Journal of Computers Vol. 34 No. 4, August 2023, pp. 119-126
doi: 10.53106/199115992023083404010

A Method for Assembly Accuracy Detection and Intelligent Error Estimation Based on Computer Vision

Dan-Dan Cui[1], Chao Xu[2], Hong-Chao Zhou[3*]

[1] School of Automation Engineering, Tangshan Polytechnic College,
Tangshan City 063299, Hebei Province, China
[2] Tangshan Caofeidian Shiye Port Co., Ltd,
Tangshan City 063200, Hebei Province, China
[3] Tangshan Polytechnic College,
Tangshan City 063299, Hebei Province, China
{cuidandan2018, xuchao2021, Hongchao20170203}@126.com

Received 1 July 2023; Revised 20 July 2023; Accepted 25 July 2023

Abstract. This article focuses on the current situation of large assembly errors, easy omissions and errors in the mechanical assembly process. Computer vision is introduced in the assembly process, and visual images are used to estimate assembly errors, thereby improving assembly accuracy. To this end, through improvements to the neural network, the addition of attention and measurement mechanisms, the network's ability to extract and distinguish features from assembly images has been improved. Finally, deep learning algorithms are used to estimate assembly features in the image. Finally, simulation experiments have shown that the algorithm proposed in this paper can achieve 94.7% improvement in assembly accuracy and error estimation accuracy.

Keywords: mechanical assembly, assembly error, error estimation, deep learning

1 Introduction

The assembly process of mechanical products has the characteristics of multiple operational links and complex assembly processes, which can easily lead to errors such as missed or incorrect assembly. In the process of assembling complex assembly components, failure to timely detect whether newly assembled components are correctly assembled can affect the quality and assembly efficiency of mechanical products.

Currently, commonly used error detection methods include laser interferometer, ball and rod instrument, R-test, etc. However, such measurement equipment has problems such as difficulty in installation and debugging, long measurement time, and difficulty in error tracing. Each measurement equipment has its own applicable range, making it difficult to achieve rapid measurement of multiple errors. Visual measurement technology has advantages such as three-dimensional measurement, fast and convenient, and low cost. Therefore, this article combines artificial intelligence recognition methods and current detection methods to achieve intelligent detection and error estimation of mechanical assembly accuracy, assisting in improving assembly accuracy. Therefore, the work done in this article is as follows:

1) The recognition network structure has been improved by incorporating attention mechanisms and measurement modules into the deep convolutional network structure, greatly improving the algorithm's recognition ability for specific features in images.

2) In the error estimation stage, the algorithm structure has been improved, and the concept of weight has been introduced. By optimizing the weight values, the estimation algorithm has been optimized.

This article consists of the following chapters: Chapter 2 mainly introduces the relevant research of relevant scholars, pointing out the direction of this article. Chapter 3 improves the feature recognition network of images. Chapter 4 introduces the concept of weight and optimizes the weight values to achieve optimization of error estimation methods. Chapter 5 is a simulation phase to prove the feasibility and effectiveness of the algorithm.

* Corresponding Author

2 Related Work

There are few relevant literature on assembly accuracy testing abroad, and domestic research is the main reference. F. Dietrich proposed an assembly accuracy analysis method and a method for improving assembly accuracy to improve mechanical assembly accuracy. At the same time, this improved method is convenient for computer control and analysis, and is an effective means of elevating traditional methods to intelligent control methods [1]. Deepak Agrawal has designed a software for detecting assembly accuracy. Through the software, prediction and simulation of various assembly stages can be achieved, which can maximize the accuracy of each assembly stage [2]. Minghua Li, proposed an improved genetic algorithm, which can predict the assembly accuracy of aeroengine rotor, and effectively overcome the problem that the assembly benchmark and Geodetic datum are difficult to keep consistent during the assembly process [3]. Wei Wu established a calibration error accumulation model for the multi sensitive axis assembly process based on the characteristics of hole axis assembly, and derived the relationship between machining error and angular deviation [4]. Yang Yi, in response to the issue of inaccurate accuracy prediction in the assembly process of complex products, proposes a complex product assembly accuracy prediction method based on the number twin method and assembly process suggestions, providing a new approach to ensure assembly accuracy [5].

