The Robot Visual Servoing Based on Transformer Model

Siting Sheng

{ gypsophila@stu.xjtu.edu.cn}

Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

Abstract. With the continuous development of industry, the ability of robot visual servoing to complete generalization tasks in complex or unknown environments has gradually become one of the focuses. However, since the Transformer model's introduction, it has shown great potential in many application fields, especially in improving generalization ability and performing global sequence calculations, where it has excellent applications. This paper starts by analyzing the characteristics and development history of the Transformer model technology and the robot visual servoing, classifies the visual servoing, and discusses the problems it is facing. It also elaborates on the contributions, applications, and problems solved by scientists when integrating the two technologies, thereby future enhancing the performance of the robot visual servoing. This paper systematically classifies and summarizes various studies on robot visual servoing based on the Transformer model, and also summarizes the current challenges and future development directions in improving generalization ability. This paper aims to provide valuable insights and references for future research and development in the field of robot visual servoing and intelligent control.

Keywords: Robot, Visual Servoing, Transformer Model.

1 Introduction

Since its inception, the Transformer model has not only achieved revolutionary success in its initial application in the field of natural language processing, but has also been successively applied in areas such as computer vision and multimodal learning, demonstrating strong application potential. Visual servoing technology, as a connection between perception and control, has gradually expanded its application from laboratories to more diversified scenarios since its concept was proposed in the last century, and has greatly promoted the development of the robotics field. However, traditional robot visual servoing relies on precise system modeling and local visual information. In the open, complex, and unfamiliar unstructured real world, it is often difficult to accurately and safely complete generalization tasks. How robots can quickly understand unfamiliar environments and planned actions has become the core of current research.

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364829

It is worth noting that the core of the Transformer model - the self-attention mechanism can achieve long-distance sequence modeling or global modeling, and has strong generalization ability and few-shot learning ability [1]. This feature is applicable to addressing the excessive reliance on the accuracy of spatial system modeling in visual servoing and the understanding and prediction of continuous-time images. When the Transformer model is applied to the robot visual servoing, it is expected to overcome the reliance of traditional methods on short-term and local feature information, thereby significantly improving the robustness and accuracy of the robot visual servoing in complex environments. The integration of the two is an inevitable trend in the development of intelligent robot vision.

Against this backdrop, this paper aims to systematically explore the classification, application potential, and potential risks and challenges that may be encountered in the current stage of visual servoing systems based on the Transformer model. This article will firstly sort out the classification framework and conduct an in-depth analysis of the characteristics and advantages of different technical routes. Furthermore, explore the huge potential in the field of robot applications and objectively analyze the possible challenges it may face in practice. Ultimately, through comprehensive comparison and outlook, it is demonstrated that the visual servo technology based on Transformer has significant value and development prospects in enhancing the performance, generalization ability and intelligence level of robot systems.

2 Transformer Model and Robotic Visual Servoing

2.1 Traditional Robotic Visual Servoing

The visual servo system, when applied in the field of robotics, is a core technology that connects robots with the environment. The visual servo system perceives the external environment through sensors, acquires image information and calculates control signals in real time, driving the robots to take corresponding actions and forming a closed-loop feedback system [2].

Image-based visual servoing takes the two-dimensional image plane as visual information, processes data using the Jacobi matrix, and continuously improves accuracy by minimizing errors. However, this technique has high requirements for the control law of minimizing errors and cannot plan a smooth behavioral paradigm. It can only correct the next behavior within a short period of time [2].

Position-based visual servoing combined with modeling of known target images, taking three-dimensional pose as the information input, and adjusting the robot behavior by calculating the error between the current three-dimensional pose and the expected pose in the virtual system modeling, is easy to control. However, this technology has high requirements for the accuracy of system modeling and can only be applied in known environments, with low stability [2].

Malis et al. proposed the renowned 2.5D visual servoing method. By extending the decoupling control of the image and the partial displacement estimation based on the homography matrix, it avoids the complex control law calculation and the pose difference caused by calibration error [3].

