Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

AI Machine Vision Automated Defect Detection System

QU, Meina [1*]

[1] City University of Seattle, USA

**** QU, Meina is the corresponding author, E-mail: qumeina@yahoo.com***

Abstract: With the rapid development of smart manufacturing technologies, automated production has become an important trend in the transformation of the industrial chain. Among various automation applications, robotic arm grasping and visual inspection systems are the most widely used. This paper focuses on unstructured stacking scenarios and workpiece defect detection, and designs two deep learning-based vision systems. In terms of theoretical research, the study focuses on the fundamental knowledge and technical methods related to robotic arm grasping in unstructured environments and workpiece defect detection. To address the issue of grasping randomly stacked objects, a 2D/3D vision-based robotic arm grasping solution is proposed. This solution employs an eye-in-hand configuration, where RGB and depth images are captured by a stereo camera, and a depth feature extraction branch is added to the Mask R-CNN network to improve the accuracy of object detection and segmentation in complex scenes. For object localization, the segmented results are mapped to a 3D point cloud through RGB-D data registration, and the RANSAC and PCA algorithms are used to extract the target plane and bounding box, thereby obtaining the 6D pose information of the target. Combined with the hand-eye calibration results, the robotic arm can accurately grasp the target. Additionally, taking an automotive one-way clutch as an example, an automated defect detection system based on deep learning is designed. Using an industrial camera to capture images, the system utilizes a semantic segmentation network and a defect classification network to detect the number of teeth, copper sleeve, semicircular piece, and chamfer of the one-way clutch, thereby achieving automatic recognition of part defects. This paper integrates 2D image and 3D point cloud information, combined with deep learning methods, to explore robotic arm grasping and workpiece detection, providing new ideas and solutions for the development of smart manufacturing.

Keywords: Smart Manufacturing, Robotic Arm Grasping, Defect Detection, 2D/3D Vision Fusion, Deep Learning.

Disciplines: Computer Science. **Subjects:** Deep Learning.

DOI: https://doi.org/10.5281/zenodo.13763253 **ARK:** https://n2t.net/ark:/40704/JCTAM.v1n4a01

1 INTRODUCTION

Automated industrial inspection technology is an indispensable component of modern industrial production, closely related to fields such as computer vision, sensors, robotics, and artificial intelligence. With the rapid growth of the industrial automation market, particularly in China, which is the world's largest industrial automation market, automated inspection technology plays an increasingly prominent role in improving product quality, production efficiency, and reducing costs[1]. In practical applications, robotic arm grasping and visual inspection are two common automated solutions. For unstructured stacking scenarios, traditional teaching-based methods fail to meet the requirements, necessitating the development of intelligent visual grasping systems that endow robots with autonomous perception and analysis capabilities for efficient unstructured grasping. To ensure product quality, more and more companies are adopting machine vision technology for fully automated inspections to detect common defects such as surface scratches, coating errors, and assembly errors, thereby improving inspection efficiency and accuracy. Automated industrial inspection technology is of great significance in enhancing product quality, production efficiency, and cost reduction, making it a field full of innovation and development opportunities. As technology continues to advance, automated inspection systems will be applied in a wider range of industrial scenarios, providing smarter, more efficient, and more reliable solutions for modern industrial production.

2 LITERATURE REVIEW OF DOMESTIC AND INTERNATIONAL RESEARCH

2.1 Literature Review on Educational Big Data and Personalized Learning

In the rapid development of automated industrial inspection technologies, we see how technological advancements are driving improvements in production efficiency and product quality. This trend is not only evident in the industrial sector but also shows similar impacts in the

education sector. With the continuous evolution of information technology and the accumulation of data, the education sector is also undergoing a technological transformation. This transformation relies heavily on advanced data analysis technologies such as educational big data and personalized learning, which are rapidly changing traditional teaching models. Just as automated inspection systems enhance the accuracy and efficiency of industrial inspections through intelligence and data-driven methods, the education sector is optimizing the learning process and outcomes through big data and personalized teaching.

Next, we will explore the research progress on educational big data and personalized learning in the context of educational informatization, and further understand how these technologies are driving transformation and innovation in the education field. Some studies have analyzed how to use machine learning methods to analyze educational big data to predict students' academic performance [2]. They proposed a modeling and training method based on student attribute data, which can identify key factors affecting academic performance and provide references for teaching optimization. They further explored how to use deep learning methods to select personalized teaching strategies [3]. By analyzing and modeling data on students' learning behaviors, cognitive levels, etc., it is possible to predict the effectiveness of different teaching strategies, recommend optimal learning paths and resource combinations for each student, and achieve tailored instruction. This research proposed a data-driven personalized teaching optimization framework that achieved significant performance improvements in real teaching scenarios.

In addition to predicting academic performance and selecting teaching strategies, educational big data is also used to analyze students' emotional states and social behaviors. For example, the EmotionQueen benchmark is used to evaluate large language models' understanding and expression of student emotions [4]. By analyzing the responses generated by the model, it is possible to measure its level of empathy and support emotional teaching.

Overall, research on educational big data and personalized learning focuses on how to use data analysis technologies to uncover patterns and characteristics in students' learning processes, providing intelligent support for teaching decisions. On one hand, modeling and predicting data such as academic performance and learning behaviors can reveal key factors affecting learning outcomes and recommend targeted teaching strategies. On the other hand, analyzing data on students' emotions and social interactions can provide insights into students' psychological states and interpersonal interactions, guiding emotional teaching and collaborative learning. In the future, further integrating multimodal educational data to construct more comprehensive and dynamic student profiles, and achieving more precise and real-time personalized teaching, remains a valuable research direction.