3 Deep Convolutional Networks for Assembly Feature Detection

During the assembly process of mechanical products, the components need to be assembled together according to the given assembly sequence. If the correctness of the newly installed components and the correct assembly position of the newly installed components are not detected in a timely manner, the assembly results will affect the quality and assembly efficiency of the mechanical products. Therefore, this article uses a deep image attention mechanism feature extraction framework for detecting assembly changes from multiple perspectives. The framework is shown in Fig. 1.

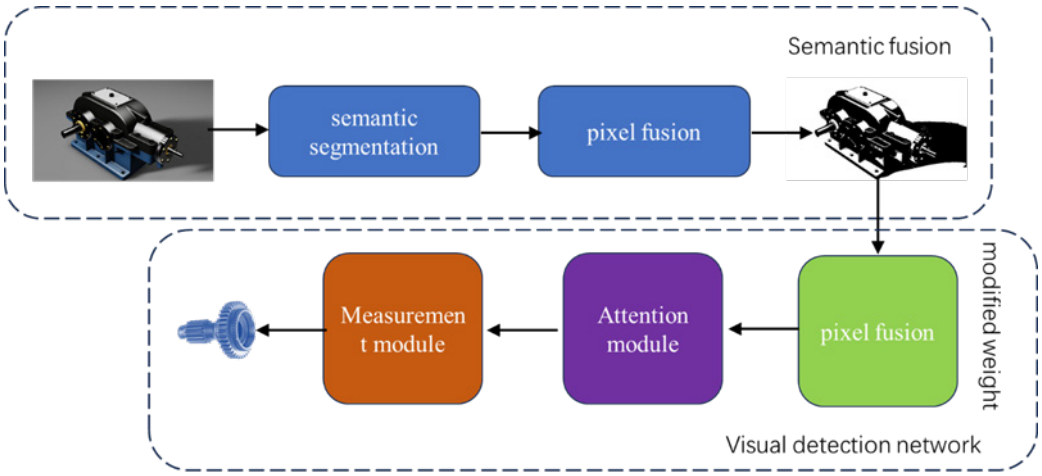

Fig. 1. Detecting network structure

The framework is divided into two networks, semantic fusion network and visual detection network, divided into five modules: semantic segmentation module, pixel fusion module, feature extraction module, attention module, and measurement module. Given the depth image and the image to be detected of the reference image from different perspectives during the assembly process. In order to detect the change position of the reference map relative to the image to be detected, a new change map is generated, and each pixel in the new change map is assigned two labels. One label indicates that there is a change in the image pixel, and one label indicates that there is no change.

3.1 Semantic Segmentation Module

The semantic segmentation network uses Deeplab as the basic network, and improves the original encoder module. The encoder structure uses deep Convolutional neural network to extract low-level details, and controls the resolution of the output feature map through hole convolution [6]; Then the initial features are transferred to the ASPP module, and richer semantic information is obtained from the convolution of holes with different expansion rates. The output channel is adjusted to 128 using convolution. Since the size of the feature map output by the encoder is one sixteenth of the original map, direct 16 times up sampling will lose information, so the decoder first uses the Bilinear interpolation method to conduct 8 times up sampling of the output feature map, Then it is spliced with the feature map of the corresponding resolution in the depth Convolutional neural network, and 8 times of up sampling is performed to make the feature map the same size as the original image, so as to gradually obtain a clear boundary of the segmented object. The network structure is shown in Fig. 2.

