Research on the Collaborative Perception and Decision Mechanism of Vision and Force Perception for Industrial Robot Grasping Tasks

Ding Cong
Zhengzhou Technical College of Commerce, Henan, China

Abstract:Aiming at the problem of low grasping accuracy and insufficient robustness of industrial robots in complex scenes, a visual and force perception collaborative perception and decision-making mechanism is proposed. By integrating the hardware of "binocular stereo vision+structured light 3D camera" and six axis force sensor, a multimodal feature fusion network based on attention mechanism (AMFF Net) is constructed to achieve deep data fusion; Design a reinforcement learning grasping point decision model and a force position hybrid dual closed-loop control strategy to form a "perception decision execution" closed-loop system. The verification results in three typical scenarios of automobile manufacturing, 3C electronic assembly, and logistics sorting show that this mechanism can control the grasping positioning error within ± 0.3mm, and improve the success rate of grasping in complex working conditions to over 95%, meeting the high-precision and high robustness requirements of flexible production.

Keywords: Industrial Robots; Collaborative Perception; Visual Sensory Fusion; Reinforcement Learning; Force Position Hybrid Control

1. Introduction

1.1 Research Background and Significance

In the wave of deep integration between Industry 4.0 and intelligent manufacturing, industrial robot grasping tasks have become the core support link of flexible production systems, widely used in key fields such as automobile manufacturing, 3C electronic assembly, logistics sorting, etc. In the current industrial scene, the forms of workpieces are becoming increasingly diverse (from standard boxes to irregular precision parts), and the working environment presents dynamic uncertainties (such as disorderly stacking of workpieces, fluctuating lighting, and random appearance of obstacles). Traditional single perception technology is gradually exposing its shortcomings: although pure visual perception can achieve target localization, it is difficult to obtain physical properties of objects (such as stiffness and surface friction coefficient) due to factors such as occlusion, reflection, and weak texture; Pure force perception can monitor the state of contact force, but lacks global spatial positioning ability and cannot independently complete target search and attitude alignment.

The collaborative perception of vision and force perception can achieve precise target positioning and adaptively adjust grasping force and posture by deeply integrating spatial geometric information and contact mechanics feedback, effectively solving grasping difficulties in complex scenes. This technology can not only enhance the adaptability of robots to unknown working conditions, reduce dependence on manual teaching, but also promote the transformation of smart factories from "fixed program execution" to "autonomous decision-making operation", providing core technical support for the large-scale implementation of "black light factories", and has important theoretical research value and significant engineering application prospects.

1.2 Research Status and Technical Bottlenecks

In recent years, visual perception technology has made rapid progress in the fields of object detection and pose estimation: algorithms such as YOLOv8 and Faster R-CNN have achieved efficient object recognition in complex scenes, while point cloud processing models such as PointNet and PointTransformer have improved the accuracy of 3D pose estimation. However, in

strong noise and high occlusion industrial environments, the robustness of visual algorithms is still insufficient, and pose estimation errors can easily exceed 1mm, making it difficult to meet precision grasping requirements. In terms of force perception, the sampling frequency and measurement accuracy of the six axis force/torque sensor continue to improve, but its data is susceptible to temperature drift and vibration interference, and cannot establish a correlation between force feedback and spatial position when used alone.

The existing collaborative perception research mostly stays at the simple superposition mode of "visual positioning+force correction", which has three core bottlenecks: firstly, the accuracy of spatiotemporal alignment of multimodal data is insufficient, and the time delay (usually 10-50ms) of visual and force perception data and spatial coordinate system deviation lead to poor fusion effect; Secondly, feature fusion lacks dynamic adaptation capability and fails to adjust the weight allocation of the two modes according to changes in operating conditions; Thirdly, the closed-loop mechanism of "perception decision execution" is not sound, and decision-making algorithms are difficult to quickly respond to the dynamic changes of perception data, resulting in a success rate of less than 90% in unstructured environments. It is urgent to achieve technological breakthroughs through architectural innovation and algorithm optimization.

2. Architecture Design of Visual and Force Sensing Collaborative Perception System

2.1 Hardware Layer: Multimodal Sensor Integration Solution

2.1.1 Visual Perception Module

Adopting a composite perception scheme of "binocular stereo vision+structured light 3D camera": the binocular camera uses Basler acA2500-14gm, baseline length 120mm, resolution 2592 × 1944, calculates disparity map based on SGMM algorithm, obtains depth information within the range of 0.5-5m, depth error ≤ 0.3mm; the structured light camera adopts Intel RealSense D455, improves the detail capture ability of weakly textured objects through Gray code encoding technology, with a frame rate of 30fps, and the pose estimation accuracy of different materials such as metal and plastic workpieces can reach ± 0.1mm. At the same time, it is equipped with an industrial grade circular LED light source (wavelength 650nm), which eliminates reflective interference through brightness adaptive adjustment, ensuring stable imaging in a wide illumination range of 0.1lux~10000lux.

