3D Reconstruction of Aircraft Structures via 2D Multi-view Images

Tianyou Zhang, Runze Fan, Yu Zhang, Guangkun Feng, Zhenzhong Wei✉
Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education and the School of Instrumentation and Optoelectronic Engineering, Beihang University, Beijing, 100191, China

ABSTRACT

Aircraft pose is of great importance to the monitor during flight test. Commonly, traditional pose measurement is based on various sensors, which has the inconvenience of repair and replacement because of damage. Two-dimension (2D) single image processing method alleviate the inconvenience, but it has the ambiguity of single image three-dimension (3D) reconstruction. To address these problems, we accomplish 3D reconstruction of the aircraft's structures via 2D multi-view images. Structures are obtained from 2D multi-view images of aircraft by a convolutional neural network (CNN) and then used to accomplish reconstruction. Structures typically represent the topological relationship between components of aircraft, reducing the self-occlusion of point features. To more precise evaluation of the experimental results, we propose a new Mean Per Structure Position Error (MPSPE) calculation for the structures position. Compared with the Mean Per Joint Position Error (MPJPE), the MPFPE takes the length of structures into account and mixes the multi-view images. Experiments show the mean error of our method is 1.47%, which shows great potential for aircraft pose estimation.

Keywords: Feature extraction, structure features, 3D reconstruction

1. INTRODUCTION

Get the 3D information of the aircraft through the captured aircraft's images, which has a good application prospect in the fields of augmented reality (AR) technology and aircraft flight attitude measurement. Structures typically represent the topological relationship between aircraft components, contain abundant 3D information of the aircraft. The further analysis of the aircraft flight attitude and process through the three-dimensional reconstruction of the structure is conducive to the real-time monitoring of the aircraft flight process and the detailed analysis afterwards.

For aircraft, obtain the 3D structure information is an important parameter reflecting the flight state of the aircraft in the air. The camera takes the flight image of the aircraft, with the non-contact and low-cost characteristics compared with the GPS system and SINS system, makes the aircraft do not need to carry additional equipment, which is convenient for ground detection. With the application of multi-view information, the ambiguity of single image will be solved. After that, the integrated information with other sensors can be used to reproduce the complete trajectory state of the aircraft flight test.

This paper takes C919 as the research target, accomplishes 3D reconstruction of the aircraft's structures via 2D multi-view images. Research the structure definitions applicable to aircraft targets, and the extracting and matching method of aircraft structures. The method of multi-view 3D reconstruction is also researched. The purpose of this paper is to reconstruct the 3D feature of the aircraft structure using the 2D images from different perspectives. Provide some other helpful information about practical application.

✉ Corresponding author;
Zhenzhong Wei, E-mail: zhenzhongwei@buaa.edu.cn;
Tianyou Zhang, Runze Fan, Yu Zhang, Guangkun Feng, E-mail: {tiuzhang520, sy1817326, by1817156, fenggk}@buaa.edu.cn;
Tel: 86-010-82338768

2. RELATED WORKS

2.1 Structure extracting method

Skeleton is commonly and widely used in human pose estimation methods. Therefore, we can get a glimpse of the current situation of feature extraction methods from human pose estimation methods. Most methods of human pose estimation are based on human skeleton features. With the development of convolutional neutral network, a series of new methods of human pose estimation have emerged, which also improves the accuracy and robustness of pose estimation. In the field of 2D pose estimation, in 2015, CPM (Convolutional Pose Machine) was the first article to combine image feature with spatial context [1], which provided an idea for end-to-end learning of pose estimation in the future. In 2016, the stacked hourglass net proposed by Alejandro Newell had a far-reaching impact on the future pose estimation framework. The image inputted into the network was down sampled to the minimum resolution through multiple groups of symmetrical neural networks, and then up sampled to the output resolution, in which the high-level and low-level features were interpolated and learned to achieve good results[2]. Then other researchers attain many results in human pose estimation area through the application of CNN[3~5], 2D pose estimation study become successful gradually.

Based on CPM and stacked hourglass net, 3D human pose estimation began to flourish after 2017. In 3D pose estimation area, there are quite a number of methods to obtain the relative 3D skeleton features of the target based on a single RGB image, which can be divided into two different methods. One is to directly regress from the 2D image to the 3D coordinates. Some article is based on this method[6]. However, it is fuzzy to predict 3D pose directly from single 2D image, because a 3D pose can correspond to different 2D poses. This kind of problem can be solved by using multi-view information, some articles take more research[7~11].