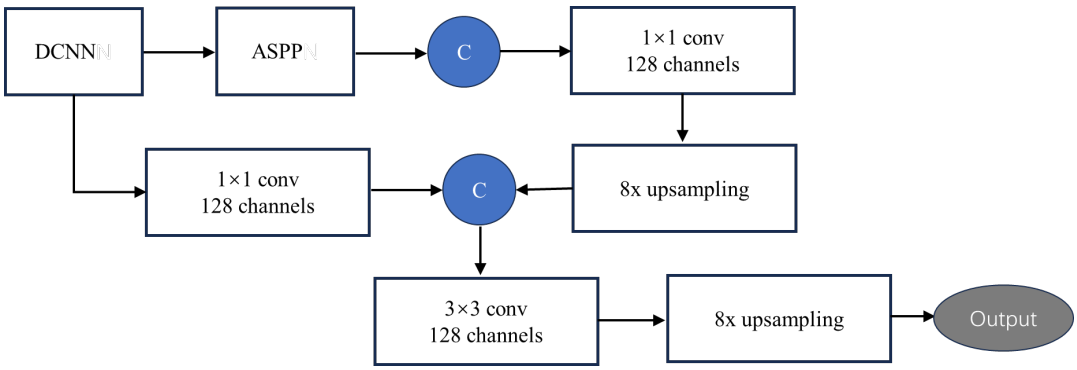

Fig. 2. Semantic segmentation module

3.2 Semantic Fusion

The Receptive field of the low level convolution part of the neural network is small, and it mainly extracts local details such as assembly corners, fillets, and chamfers. In higher-level networks, as the number of layers deepens, abstract information such as assembly texture and geometric shape is further extracted [7]. This paper designs a feature fusion module, which consists of four feature preprocessing units, an average pooling layer, and a Flatten layer.

1) In order to avoid the phenomenon of parameter multiplication caused by directly using convolutional layer sampling, this article uses convolutional layer for feature channel changes, and then uses pooling layer sampling for feature map size change processing.

2) By concatenating dimensions and reducing dimensionality through the average pooling layer and the Flatten layer, the multi semantic fusion features of the assembly image are obtained, improving the network's ability to read assembly features.

3.3 Semantic Fusion

In order to effectively detect the changing areas of the assembly process, the multi perspective change detection network consists of a feature extraction module, an attention module, and a measurement module. The specific introduction is as follows:

1) The feature extraction module, based on Rep VGG, removes the final global pooling layer and fully connected layer of Rep VGG, and embeds attention mechanisms after each step of feature extraction. The attention mechanism can capture global positional dependencies and quickly locate target information. The structure of the feature fusion module is shown in Fig. 3. The Rep VGG network includes two residual structures. One only adds parallel 1×1 convolutional integral branches to the 3×3 convolutional layer, and the other not only adds

1×1 convolutional branches, but also adds Identity branches. Due to its simple and single structure, it greatly improves memory utilization and facilitates model inference and acceleration.

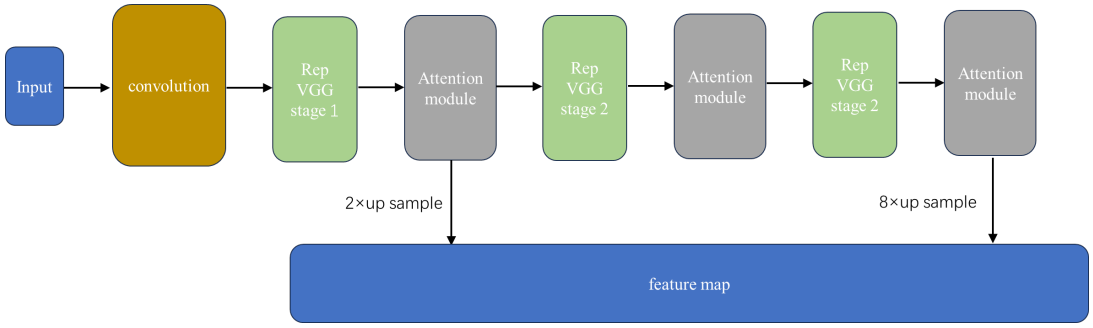

Fig. 3. Schematic diagram of feature extraction module

2) The attention mechanism module first obtains the initial attention feature map through Self attention 1, and collects contextual information in both horizontal and vertical directions while preserving the original information; Then, the generated attention feature map is fed to Self Attention 2, and other contextual information is obtained again from the horizontal and vertical intersecting paths; The structural diagram of the module is shown in Fig. 4.

