This item was submitted to [Loughborough's Research Repository](#) by the author.
Items in Figshare are protected by copyright, with all rights reserved, unless otherwise indicated.

FADN: Fully connected attitude detection network based on industrial video

PLEASE CITE THE PUBLISHED VERSION

<https://doi.org/10.1109/tii.2020.2984370>

PUBLISHER

Institute of Electrical and Electronics Engineers (IEEE)

VERSION

AM (Accepted Manuscript)

PUBLISHER STATEMENT

Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

LICENCE

All Rights Reserved

REPOSITORY RECORD

Yang, Jiachen, Meng Xi, Bin Jiang, Jiabao Man, Qinggang Meng, and Baihua Li. 2020. "FADN: Fully Connected Attitude Detection Network Based on Industrial Video". Loughborough University.
<https://hdl.handle.net/2134/12965063.v1>.

FADN: Fully Connected Attitude Detection Network Based on Industrial Video

Jiachen Yang, *Member, IEEE*, Meng Xi, Bin Jiang, Jiabao Man, Qinggang Meng, *Senior Member, IEEE*, Baihua Li, *Member, IEEE*

Abstract—In 3D attitude angle estimation, monocular vision-based methods are often utilized for the advantages of short-time and high efficiency. However, the limitations of these methods lie in the complexity of the algorithm and the specificity of the scene, which needs to match the characteristics of the cooperation object and the scene. We propose a fully connected attitude detection network (FADN) which combines neural network and traditional algorithms for 3D attitude angle estimation. FADN provides a whole process from the input of a single frame image in the industrial video stream to the output of the corresponding 3D attitude angle estimation. Benefiting from the end-to-end estimation framework, FADN avoids tedious matching algorithms and thus has certain portability. A series of comparative experiments based on the rendering software 3D Studio Max (3d Max) have been carried out to evaluate the performance of FADN. The experimental results show that FADN has high estimation accuracy and fast running speed. At the same time, the simulation results reliably prove the feasibility of FADN, and also promote the research in real scenarios.

Index Terms—Attitude angle estimation, convolutional neural network, deep learning, fully connected network, monocular vision

I. INTRODUCTION

AS a prerequisite for plenty of application scenarios such as intelligent robot, aerospace, industry, the real-time and accurate 3D attitude angle estimation has attracted tremendous attention [1], [2]. Recent years have witnessed the flourishing of attitude angle estimation, which could be divided into the monocular vision based, the binocular vision based and the multi-vision based methods, according to the number of cameras [3], [4]. The monocular vision based methods are widely used for their advantages of simple equipment, low cost and real-time.

At the same time, the attitude angle estimation could also be divided into the visual correspondence based algorithms [5] and the deep learning based methods [6]. Currently, the former is the mainstream method in industrial applications. However, the inevitable disadvantage is that it needs to design the corresponding algorithm according to the application scenario. Typically, the following processes are needed: feature design such as point feature and line feature [7]; image process

for feature matching; algorithm solving. Moreover, accurate feature matching mostly relies on additional auxiliary equipment such as laser, which puts higher demands on equipment. The deep learning based method can achieve an end-to-end detection, which automatically fits the camera imaging model and outputs the 3D attitude angle directly. Therefore, a series of cumbersome tasks such as manual calibration, feature extraction and matching are avoided. However, most of the methods based on deep learning are in the stage of synthetic simulation and theory. Although good results have been achieved on manual datasets, such as LINEMOD [8], OCCLUSION [9], and T-LESS [10], they have not been widely adopted in industry.

Many deep learning based methods utilize convolution neural network (CNN) to estimate 3D attitude angle. Thus, through large-scale deep learning, an end-to-end estimation can be directly implemented from the input of the RGB image to the output of the attitude angle. However, the CNN frameworks they rely on are mainly designed for targets detection and image segmentation. They fail to fully analyze the characteristics of 3D attitude angle estimation and design a specialized and suitable estimation network accordingly.

In this paper, we propose a fully connected attitude detection network (FADN) to estimate 3D attitude angle. To verify its validity, we also build a dataset on the rendering software 3D Studio Max (3d Max). And subsequent experiments are conducted on our own dataset. The ideal results prove that FADN can achieve fast and accurate estimation on the one hand, and on the other hand promote the industrialization of FADN and lay the foundation for future work. The main contributions of this paper can be summarized as follows.

1. For the first time, we use a fully connected neural network architecture to estimate 3D attitude angle. Different from the popular CNN based methods, it combines the characteristics of estimation with deep learning. The relationship between 2D image and 3D attitude angle can be automatically learned, and thus the manual workload in monocular visual estimation is greatly simplified.

2. We construct a complete set of feature points extraction and matching algorithm, which provides accurate pixel coordinates as the input of FADN.