2.2 Trinocular stereo vision model

The basic measurement principle of the trinocular stereo vision system is based on the principle of the monocular stereo vision system and the binocular stereo vision measurement system. The biggest difference is that it breaks through the field of vision limitations of the monocular stereo vision system and the binocular stereo vision system. Therefore, the application of trinocular stereo vision technology is becoming more and more extensive. In the commonly used binocular stereo vision system, the depth information of the target can be obtained according to the parallax of two cameras with different viewing angles, and the essential matrix or basic matrix between the two images can be obtained through feature matching, and then by the linear triangle method, The three-dimensional point coordinates projected into two images can be reconstructed.

The trinocular stereo vision system is not much different from this. The stereo matching method in trinocular vision can be regarded as three groups of two-by-two binocular stereo matching[12]. The current commonly used method is still based on key point feature, which is first key points matching is performed in the images obtained by two cameras, and then under certain conditions (such as: satisfying the least square method) to achieve trinocular stereo matching[13]. However, the trinocular stereo vision system has a more effective matching principle for linear features than the binocular stereo vision system. A three-dimensional straight line in the space is projected into three pictures. Two of the pictures can be used to back-project a plane, plus the information of the third angle of view to determine a three-dimensional straight line. Relative to the basic matrix, the trifocal tensor can be used in the trinocular stereo vision system to achieve the feature line matching between the three views. The multi-view 3D measurement method of structure features can be matched based on the combination of key points into straight lines, but straight-line matching can also be applied to multi-view matching, which has application flexibility.

3. OUR METHOD

3.1 Define structure feature

First of all, it is necessary to define the features, which is also the prerequisite for the extraction algorithm to proceed, so the definition of aircraft features that are particularly important for the study. However, if the method of key points commonly used in object pose estimation is used alone, the key points will be missing when the aircraft is partially occluded, so the structure feature can improve the accuracy of 3D reconstruction based on the key points. In the currently human pose estimation, the definition of structure features is very natural, because the human body is based on the skeleton, and there are obvious joint points between the skeletons. Similar to human skeletons, component of aircraft can be seen as

several rigid structures connect in joints. The largest challenge is the joint of aircraft structure is not as specific as human skeletons, which leads to cannot locate the structure highly accurate. Through the analysis of the human skeleton, the posture can be fully estimated. The purpose of estimation and motion analysis.

Since each part of the aircraft can be regarded as the characteristics of a rigid body, and taking the above factors into consideration, there is no relative movement inside each part of the aircraft, so each part of the aircraft can be regarded as a rigid body without relative deformation, so that the definition of the aircraft structure is more natural: the fuselage, left and right wings, left and right horizontal tail, vertical tail, at total of 6 structures[14].

Figure 1. The structures of C919 aircraft.

3.2 Extract structures

To predict the aircraft's structure, Runze Fan proposed a network based on the stacked hourglass net[14]. On the basis of this study, we apply the method to structure reconstruction, the net is shown in Figure 2. The method first beneficiates the input image: crop it to a suitable size and handle with a convolution layer. The max-pool layer is widely used in this net, actually it is also applied into the residual block. As shown in Figure 3. to match the max-pool, up sample layer is used followed, which can increase the resolution of the output to the same level with the input. The components of the residual block are shown in Figure 4.

Based on this network framework, we use the mean square error (MSE) function to calculate the error. Assuming we have annotated heat maps of S structures, and use MSE loss to calculate the predicted heatmap and the annotated heatmap. Similar to the loss function of key points, the loss function of the structure is defined as:

$$L_{structure} = \frac{1}{S}\sum_{i=1}^{S} ||y_i - g_i(x)||^2$$

where $g_i(x)$ represents the i-th predicted heatmap. For each structure, the least squares method is used to fit: firstly, select pixels larger than a certain threshold, and then use the least squares method to fit the most suitable structure. In addition, the reciprocal of the average distance from each pixel to the structure is recorded as the confidence probability of this structure.

