews the research related to the structures and characteristics of each type.

2.1. Industrial Grippers for High Precision and Heavy Payload Grasping

Industrial grippers are mainly composed of rigid links and joints, providing strong load capacity and environmental adaptability. In recent years, research on industrial robotic grippers has made significant progress in positioning and grasping accuracy. The introduction of new algorithms and intelligent control methods enables grippers to handle higher-mass objects and achieve higher grasping accuracy, providing solid technical support for efficient industrial operations.

For example, the Berkeley AutoLab proposed Dex-Net 3.0. The method combines a compliant suction contact model with large-scale synthetic point cloud data to train a GQ-CNN. On an ABB YuMi, the success rate for novel objects reached 98% on the basic set. On the typical and adversarial sets, the success rates reached 82% and 58% respectively. After targeted training on the adversarial set, the success rate reached 81%. The approach improves localization robustness and the success rate of suction grasping [7]. In addition, the NVIDIA Research team proposed DOPE. It achieves real-time estimation of the 6-DoF pose of objects using a monocular camera and training on purely synthetic data. This method provides precise pick and place control for robotic grippers. Its real-time output is sufficient for grasping, placement, and alignment in real-world settings. Experimental results show that the gripper's placement error is at the centimeter scale. The method effectively improves the localization and grasping accuracy of a known rigid robotic gripper [8].

Furthermore, the QUT team developed GG-CNN. It improves closed-loop grasping control for robotic grippers. It operates at 50 Hz. In dynamic environments and in the presence of control errors, the success rate remains between 81% and 88%. The system outperforms traditional open-loop control methods. This result markedly improves the adaptability of robotic grippers in complex and unstable environments. It provides a more stable and reliable solution for real-time localization and grasping with parallel-jaw grippers. With this optimization, the gripper maintains high grasp accuracy and success rates under rapidly changing operating conditions. This progress further strengthens the application potential of industrial automation [9].

In summary, recent research has shown significant advancements in the accuracy, robustness, and adaptability of industrial robotic grippers' localization and grasping. Besides, progress in algorithms, intelligent control, and large-scale data training continues to enhance the grasping precision and payload capacity of robotic grippers. For dynamic environments and high-load conditions, these methods enhance the robustness of localization, allowing the robotic gripper to maintain system stability under constantly changing operating conditions. In the future, continued algorithm improvements and better hardware will expand the use of localization for grasping. The technology will support efficient industrial automation.

2.2. Medical Robotics Grippers for Safe Grasping and Precise Manipulation.

Medical robotic grippers are used mainly for surgical assistance and rehabilitation. They are also used for microscale manipulation. It usually consists of precision mechanical components and a human-machine interface control system. It requires a high level of safety, flexibility, and dexterous manipulation.

In medical applications, the gripper's localization and grasping capability are central. The workflow of localization-based grasping can be refined into a three-level closed-loop control architecture. The perception layer employs microscopy or laparoscopic vision, using deformation tracking, boundary tracking, and tool-tip marking to output a stable relative pose between the target and end-effector. When needed, it incorporates depth information from optical coherence tomography or ultrasound. The planning layer primarily operates under constraints such as anatomical boundaries, sterile fields, and instrument work cones to generate approach and withdrawal paths. The execution layer typically uses impedance control, hybrid force, or position control to maintain the desired contact force and envelope.

To ensure precise operation, the system often uses closed-loop control so that it maintains the desired contact geometry and mechanical boundaries. The common end-effector designs include micro-forceps, needle drivers, clip appliers, retractors, and micro grippers with compliant fingertip pads. In practice, the STAR autonomous suturing platform from Johns Hopkins University and the Children's National Medical Center has achieved autonomous intestinal anastomosis for soft tissue. Its performance matched or exceeded manual surgery and standard robot-assisted approaches on multiple metrics of suture quality and consistency. The results show precise registration and stable operation in deformable tissue environments [10, 11]. Besides, the JHU ophthalmic robotics team integrated a micro-force sensing system into the Steady-Hand Eye Robot. It achieves sub-Newton force sensing and supports force-guided cooperative control. The system uses a precise force feedback mechanism that enhances the robot's perception of delicate tissue in minimally invasive procedures. Significantly reducing the risk of tissue damage in microsurgery. Compared with conventional microsurgical methods, this technology improves surgical precision while it reduces intraoperative uncertainty and is most effective when precise manipulation of small tissues is required during the procedure [12].