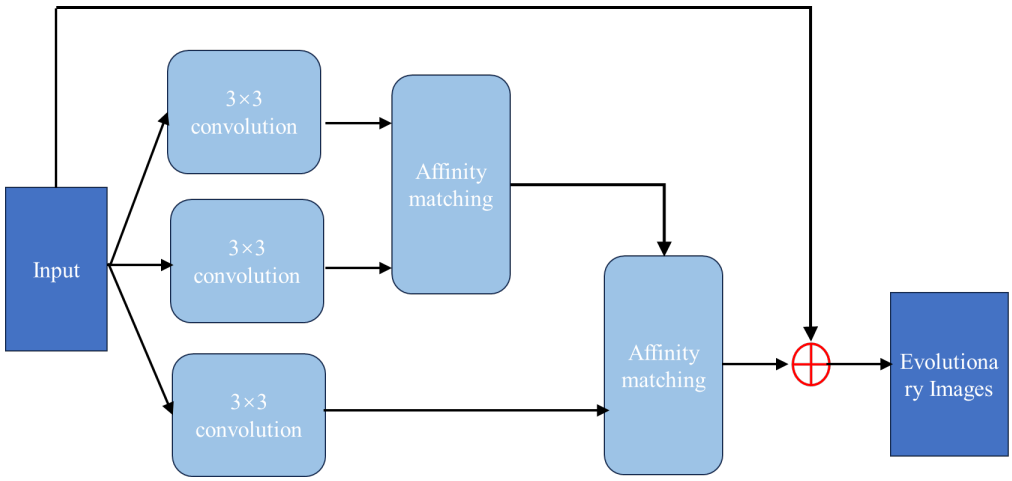

Fig. 4. Attention module schematic diagram

The specific process is to give a local feature map $P \in R^{C\times W\times H}$ and use two M convolutional layers in 1×1 to generate feature maps K and $\{M, K\} \in R^{C\times W\times H}$, with C and $A \in R^{(H+W-1)\times W\times H}$ being the number of channels in the feature map. Then, an attention map H is generated using an affinity matching algorithm, which is defined as:

$$d_{i,u} = M_u M_{i,u}^T. \tag{1}$$

Among them, M_u is the set of feature vectors, $M_{i,u}$ represents the i -th element of M_u , and $d_{i,u} \in D$ represents the degree of correlation between M_u and $M_{i,u}$. To achieve the aggregation of contextual information, aggregation operations are used:

$$H_u^{'} = \sum_{i \in |\Phi_u|} A_{i,u} \Phi_{i,u} + H_u . \tag{2}$$

In the formula, $H_u^{'}$ represents the feature vector of the output feature map $H^{'} \in R^{C \times W \times H}$ at u, $A_{i,u}$ is the scalar value of channel i and position u, and $\Phi_{i,u}$ is the set of feature vectors in the convolutional layer.

$$L = \frac{1}{2N} \sum_{n=1}^{Z} xd^2 + (1-x) \max(m \arg in - d, 0)^2 . \tag{3}$$

$$d = \left\| a_n - b_n \right\|_2 . \tag{4}$$

Formula 4 is the Euclidean distance between two sample vectors, x is the matching degree label, Z is the number of samples, margin is the set threshold, and is set to 1.

The change mapping $P_{i,j}$ is represented as:

$$P_{i,j} = \begin{cases} 1 & D_{i,j} > 0 \\ 0 & else \end{cases} . \tag{5}$$

$i, j (1 \le i \le W_0, 1 \le j \le H_0)$ represents the index of the width and height of the feature vector. Based on the threshold set in the change mapping P, the image area is divided into changing regions and invariant regions to obtain the final changed image.

4 Error Estimation Method Based on Deep Learning

The assembly error estimation of deep learning is divided into: 1. dataset collection step, 2. image input step, 3. pose estimation learning step, 4. model judgment step, and 5. perspective update step.