3. We adjust the joint output mode to independent output mode, which better adapts to the difference in the three attitude angles and further improves the output accuracy.

4. For industrialization, we build a specialized dataset using rendering software. It contains one million training samples and 1,000 test samples.

This work was partially supported by National Natural Science Foundation of China (No. 61871283), the Foundation of Pre-Research on Equipment of China (NO.61400010304) and Major Civil-Military Integration Project in Tianjin, China (NO.18ZXJMTG00170). (*Corresponding author: Bin Jiang.*)

J. Yang, M. Xi, B. Jiang, J. Man are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mails: {yangjiachen, ximeng, jiangbin, manjiabao}@tju.edu.cn).

Q. Meng and B. Li are with the Department of Computer Science, Loughborough University, Loughborough, UK (emails: {q.meng, b.li}@lboro.ac.uk).

The rest of this paper is organized as follows. In Section II, we introduce the related work, and describe the methodology of our method in Section III. Section IV explains our dataset and the experiments in detail. In Section V, we make a conclusion and discuss the possible problems in real scenarios.

II. RELATED WORK

The method based on monocular vision has the advantage of simple equipment, only utilizing RGB images. And the method based on deep learning is fast and convenient, realizing an end-to-end estimation. Therefore, in order to inherit the superiority of both, we mainly focus on the deep learning based estimation method with RGB images. It is necessary to review the basic flow of monocular visual attitude angle estimation, the deep learning method based on CNN framework, and the related properties of fully connected network.

A. Monocular Vision System

The monocular vision system uses RGB images to inversely solve the 3D attitude angle, which is a complex task due to the lack of depth information. Generally, more information is needed to assist the estimation. And there are many methods currently available through monocular vision, such as stereoscopic monocular odometry [11], [12]. The essence of attitude estimation based on monocular vision is to establish a camera imaging model which reflects the 2D/3D mapping relationship. The model building process involves camera calibration, lens distortion correction and so on [13]. In addition, the selection of features in cooperative object also involves many aspects. Standard algorithms typically utilize point feature, line feature, and surface feature [7]. All the time, people have been studying some links in these processes to improve the accuracy of estimation. For example, Tian *et al.* proposed a new calibration approach to relocate the camera's pose [14]. Konigseder *et al.* proposed an attitude estimation strategy based on extended Kalman filter (EKF) and unscented Kalman filter (UKF) [15].

B. Convolutional Neural Network

Recently, CNN based deep learning methods have been widely used and achieved excellent results in detection, classification, and regression [16]–[18]. CNN frameworks with superior performance have been proposed continuously, such as GoogleNet, R-CNN, VGG and so on [19]–[21]. At the same time, the research on the role of network layer is also carried out in depth. Zeiler *et al.* had a deep understanding of the functions of the middle feature layer and the operation of the classifier [22]. Bojarski *et al.* proposed a new method for visualizing CNN [23].

Deep learning also provides a new possibility for attitude angle estimation. Many methods using CNN have emerged. Brachmann *et al.* developed a framework to reduce the uncertainty in the estimation [24]. Tekin *et al.* proposed an approach to predict the pose of a seamless object [25]. Kehl *et al.* presented a method based on SSD [26] to detect 3D model instances and estimate their attitudes [27]. Most of the deep learning based methods are evaluated on the existing datasets

such as LINEMOD [8], OCCLUSION [9] and T-LESS [10]. However, due to limited conditions, the current deep learning based methods have not been widely adopted in practical industrial applications. Therefore, we are trying a method that can be applied in real industrial scenarios.

C. Full Connected Network

The fully connected network is based on dense layer, which is completely different from CNN. In essence, it is a switch that connects all inputs and outputs, with high throughput, high reliability and low latency. What CNN learns is a generalization feature [22]. The shallow convolutional layer extracts intuitive features which are sensitive to the region of the image, and the higher layer extracts more abstract features. Whereas the dense layer flattens the learned features into a one-dimensional vector for integration and adjusts the weight for specific tasks. Therefore, the dense layer is more targeted, more stable and more resistant to interference. For example, transfer learning is an emerging technology that typically takes advantage of dense layers. It replaces original dense layers at the end of the previously trained CNN models to fine tuning for other application tasks [28], [29].

According to the excellent characteristics of the fully connected network, we creatively construct FADN for attitude angle estimation. The task of the dense layer is to establish a parameter mapping from low-dimensional vectors to 3D attitude angles. Therefore, FADN can directly fit the camera imaging model after a large scale deep learning and achieve a fast and end-to-end estimation.