Furthermore, the Micron active tremor suppression tool from CMU combines high bandwidth optical tracking with piezoelectric micro-actuation to suppress hand tremor in ophthalmic microsurgery. This tool tracks hand motion in real time and controls the actuator response so that hand tremor is reduced and positional error decreases by 32–52%. It achieves high accuracy and stability in submillimeter ophthalmic microsurgery, where surgeons require flexible control, providing a stable and reliable surgical platform [13]. These examples show that localization-based grasping for medical robotic manipulators has achieved significant advances in accuracy and stability while delivering application breakthroughs across multiple clinical domains.

Overall, medical manipulators are extending the paradigm of localization, constraint, and grasping to more soft-tissue and microsurgical settings, where perception evolves from single-view vision to multimodal fusion while control shifts from open-loop teaching to closed-loop control that combines tactile and vision, and where the system moves from a stand-alone device toward deep integration with intraoperative navigation, hospital information systems, and quality assurance. Therefore, grasping and localization represent a comprehensive process engineering challenge, requiring precise alignment and tissue protection capabilities, while ensuring repeatability, auditability, and portability.

2.3. Adaptive and Bioinspired Grasping with Soft Grippers

To provide a unified description of soft gripper design and use under the localization-based grasping paradigm, we first explain the operating principle at the configuration level and the system level. The soft material forms a contact envelope through passive deformation during contact. The end effector achieves compliant contact under low-pressure actuation.

Subsequently, tactile and visual feedback correct the relative pose between the end effector and the target in real-time, ensuring that the grasping process, from approach to closure, remains within safe limits. The end effector of a soft gripper is often a monolithic compliant fingertip or a membrane chamber. The core components usually use elastomers and fiber-reinforced materials. In terms of the driving method, low-pressure pneumatic or tendon drive is commonly used, and it is coordinated with the underactuated mechanism. At the same time, variable stiffness is superimposed, and visual and tactile feedback are combined to achieve closed-loop control. In localization-based grasping with soft grippers, the workflow usually follows estimation, alignment, envelopment, and holding. First, external vision and fingertip tactile sensing estimate the relative pose between the target and the fingertip. Next, the system plans a safe approach path under anatomical, procedural, or environmental constraints. Then it forms a stable contact geometry through enveloping contact and maintains the desired contact force and pose with closed-loop control. This workflow enables soft grippers to manage complex grasping tasks, ensuring high efficiency and accuracy even in dynamic environments.

The system must tolerate pose error and keep contact stress low so that fragile objects are protected, while it adapts to irregular or unknown targets and maintains stable enveloping grasping under weak sensing and unstructured environments [14, 15]. To achieve these capabilities, implementations often use low-dimensional control in conjunction with underactuated mechanisms so that each joint allocates its travel adaptively according to the order of contact. Brown et al. proposed a localization and grasping method that uses granular jamming to realize variable stiffness. The method executes a sequence of approaches, including gentle touch, enclosure, stiffness increases by negative pressure, hold, and fine adjustment. During the grasp, the gripper incrementally adjusts stiffness so that it can accommodate different object shapes and material properties. Unlike traditional methods, this approach does not rely on precise geometric models or strict calibration, exhibiting high robustness and adaptability.

When handling irregular or unknown targets, the method achieves robust grasping and markedly improves system fault tolerance [16]. The variable stiffness module switches between compliant conformity and stable holding. Visuospatial sensing provides contact surface geometry and incipient slip cues during the enclosure phase, which facilitates online fine adjustment and necessary regrasping.