For the defined 6 aircraft structures, the network will output 6 corresponding heatmaps, and each loss function needs to be calculated for learning and optimization. The final output is the result of the integration of the 6 heatmaps.

3.3 Trinocular stereo measurement method

The trinocular multi-view measurement is completed by the images taken by a trinocular stereo vision system composed of three cameras or a trinocular system composed of a single camera at different positions. Regardless of the composition, it can be regarded as three groups of binocular stereo vision system.

For straight line features, tri-focal tensors can be used for research in the trinocular stereo vision model. The tri-focal tensor in the trinocular system is similar to the basic matrix in the binocular system. It plays an important role in expressing the essential geometric relationship. It reveals the basic affine geometric relationship between the three views, which is very important for trinocular stereo matching. According to the tri-focal tensor, the relationship between points and straight lines among the three views can also be obtained[11].

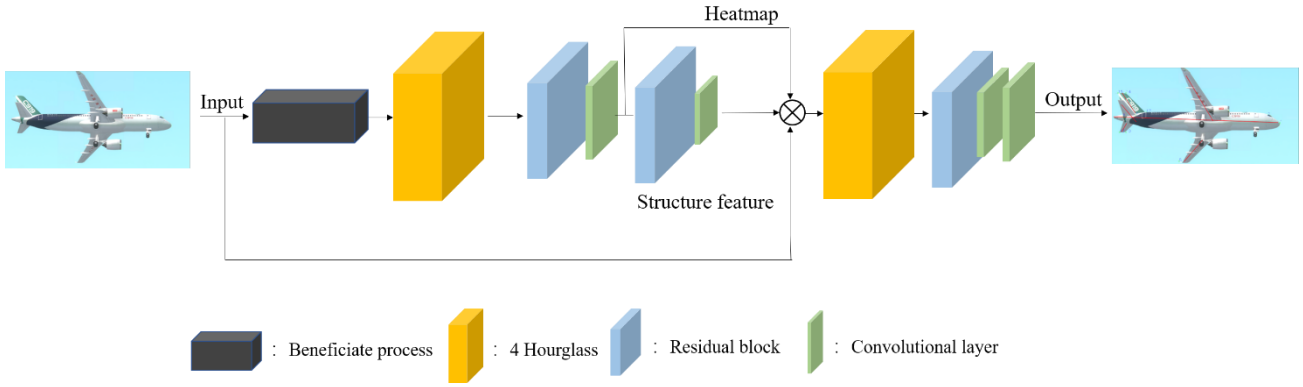


Figure 2. The architecture of prediction net[14]. The forward part generates the heatmaps of features, with the original image and the predict location input into the backward part together, which to refine the prediction.

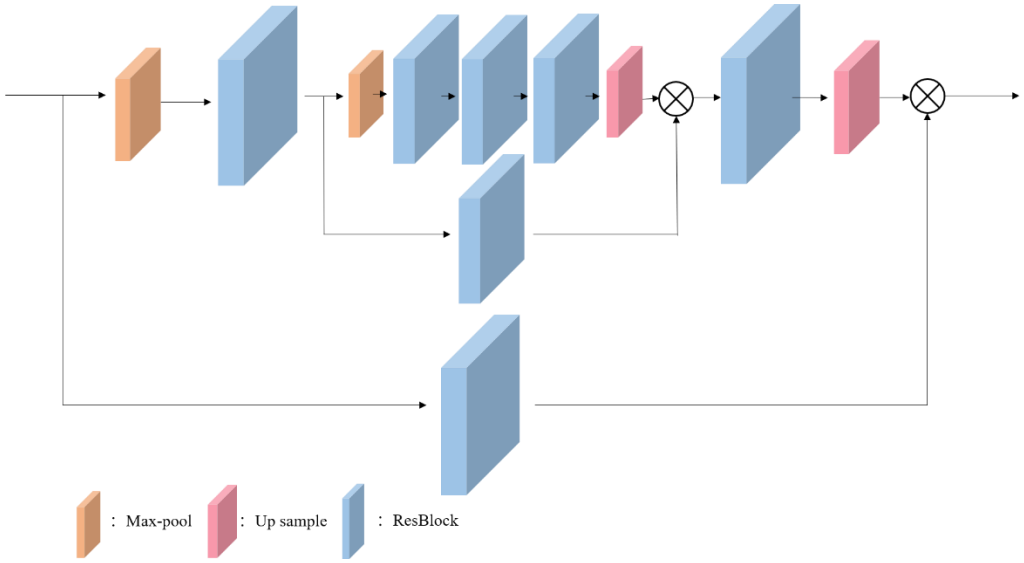


Figure 3. 2 stacked hourglass net.