III. METHODOLOGY

Our goal is to accurately estimate the 3D attitude angle of the cooperative object using RGB image, subsequently providing possibilities for real scenarios. Fig. 1 shows an overview of our pipeline that consists of three stages: Marker Process, FADN Structure, and Attitude Output. In the first stage, the RGB image of the marker is processed to obtain feature point coordinates as the input of the next stage. In the second stage, a fully connected network based on the dense layer is constructed to fit 2D/3D mapping relationship. And the network output which matches the dimension of the 3D attitude angle provides an accurate estimation in the last stage.

A. Marker Process

To extract the feature points effectively and steadily, we design the feature marker which is attached to the cooperative object. The marker has eight different sizes of graphics, including circles and rectangles. At the same time, we design a stable algorithm to extract feature points, which are the centroids of graphics. Firstly, we process the input image by graying and filtering. Then, we use Canny operator to get the edge contours of the eight graphics. Finally, the centroid of each graphic is obtained by calculating the moment of the contour. After marker processing, we get eight stable coordinates of the centroids and integrate them into a 16-dimensional array as the input of FADN.

Fig. 1: The architecture of FADN

The graphics with different sizes and shapes in the feature marker can improve the accuracy of the coordinates extraction and avoid the extraction offset effect caused by the similarity of graphics.

B. Feasibility Analysis of FADN

1) *2D/3D Mapping*: RGB images are utilized to estimate attitude angles. Therefore, it is necessary to clarify the basic mapping between 2D images and 3D information [30], [31]. We briefly summarize the process by a matrix multiplication as follow.

$$(u, v, 1)^T = M \begin{bmatrix} I & 0 \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} R & T \\ 0^T & 1 \end{bmatrix} (X_W, Y_W, Z_W, 1)^T \quad (1)$$

For the convenience of calculation, they are all in the form of homogeneous coordinates. $(u, v, 1)$ represents the pixel coordinate, and $(X_W, Y_W, Z_W, 1)$ is in the world coordinate system. M represents the relevant parameters of the camera. R is a matrix of 3 times 3 for rotation, and T is a matrix of 3 times 1 for translation.

However, in real scenarios, there are factors that can cause nonlinear effects, such as camera distortion and lighting condition. Some of them can be represented by mathematical models, while others cannot even be quantified. In order to visually reflect these nonlinear effects, we quantify one of the camera distortions by mathematical models as an example. In (2), (x, y) represents the physical image coordinate. And k , p_1 , p_2 , and s represent the corresponding parameters of radial distortion, centrifugal distortion and lens distortion respectively. Obviously, these factors bring estimated errors and cannot be ignored.

$$\delta(x, y) = kx(x^2 + y^2) + [p_1(3x^2 + y^2) + 2p_2xy] + s(x^2 + y^2) \quad (2)$$

Due to the presence of nonlinear effects, the 2D/3D mapping is no longer a simple linear matrix regression relationship. Therefore, the network needs excellent nonlinear expression ability to learn and generalize the mapping relationship, so as to further meet the requirements of practical applications.

2) *Expression Ability of FADN*: FADN is a fully connected network architecture. The forward parameter propagation of each layer can be represented by the following mathematical model.

$$\begin{bmatrix} \vartheta_1^{k+1} \\ \vartheta_2^{k+1} \\ \vdots \\ \vartheta_{n^{k+1}}^{k+1} \end{bmatrix} = \begin{bmatrix} \hat{\theta}_{11}^k & \hat{\theta}_{12}^k & \cdots & \hat{\theta}_{1n^k}^k \\ \hat{\theta}_{21}^k & \hat{\theta}_{22}^k & \cdots & \hat{\theta}_{2n^k}^k \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_{n^{k+1}1}^k & \hat{\theta}_{n^{k+1}2}^k & \cdots & \hat{\theta}_{n^{k+1}n^k}^k \end{bmatrix} \begin{bmatrix} \vartheta_1^k \\ \vartheta_2^k \\ \vdots \\ \vartheta_{n^k}^k \end{bmatrix} + \begin{bmatrix} \varsigma_1^k \\ \varsigma_2^k \\ \vdots \\ \varsigma_{n^k}^k \end{bmatrix} = \varpi^k \begin{bmatrix} \vartheta_1^k \\ \vartheta_2^k \\ \vdots \\ \vartheta_{n^k}^k \end{bmatrix} + \begin{bmatrix} \varsigma_1^k \\ \varsigma_2^k \\ \vdots \\ \varsigma_{n^k}^k \end{bmatrix} \quad (3)$$

Here, $\vartheta^k = (\vartheta_1^k, \vartheta_2^k, \dots, \vartheta_{n^k}^k)^T$ denotes n^k neuronal nodes of layer K , and $\vartheta^{k+1} = (\vartheta_1^{k+1}, \vartheta_2^{k+1}, \dots, \vartheta_{n^{k+1}}^{k+1})^T$ denotes n^{k+1} neuronal nodes of layer $K+1$. So, we can deduce that ϖ^k is a matrix of n^{k+1} times n^k . $\sigma^k = (\varsigma_1^k, \varsigma_2^k, \dots, \varsigma_{n^k}^k)^T$ is the weighting bias term of neuron nodes. Each layer has such a nonlinear factor, and all of them are different. Therefore, the weighting bias gives layer K nonlinear expression ability. After multiple layers of iteration, FADN has sufficient nonlinear expressive ability.