Despite these advantages, the design and deployment of soft grippers still face several challenges. These include material fatigue and air-tightness issues, membrane wear and contamination, tactile calibration and drift compensation, and delays in actuation and sensing. These factors may compromise long-term stability and maintainability. Addressing these issues is critical for improving the practical reliability of soft grippers, and the Soft Hand proposed by the Pisa and IIT team uses a 19-joint single-actuator adaptive synergy design that reduces control complexity while maintaining soft, safe, and robust grasping performance, thereby enabling stable enveloping grasps across diverse objects [17]. In addition, the TU Berlin team developed RBO Hand 2, which uses highly compliant pneumatic chambers and an underactuated structure and reproduces the Feix grasp set. With a mass of 178 g, it carries about 0.5 kg, which broadens compatibility with irregular targets [18]. Overall, the two applications show complementary strengths. The former highlights the simplicity and flexibility of synergies and underactuation for low-dimensional control and stable envelopment. The latter emphasizes the advantage of highly compliant pneumatic structures in grasp coverage and target adaptability. Both applications indicate that in localization for grasping, soft grippers use passive compliance and closed-loop fine adjustment to reduce upstream modeling and calibration requirements. This provides safety margins and fault tolerance during contact. To further improve robustness in complex environments, the operation strategy can adopt a phased procedure that includes approach, gentle contact, assessment, realignment, and closure. The strategy also integrates slip detection and end effector micromotion so that the system maintains stable contact on unknown shapes and uncertain surfaces.

3. Perception-Based Localization and Grasping Control Strategies

Once a gripper has basic grasping capability, perception of the target and action adjustment during contact become critical. These factors determine whether localization-based grasping attains a stable closed loop in complex and uncertain environments. This section focuses on three themes: prior localization and vision guidance, contact-based force feedback, and multi-modal fusion. It summarizes representative advances in recent years and reusable engineering patterns.

3.1. Prior Localization and Vision-Guided Grasping

In robotic grasping tasks, precise target localization is a prerequisite for successful operation. However, in complex and dynamic environments, a single image or sensor often cannot support high-precision grasping. Therefore, it has become crucial to utilize prior knowledge for initial localization and to employ vision guidance for dynamic adjustment, thereby enhancing the stability and adaptability of grasping systems. Prior localization utilizes calibrated fiduciary markers and sensors, exploiting known geometric and physical information to provide a reliable initial pose for grasping. Vision guidance utilizes real-time image recognition and pose estimation, enabling the system to respond to changes in the target and environment, and achieve precise grasp control.

As the field shifts from calibrated priors to generalizable recognition and tracking, visual fiducials remain the most common prior for localization in industrial settings. For example, planar markers such as AprilTag rely on a known size and a calibrated camera model to output a stable 6D pose under low-texture, intense light, and low-light conditions, and they provide an auditable base coordinate frame for localization and grasping.

These methods are an ideal choice for a common reference layer between mobile platforms and manipulators because they keep computation low while maintaining robustness and enabling online recalibration [19]. For 6D pose estimation of known rigid objects, end-to-end learning has produced four mainstream paradigms based on perception and fusion strategies. They include synthetic-data-driven monocular networks (DOPE), decoupled regression (PoseCNN), RGB-D pixel-level fusion with iterative refinement (DenseFusion), and multi-view consistency optimization (CosyPose). These methods address core pain points, including simplified mapping, robustness under occlusion and clutter, reduced data and annotation costs, and cross-view consistency. They now serve as the pose front end in the perception-to-grasp pipeline. PoseCNN simplifies the mapping from image to pose by decoupling semantic segmentation, translation, and rotation regression [20]. DenseFusion fuses RGB and depth features at the pixel level and employs iterative refinement to maintain stable registration, even in the presence of clutter and occlusion. DOPE utilizes a monocular network trained exclusively on synthetic data, which significantly reduces annotation and data collection costs. CosyPose, which utilizes multi-view consistency optimization, achieves leading results on standard datasets such as YCB Video and T-LESS, providing a practical solution for engineering applications of perception in grasp tasks [21].