2.1.2 Force perception module

Integrate ATI Nano17 six axis force/torque sensors at the end of the robotic arm (such as ABB IRB 1200), with a range of ± 17N (force) and ± 0.85N · m (torque), a sampling frequency of 1kHz, a measurement accuracy of ± 0.1% FS, and a resolution of 0.01N (force) and 0.001N · m (torque). To solve the problem of temperature drift, a polynomial fitting temperature compensation algorithm is adopted to control the zero drift within ± 0.05N at an ambient temperature of -10~60 ℃; Through FPGA triggered synchronization technology, microsecond level time alignment of visual and force sensing data is achieved to ensure temporal consistency of multi-source data.

2.2 Algorithm Layer: Multi Source Data Fusion Framework

2.2.1 Spatiotemporal Calibration Techniques

Establish a joint calibration model of "camera robotic arm force sensor": use Zhang's calibration method to solve the camera internal parameters (focal length, principal point coordinates, distortion coefficient), collect 20 sets of different pose images through a checkerboard calibration board, and calibrate the error ≤ 0.2 pixels; Based on Eye in Hand calibration, the robotic arm was moved to 15 different poses to collect calibration plate images captured by the camera and joint angle data of the robotic arm. The least squares method was used to solve the transformation matrix between the camera coordinate system and the end coordinate system of the robotic arm, ultimately achieving an alignment error of ≤ 0.5mm between visual perception and the robotic arm base coordinate system. For the force sensor, 5 sets of output data under different loads were collected through static loading experiments, and a zero offset compensation matrix was constructed to eliminate the influence of installation errors and initial deviations on the force feedback signal.

2.2.2 Feature Fusion Algorithm

Propose a multimodal feature fusion network based on attention mechanism (AMFF Net), which consists of a visual branch, a force

sensing branch, and an attention fusion layer. The visual branch uses ResNet50 as the backbone network to extract 2D texture features of objects and shape and pose features of 3D point clouds (with an output dimension of 512 dimensions); The force sensing branch encodes the temporal data of the force sensor (100 sampling points within a 100ms window) through a 3-layer 1D convolutional network, extracting features such as contact force distribution and stiffness characteristics (output dimension 256 dimensions); The attention fusion layer calculates the mutual information between two modal features and dynamically assigns weight coefficients (visual weight range 0.3~0.8, force weight range 0.2~0.7) to achieve adaptive fusion of multi-source features. Experimental verification shows that in cluttered stacking scenarios, the object recognition accuracy of this algorithm reaches 92%, which is 18% higher than traditional serial fusion methods, providing more comprehensive environmental and target representations for subsequent decision-making.

3. Collaborative decision-making mechanism and control strategy

3.1 Capture Planning Stage: Intelligent Decision Model

3.1.1 Optimization algorithm for grasping points

Build a reinforcement learning based grasping point selection model (GrassRL) to solve the optimal grasping posture decision-making problem for objects of different materials and shapes. The input of the model is the fused features output by AMFF Net, including visually extracted object surface normal vectors, curvature distributions, and contact stiffness pre estimated by force perception; Using Deep Q-Network (DQN) as the decision network, a multi-objective reward function is designed:

$$R = \alpha R_{ssuccess} + \beta R_{force} + \gamma R_{efficiency}$$

where $R_{ssuccess}$ is the grasping success rate, R_{force} is the contact force deviation (difference from the optimal force threshold), $R_{efficiency}$ is the grasping time, and α=0.6, β=0.3, and γ=0.1 are the weight coefficients. By conducting 100000 training sessions in a virtual simulation environment (including 100 typical industrial workpieces), the model converged and controlled the gripper positioning error within ± 0.3mm in the grabbing task of new energy battery pole pieces (thickness 0.1mm, width 50mm), which improved the decision-making efficiency by 40% compared to traditional heuristic algorithms (such as geometric center method) and effectively avoided pole piece wrinkles or fractures.