2.2 Literature Review on Multimodal Learning and Cross-Modal Alignment

Multimodal learning aims to leverage information from different modalities (such as vision, language, and audio) to build more comprehensive and accurate artificial intelligence models. However, a key challenge is how to achieve cross-modal information alignment and integration due to differences in feature distributions and representations across modalities. For instance, multimodal preference alignment methods are used to address the regression issues of language models with visual instructions [5]. It was found that when visual and language information are inconsistent, models often suffer from overfitting and inadequate generalization. By incorporating techniques like adversarial learning and contrastive learning, features from different modalities can be mapped into a shared semantic space, aligning preferences and enhancing the robustness and generalization of the model. Additionally, some studies have constructed a large-scale scientific chart question-answering dataset, SciGraphQA, to evaluate model performance in cross-modal reasoning tasks [6]. This dataset includes numerous scientific concepts, entities, and relationships, requiring models to simultaneously understand text, images, and charts to answer complex questions. Such cross-modal reasoning capabilities are crucial for scientific education and intelligent tutoring systems.

In addition to vision-language alignment, speech-text alignment is also a significant direction in multimodal learning. Chen, Y. et al. (2024) studied how to use large language models to assess the empathetic capabilities of speech assistants. By converting speech to text and analyzing the sentiment and semantics of the text, it is possible to evaluate whether the speech assistant's responses meet the user's emotional needs. This cross-modal emotional alignment plays an important role in improving the naturalness and friendliness of human-computer interactions.

Research in multimodal learning and cross-modal alignment focuses on how to integrate and coordinate information from different modalities to build more intelligent and comprehensive AI systems. On one hand, joint modeling and representation learning of visual, language, and audio data can enable cross-modal information transfer and enhancement, improving the model's understanding and generation capabilities. On the other hand, introducing alignment and consistency constraints can mitigate differences and conflicts between modalities, enhancing the model's robustness and generalization. Future research will need to explore deeper and more granular cross-modal alignment by further uncovering high-level semantic associations between modalities while reducing computational and annotation costs, which remains a challenging research direction.

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

2.3 Literature Review on the Application of Intelligent Optimization Algorithms in Engineering

Intelligent optimization algorithms, such as evolutionary algorithms and swarm intelligence algorithms, efficiently solve complex engineering optimization problems by simulating natural optimization mechanisms. In recent years, these algorithms have been widely applied in fields such as engineering design, manufacturing, and control, achieving significant results.

Some studies have explored how to use an improved snow leopard optimization algorithm and the Inception-V4 network to optimize the detection of diabetic retinopathy [7]. By adaptively adjusting algorithm parameters, the model's convergence speed can be accelerated and the accuracy of lesion area identification improved. This research exemplifies the combination of intelligent optimization and deep learning, demonstrating significant application value in medical image analysis. Zhang, Y. et al. (2024) investigated the application of BIM technology in smart buildings [8]. By integrating Building Information Modeling with IoT and big data technologies, digital management and optimization control throughout the building's lifecycle can be achieved. Intelligent optimization algorithms can be used for tasks such as building energy consumption analysis and equipment scheduling, enhancing the level of building intelligence and operational efficiency.

Researchers in the aerospace field have studied how to use machine learning methods to predict hazardous flight weather conditions [9]. By analyzing and modeling historical weather and accident data, models that can provide early warnings for severe weather can be trained, improving flight safety. Furthermore, they explored how to use backpropagation neural networks to predict the occurrence of flight accidents [10]. By training on flight and accident data, neural network models capable of assessing flight risk can be established, providing decision support for safety management. These studies highlight the important applications of intelligent optimization algorithms in aviation safety.

In the industrial manufacturing sector, intelligent optimization algorithms are used for tasks such as production scheduling and process parameter optimization. For instance, machine learning methods are applied to optimize credit risk assessment [11]. By modeling and predicting borrower attribute data, high-risk customers can be more accurately identified, reducing bad debt losses. Huang, D. et al. (2024) explored the use of the Louvain algorithm for genomic data identification and analysis [12]. By clustering and modularizing gene sequences, functional modules and regulatory relationships can be discovered, providing insights for biomedical research. The authors also studied the use of image enhancement methods for tumor segmentation [13]. By preprocessing medical images through denoising and contrast enhancement, the visibility of tumor regions can be improved, providing clearer image data for subsequent automatic segmentation and diagnosis. Deep neural networks are used in medical image classification research, such as pneumonia detection with AlexNet and InceptionV3 models [14]. These studies showcase the extensive application of intelligent optimization algorithms in industrial manufacturing and biomedical fields [15].

Research on intelligent optimization algorithms in engineering focuses on utilizing artificial intelligence technologies to solve complex optimization problems, enhancing system efficiency and performance. On one hand, modeling and algorithm design can automate the search and optimization of design parameters and control strategies, reducing the cost of manual tuning. On the other hand, integrating with deep learning and big data technologies allows for the extraction of hidden patterns and rules from massive engineering data, enabling intelligent decision-making and prediction. Future research will need to address how to further improve the convergence speed, stability, and interpretability of intelligent optimization algorithms, while expanding their application in more engineering scenarios, making this a promising and challenging research direction.

3 AI-RELATED CONCEPTS

3.1 Time Series Anomaly Detection

Time series anomaly detection aims to identify anomalous points or segments that deviate from the normal patterns in sequential data. This has significant applications in fields such as industrial fault diagnosis and financial fraud detection. Traditional anomaly detection methods primarily rely on statistical models, such as Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), which model the probabilistic distribution of time series data to determine whether new incoming data points are anomalous. These methods typically assume that data follows a specific distribution and struggle to model complex, nonlinear, and non-stationary time series effectively.