However, in practical vision-guided grasping, the target must be maintained and tracked over long durations, across viewpoints, and under large displacements. The system must also recover the target when occlusion occurs, so the perception process needs longer-term temporal memory. Ho Kei Cheng and collaborators proposed XMem, which adopts the Atkinson and Shiffrin human memory model and builds a unified feature memory for short-term and long-term information so that mask propagation remains stable and efficient in memory when processing long videos on the minute scale. The method achieves state-of-the-art results on multiple long video benchmarks and provides sustained visual tracking for robotic manipulators during slow operation cycles and long-range motion. Subsequently, the Tracking Anything approach encapsulates prompt-based and target-agnostic tracking capabilities into a modular module, allowing it to integrate with upstream detectors or semantic prompts [22, 23]. These techniques provide modular components for a three-stage workflow that covers prior localization, perception-based alignment, and persistent tracking. When prior information is lost or drifts, long-term memory-based segmentation and tracking serve as soft constraints, maintaining the visibility of the target and the grasp point. This reduces the frequency of

relocalization and replanning. When prior localization aligns the world frame, and 6D pose estimation aligns the object. At the same time, long-term memory maintains temporal alignment, and the system gains a solid technical foundation for vision-guided grasping.

3.2. Contact-Based Force Feedback and Adaptive Grasping

In robotic grasping, vision provides a coarse target location, while tactile sensing is crucial for securing the grasp and maintaining stability throughout the operation. Traditional grasp control often relies on a binary decision of whether contact occurs. As technology advances, modern robots require precise perception of contact geometry, small displacements, and surface slip. High sensitivity to these contact details is driving continued innovation in tactile sensing.

In recent years, tactile sensing has evolved from sensor arrays and strain gauges to visuo-tactile sensing. It uses a transparent elastic skin with an internal camera and illumination. It reconstructs contact surface geometry, displacement, and micro-slip. These new tactile sensing technologies provide precise tactile information and deliver real-time feedback in diverse and complex environments, thereby enhancing the perceptual capability for robotic grasping. These advances enable robots to adapt to varying contact conditions in uncertain and dynamic environments, thereby significantly improving the reliability and stability of grasping tasks.

For example, MIT's GelSlim builds a finger-shaped, visuo-tactile resolution module. It emphasizes a thin form factor, durability, and repeatable calibration. During enveloping contact, it reconstructs the contact geometry, as well as the normal and tangential micro-displacements. It detects a micro slip and triggers regrasp. It supports a small, closed loop for localization and regrasping in practice [24].

Furthermore, Meta AI and GelSight's DIGIT targets low cost, small size, and mass production. It provides open-source hardware and software, along with standard interfaces. It fits multi-finger end effectors and high-speed policy learning. It also supports large-scale tactile data collection and evaluation of slip and stability. These features reduce integration and maintenance barriers [25]. Besides, UC Berkeley's OmniTact uses a multi-camera hemispherical layout to achieve omnidirectional contact sensing. It improves error tolerance and relocalization in alignment-sensitive tasks such as edge and hole search, insertion, and peg-in-hole. It enables closed-loop grasping under pose uncertainty and occlusion [26]. Together, these systems form a complementary chain that provides a reusable tactile basis for the closed loop of localization and grasping, as GelSlim delivers high-resolution contact imaging, DIGIT reduces costs and enables large-scale deployment through standard interfaces, and OmniTact increases coverage in all directions and tolerance to alignment errors.

In the grasping control of robotic manipulators, tactile sensing remains the basic method at the execution level, as it can integrate contact state estimation, slip detection, and readjustment into a fast feedback loop. MIT's GelSlim builds a finger-shaped visual-touch module. On contact, it produces a high-resolution tactile image field. It computes slip entropy from optical flow and shear vectors, as well as from the growth rate of the contact patch area. It maps events such as first contact, local slip, and changes in the contact surface to state machine triggers that represent approach, light touch, envelopment, holding, and release. The controller utilizes these triggers for realignment or recrawl, as well as for coordinated force and position control. This reduces blind grasping that relies on vision alone [24]. DIGIT from Meta AI and GelSight emphasizes low cost, a compact form, and mass production. It utilizes stable illumination and unified calibration, enabling the system to compute tactile optical flow and texture similarity in real-time. It connects slip and stability thresholds to the end effector state machine, which then triggers fine pose adjustments or recrawl. This design facilitates integration with multi-finger end effectors and enables online closed-loop control for high-speed policy learning [25]. Furthermore, BioTac uses a skin with three modalities that sense force, micro-vibration, and heat flow. It maps raw multimodal signals into a dense deformation and stress field by utilizing physical priors and data-driven models, making contact classification, slip detection, and material and texture recognition more interpretable. It also incorporates stability metrics into the

safety bounds of impedance and admittance control, which support adaptive realignment and regrasp for fragile and unknown targets.

