Small Target Detection in Human-Robot Interaction: A Research and Application Analysis

Peng Pan

Beijing 21 Century School, Beijing, China

Abstract. The advancement of computer vision and robot vision techniques has revolutionized human-robot interaction by enabling more precise detection of small targets, particularly in challenging environments. This paper presents a comprehensive analysis of remote gesture recognition systems that address the limitations of traditional visual inspection methods, such as interference from cluttered backgrounds, low-resolution inputs, and partial occlusions. A detailed investigation of state-of-the-art algorithms—including Faster R-CNN, Mask R-CNN, SSD, RetinaNet, and YOLO variants—reveals persistent challenges in feature loss, scale sensitivity, and computational inefficiency during gesture detection in complex scenarios. The research explores innovative strategies that integrate multi-scale progressive fusion methods, such as the Adaptive Feature Pyramid Network (AFPN) and adaptive spatial fusion techniques, to enhance detection performance. Enhancements to YOLO-v8 combined with a shape-IoU loss function further improve the accuracy and robustness of small gesture recognition. The approach employs advanced deep learning, sensor fusion, and edge computing techniques to refine the gesture recognition process, from data acquisition by various sensors—including monocular cameras, multi-ocular cameras, and depth sensors—to sophisticated preprocessing, segmentation, and feature extraction. Comparative experimental results indicate that high-precision models like Mask R-CNN deliver superior accuracy, while optimized lightweight frameworks such as YOLO-v8 achieve real-time performance at 30+ frames per second, making them highly suitable for dynamic applications.

1 Introduction

The rapid advancement of computer vision and robot vision techniques has paved the way for transformative changes in human-robot interaction, enabling systems to perform complex tasks with higher precision in challenging environments. The detection of small targets, particularly in gesture recognition, has emerged as a critical area of research due to its significant impact on application domains such as medical robotics, industrial automation, and consumer devices [1-3]. Given the increasing reliance on remote gesture recognition for

panpeng@uok.edu.gr

operational control and communication in these fields, addressing the limitations of traditional visual inspection systems is imperative.

Conventional approaches often encounter difficulties when processing images with complex background interference, low-resolution inputs, and partial occlusions, which can lead to information loss and misinterpretation of subtle hand movements [4-5]. Such challenges are especially prominent in scenarios where even minor inaccuracies can trigger significant consequences, for instance, the inadvertent activation of surgical tools or erroneous commands in automated industrial processes. This necessitates the exploration of innovative methods that can effectively overcome these obstacles while maintaining high detection accuracy and real-time performance.

Recent developments in deep learning have spurred the emergence of robust models such as Faster R-CNN and Mask R-CNN, which combine convolutional networks, regional proposal networks, and region-of-interest pooling to achieve reliable detection. Nevertheless, despite their proven strengths, these models still struggle with small target detection due to issues related to feature loss, scale sensitivity, and computational inefficiency. To overcome these deficits, advanced techniques incorporating multi-scale progressive fusion have been proposed. For example, the Adaptive Feature Pyramid Network (AFPN) and adaptive spatial fusion (ASF) methods have demonstrated the potential to harness multi-level feature information effectively, thereby significantly enhancing detection performance while mitigating computational overhead.

In parallel, lightweight detection frameworks like the improved YOLO-v8 have been developed to provide a balance between precision and speed. By integrating specialized loss functions such as the shape-IoU loss, these frameworks can achieve accurate small gesture recognition in real time, even under adverse conditions. Moreover, the deployment of diverse sensor technologies—including monocular cameras, multi-ocular cameras, and depth sensors—further enriches the data acquisition process. This multimodal sensor fusion not only improves the quality of input data but also bolsters the subsequent processes of noise reduction, segmentation, and feature extraction, thereby enhancing overall system robustness.

The primary objective of this study is to critically analyze the existing limitations inherent in small target detection and to propose targeted algorithmic optimizations coupled with strategic hardware integrations. Emphasis is placed on bridging the gap between high-precision but computationally intensive detection models and real-time, resource-efficient frameworks. Comparative experimental evaluations are conducted using standard metrics such as Average Precision (AP), accuracy, and processing speed, which demonstrate that advancements in both theoretical designs and practical implementations are essential for elevating the performance of human-robot interaction systems. Ultimately, this research lays the groundwork for safer, more efficient, and highly adaptive gesture recognition systems, thereby redefining the boundaries of modern human-machine collaboration while addressing the persistent challenges associated with small target detection.

2 Relevant Theories

2.1 Overview of Computer Vision and Robot Vision

2.1.1 Faster R-CNN

Faster-R-CNN mainly consists of four parts: Convolutional network layer, (RPN), ROI Pooling layer and classifier. It can achieve fast and accurate detection of the target. And it's based on Fast R-CNN, an object detection method is proposed after introducing Regional Suggestion Network (RPN) [6].