In recent years, with the advancement of deep learning technology, more research has focused on using neural networks for time series anomaly detection. For example, the framework for finding faithful time filters (FTS) [16] uses decomposition and reconstruction of time series data to automatically learn multi-scale periodic patterns, constructing anomaly scores to enable real-time anomaly detection. Li, B. et al. (2019) proposed an adaptive ensemble empirical mode decomposition (AETA) method for daily cyclic time series feature extraction [17]. This method performs time-frequency analysis and periodicity detection on electromagnetic interference data to extract key features that reflect equipment status and fault precursors. These studies demonstrate the advantages of deep learning in time series anomaly detection, as it can automatically learn complex patterns from data without strong assumptions about

data distribution.

In addition to deep learning, other machine learning methods such as Support Vector Machines (SVMs) and Isolation Forests are also used for time series anomaly detection. These methods learn the boundary or density distribution of normal data to identify points that deviate from these boundaries or have lower density as anomalies. For instance, a time series anomaly explanation method based on Factorization Machines [18] decomposes time series into multiple latent factors and uses interactions between these factors to characterize the root causes of anomalies, providing an interpretable anomaly analysis framework.

Research in time series anomaly detection focuses on how to automatically identify abnormal patterns from sequential data, providing a basis for system monitoring and diagnosis. On one hand, advanced modeling methods like deep learning can uncover complex nonlinear relationships within time series data, improving the accuracy and robustness of anomaly detection. On the other hand, integrating technologies such as causal inference and explainability can trace the origins and impacts of anomalies, offering richer information for anomaly analysis and decision-making. Future research will need to address improving the real-time capabilities and adaptability of time series anomaly detection while expanding its deployment across more application scenarios, which remains a promising and challenging area of study.

3.2 Knowledge Graphs and Inference in Natural Language Processing

Knowledge graphs are structured knowledge bases that represent entities and their relationships in the form of a graph. In natural language processing (NLP), knowledge graphs provide rich background knowledge and common-sense support for tasks such as text understanding, question answering, and reasoning. However, automatically constructing high-quality knowledge graphs from text and utilizing these graphs for complex semantic reasoning remain challenging issues.

Li, B. et al. (2024) explored how to use large language models to handle large-scale structured and unstructured data [19]. By unifying the representation and encoding of data in different formats, efficient information extraction, retrieval, and generation can be achieved. This method can be used for the automatic construction and completion of knowledge graphs, mining entities and relationships from vast amounts of text data, and expanding the scale and quality of knowledge bases. They also explored how large language models can optimize travel route planning [20]. By understanding and reasoning about user preferences, attraction attributes, and other information, personalized travel route recommendations can be generated to enhance user experience. This knowledge-based reasoning capability has significant value for intelligent question answering, decision support, and other applications.

In recent years, the integration of knowledge graphs with deep learning has become a hot topic in NLP research. On one hand, knowledge graphs can provide structured prior knowledge to guide deep learning models in making reasonable inferences and decisions. On the other hand, deep learning models can automatically learn the representation and update rules of knowledge graphs from large-scale text data, improving the efficiency of knowledge acquisition and reasoning. Meta-learning-based adversarial defense methods, which learn attack and defense strategies across multiple tasks, can quickly adapt to new types of attacks and enhance model robustness [21]. This meta-learning paradigm can be applied to the continuous learning and updating of knowledge graphs, enabling them to adapt to the ever-changing real-world application environment.

Despite the promising progress in combining knowledge graphs with deep learning, several issues remain to be addressed. Among these, the interpretability and credibility of knowledge graphs are key challenges. Since knowledge graphs are often automatically constructed by machines, they inevitably contain noise, errors, and biases, making the evaluation and assurance of knowledge quality an important research topic. Furthermore, integrating knowledge graphs with causal reasoning, logical reasoning, and other advanced cognitive abilities to achieve deeper and more comprehensive semantic understanding is also a valuable direction for exploration.

Research on knowledge graphs and inference in NLP focuses on how to enhance language understanding and generation capabilities using structured knowledge. On one hand, automated knowledge acquisition and representation learning methods can build large-scale, high-quality knowledge bases, providing rich background knowledge for various language tasks. On the other hand, combining knowledge graphs with deep learning, meta-learning, and other technologies can enable more efficient and robust knowledge reasoning, improving the generalization and adaptability of language models. Future research will need to break through the bottlenecks in knowledge acquisition and reasoning, achieving more interpretable, credible, and intelligent language understanding systems, which remains a field full of imagination and innovation.

3.3 Applications of Reinforcement Learning in Robot Control

Reinforcement learning (RL) is a machine learning paradigm where an agent interacts with its environment and optimizes its decision-making strategy based on feedback signals. In recent years, RL has gained significant attention in the field of robot control, offering new ideas and methods for autonomous learning and adaptive control of robots.

Traditional robot control methods rely mainly on manually designed controllers and rules, which can struggle to handle complex and dynamic environments. In contrast, RL enables robots to autonomously learn and adapt control

strategies for different tasks and conditions by continuously trying and optimizing in their environment. For example, models considering the effect of hole collapse on wellbore stability, by coupling mechanical and seepage effects, can more accurately predict wellbore instability risks and guide drilling operations [22]. This physics-based approach can provide prior knowledge and constraints for RL, improving the efficiency and convergence of policy search.