However, traditional visual servoing methods still face many problems, such as the need for manual extraction of measurement environment information, weak generalization ability in unknown environments, and weak adaptability to dynamic environments. There are still many technical challenges waiting to be overcome Transformer model.

2.2 Transformer Model

The core of the Transformer model is the self-attention mechanism, which abandons the loop and convolutional structures, endowing it with powerful sequence modeling capabilities and global information processing capabilities [1]. Its multi-head attention mechanism and other features enhance its own modeling and representation capabilities, and it has great potential for application in the field of visual servoing.

Dosovitskiy et al. proposed the concept of Vision Transformer (ViT) in 2020, and for the first time proved that the pure Transformer model could surpass the then most advanced CNN Model in image classification applications. It laid the foundation for modeling images as sequences and is the basis for the application of the visual Transformer model [4].

Carion et al. proposed DETR in 2020, applying the Transformer Model to the field of object detection to achieve end-to-end detection, demonstrating the development potential of the Transformer Model in handling set prediction problems, and it is applicable to the modeling problem of multiple objects in visual servoing systems [5].

The Robotics Transformer 1(RT-1) established by Brohan et al. in 2022 demonstrated on a large scale for the first time that the pure Transformer architecture model can be applied as an efficient end-to-end visual motion strategy in the field of robotics. It demonstrates a significant improvement in the generalization ability of RT-1 in diverse task scenarios [6].

Robotics Transformer 2(RT-2) established by Brohan et al. in 2023 successfully transformed the visual-language model (VLM) pre-trained on Internet-scale data into efficient robot control strategies for the first time based on RT-1. It also marks a revolutionary breakthrough for RT-2 in semantic understanding and reasoning capabilities [7].

The Transformer Model has been applied in fields such as visual servoing and robotics, demonstrating great potential in task generalization and enhancing robustness. It helps to improve the behavioral paradigm capabilities of robot visual servoing systems and enrich their application scenarios.

3 Classification

3.1 Classify by Task Types

People can divide the current visual servoing tasks into two major categories, namely the known task object environment and the unknown task object environment.

The traditional location-based visual servo system, due to the fact that its operation requires spatial modeling of environmental objects, can only be applied in the environment of known task objects, with high limitations. It is mostly used in simple and structured tasks. This kind of model has accurate object pose information, and thus often has extremely high accuracy and performs well in simple and structured tasks.

Image-based visual servo systems need to obtain photos of the surrounding environment and perform recognition during task operation. When applied in a known environment, they often have good performance due to a thorough understanding of the environment. However, when obtaining image information in an unknown environment, they often have limitations and it is difficult to have a comprehensive understanding, thereby reducing the accuracy of task completion.

With the development of industry, the application of robots in unknown environment tasks has gradually become a new focus of robot visual servo tasks. Completing generalization tasks in an environment of unknown task objects has also become an important evaluation index for the performance of robot visual servo systems.

In traditional robot servo systems, position-based ones are difficult to apply, and image-based ones have certain limitations. People obtain environmental information through different sensing methods and also introduce deep learning and various large models to continuously enhance the generalization ability of robot visual servo systems to complete training tasks. The introduction of deep learning and large models is an inevitable choice for the development of the industrial age and has also brought new improvements to the robot visuual servo system.

3.2 Classified by Learning Paradigms

In robot learning, imitation learning is the most fundamental learning paradigm. Robots, through behavior cloning, observe and imitate the behaviors of experts, thereby obtaining action instructions and parameters and completing basic behaviors. However, minor errors in the observation of imitative behavior may lead to significant deviations in the final behavior. Moreover, robots cannot understand the specific meaning of the behavior, lack generalization ability, and have limited capacity to handle tasks.

Reinforcement learning is a process in which robots continuously interact with the environment, receive feedback from the environment, and constantly improve and optimize their behavior to reduce errors. During the process, the robot continuously reduces deviations through task training, stabilizes the behavior of the robot system, and enhances the working performance of the robot.