2.1.2 Mask R-CNN

It's a deep neural network model, which performs well in the task of identifying and segmenting objects in a single image. It also comes up with the advantages of high processing speed and high precision, it is a very flexible framework, which can add different branches to complete different tasks, such as target classification, target detection, semantic segmentation, instance segmentation and human posture recognition [6].

2.1.3 Choose two (table comparison)

The speed is 71 frames per second, ranking third among all comparison methods, as shown in Table 1. Through comparative experiments, it can be verified that the method proposed in this paper has good performance.

Table 1 Comparative experimental results on the Egolands dataset (bolded for the best effect)

Method	AP[IoU=0.50:0.95]	Acc(%)	Speed(frame/s)
Method	0.426	92.7	5
Faster R-CNN	0.392	93.4	63
SSD	0.421	94.6	14
RetinaNet	0.445	97.8	43
YOLOv5	0.446	93.3	87
HandDetNet	0.427	96.1	76
Mask R-CNN	0.437	95.1	14
RestNet 101-FPN	0.449	97.2	71
This article's method	0.426	92.7	5

2.2 Small Target Detection Techniques

2.2.1 The first improvement method

AFPN: AFPN is an improved feature pyramid network designed to significantly improve target detection performance through multi-level design and adaptive feature fusion mechanism. The advantages of the AFPN module are not only confined to the efficient extraction and fusion of features but also extend to its adaptability across different network depths. By effectively leveraging multi-scale information, the AFPN module provides robust support for YOLOv8's object detection tasks in complex scenarios, enabling it to excel when dealing with challenges at various scales. This flexibility and efficiency allow YOLOv8 to maintain detection accuracy while reducing computational resource requirements, significantly enhancing its practicality [7].

2.2.2 The second improvement method

Shape-IoU loss function: It can improve the accuracy and robustness of multi-scale gesture recognition, especially small gesture recognition [8].

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \qquad (1)$$

$$\omega\omega = \frac{2 \times (\omega^{gt})^{scale}}{(\omega^{gt})^{scale} + (h^{gt})^{scale}} \tag{2}$$

$$hh = \frac{2 \times (h^{gt})^{scale}}{(\omega^{gt})^{scale} + (h^{gt})^{scale}} \tag{3}$$

$$d^{shape} = hh \times \frac{(x_c - x_c^{gt})^2}{c^2} + \omega\omega \times \frac{(y_c - y_c^{gt})^2}{c^2} \tag{4}$$

$$\Omega^{shape} = \sum_{t=w,h} (1 - e^{-w_t})^{\theta}, \theta = 4 \tag{5}$$

$$\omega_{\omega} = hh \times \frac{|\omega - \omega^{gt}|}{\max(\omega, \omega^{gt})} \tag{6}$$

$$\omega_h = \omega\omega \times \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{7}$$

2.3 Gesture Recognition Systems

The complete gesture recognition process, including sensor technology data and processing gesture detection feature extraction gesture classification, and finally human-computer interaction. The complete process requires data collection and subsequent processing.

2.3.1 Sensor

For sensor technology, there are monocular cameras, multi-ocular cameras and depth sensors, depth sensors are also more popular and well-known in recent years, depth sensors are also called non-three-dimensional depth sensing equipment, it is more than the traditional sensor capture speed, accuracy and so on. Legend also includes data gloves wireless models and so on they can be more accurate to achieve medium or long distance transmission, such as data gloves, its induction speed will be very fast, positioning is more accurate, and wireless signals it can avoid contact with the human body to achieve long-distance detection.

2.3.2 Data processing

Data processing refers to the subsequent calculation and processing of the detected and received data. For example, the acquired data needs to be processed in batches first, such as the detection of whether the picture is a high-definition picture without occlusion. If there is such a problem, it needs to be processed first, such as sampling, noise reduction, enhancement and so on. Noise reduction allows the picture to remove some noise, enhancement can improve its contrast, magnify key information and other functions.

2.3.3 Gesture detection

Gesture detection refers to the segmentation, gesture analysis and gesture recognition of processed information. Segmentation is also the most critical step, which will affect the accuracy of the next analysis and recognition. Segmentation methods also include gesture segmentation based on monocular visual gesture style or stereo vision. Gesture analysis is to obtain the shape features or motion trajectories of gestures [9,10]. The main methods of gesture analysis include edge contour extraction, centroid finger, multi-characteristic combination, and knuckle tracking. Gesture recognition is the process of classifying tracks in model parameters into a subset of space.