In summary, the three systems form a complementary chain that operates in a bottom-up manner. GelSlim triggers a recrawl through event classification. DIGIT enables low-barrier closed-loop integration. BioTac reconstructs an interpretable contact field. They jointly provide a reusable tactile foundation and auditable safety bounds for the closed-loop localization and grasping system. This analysis shows that tactile sensing turns grasping from a single contact event into a continuous process that maintains a stable hold, and it extends the binary question of whether the object is grasped into a combined decision about how much force to apply, where to apply it, and whether slip occurs.

3.3. Multimodal Sensor Fusion and Intelligent Grasping

As grasping tasks become increasingly complex, a single sensing modality can no longer meet operational requirements across diverse environments.

The fusion of vision, tactile, and force sensing provides robots with richer environmental information, improving the robustness and adaptability of grasping. Therefore, relying on tactile and other sensors to stabilize the grasp when visual information degrades has become an important research direction.

Calandra et al. proposed an end-to-end action-conditioned model that utilizes visual recognition and GelSight tactile sensing to learn regrasp strategies. In transparent and reflective scenes where vision degrades, the model reduces the number of trials and the required gripping force, thereby improving the grasp success rate.

The method does not rely on tactile calibration or analytical mechanics models, but instead predicts and decides directly from grasp outcomes, establishing a new closed-loop learning paradigm that integrates perception, prediction, and decision-making [26]. However, in real-world settings, transparent objects, reflective surfaces, severe occlusion, and cluttered backgrounds often introduce uncertainty in vision-based grasp point estimation, which significantly increases the difficulty of grasping. In this case, a fusion framework that combines vision and tactile sensing is essential. The system first uses vision for detection and coarse pose estimation. During envelopment, it utilizes tactile sensing to refine the grasp pose and force magnitude, thereby mitigating the effects of visual degradation. For grasping transparent objects, vision and tactile approaches that utilize synthetic data and domain randomization significantly improve efficiency and stability in complex environments, demonstrating the value of complementary multimodal sensing [27].

For higher-level understanding tasks, vision, language, and action (VLA) models use visual and language priors to improve task generalization and semantic understanding. They show strong potential for further development. For example, the RT-2 model produces semantic goals, grasp regions, and tool choices at the high level, and it integrates pose estimation and tactile servoing at the low level to achieve alignment and stable holding. In grasping tasks, a layered scheme that organizes semantics, geometry, and contact is a key organizational approach for operating in an unstructured environment. Meanwhile, Bauza et al. proposed SimPLE and built an end-to-end sim-to-real pipeline that unifies vision–tactile task perception, grasp policy learning, and regrasp planning in one workflow. The method employs domain randomization, tactile rendering, and policy distillation, with a small amount of real-world calibration, which reduces the trial-and-error cost of deployment. It also provides reusable data formats, training recipes, and evaluation benchmarks that support scalable training and practical deployment [28].

These advances demonstrate that multimodal sensor fusion does not aim to stack sensors, but rather utilizes layered and complementary roles, where vision provides priors, tactile sensing supplies evidence, and the policy generates actions. This fusion approach demonstrates strong cross-domain generalization and engineering reusability in scenarios involving transparent objects, soft objects, and severe occlusions. Despite significant progress, key challenges remain in the efficient processing of multimodal data, learning to weight the contributions of different modalities in complex scenes, and enhancing robot adaptability in multitask settings. Therefore, future research should optimize

multimodal fusion algorithms. It should also improve system-level robustness. These steps will advance the pursuit of higher levels of intelligence and automation.

4. Conclusion

Within the closed-loop framework of localization, perception, and grasping, this paper provides a systematic review of sensor-based grasping for robotic grippers, analyzing the current progress, bottlenecks, and future directions across industrial, medical, and soft applications. Despite significant improvements in accuracy and stability, substantial challenges persist in complex environments, under dynamic conditions, and during high-precision task execution.