Deep reinforcement learning (DRL) integrates deep neural networks into RL, allowing robots to extract features and learn strategies directly from raw sensor data, significantly enhancing control flexibility and adaptability. For instance, black-box attack methods based on sequence queries can infer the decision boundaries and key features of a target model by conducting a series of carefully designed queries without knowing its internal structure [23]. Such black-box optimization methods can be used in RL for policy search, finding optimal control strategies without relying on environmental models.

However, applying RL to practical robotic systems presents several challenges. First, the complexity and uncertainty of real environments far exceed those of simulation environments. Ensuring that learned strategies can safely and effectively transfer to real systems is a critical issue. Second, RL typically requires a large amount of trial-and-error and exploration, and the exploration cost in robotic systems is relatively high. Learning good strategies within limited interaction times is a major challenge. Additionally, designing appropriate reward functions to guide robot learning and balancing exploration with exploitation to avoid local optima are important research areas.

Research on RL in robot control focuses on how robots can master control strategies for complex tasks and environments through autonomous learning and optimization. On one hand, advanced modeling methods like deep learning can enhance robots' perception and decision-making capabilities, achieving end-to-end adaptive control. On the other hand, integrating prior knowledge and physical models can accelerate policy search and improve learning efficiency, overcoming the limitations of purely data-driven approaches. The transformative role of artificial intelligence (AI), particularly large language models, in improving government operations and detecting AI-generated content [24], demonstrates how AI applications in administrative automation, public safety, resource management, and citizen services can optimize operational efficiency and decision quality. Case studies from the U.S. IRS and Social Security Administration highlight successful AI implementations in these areas [25]. Looking ahead, improving the sample efficiency, safety, and interpretability of RL, while expanding its deployment in more robotic applications, remains an attractive and impactful research direction.

3.4 Explainability and Fairness in Artificial Intelligence

With the rapid development and widespread application of artificial intelligence (AI) technologies, the issues of explainability and fairness have become increasingly prominent. Explainability refers to the ability of AI systems to elucidate the reasons and processes behind their decisions, thereby enhancing transparency and understandability. Fairness involves ensuring that AI systems' decisions are not influenced by biases or discrimination, and that different individuals or groups are treated equitably. Both issues are crucial for the trustworthiness and social acceptability of AI systems.

Traditional AI models, especially deep learning models, are often considered "black boxes," making it difficult to explain their internal mechanisms and decision bases. This limitation not only restricts users' understanding and trust in the system but also complicates debugging and improving the system. To enhance AI explainability, researchers have proposed various methods, such as attention mechanisms, causal reasoning, and rule extraction, to uncover key features and decision pathways of models. Wang, H. et al. (2024) explored how deep learning algorithms like BERT can detect and classify AI-generated text [26]. By analyzing grammatical, semantic, and stylistic features of text, they can identify machine-generated content, thereby maintaining content authenticity and credibility. This adversarial generation and detection process helps improve the robustness and explainability of AI systems.

The fairness issue in AI arises from potential biases and imbalances in training data and algorithm design. For example, if the training data contains fewer samples from certain groups or attributes, or if there is historical bias, the learned model may discriminate against or negatively impact these groups or attributes. To enhance AI fairness, researchers have proposed various methods such as data debiasing, algorithmic fairness constraints, and model post-processing, aimed at eliminating or mitigating biases and disparities in models.

Although progress has been made in the research on explainability and fairness, numerous challenges and issues remain. First, there is often a trade-off between explainability and performance, making it difficult to improve transparency and bias while maintaining model effectiveness. Second, standards and methods for evaluating explainability and fairness are not yet standardized, and different application scenarios and tasks may require different considerations. Additionally, how to integrate explainability and fairness into the design and development process of AI systems, and how to align these with ethical, legal, and societal norms, are also pressing questions that need to be explored.

4 SYSTEM FRAMEWORK AND SYSTEM CALIBRATION

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

4.1 Robotic Arm Grasping System Framework and Hardware Selection

4.1.1 System Requirements

For unordered stacking grasping environments (as shown in Figure 4-1), the robotic arm grasping system needs to automatically perform the following tasks: utilize deep learning algorithms to integrate 2D and 3D visual information for object recognition and segmentation. A curvature-based region-growing segmentation method is used to extract the flat surfaces of target objects. Then, through calibration techniques between the robotic arm and camera, and using system coordinate transformations, the pose information of the objects relative to the robotic arm is calculated. Based on the obtained accurate pose information, the robotic arm actively performs the target grasping operation. Completing this series of tasks will enable the robotic arm to achieve automated grasping in unordered stacking scenarios.

Figure 4-1 Unordered Stacking Scene

4.1.2 System Framework

This paper addresses the problem of object sorting in unordered stacking scenarios by designing a robotic arm grasping system based on an eye-in-hand structure. The system captures RGB images and depth information of target objects using a stereo camera. Improvements are made on the Mask R-CNN network to achieve precise localization and segmentation of the targets. The depth images are used to measure the distance between the target and the camera[27]. Through a hand-eye calibration process, a transformation relationship between the camera coordinate system and the robotic arm coordinate system is established, ultimately calculating the 6D pose information of the target object relative to the robotic arm. Based on the obtained pose information, the robotic arm can accurately perform grasping operations and complete the automatic sorting task of scattered and stacked objects.