3.1.2 Path planning and obstacle avoidance strategy

Design a fusion path planning algorithm of dynamic window method (DWA) and artificial potential field method (APF) to achieve safe and efficient movement of robotic arms. The global path planning adopts APF and is based on the visual system to construct an environmental point cloud map. The target position is set as the gravitational source and obstacles as the repulsive source to generate a collision free initial path; DWA is used for local path correction. During the movement of the robotic arm, the visual system updates the obstacle position at a frequency of 10Hz, and the force sensor monitors the contact force in real time. When a sudden change in force is detected exceeding the threshold (5N), it is judged as a potential collision and immediately triggers the local obstacle avoidance strategy: by adjusting the joint angular velocity of the robotic arm (maximum deceleration of 0.5rad/s ²), the local path within 300ms is re planned to ensure safe and compliant control at 0.5m/s high-speed movement. The experiment shows that the obstacle avoidance response time of this algorithm is ≤ 50ms, and the obstacle recognition rate reaches 98%, which is 60% lower than the collision rate of a single DWA algorithm.

3.2 Grabbing Execution Stage: Force Position Hybrid Control

3.2.1 Impedance Control Model

Establish impedance control equations based on visual target pose to achieve compliant contact control at the end of the robotic arm. The core of impedance control is to simulate the mechanical characteristics of the end effector of a robotic arm by adjusting the stiffness matrix K, damping matrix B, and inertia matrix M. The equation is as follows:

$$M\Delta\ddot{x}+B\Delta\dot{x}+K\Delta x=F_{des}-F_{meas}$$

where Δx is the end effector position deviation, F_{des} is the expected contact force, and F_{meas} is the measured force of the sensor. Based on the characteristics of different workpieces, preset

adaptive stiffness matrices: metal workpiece K=diag ([100, 100, 80, 50, 50, 30]) N/m, plastic workpiece K=diag ([80, 80, 60, 40, 40, 20]) N/m, glass workpiece K=diag ([60, 60, 40, 30, 30, 15]) N/m. When the force sensor detects that the contact force exceeds the preset threshold (such as Fz>15N for glass products), the system automatically switches to force control mode, and adjusts the gripper driving force through a proportional integral (PI) controller to ensure that the contact force is stable within a fluctuation range of ± 2N and avoid damage to the workpiece.

3.2.2 Closed loop feedback regulation

Design a dual closed-loop control architecture to achieve dual guarantees of pose accuracy and force stability: the outer loop is a visual pose loop, using a PID controller (parameters Kp=0.8, Ki=0.2, Kd=0.1), based on real-time acquisition of pose deviations (position deviation ≤ 0.5mm, posture deviation ≤ 0.3 °) between the end effector and the target object by the visual system, and outputting joint position correction quantities; The inner loop is a force feedback loop, using sliding mode control algorithm (SMC), designing a switching function s=e ·+λ e (λ=5 is the adjustment coefficient), and adjusting the joint torque in real time through the control law u=- Ks sign (s) (Ks=10 is the sliding mode gain), quickly compensating for force deviations caused by changes in workpiece stiffness and environmental disturbances. In 3C electronic precision assembly tasks (such as mobile phone camera module insertion), this dual closed-loop control architecture increases the success rate of insertion from 85% to 97%, with a position repetition positioning accuracy of ± 0.05mm, meeting micron level assembly requirements.

4. Typical Industrial Scenario Application Verification

4.1 Automotive Manufacturing: Flexible Grasping of Engine Cylinder Blocks

In a new energy vehicle engine production line, the collaborative perception and decision-making system proposed in this paper is applied to meet the flexible grasping requirements of aluminum alloy cylinder bodies (size 500mm × 300mm × 200mm, weight 8 ± 0.5kg). The visual system uses ICP point cloud registration technology to achieve cylinder pose estimation (position error ≤ 0.3mm, pose error ≤ 0.3 °), effectively solving the problem of positioning deviation caused by reflection on the cylinder surface; The force feedback system monitors the contact force between the gripper and the cylinder in real time, dynamically adjusts the gripper closure force (150 ± 10N), and avoids scratches on the cylinder surface caused by uneven force. Compared with traditional teaching programming methods, the grasping cycle of this system has been shortened from 6s/time to 4.2s/time, with an efficiency improvement of 30%. It can operate continuously for 1000 times without damage to the workpiece, meeting the high cycle and high reliability requirements of the production line.