First, an integrated design that combines sensors with the mechanical structure is a key direction for future gripper development. Such integration forms a conformal sensing skin. The skin is replaceable and sterilizable, and it supports self-calibration. It meets the needs of sterile medical operations and high-throughput industrial production. Variable stiffness designs, such as granular jamming and coupled tendon and pneumatic actuation, allow reliable switching between compliance and stiffness. They support fine manipulation in practical applications. To ensure durability and maintainability, the system should include low latency electrical, pneumatic, and optical communication interfaces. It should also include health monitoring modules to support engineering practice better. Second, improving grasping capability requires high-precision sensing and robust decision-making under uncertainty. This paper proposes a closed-loop design for the vision and tactile pipeline. It integrates vision modules such as AprilTag 3 and PoseCNN with tactile sensing technologies such as GelSlim, DIGIT, and BioTac. The design provides higher robustness and real-time performance for grasping in complex environments. The integration of visual and tactile sensing enables robots to make more precise decisions and take more accurate actions in the presence of occlusion, drift, and irregular objects, thereby achieving reliable alignment and grasping in dynamic environments. Integrating visual and tactile sensing enables robots to make more precise decisions and take more accurate actions when encountering occlusion, drift, and irregular objects, thereby achieving reliable alignment and grasping in dynamic environments.

Future research should further optimize data processing and the pipeline from learning to deployment. Building on Calandra's vision and tactile fusion strategy, this approach combines policy distillation with a small amount of real-world calibration, supporting efficient and low-cost deployment. In particular, fusing synthetic data with domain randomization will further improve grasping accuracy and robustness in complex scenes with transparent or reflective objects. In addition, the combination of high-level planning methods such as RT-2 VLA and tactile servo execution will enable more flexible task execution by robotic grippers. In addition, combining high-level planning techniques such as RT-2 VLA with tactile servo execution enables robotic grippers to perform tasks with greater flexibility. In the long term, research and development should focus on several directions. One direction is to enhance semantic understanding and autonomous decision-making, enabling robotic grippers to adapt more effectively to complex and dynamic environments, particularly in medical, warehousing, and flexible manufacturing applications. Another direction is to pursue cross-disciplinary integration that improves coordination among sensors, actuation, and control, which yields more compact and maintainable integrated systems. A further direction is to increase generality and scalability, allowing grippers to be deployed flexibly and execute tasks effectively across diverse scenarios.

In summary, the localization-based grasping techniques described in this paper combine advanced visual and tactile sensing with multimodal fusion and variable stiffness control, moving toward higher levels of intelligence and autonomy. As technology advances and cross-domain collaboration deepens, robotic grippers will play a more significant role in both industry and medicine, driving broader applications of robotics.

References

[1] Volodymyr Tonkonogyi, Vitalii Ivanov, Justyna Trojanowska, et al. Advanced Manufacturing Processes Selected Papers from the Grabchenko's International Conference on Advanced Manufacturing Processes (InterPartner-2019), September 10-13, 2019, Odessa, Ukraine.

[2] Y. Hao, H. Zhang, Z. Zhang, C. Hu and C. Shi. Development of Force Sensing Techniques for Robot-Assisted Laparoscopic Surgery: A Review, in IEEE Transactions on Medical Robotics and Bionics, vol. 6, no. 3, pp. 868-887, Aug. 2024

[3] Kleeberger, K., Bormann, R., Kraus, W. et al. A Survey on Learning-Based Robotic Grasping. Curr Robot Rep 1, 239–249 (2020).

[4] Del Bianco E, Torielli D, Rollo F, Gasperini Det al. A High-Force Gripper with Embedded Multimodal Sensing for Powerful and Perception-Driven Grasping. 2024 IEEE-RAS 23rd International Conference on Humanoid Robots (Humanoids), Nancy, France, 2024, pp. 149-156,

[5] Tai, K, El-Sayed, A.-R., Shahriari, M., Biglarbegian, M., Mahmud, S. State of the Art Robotic Grippers and Applications. Robotics 2016, 5, 11.