4.1.3 RGB-D Camera Selection

The RGB-D camera is capable of simultaneously capturing color images and depth images, using physical methods to measure the distance from the object to the camera, enabling three-dimensional spatial reconstruction and perception. There are several types of RGB-D cameras:

Structured Light Method: Structured light cameras use triangulation principles by projecting known infrared light patterns onto a scene. Depth is calculated based on the received infrared light images. The advantage of structured light cameras is their high accuracy and immunity to ambient light and object texture, making them suitable for close-range measurements.

Time of Flight (TOF) Method: TOF cameras measure object distance using the time-of-flight principle. They emit modulated light pulses and receive the reflected light from the object. Depth information is calculated based on the round-trip time or phase difference of the light, providing complete geometric information of the 3D scene. TOF cameras are divided into indirect TOF (i-ToF), which measures distance by analyzing phase differences of sine or pulse waves, and direct TOF (d-ToF), which measures distance by the time it takes for light pulses to travel.

Line Laser Method: Line laser cameras also use triangulation principles by projecting a laser line onto the object's surface. The image sensor receives the reflected laser line, and depth information is computed based on the position, shape of the laser line in the image, and the geometric relationship between the camera and the laser.

Based on practical requirements, this paper selects the Graphy FS-820 stereo structured light camera (as shown in Figure 4-2), with its performance parameters listed in Table 4-1. This camera can simultaneously capture high-quality color images and depth images, providing reliable perceptual information for object localization and grasping in unordered stacking scenarios.

Figure 4-2 Tuya FS-820

Table 4-1 Performance Parameters

Indicator	Parameter
Working Distance	0.3m-1.4m
FOV (H/V)	66°/44°
Accuracy (X, Y)	4.88mm@700mm
Accuracy (Z)	0.14mm@400mm; 1.53mm@700mm
Depth Resolution	1280*800
RGB Resolution	1920*1080
Data Interface	Gigabit Ethernet (RJ45 Aviation Interface)

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

4.1.4 Robotic Arm Selection

Industrial six-axis collaborative robots are advanced robotic technologies that can safely perform various tasks while sharing the workspace with humans. These robots have six degrees of freedom, allowing for flexible movement and operation in three-dimensional space. Industrial six-axis collaborative robots typically consist of several main parts, including the base, rotation axis, lower arm, upper arm, wrist, and end effector. The rotation axis, lower arm, upper arm, and wrist correspond to the S, L, U, and R axes, respectively, while the wrist also includes the B and T axes. The coordinated movement of these six axes enables precise positioning, posture adjustment, and directional control of the robot in space.

The operation principle of a six-axis collaborative robot arm relies on the coordinated work of driving components such as motors, reducers, and encoders, along with control components like sensors and controllers. The driving components are responsible for moving each joint, while the control components ensure the coordination and precise control of these joints to achieve predetermined motion trajectories and task goals.

In this project, based on actual requirements, we selected the six-axis collaborative robot arm produced by Fairino Company, as shown in Figure 4-3. This collaborative arm offers excellent performance and reliability, capable of meeting our application needs.

FIGURE 4-3 FAIRINO FR3 COLLABORATIVE ROBOT ARM

4.1.5 End Effector

Considering that the workpieces to be handled are primarily flat parts, we decided to use a vacuum suction cup as the robot's end effector, as shown in Figure 4-4. This vacuum suction cup uses an industrial pump as its power source and controls the suction and release processes through an electromagnetic valve. When the valve is opened, the industrial pump extracts air, creating a vacuum between the suction cup and the surface of the workpiece, generating suction force that allows the suction cup to securely grasp the workpiece. When the workpiece needs to be released, the valve switches states, allowing air to re-enter the suction cup, eliminating the vacuum and releasing the workpiece. This vacuum suction cup has a simple structure, is easy to control, and is well-suited for gripping flat parts.

FIGURE 4-4 END FLANGE AND VACUUM SUCTION CUP

4.2 FRAMEWORK AND HARDWARE SELECTION FOR AUTOMOTIVE ONE-WAY CLUTCH DEFECT DETECTION SYSTEM

To meet the requirements for workpiece inspection, we selected an automotive part commonly used in actual production as the inspection object. Based on production inspection standards, we will use a vision system to perform "misassembly and omission" detection to identify and eliminate defective parts.

The one-way clutch is a crucial component in automotive parts, primarily functioning to ensure that the engine's power is transmitted to the transmission in only one direction, thereby maintaining the smooth operation of the vehicle. The one-way clutch typically consists of several key parts, including the housing, ball bearings, springs, and the inner sleeve. Among these, the ball bearings are the key components that can freely roll or lock between the housing and the inner sleeve, enabling the one-way transmission function.

The working principle of the one-way clutch relies on the friction between the ball bearings and the housing and inner sleeve. When the housing rotates clockwise relative to the inner sleeve, the ball bearings are pressed into the grooves between the housing and inner sleeve by the force of the spring, creating a locked state, allowing effective power transmission between the housing and inner sleeve. Conversely, when the housing rotates counterclockwise relative to the inner sleeve, the ball bearings are pushed out of the grooves by the spring, creating a rolling state, preventing power transmission between the housing and inner sleeve, thus achieving the one-way transmission function.

4.3 SYSTEM CALIBRATION

A stereo camera uses the principle of triangulation to obtain depth information of a target object. The basic idea of this principle is to capture the same scene from two different angles using two cameras placed at different positions. By analyzing the disparity between the two images, the distance to the target object can be calculated. This method mimics the way human binocular vision works, where the difference

Journal of Computer Technology and Applied Mathematics
Journal Home: http://jctam.suaspress.org | CODEN: JCTAAD
Vol. 1, No. 4, 2024 | ISSN 3007-4126 (Print) | ISSN 3007-4134 (Online) | ISSN 3007-4142 (Digital)

between the images seen by the left and right eyes is compared to determine the distance of objects.