Self-supervised learning is usually pre-training. Before the actual task training, a large amount of task data is learned in the pre-task to understand the model and the characteristics of the task scenario, and the behavior is improved through prediction. After establishing a certain initial cognitive model, a large amount of training is invested. In self-supervised learning, robots can acquire a large amount of general knowledge, have a basic understanding of the environment, and provide underlying information input for higher-level decision-making.

4 Application and Implementation Scenarios

4.1 Different Application Scenarios of Rrobot Visual Servo

The robot visual servo system originated in industry and is the most mature field of this technology. The robot visual servo system has been maturely applied in grasping and assembly, and has already created certain industrial value. Peng et al. Conducted an in-depth comparative study on the performance of position-based and image-based visual servos in the assembly tasks of large composite panels in 2020. The comparison concluded that the two models have different advantages in different industrial assembly tasks and can better achieve the corresponding experimental goals [8].

The robot visual servo system also has certain applications in medical robots. The application of robot visuual servo systems in medical imaging fields such as endoscopes and laparoscopes demonstrates greater robustness compared to basic model-based or electromagnetic tracking methods in dynamic environments, achieving a progress from "completely manual" to

"automatic" [9]. Saeidi et al. developed the STAR system, applying the visual servo system to medical robots to track positions in real time and plan paths. It even has the potential to surpass human experts in complex soft tissue surgeries, providing a strong technical basis for the autonomous treatment of future medical robots [10].

In fields such as human-machine interaction, service robots and agricultural robots that interact with the planting environment, visual servo systems are also applied. They help robots better identify the environment and provide feedback in their application areas, thereby improving work efficiency and better fulfilling work tasks.

4.2 Robot Visual Servoing System Based on Transformer Model

The RT-1 and RT-2 models established by Brohan et al. Apply the Transformer model to the robot system to solve the problems of the weak generalization ability of robots and the high cost of database collection.

RT-1 does not directly apply the Transformer architecture. It is applied via FiLM layers to condition an EfficientNet to pretrained tasks with a USE embedding. It results in the vision-language tokens reduced by the TokenLearner and fed into a decoder-only Transformer, which outputs tokenized actions. This reduces the computational complexity and meets the requirements of real-time control of robots. RT-2 is an extension of the VLM large model, enhancing the robot's semantic processing capabilities, building a single-end end-to-end model, and simultaneously improving generalization ability and semantic reasoning ability.

RT-1 is an imitation learning method that performs the "behavior" in the training data upon receiving "instructions". It is a standard behavior clone and has achieved a task success rate of up to 92% in the seen tasks, which is higher than the 75% data of other baseline models. It demonstrates strong anti-interference ability, stability and robustness, but it can only generalize "known skills and known objects". Unable to learn new movements by oneself, and completing tasks mainly relies on basic operations, with insufficient dexterity and accuracy.

RT-2, on this basis, introduces the pre-trained VLM to enhance the robot's semantic understanding ability, expand the recognition range of instructions, and further improve the generalization ability of the robot's task processing. It demonstrates comparable capabilities to RT-1 in unknown environment already seen tasks, and achieves an average success rate of 62% in tasks with unknown objects. It is approximately twice as much as RT-1 (32%) and performs better in complex generalization scenarios. And it has a more excellent semantic understanding ability, with an average index of 60%, far exceeding the 17% of RT-1. However, it is still impossible to expand new behaviors. Only the existing skills in the learning library can be applied, and the reasoning frequency is low, making it difficult to implement dynamic visual servo applications for robots. At the same time, it is highly dependent on closed-source VLMs, and more open-source VLMs need to be developed to lower the application threshold as Table 1.

Table 1. VC-1、RT-1 and RT-2[7]

Model	Seen Task	Unseen Average	Language Understanding
VC-1	63	10	11
RT-1	92	32	17
RT-2-PaLI-X-55B	92	62	60

5 Challenges and Future Developments

5.1 Challenges

The dynamic computational nature of large Transformer models can lead to the same input being directed to completely different outputs, which is unstable. Moreover, large models sometimes generate AI illusions during operation, and the input of interfering information can cause deviations in large models. All these will have a certain impact on the stable application of robot visual servoing.