[6] T. M. Huh et al. A Multi-Chamber Smart Suction Cup for Adaptive Gripping and Haptic Exploration, 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 2021, pp. 1786-1793.

[7] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy and K. Goldberg. Dex-Net 3.0: Computing Robust Vacuum Suction Grasp Targets in Point Clouds Using a New Analytic Model and Deep Learning, 2018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 2018, pp. 5620-5627.

[8] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, et al. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects

[9] Douglas Morrison, Peter Corke, Jürgen Leitner. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach

[10] Shademan A, Decker RS, Opfermann JD, Leonard S, Krieger A, Kim PC. Supervised autonomous robotic soft tissue surgery. Sci Transl Med. 2016, 8(337):337ra64.

[11] Saeidi, H., Opfermann, J. D., Leonard, S., et al. Autonomous robotic laparoscopic surgery for intestinal anastomosis in pigs. Science Robotics.2022.

[12] Uneri A, Balicki MA, Handa J, Gehlbach P, Taylor RH, Iordachita I. New Steady-Hand Eye Robot with Micro-Force Sensing for Vitreoretinal Surgery. Proc IEEE RAS EMBS Int Conf Biomed Robot Biomechatron. 2010, (26-29):814-819.

[13] Maclachlan RA, Becker BC, Tabarés JC, Podnar GW, Lobes LA Jr, Riviere CN. Micron: an Actively Stabilized Handheld Tool for Microsurgery. IEEE Trans Robot. 2012 Feb 1;28(1):195-212.

[14] Rus, D., Tolley, M. Design, fabrication and control of soft robots. Nature 521, 2015, 467–475.

[15] Shintake J, Cacucciolo V, Floreano D, Shea H. Soft Robotic Grippers. Adv Mater. 2018 May 7: e1707035.

[16] E. Brown, N. Rodenberg, J. Amend, A. Mozeika, E. Steltz, M.R. Zakin, H. Lipson, and H.M. Jaeger. Universal robotic gripper based on the jamming of granular material, Proc. Natl. Acad. Sci. U.S.A. 107 (44) 18809-18814.

[17] Catalano MG, Grioli G, Farnioli E, Serio A, Piazza C, Bicchi A. Adaptive synergies for the design and control of the Pisa/IIT SoftHand. The International Journal of Robotics Research.

[18] Deimel R, Brock O. A novel type of compliant and underactuated robotic hand for dexterous grasping. The International Journal of Robotics Research. 2015;35(1-3):161-185.

[19] J. Wang and E. Olson, "AprilTag 2: Efficient and robust fiducial detection," 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Daejeon, Korea (South), 2016, pp. 4193-4198.

[20] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes

[21] Labbé, Y., Carpentier, J., Aubry, M., Sivic, J., CosyPose: Consistent Multi-view Multi-object 6D Pose Estimation. Springer, 2020, 574–591.

[22] Cheng, H.K., Schwing, A.G. (2022). XMem: Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds) Computer Vision ECCV 2022. Lecture Notes in Computer Science, vol 13688. Springer, Cham.

[23] Ho Kei Cheng et al. Tracking Anything with Decoupled Video Segmentation. ICCV 2023.

[24] Elliott Donlon, Siyuan Dong, Melody Liu, et al. GelSlim: High-Resolution Tactile-sensing Finger.

[25] M. Lambeta *et al*., "DIGIT: A Novel Design for a Low-Cost Compact High-Resolution Tactile Sensor With Application to In-Hand Manipulation," in *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838-3845, July 2020.

[26] Roberto Calandra, Andrew Owens, Dinesh Jayaraman et al. More Than a Feeling: Learning to Grasp and Regrasp using Vision and Touch. RA-L, 2018.

[27] Shoujie Li, Haixin Yu, Wenbo Ding *et al*., "Visual–Tactile Fusion for Transparent Object Grasping in Complex Backgrounds," in *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3838-3856, Oct. 2023,

[28] Maria Bauza *et al.* SimPLE, a visuotactile method learned in simulation to precisely pick, localize, regrasp, and place objects.Sci. Robot.9, eadi8808, 2024.