We can describe the triangulation principle in stereo cameras using mathematical formulas. Let the optical centers of the two cameras be OL and OR, with their image planes being IL and IR, respectively. The baseline length between the two cameras is denoted as b. Suppose the target object P has coordinates (x, y, z) in 3D space, and its projection points on IL and IR are PL and PR, with coordinates (xL, yL) and (xR, yR), respectively. Using the properties of similar triangles, we derive the following relationships:

$$xL / f = x / z \quad (4\text{-}1)$$

$$xR / f = (x-b) / z \quad (4\text{-}2)$$

where f represents the focal length of the cameras. From these two equations, we can solve for the distance z from the target object P to the camera plane:

$$z = b * f / (xL - xR) \quad (4\text{-}3)$$

This formula shows that the depth information z of the target object depends only on the baseline length b, the focal length f, and the disparity (xL - xR). The larger the disparity, the closer the target object is to the camera; conversely, the smaller the disparity, the farther the object is. Therefore, by measuring the disparity between the left and right images, the depth information of the target object can be obtained.

In practical applications, some errors are inevitable in stereo imaging:

Algorithm Factors: The performance and complexity of the matching algorithm directly affect the accuracy and speed of disparity calculation. Matching algorithms are typically divided into local and global methods. Local methods are based on pixel or block matching, offering faster speed but are susceptible to noise and lighting interference, and struggle with low-texture or occlusion areas. Global methods, based on energy minimization optimization, consider overall image consistency and can produce more accurate and smoother disparity maps, but at the cost of higher computational demand and slower speed.

Hardware Factors: The hardware for stereo imaging mainly includes two cameras and a baseline. Parameters such as camera focal length, resolution, distortion, and baseline length all affect the errors in stereo imaging. Generally, greater focal length, higher resolution, lower distortion, and a longer baseline improve stereo imaging accuracy. However, these parameters involve trade-offs; for instance, too large a focal length reduces the field of view, higher resolution increases computational burden, and a longer baseline exacerbates occlusion issues.

Target Object Factors: The characteristics of the target object itself can also affect stereo imaging errors. For example, the distance of the target object from the cameras impacts the disparity. Generally, the farther the object is, the smaller the disparity and the greater the error.

5 INSTANCE SEGMENTATION ALGORITHM BASED ON MASK R-CNN

5.1 MASK R-CNN NETWORK FRAMEWORK

Mask R-CNN is a two-stage neural network architecture, and its workflow can be divided into two main steps. The network first generates a series of candidate regions that may contain target objects, a step known as Region Proposal. In the second step, the network further processes these candidate regions by classifying them, determining the object category to which they belong, and performing regression to refine their position and size for more accurate bounding boxes. Through these two steps, Mask R-CNN effectively detects and segments target objects in images, producing high-quality segmentation masks. Figure 5-1 shows the Mask R-CNN network framework.

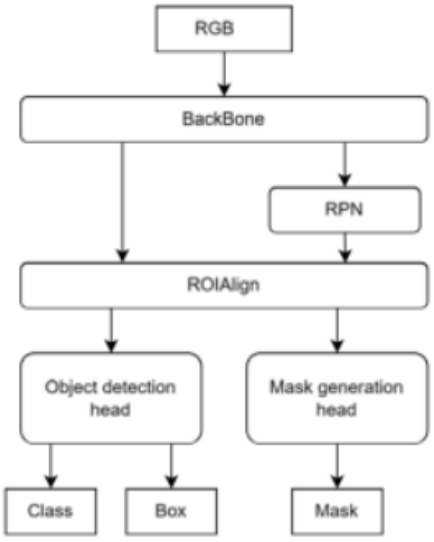

FIGURE 5-1 MASK R-CNN NETWORK FRAMEWORK

5.2 2D/3D FUSION DUAL-MODALITY FUSION MODEL

The early fusion of RGB images and depth images is a simple and effective multimodal data fusion method. It combines RGB images and depth images into a new input data before feeding it into the model. There are two common methods for early fusion. Channel Concatenation: Directly concatenating the three channels of the RGB image with the one channel of the depth image to form a four-channel input. This method retains all the original information but increases the input dimensionality. Weighted Fusion: Performing a weighted sum of the pixel values from the RGB image and the depth image to generate a new grayscale or color image as input. The weights can be fixed or adaptive. This method can highlight certain information but may lose some other information. Encoding Mapping: Depth values are mapped to color space using specific encoding rules, and then fused with

the RGB image to create a pseudo-color image. Common encodings include rainbow colors and heatmaps. This method provides good visualization effects but introduces prior assumptions.

Data resulting from early fusion can be directly input into existing CNNs or other models for training and inference without modifying the network structure. However, early fusion may overlook some associations and complementary information between RGB and depth. As a convenient multimodal fusion approach, RGB-D early fusion can quickly enhance model performance in tasks such as semantic segmentation and object detection, but it is not always the optimal fusion strategy. Effectively leveraging the appearance information from RGB images and the geometric information from depth images is key to RGB-D fusion.