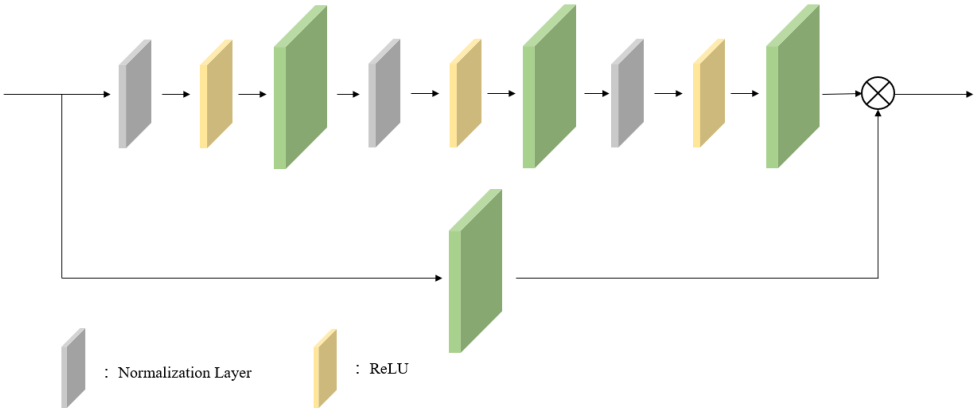


Figure 4. Residual block.

4. EXPERIMENTS

4.1 Dataset

In response to the problem of insufficient multi-view aircraft data, our laboratory established the C919 large aircraft data set. This dataset includes actual dataset and virtual dataset. The real dataset is the pictures taken during the actual flight of the plane, but the real dataset only has about 3,000 images, which is far from enough for network training. It is difficult to obtain sufficient resolution for the images, only the pictures during take-off and landing are clearer. Therefore, using virtual image generation software, the virtual 3D model of the C919 aircraft is nested into different backgrounds to obtain multiple virtual images. The attitude of the aircraft model can also be adjusted arbitrarily, and the background can also be selected from various simple or complex situations. In this way, 216,000 images of the C919 aircraft can be obtained. At the same time, there are corresponding labeled attitudes and camera parameters for the next network training.

(a) Render image in real environment (b) Render aircraft image

Figure 5. The render picture dataset of aircraft. The left image obtained by rendering the right model in real environment.

In addition, this article also makes enhanced operation of the dataset. Perform operations such as rotation, scaling, and cropping on the image to reduce the possibility of overfitting. And in order to reduce the difficulty of learning, the rectangular outline of the aircraft in the image is also drawn. The actual operation will be performed within the rectangular outline, which greatly reduces the pressure on the network and helps to reduce the error of structures.

4.2 Training setting

To prevent over-fitting on the real dataset, we first train on the virtual dataset. The stochastic gradient descent method is used to optimize the training. The initial learning rate is set to 0.001. After reaching a certain number of trainings, the learning rate is reduced. After 160 epochs of training on the virtual data set, satisfied results can be obtained. It can be further optimized on the real dataset, and the training can be performed on the real dataset under the condition of freezing the key points and the structures. The initial learning rate is set to 0.00025. The number of each batch is set to 4 in consideration of the hardware conditions. The output of the network is shown in the figures below. The red straight line in the figures represents the structures, and the blue circles and numbers represent structure's related key points and key point labels.

Analysis the final result, the stacked hourglass network built in this paper can complete the tasks of structures estimation, and can still obtain better results in some severe weather conditions (shown in Figure 6.), which is sufficient to support further structures 3D reconstruction.

(a) Dark environment (b) Occlusion condition

Figure 6. Experiments in bad weather.