In addition, in human-computer interaction, robots often need to provide immediate feedback. However, the training time of Transformer models is long, and it takes a considerable amount of time to train a model that can be stably trained. Moreover, after the training is completed, it still requires a certain amount of computing time, making it difficult to improve the performance of robots.

The sensor data of robots is multimodal, including sound, image, electromagnetic and other signals. Integrating and calculating the multimodal data of robots will increase the calculation sequence of the Transformer model, further increase the calculation time, and the instability will increase accordingly.

In conclusion, the practical application of the Transformer model in the field of robot visual servo still faces considerable challenges.

5.2 Future Development Direction

There are still many directions that can be explored in the future when applying the Transformer model to the field of robot visual servo.

First of all, people should develop the Transformer large model itself, enhance the stability of model computation, and improve the computational performance of the model, and research and seek large models with higher performance.

Meanwhile, when people apply the Transformer model to the field of robot visual servo, people need to continuously refine and adjust it during task training, so as to integrate and complement the advantages of both.

Finally, in the field of robot visual servo, people will gradually realize the process from model establishment to task training, and ultimately apply it to real scenarios to solve practical problems and improve the efficiency of production and life.

6 Conclusion

This paper first introduces the characteristics of the robot visual servo system, such as weak generalization ability and difficulty in application in unfamiliar and complex environments. Then, it presents the features of the Transformer model, which has strong generalization ability and global sequence modeling ability. Subsequently, it elaborates on the problems faced by classifying the robot visual servo system in detail. The effect of applying the Transformer model to the robot visual servo system is introduced. Finally, the challenges such as computing power faced by the robot visual servo system based on the Transformer model and the potential for solving problems in future development are proposed. In conclusion, this study provides useful insights and references for future research and technological development in the field of robot

visual servoing, helping to promote continuous innovation and broader applications of intelligent control technologies in robotics.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. Advances in Neural Information Processing Systems 30, 1–11 (2017)

[2] Hutchinson, S., Hager, G.D., Corke, P.I.: A tutorial on visual servo control. IEEE Transactions on Robotics and Automation 12(5), 651–670 (2002)

[3] Malis, E., Chaumette, F., Boudet, S.: 2½ D visual servoing. IEEE Transactions on Robotics and Automation 15(2), 238–250 (2002)

[4] Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al.: An image is worth 16×16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

[5] Carion, N., Massa, F., Synnaeve, G., et al.: End-to-end object detection with transformers. In: European Conference on Computer Vision (ECCV 2020), pp. 213–229. Springer, Cham (2020)

[6] Brohan, A., Brown, N., Carbajal, J., et al.: RT-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)

[7] Zitkovich, B., Yu, T., Xu, S., et al.: RT-2: Vision-language-action models transfer web knowledge to robotic control. In: Conference on Robot Learning (CoRL 2023). PMLR, pp. 2165–2183 (2023)

[8] Peng, Y.C., Jivani, D., Radke, R.J., et al.: Comparing position- and image-based visual servoing for robotic assembly of large structures. In: IEEE 16th International Conference on Automation Science and Engineering (CASE 2020), pp. 1608–1613. IEEE, Piscataway (2020)

[9] Azizian, M., Khoshnam, M., Najmaei, N., et al.: Visual servoing in medical robotics: A survey. Part I: Endoscopic and direct vision imaging – techniques and applications. The International Journal of Medical Robotics and Computer Assisted Surgery 10(3), 263–274 (2014)

[10] Saeidi, H., Opfermann, J.D., Kam, M., Wei, S., Léonard, S., Hsieh, M.H., Krieger, A.: Autonomous robotic laparoscopic surgery for intestinal anastomosis. Science Robotics 7(62), eabj2908 (2022)