Using the corresponding point matching algorithm to fuse color feature information and geometric feature information, pixel level feature embedding is carried out. The pose iterative refinement algorithm is used to map it to a spatial feature vector, and finally, this spatial feature vector is used for pose prediction, which can effectively solve the problem of part occlusion. The Loss function is defined and expressed as follows:

$$L_i^p = \frac{1}{Q} \sum\nolimits_j \left\| \left(R y_j + t \right) - \left(\hat{R}_i y_j + \hat{t}_i \right) \right\| . \tag{6}$$

Where, Q is a random point cloud, y_j is the j point cloud in the Q point cloud, $p = [R|t]$ is the ground truth pose, $p = \left[\hat{R} \middle| \hat{t} \right]$ is the predicted pose, and the Loss function is:

$$L_i^p = \frac{1}{Q} \sum\nolimits_j \min_{0<k<Q} \left\| \left(R y_j + t \right) - \left(\hat{R}_i y_k + \hat{t}_i \right) \right\| . \tag{7}$$

At the same time, when calculating the Loss function for each pixel, add a weight to judge the confidence degree of the predicted attitude. The weight L is defined as follows:

$$L = \frac{1}{N} \sum\nolimits_i \left(L_i^P c_i - \omega \log(c_i) \right) . \tag{8}$$

In the equation, N represents the number after fusion, and ω is a hyperparameter. After K iterations, the algorithm concatenates the pose estimates for each iteration as the median pose estimation.

$$\hat{p} = \left[R_K \middle| t_K \right] \cdot \left[R_{K-1} \middle| t_{K-1} \right] \cdots \left[R_0 \middle| t_0 \right]. \tag{9}$$

The pose estimation error during the initial stage of training is too large, so in the specific training process, joint training begins after the main network training converges.

In order to determine whether the pose of the parts during the mechanical assembly process is accurate, it is necessary to build a data acquisition system, select a camera for visual acquisition, and use Intel RealSense D415 as the image acquisition device. In the dataset, a two-stage reducer is selected as the assembly body, which includes shafts, large bevel gears, large spur gears, bearings, etc. Selecting a horizontal plane as the benchmark and rotating around the target, 1777 photos including color and depth maps were collected. Calculate the matrix conversion between each frame of the collected image at a specified interval and save it in an array format file; Then register the file scene, convert it into a registered point cloud, and use Meshlab software to crop the registered point cloud to remove scene noise in the registered point cloud; After only including the target object point cloud scene, use Meshlab software to connect the point cloud data again, connecting each point to the nearest surrounding point, and generating a 3D model in the form of a surface with the connected overall data. According to the position of the part 3D model in each image, the projection segmentation mask image is established, and the standard Rotation matrix and offset matrix of the point cloud in the part 3D model and other pose labels are obtained at the same time.

5 Simulation Experiment and Result Analysis

The experimental platform is on the Ubuntu 18.04.4 LTS operating system, with a 3.000GHz Intel (R) Core (TM) i7-9700 CPU model, 32GB memory, GeForce RTX 2060 graphics card model, 6144MB total memory, and NVIDIA Driver 440.100 graphics card driver. The network program adopts PyTorch deep learning framework and Python programming language. Firstly, the assembly accuracy of the assembly is tested. To demonstrate the effectiveness of the proposed method, the algorithm proposed in this paper is compared with other algorithms as shown in Fig. 5.

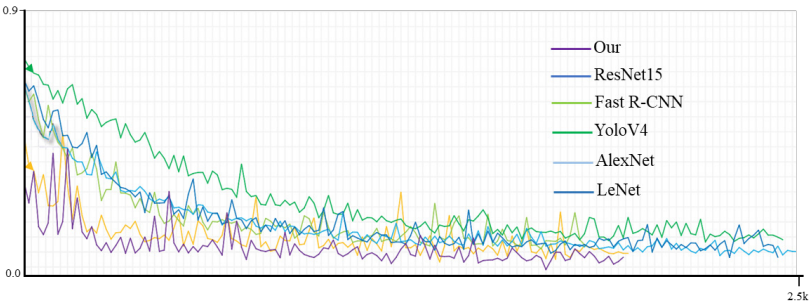

Fig. 5. Attention module schematic diagram

From the figure, it can be seen that the algorithm proposed in this article converges faster. This is mainly because the pre training method is used, and the pre training weight has saved the model learning parameters for the original dataset. When learning a new similar dataset again, the learning task can be completed with relatively few iterations. To demonstrate the universal applicability of the algorithm under different data sets, the algorithm was validated on three different datasets. The validation results are shown in Table 1.