4.3 Trinocular stereo structures reconstruction

Because of the lack of three-view images of the real aircraft, this paper uses virtual image generation software to render the same scale C919 aircraft model to obtain virtual images. In the virtual image generation, the positions of the three camera coordinate systems are placed 100 meters apart in a certain direction. The coordinates in the world coordinate system are expressed as (300, 0, 5), (300, 100, 5) and (300, 200, 5) (unit: m), the focal length of the camera is 55 mm, the image size is 1920×1080, and the pixel size is 5 microns. The aircraft is placed in an unknown position in the world coordinate system. Since the aircraft is viewed from the ground, the setting can only capture the bottom view of the aircraft, but the optical axis of the camera always points to the center of the aircraft. The picture below shows the aircraft images taken by the three cameras on the left, middle and right.

(a) Left view (b) Middle view (c) Right view

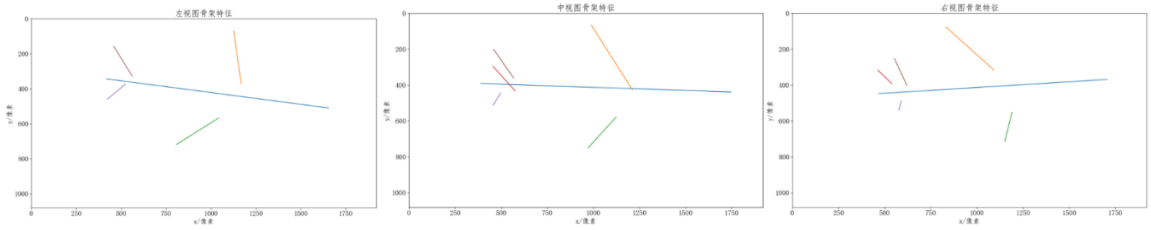
Figure 7. Images captured by the left, middle and right camera.

Input each image into the trained network to get the and structures of the aircraft:

(a) Structures of left view (b) Structures of middle view (c) Structures of right view

Figure 8. Obtained Structures of the output aircraft.

Draw the extracted aircraft structures separately:



(a) Left image of structures (b) Middle image of structures (c) Right image of structures

Figure 9. Images of structures.

After obtaining the two-dimensional optimization data of the structures, the two sets of three cameras are combined into a binocular stereo imaging system to calculate their respective depth information. The specific algorithm for obtaining point depth information based on the principle of binocular stereo system measurement. Observing the Figure 11., obviously the three groups of binocular systems have a large reconstruction error for the left wing where the complete part is not observed, and the most complete fuselage structure has the smallest error. Subsequently, the three groups of reconstruction images

are combined through an optimization algorithm to obtain the final aircraft three-dimensional reconstruction structures, as shown in Figure 11(a). Comparing with the middle view skeleton, it can be observed that although the overall skeleton features can be obtained, there are still some errors in the final result, indicating that there is still room for improvement in the reconstruction algorithm.

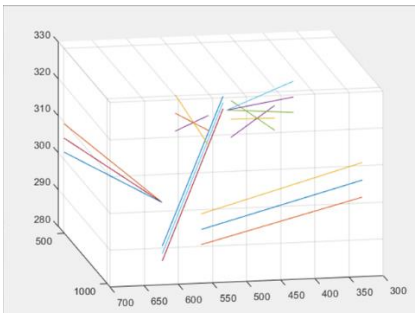
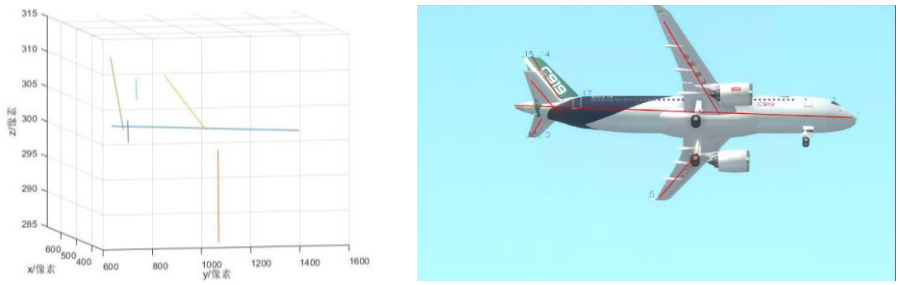


Figure 10. Reconstructed structures of three views.



(a) Optimized reconstructed image (b) Structures of middle view

Figure 11. Analysis of reconstruction result.