5.3 Dataset Creation

This section introduces an RGB-D instance segmentation algorithm based on Mask R-CNN. The Mask R-CNN has been improved into a dual-modal fusion network by adding a parallel depth branch. The depth image is first re-encoded into an HHA map and then input into the depth branch to extract features. These features are fused with those from the RGB branch after the RoI Align layer, and are used for subsequent classification, bounding box regression, and mask prediction. For unordered stacking scenes, a total of 200 pairs of RGB-D images with random positions, quantities, and stacking states of objects were collected using a Graphy FS-820 camera. The RGB images are color images, while the depth images are obtained through stereo disparity computation. The dataset was semantically annotated using the LabelMe tool. During annotation, attention was given to marking only the visible regions of objects, checking the quality of bounding boxes, and selectively discarding low-quality images based on requirements. Data augmentation was performed on the annotated data, including transformations such as rotation, cropping, exposure changes, and adding noise, to enhance the diversity of the data and improve the model's generalization and robustness. The work in this section prepares the data for subsequent RGB-D instance segmentation model training and evaluation. A well-designed data collection, annotation, and augmentation strategy is fundamental to obtaining high-performance segmentation models. Additionally, specifically improving the segmentation network structure and integrating the appearance information from RGB images with the geometric information from depth images is a key approach to enhancing segmentation accuracy in complex scenes.

6 System Algorithm and Experimental Analysis

6.1 Robotic Arm Grasping System

The RGB-D instance segmentation algorithm based on Mask R-CNN focuses on improving the original Mask R-CNN by transforming it into a dual-modality fusion network, adding a parallel depth branch. The depth images are first re-encoded into HHA images, which are then input into the depth branch to extract features. These features are fused with those from the RGB branch after the RoI Align layer and are used for subsequent classification, bounding box regression, and mask prediction. For unstructured stacking scenarios, 200 sets of RGB-D image pairs with random object positions, quantities, and stacking states were captured using a FS-820 camera. The RGB images are in color, while the depth images are obtained via stereo disparity calculation. Semantic labeling of the dataset was performed using the labelme tool. During labeling, special attention was given to marking only the visible regions of the objects, checking the quality of the bounding boxes, and filtering out low-quality images based on the specific requirements. The labeled data was augmented through transformations such as rotation, cropping, exposure variation, and noise addition to enhance data diversity and improve the model's generalization and robustness.

6.2 One-Way Clutch Defect Detection System

This section will describe the experimental platform setup and software algorithm implementation for the robotic arm grasping system and the "wrong/loose assembly" detection system. In the grasping system, we utilized the 2D/3D multimodal fusion Mask R-CNN network proposed in the previous chapter. By comparing the network performance before and after improvements, we found that the improved network exhibits better adaptability in various environments. We integrated point cloud algorithms to achieve precise localization of target objects and tested the system's positioning capabilities. The results indicate that the system meets actual application requirements.

In the "wrong/loose assembly" detection system, we provide a detailed description of the algorithm process and detection results. By using an industrial camera and lens to form an image acquisition system, we collected images of unidirectional devices with different models and defects on a specific experimental platform. Semantic segmentation was performed on the unidirectional gear, and by comparing combinations of different backbone networks and semantic segmentation networks, we found that using a combination of vgg16 and unet networks achieved the highest segmentation accuracy. Based on this, we located the detection ROIs for the copper sleeve, chamfer, and semicircle according to the gear segmentation results, cropped the images, and created a dataset for subsequent classification detection. Testing different models and sizes of unidirectional devices with the trained model showed that the algorithm has strong generalization ability, with an overall detection accuracy of over 98%, meeting the practical production needs of enterprises.

7 CONCLUSION

This paper proposes an innovative 2D/3D vision fusion robotic arm grasping solution, which uses a stereo camera to identify and locate randomly stacked target objects, providing their 6D pose information and enabling precise robotic arm grasping. To meet the requirements of workpiece quality control, the detection of automotive one-way clutch components was used as a case study. A deep learning model was developed to detect component defects, improving the efficiency of workpiece inspection. A comprehensive system framework was designed, and an experimental platform was built. For unstructured stacking scenarios, the Mask R-CNN network was improved to enhance workpiece segmentation performance. Point cloud processing algorithms were used to determine workpiece pose, enabling robotic arm grasping. For detecting "misassembly" of the one-way clutch, an automated detection system was designed, utilizing segmentation and classification networks to inspect the various parts of the clutch. Experimental results demonstrate that the proposed solution exhibits high robustness, positioning accuracy, and defect detection accuracy, meeting the needs of practical applications.

In future work, we will explore the use of multithreading parallel computing to accelerate the depth map processing, improving system real-time performance. Additionally, coupling pose estimation with deep learning networks will be investigated to simplify the algorithm process and achieve end-to-end predictions. Another direction for development includes using lightweight networks to reduce model parameters and optimize deployment. Moving forward, we will further extend this system to other industrial scenarios, such as part assembly and complex defect detection, providing more intelligent solutions to enhance the level of industrial automation.

ACKNOWLEDGMENTS

The authors thank the editor and anonymous reviewers for their helpful comments and valuable suggestions.

FUNDING

Not applicable.

INSTITUTIONAL REVIEW BOARD STATEMENT

Not applicable.

INFORMED CONSENT STATEMENT

Not applicable.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

CONFLICT OF INTEREST

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

PUBLISHER'S NOTE

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

AUTHOR CONTRIBUTIONS

Not applicable.

ABOUT THE AUTHORS

QU, Meina

City University of Seattle, USA.

REFERENCES

[1] Wang, D. (Ed.). (2016). Information Science and Electronic Engineering: Proceedings of the 3rd International Conference of Electronic Engineering and Information Science (ICEEIS 2016), January 4-5, 2016, Harbin, China. CRC Press.

[2] Wang, C., Chen, J., Xie, Z., & Zou, J. (2024). Research on Personalized Teaching Strategies Selection based on Deep Learning.