Table 1. Comparison of recognition results under different datasets

	Dataset 1	Dataset 3	Dataset 3
Accuracy	98.21%	90.27%	95.36%

From the above table, it can be seen that the deep convolutional network proposed in this chapter can achieve high accuracy in identifying assembly accuracy images, whether it is the number of assembly steps or the synthetic and real scenes.

In order to verify the performance of the algorithm in this article for error estimation during the assembly process, a dense voting method was used to achieve the pose estimation of the target object. The test results of the dataset are shown in Table 2.

Table 2. Test results of Our Net pose estimation network on the dataset of this chapter

	Axle	Bearing	Shaft sleeve	Bevel gear wheel	Bevel pinion
Accuracy	93.41%	97.82%	98.31%	96.43%	97.43%

From the table, it can be seen that the pixel level fusion method used in this article can effectively infer the local appearance and geometric information of the parts, and effectively handle the occlusion problem of the parts. The pose iterative refinement algorithm used in this article can iteratively extract the initial predicted pose, which has a certain improvement in the performance of the network model. To verify the effectiveness of the algorithm in the process of true accuracy detection and estimation, a two-stage reducer was used as the experimental object, and the experimental results are shown in Fig. 6.

Fig. 6. Identify network test results

From the test results, it can be seen that the red bounding box represents the predicted value of the pose. From the figure, it can be seen that for helical gears, the angle and position of the red bounding box envelope are basically close to the true pose of the helical gear, and the key points are accurate. Therefore, overall, the prediction and evaluation ability of the algorithm can meet the practical application requirements.

6 Conclusion

This article focuses on the current situation of large assembly errors, easy omissions and errors in the mechanical assembly process. Computer vision is introduced in the assembly process, and visual images are used to estimate assembly errors, thereby improving assembly accuracy. To this end, through improvements to the neural network, the addition of attention and measurement mechanisms, the network's ability to extract and distinguish features from assembly images has been improved. Finally, deep learning algorithms are used to estimate assembly features in the image. Finally, simulation experiments have shown that the algorithm proposed in this paper can achieve 94.7% improvement in assembly accuracy and error estimation accuracy.

At the same time, there are still many shortcomings in this article, and improvements will be made in further research. Firstly, the recognition accuracy of the algorithm for assembly features needs to be improved. In addition, in the pose estimation process, the predicted pose is almost close to the actual pose, but there are still certain errors. Therefore, future work will improve the accuracy of pose estimation.

References

[1] F. Dietrich, A. Glodde, S. Solmaz, Accuracy analysis and improvement method for continuous web-based precision assembly, CIRP Annals - Manufacturing Technology 69(1)(2020) 5-8.
[2] D. Agrawal, S. Kumara, D. Finke, Automated Assembly Sequence Planning and Subassembly Detection, in: Proc. Industrial and Systems Engineering Research Conference, 2014.
[3] M.-H. Li, Y.-L. Wang, Q.-C. Sun, X.-K. Mu, Assembly accuracy prediction and optimization of aero-engine rotor under the separation condition of assembly and measurement, The International Journal of Advanced Manufacturing Technology 120(5-6)(2022) 3103-3112.
[4] W. Wu, Z. Deng, J.-Z. Shang, Z.-R. Luo, Y.-J. Cao, Assembly Accuracy Analysis and Process Optimization of Multi-sensitive Axes for Precision Optical Mechanical System, Transactions of the Chinese Society for Agricultural Machinery 52(4)(2021) 418-426.
[5] Y. Yi, J.-D. Feng, J.-S. Liu, C.-J. Chen, X.-J. Liu, Z.-H. Ni, model expression and accuracy prediction method of digital twin-based assembly for complex products, Computer Integrated Manufacturing Systems 27(2)(2021) 617-630.
[6] B.-W. Wei, H.-Y. Quan, Semantic Segmentation Network Based on Semantic and Morphological Feature Fusion, Acta Electronica Sinica (11)(2022) 2688-2697.
[7] Z.-L. Wang, L. Shen, W.-G. Xu, Q. Li, Finger vein recognition based on Arcface loss and multi-semantic fusion network, Jounal of Hangzhou Dianzi University (Natural Science) 42(1)(2022) 53-59.