In order to conduct error analysis more appropriately, we define Mean Per Structure Position Error (MPSPE), which is similar to Mean Per Joint Position Error (MPJPE), based on the structures obtained by the network.

$$MPSPE \stackrel{\text{def}}{=} \frac{1}{L} \cdot \sqrt{\frac{1}{S}\sum_{i=1}^{S} \|\max\{f_i(x) - x_i\}\|^2} \times 100\%$$

Where $f_i(x)$ is the result of $i-th$ estimated structure, x_i is the groundtruth of the $i-th$ structure , $\max\{f_i(x) - x_i\}$ is the biggest distance between estimated structure i with the groundtruth i. S is the number of structures, L is the scale of aircraft.

The ratio of the distance from the structure in the groundtruth image to the scale of the aircraft is recorded as the average structure position error. The annotation of the structures is taken as the groundtruth position of the structure. Since the structures of various parts of the aircraft are not as obvious as the human body, the key point of the structure may appear at an uncertain point on the edge of the structure. For example, the point of the root of the right wing may appear at the plane where the root of the right wing of the aircraft intersects the fuselage. Therefore, the distance between the estimated structure and the groundtruth structure is determined by the intersection of the end point of the structure and each structure of the aircraft. The calculation error of the 3D structures reconstructed from the network estimation and the 3D structure reconstructed from the groundtruth image.

Table 1. 3D reconstruction error results.

Structure	MPSPE（%）
fuselage	0.608
right wing	3.311
left wing	2.329
right horizontal tail	1.678
left horizontal tail	0.450
vertical tail	0.442
mean	1.469

5. CONCLUSION

This article mainly studies the three-dimensional measurement method of aircraft structures. We build a training network based on the Stacked Hourglass Net, reconstruct the structure's 3D position after two-stage hourglass network. Compared with other methods, this method uses the multi-view information to carry out the ambiguity of structures reconstruction. This article provides another method for aircraft attitude estimation, flight monitoring and other aspects. Compared with the traditional three-dimensional measurement method of aircraft, this method proposed is more convenient and robust, reduces the number of airborne equipment, and provides the possibility for real-time measurement and monitoring of aircraft.

ACKNOWLEDGMENTS

This work was supported by the National Science Fund for Distinguished Young Scholars of China (51625501).

REFERENCE

[1] Wei, S. E., et al. "Convolutional Pose Machines." IEEE, 4724-4732 (2016).
[2] Newell, A., K. Yang, and D. Jia. "Stacked Hourglass Networks for Human Pose Estimation." European Conference on Computer Vision Springer International Publishing, (2016).
[3] Chen, Y., et al. "Cascaded Pyramid Network for Multi-person Pose Estimation." Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, (2018).
[4] He, K., et al. "Deep Residual Learning for Image Recognition." IEEE (2016).
[5] Sun, K., et al. "Deep High-Resolution Representation Learning for Human Pose Estimation." arXiv e-prints, 5686-5696(2019).
[6] Pavlakos, G., et al. "Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations." IEEE (2017).
[7] Martinez, J., et al. "A simple yet effective baseline for 3d human pose estimation." IEEE Computer Society (2017).
[8] Chen, X., et al. "Weakly-Supervised Discovery of Geometry-Aware Representation for 3D Human Pose Estimation." (2019).
[9] Pavlakos, G., et al. "Coarse-to-Fine Volumetric Prediction for Single-Image 3D Human Pose." Conference on Computer Vision & Pattern Recognition(CVPR) IEEE, (2017).
[10] Qiu, H., et al. "Cross View Fusion for 3D Human Pose Estimation." University of Science and Technology of China; Microsoft Research Asia, (2019).
[11] Remelli, E., et al. "Lightweight Multi-View 3D Pose Estimation through Camera-Disentangled Representation." IEEE. 6039-6048(2020).
[12] Ayache, Nicholas, and Lustman, F. "Fast and reliable passive trinocular stereovision." proc ICCV. 186–191 (1987).
[13] Shen, J., Castan, S., and Jian, Z. "A new passive measurement method by trinocular stereo vision." Industrial Metrology 1.3, 231-259 (1990).
[14] Fan, R., Xu, T., and Wei, Z. "Estimating 6D Aircraft Pose from Keypoints and Structures." Remote Sensing, 663 (2021).