3 Challenges in Small Target Detection

3.1 Cluttered Backgrounds

In many human-computer interaction scenarios, the environments in which gestures must be detected are far from ideal. Cluttered backgrounds introduce a significant challenge because they involve a variety of dynamic elements that change over time. Variations in light intensity, reflective surfaces, and the presence of shadows contribute to a noisy background that can obscure the fine details of small targets such as hand gestures. For example, industrial settings might include multiple moving objects, such as robotic arms or conveyors, that generate reflective surfaces and variable lighting conditions. These factors result in blurred boundaries between the target and the background, making it difficult for segmentation algorithms to distinguish the hand from its surroundings accurately.

The presence of ambient noise in the environment further complicates this issue. Shadows created by uneven lighting or sudden changes in light intensity can alter the appearance of key features, such as edges and contours, which are essential for gesture detection. Reflective surfaces may produce highlights that mimic the features of a hand, leading to false positives or misclassification. Adaptive algorithms must therefore incorporate robust methods to filter out these interferences. Techniques like background subtraction, dynamic thresholding, and enhanced edge detection have been explored, yet they often struggle to achieve consistent accuracy under rapidly changing conditions. Improving performance in cluttered backgrounds remains a significant area of research, requiring the development of algorithms that can intelligently differentiate between relevant gesture features and irrelevant background noise.

3.2 Low-Resolution Inputs

Low-resolution inputs are another formidable challenge in small target detection. High-resolution imagery is crucial for capturing detailed information, yet many practical systems are constrained by the use of cameras or sensors that provide only limited spatial resolution. This situation is common in low-cost devices or scenarios where cameras are positioned at a considerable distance from the subject. When resolution is compromised, the pixel density of the target decreases, which in turn reduces the amount of detailed data available for feature extraction. Small gestures, which rely on subtle variations in pixel intensity to convey critical information, become increasingly difficult to detect under these conditions.

The reduction in spatial resolution directly impacts the ability to accurately localize and classify key points of the hand. In situations where high precision is required, such as in medical robotics or fine-grained gesture control, the insufficiency of pixel data can lead to erroneous interpretations of a gesture. Methods such as image interpolation or super-resolution techniques have been developed to counteract low resolution; however, they introduce additional computational complexity and often cannot fully recover the lost detail. Consequently, optimizing detection algorithms to perform reliably on low-resolution inputs remains a critical research focus, as improvements in this area directly translate to enhanced operational accuracy and safety in real-world applications.

3.3 Partial Occlusions

Partial occlusions present a unique set of challenges in small target detection, particularly when key parts of the hand are hidden from view. In many practical scenarios, the target gesture may be partially obscured by obstacles such as clothing, tools, or even other hands. For instance, in a collaborative work environment, the overlapping of multiple users' hands can lead to significant information loss regarding the intended gesture. In medical applications, instruments or equipment might block a surgeon's hand, making it challenging to capture the entire gesture accurately. Such occlusions result in incomplete feature sets, making it difficult for detection algorithms to maintain high levels of precision in identification and classification.

The complexity of handling partial occlusions lies in the need for sophisticated interpretation of incomplete data. Traditional algorithms often fail to compensate for the missing details, resulting in misidentification or reduced confidence in the detection outcome. To address this issue, researchers have explored multi-view and multi-sensor fusion techniques, which combine information from various angles or modalities to reconstruct a more complete representation of the gesture. Machine learning approaches, particularly those employing deep neural networks, have also shown promise in learning robust representations that can tolerate partial occlusions. Nevertheless, achieving a consistently high detection rate in the presence of occlusions is an ongoing challenge that requires further innovation in feature reconstruction and robust classification methods.

3.4 Limitations of Traditional Visual Inspection Systems

Traditional visual inspection systems, while effective in controlled conditions, encounter substantial limitations when applied to the detection of small targets in dynamic and uncontrolled environments. These systems typically rely on conventional image processing techniques that assume a relatively static background and high-quality input images. As discussed in earlier sections, real-world scenarios rarely meet these ideal conditions. For instance, overlapping objects, dynamic changes in lighting, and fluctuations in sensor quality all contribute to performance degradation. Traditional systems also lack the adaptive capabilities required to handle the diversity and unpredictability inherent in human-robot interaction scenes.

One specific limitation is the inability of classical algorithms to adapt to rapid changes in the environment. Many traditional methods depend on predetermined thresholds and static models that do not account for real-time variations in input conditions. This rigidity leads to increased susceptibility to errors such as false detections or missed gestures. Additionally, these systems often struggle with the task of distinguishing between important features and background noise when the target is as small as a subtle hand movement. Enhanced detection performance in such cases requires dynamic calibration and adaptive learning techniques, which are often absent in older models.

3.5 Limitations of Traditional Visual Inspection Systems

Partial occlusion includes missing key information when the hand is partially occluded by clothes, other objects and tools, etc., which will lead to identification errors. For example, in the medical scene, some instruments block the doctor's hand movements. Or in a collaborative scenario, the hands of different users overlap each other.