4.2 3C Electronics: Precision Operation of Chip Substrates

In the assembly process of smartphone camera module, it is necessary to precisely grasp and place the 0.2mm thick borosilicate glass substrate, with a position error of ≤± 0.1mm and no substrate breakage. The system recognizes the edge feature points of the substrate through the visual module, achieving a positioning accuracy of ± 0.05mm; The force sensing module adopts a micro force control strategy, with the contact stiffness between the gripper and the substrate controlled at 0.5N/mm and a contact force resolution of 0.1N, effectively avoiding substrate bending or cracking. Simultaneously integrating vibration suppression algorithm (cut-off frequency 50Hz) to compensate for residual vibrations during the movement of the robotic arm, and controlling the substrate placement position error within ± 0.1mm. The application results show that the yield rate of substrate assembly has increased from 92% to 99.2%, and the daily production capacity has increased by 20%, significantly reducing production costs.

4.3 Logistics Sorting: Adaptive Grabbing of Unordered Packages

At the e-commerce logistics automatic sorting center, facing flexible packaging packages with varying surface textures and a weight range of 0.1~5kg, the system uses the YOLOv8m object detection algorithm (detection speed 30fps, accuracy 95%) to identify the package grasping area in real time, and the force sensor monitors the contact force change rate through a 50ms sliding window algorithm. When package sliding is detected (force change rate>2N/ms),

the system automatically adjusts the gripper pressure (increment 5N) and dynamically optimizes the gripping posture based on the weight of the package. Compared with traditional visual sorting systems, the success rate of irregular object grasping in this solution has increased from 78% to 95%, the package damage rate has decreased from 3% to 0.5%, and the sorting efficiency has reached 1200 pieces/hour, meeting the high-speed sorting needs of the logistics industry.

5. Challenges and Future Prospects

5.1 Technical Challenges

The current collaborative perception technology of vision and force still faces two core challenges: firstly, insufficient adaptability to extreme industrial environments. In environments with strong vibrations (such as stamping workshops, vibration frequencies of 50-200Hz) and high dust (such as casting factories, dust concentration>10mg/m ^{3}), the quality of visual imaging decreases and the signal noise of force sensors increases (noise amplitude increases by 30%), resulting in a significant decrease in perception accuracy; The second is the contradiction between computing power and real-time. The reasoning delay of deep learning models such as AMFF Net and GraspRL on edge computing platforms (such as NVIDIA Jetson AGX Orin) is about 80ms, which is difficult to meet the real-time decision-making needs of high-speed production lines (beat<2s), and the model complexity needs to be further optimized.

5.2 Development Trends

Future technological development will focus on three major directions: firstly, upgrading anti-interference sensing technology, developing integrated sensor packaging solutions for dust and vibration resistance, and combining adaptive noise suppression algorithms to enhance sensing robustness in extreme environments; Secondly, lightweight models and computational power optimization are used to reduce the computational complexity of deep learning models through techniques such as model pruning, quantization, and knowledge distillation, achieving real-time inference at the edge (latency<30ms); The third is the integration of digital twins and human-computer interaction, constructing a digital twin for grasping tasks, and enhancing pre trained models through virtual environment data augmentation (such as lighting changes, workpiece deformation, noise injection) to reduce dependence on real industrial data; Develop a safety collision detection algorithm based on force feedback (response time ≤ 10ms), combined with visual human pose recognition technology, to achieve safe and efficient grasping of robots in human-machine collaboration scenarios, and promote the large-scale application of "fence free" production units.

6. Conclusion

This article proposes a visual and force perception collaborative perception and decision-making mechanism for the complex scene adaptation problem in industrial robot grasping tasks. Through the hardware integration of "binocular stereo vision+structured light 3D camera" and six axis force sensor, the problem of multimodal data spatiotemporal alignment has been solved; AMFF Net based on attention mechanism achieves dynamic adaptive fusion of visual and force features; The reinforcement learning grasping point decision model and force position hybrid dual closed-loop control strategy have constructed a complete closed-loop of "perception decision execution". The application verification of three typical industrial scenarios shows that this mechanism can effectively improve the grasping accuracy (positioning error within ± 0.3mm) and robustness (success rate of over 95%) under complex working conditions, making it flexible

References

[1] Wang Yaonan, Jiang Yiming, Jiang Jiao, etc Key Technologies of Robot Perception and Control and Their Intelligent Manufacturing Applications [J]. Automation Expo, 2023, 40 (10): 50-66

[2] Gu Xin Research on Multi modal Perception and Grasping Detection of Robots Based on Spatiotemporal Attention Mechanism [D]. Guangdong University of Technology [2021-12-14]

[3] Xue Songdong Research on Coordinated Control and Simulation of Target Search Oriented Swarm Robots [D]. Lanzhou University of Technology [2021-12-14]

[4] Jie Yinggang, Lanjiang Rain A review of

research on collaborative robots and their motion planning methods [J]. Computer Engineering and Applications, 2021, 57 (13): 16