[3] Wang, C., Chen, J., Xie, Z., & Zou, J. (2024). Research on Education Big Data for Students Academic Performance Analysis based on Machine Learning. arXiv preprint arXiv:2407.16907.

[4] Chen, Y., Yan, S., Liu, S., Li, Y., & Xiao, Y. (2024, August). EmotionQueen: A Benchmark for Evaluating Empathy of Large Language Models. In Findings of the Association for Computational Linguistics ACL 2024 (pp. 2149-2176).

[5] Li, S., Lin, R., & Pei, S. (2024). Multi-modal preference alignment remedies regression of visual instruction tuning on language model. arXiv preprint arXiv:2402.10884.

[6] Li, S., & Tajbakhsh, N. (2023). Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. arXiv preprint arXiv:2308.03349.

[7] Yang, J., Qin, H., Por, L. Y., Shaikh, Z. A., Alfarraj, O., Tolba, A., ... & Thwin, M. (2024). Optimizing diabetic retinopathy detection with inception-V4 and dynamic version of snow leopard optimization algorithm. Biomedical Signal Processing and Control, 96, 106501.

[8] Zhang, Y., Qu, T., Yao, T., Gong, Y., & Bian, X. (2024). Research on the application of BIM technology in intelligent building technology. Applied and Computational Engineering, 61, 29-34.

[9] Liu, H., Xie, R., Qin, H., & Li, Y. (2024). Research on Dangerous Flight Weather Prediction based on Machine Learning. arXiv preprint arXiv:2406.12298.

[10] Liu, H., Shen, F., Qin, H., & Gao, F. (2024). Research on Flight Accidents Prediction based Back Propagation Neural Network. arXiv preprint arXiv:2406.13954.

[11] Zhang, X., Xu, L., Li, N., & Zou, J. (2024). Research on Credit Risk Assessment Optimization based on Machine Learning.

[12] Huang, D., Xu, L., Tao, W., & Li, Y. (2024). Research on Genome Data Recognition and Analysis based on Louvain Algorithm.

[13] Huang, D., Liu, Z., & Li, Y. (2024). Research on Tumors Segmentation based on Image Enhancement Method. arXiv preprint arXiv:2406.05170.

[14] Liu, H., Li, I., Liang, Y., Sun, D., Yang, Y., & Yang, H. (2024). Research on Deep Learning Model of Feature Extraction Based on Convolutional Neural Network. ArXiv, abs/2406.08837.

https://doi.org/10.5281/zenodo.11124440

[15] Hao Qin, & Li, Z. (2024). Precision in Practice: Enhancing Healthcare with Domain-Specific Language Models. Applied Science and Engineering Journal for Advanced Research, 3(4), 28–33. https://doi.org/10.5281/zenodo.13253336

[16] Lai, S., Feng, N., Sui, H., Ma, Z., Wang, H., Song, Z., ... & Yue, Y. (2024). FTS: A Framework to Find a Faithful TimeSieve. arXiv preprint arXiv:2405.19647.

[17] Li, B., Zhang, X., Wang, X. A., Yong, S., Zhang, J., & Huang, J. (2019, April). A Feature Extraction Method for Daily-periodic Time Series Based on AETA Electromagnetic Disturbance Data. In Proceedings of the 2019 4th International Conference on Mathematics and Artificial Intelligence (pp. 215-219).

[18] Tao, Y., Jia, Y., Wang, N., & Wang, H. (2019, July). The fact: Taming latent factor models for explainability with factorization trees. In Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval (pp. 295-304).

[19] Li, B., Jiang, G., Li, N., & Song, C. (2024). Research on Large-scale Structured and Unstructured Data Processing based on Large Language Model.

[20] Li, B., Zhang, K., Sun, Y., & Zou, J. (2024). Research on Travel Route Planning Optimization based on Large Language Model.

[21] Tao, Y. (2023, August). Meta Learning Enabled Adversarial Defense. In 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE) (pp. 1326-1330). IEEE. Chicago

[22] Liu, M., Jin, Y., Lu, Y., Chen, M., Hou, B., Chen, W., ... & Yu, X. (2016). A wellbore stability model for a deviated well in a transversely isotropic formation considering poroelastic effects. Rock Mechanics and Rock Engineering, 49, 3671-3686.

[23] Tao, Y. (2023, October). SQBA: sequential query-based blackbox attack. In Fifth International Conference on Artificial Intelligence and Computer Science (AICS 2023) (Vol. 12803, pp. 721-729). SPIE.

[24] Hao Qin, & Zhi Li. (2024). A Study on Enhancing Government Efficiency and Public Trust: The Transformative Role of Artificial Intelligence and Large Language Models. International Journal of Engineering and Management Research, 14(3), 57–61. https://doi.org/10.5281/zenodo.12619360

[25] Mo, Y. ., Qin, H., Dong, Y., Zhu, Z., & Li, Z. (2024). Large Language Model (LLM) AI Text Generation Detection based on Transformer Deep Learning Algorithm. International Journal of Engineering and Management Research, 14(2), 154–159.

[26] Wang, H., Li, J., & Li, Z. (2024). AI-Generated Text Detection and Classification Based on BERT Deep Learning Algorithm. arXiv preprint arXiv:2405.16422.

[27] Qu, M. (2024). High Precision Measurement Technology of Geometric Parameters Based on Binocular Stereo Vision Application and Development Prospect of The System in Metrology and Detection. Journal of Computer Technology and Applied Mathematics, 1(3), 23–29. https://doi.org/10.5281/zenodo.13366612