4 Application Analysis: Remote Gesture Recognition

Remote gesture recognition, as a revolutionary technology for human-computer interaction, is moving from the laboratory to a diversified reality scene.

- In terms of smart home control, the demand is for fine gestures of about 3-5 meters, and it is necessary to solve the scale reduction and light (reflective) interference caused by distance for small targets. It can be applied to the intelligent central control system to control lights and curtains through gestures.
- In-car gesture interaction: stably recognize the driver's single-finger commands (such as volume adjustment, map scaling) under complex light and shadow environments in the cockpit (such as direct sunlight, reflection from the instrument panel at night). When the driver can't easily distract himself from controlling the display or the buttons don't satisfy all the shortcuts.
- Medical remote control: The surgeon remotely controls the surgical robot by gesture, which needs to identify the finger joint Angle with sub-millimeter accuracy. Mask R-CNN can be used in combination with key point detection branch (HRNet architecture) and multi-modal calibration: The RGB image is fused with the Leap Motion sensor data to build a 3D pose model of the hand, and the Angle error is reduced.

5 Conclusion

This study has provided an in-depth analysis of small target detection techniques within the realm of human-robot interaction and remote gesture recognition. The research critically examined traditional visual inspection methods, which often suffer from significant limitations such as cluttered backgrounds, low-resolution inputs, and partial occlusions, resulting in considerable challenges in accurately detecting subtle hand movements. By comparing state-of-the-art algorithms—including Faster R-CNN, Mask R-CNN, SSD, RetinaNet, and various YOLO versions—a number of persistent issues, such as feature loss, scale sensitivity, and computational inefficiency, have been identified. These challenges underscore the pressing need for innovative solutions that can overcome the shortcomings of conventional methods while maintaining high detection precision.

The proposed approach in this work emphasizes advanced strategies such as multi-scale progressive fusion techniques. In particular, the integration of an Adaptive Feature Pyramid Network (AFPN) and adaptive spatial fusion mechanisms has shown considerable promise in enhancing detection performance. These strategies effectively leverage multi-level feature information to improve the precision of target detection, thereby mitigating the negative impact of complex and dynamic environments. Furthermore, the adoption of a shape-IoU loss function in enhanced YOLO-v8 models contributes to the improvement of both accuracy and robustness in recognizing small gestures in real time.

References

1. Wu, X., Ma, Y., Zhang, S., et al.: 'Yo3RL-Net: A fusion of two-phase end-to-end deep net framework for hand detection and gesture recognition', Alexandria Engineering Journal, 2025, 12, pp. 177-89
2. Zheng J, Yang C, Zhang T, et al. Dynamic Spectral Graph Anomaly Detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39(12): 13410-13418.
3. Gao M, Sun J, Li Q, et al. Towards trustworthy image super-resolution via symmetrical and recursive artificial neural network[J]. Image and Vision Computing, 2025: 105519.

4. Zhu, X., Liu, Z., Cambria, E., Yu, X., Fan, X., Chen, H., Wang, R.: 'A client–server based recognition system: Non-contact single/multiple emotional and behavioral state assessment methods', Comput. Methods Programs Biomed., 2025, 260, pp. 108564.
5. Wang, R., Zhu, J., Wang, S., Wang, T., Huang, J., Zhu, X.: 'Multi-modal emotion recognition using tensor decomposition fusion and self-supervised multi-tasking', Int. J. Multimed. Inf. Retr., 2024, 13, (4), pp. 39.
6. Wang, F., Ju, M., Zhu, X., Zhu, Q., Wang, H., Qian, C., Wang, R.: 'A Geometric algebra-enhanced network for skin lesion detection with diagnostic prior', J. Supercomput., 2025, 81, (1), pp. 1-24.
7. Zhao, Z., Zhu, X., Wei, X., Wang, X., Zuo, J.: 'Application of Workflow Technology in the Integrated Management Platform of Smart Park', IEEE Adv. Inf. Manag., Communicates, Electron. Automat. Control Conf., 2021, 4, pp. 1433-1437.
8. Zhang, Y., Zhao, H., Zhu, X., Zhao, Z., Zuo, J.: 'Strain Measurement Quantization Technology based on DAS System', IEEE Adv. Inf. Manag., Communicates, Electron. Automat. Control Conf., 2019, pp. 214-218.
9. H., Chen, X., et al.: 'Fusion progressive spatial pyramid YOLOv8 mango target detection algorithm research', Laser & Optoelectronics Progress, 2025, 1-15
10. Shi, H., Liu, M., Du, W., et al.: 'Lightweight gesture recognition algorithm in complex indoor environment', Foreign Electronic Measurement Technology, 2024, 43, (11), pp. 27-34