In summary, there are two main reasons why we confirm that FADN can achieve attitude angle estimation. First and foremost, in practical application scenarios, the 2D/3D mapping is no longer a simple linear matrix regression relationship. Accordingly, the network needs to have adequate nonlinear expression ability, which is precisely the characteristic of FADN. On the other hand, (1) and (3) are similar in expression form, describing the relationship by matrix multiplication. Naturally, FADN can regress the mapping relationship after learning.

C. Analysis of FADN and CNN

CNN shows good performance in extracting image features and information. And in the field of attitude detection, checkerboard is widely used as a marker for camera calibration because of its excellent topological characteristics. However,

Fig. 2: In (a), the point is in the 3D coordinate system. In (b), (c), (d), the rotation is projected to three 2D plane coordinate systems.

CNN has some deficiencies in checkerboard based detection, which will result in a decrease in detection accuracy and increase of training time.

Next, we will further explain the causes of these problems and give theoretical supports.

The rotation of an object is reflected on its feature point. For simplicity, we analyze the rotation of an object with its feature point. The rotation could be broken down into three directions, around the X -axis, the Y -axis, the Z -axis, and projected to three 2D planes. In Fig. 2(a), after rotation the point (x_0, y_0, z_0) is to (x, y, z) . Without loss of generality, we assume that the object rotates θ degree around the X -axis, which is represented by $Rot(X, \theta)$. Project it to the plane of $Y-O-Z$ coordinate system, as shown in Fig. 2(b). The x_0 remains unchanged, and the (y_0, z_0) turns to (y, z) .

$$y = \|(y_0, z_0)\|_2 \times \cos Rot(X, \theta) \quad (4)$$

$$z = \|(y_0, z_0)\|_2 \times \sin Rot(X, \theta) \quad (5)$$

$$\begin{aligned} \Delta y &= |y - y_0| = \|(y_0, z_0)\|_2 \times \cos Rot(X, \theta) - y_0 \\ &= |\xi_{(y_0, z_0)} \cos(\theta + \theta_0) - \xi_{(y_0, z_0)} \cos \theta_0| \end{aligned} \quad (6)$$

$$\begin{aligned} \Delta z &= |z - z_0| = \|(y_0, z_0)\|_2 \times \sin Rot(X, \theta) - z_0 \\ &= |\xi_{(y_0, z_0)} \sin(\theta + \theta_0) - \xi_{(y_0, z_0)} \sin \theta_0| \end{aligned} \quad (7)$$

$\|(y_0, z_0)\|_2$ is rewritten as $\xi_{(y_0, z_0)}$, which represents the 2-norm of vector. Comparing with imaging distance, Δz is much smaller than z_0 . Besides, according to the camera imaging model, the distance remains unchanged by default. Therefore, we ignore Δz later. This rotating motion can be decomposed into the effect of Δy , which is only related to y_0 and θ .

In the same way, if the object rotates around the Y -axis, and is projected to the plane of $X-O-Z$ coordinate system,

as shown in Fig. 2(c), the rotating motion can be decomposed into the effect of Δx , which is only related to x_0 and ψ .

$$x = \|(x_0, z_0)\|_2 \times \sin Rot(Y, \psi) \quad (8)$$

$$z = \|(x_0, z_0)\|_2 \times \cos Rot(Y, \psi) \quad (9)$$

$$\begin{aligned} \Delta x &= |x - x_0| = \|(x_0, z_0)\|_2 \times \sin Rot(Y, \psi) - x_0 \\ &= |\xi_{(x_0, z_0)} \sin(\psi + \psi_0) - \xi_{(x_0, z_0)} \sin \psi_0| \end{aligned} \quad (10)$$

$$\begin{aligned} \Delta z &= |z - z_0| = \|(x_0, z_0)\|_2 \times \cos Rot(Y, \psi) - z_0 \\ &= |\xi_{(x_0, z_0)} \cos(\psi + \psi_0) - \xi_{(x_0, z_0)} \cos \psi_0| \end{aligned} \quad (11)$$

Points on the same horizontal line have the same y_0 and Δy , so they can reflect the same characteristics around the X -axis. Points on the same vertical line have the same x_0 and Δx , so they can also reflect the rotation around the Y -axis. In summary, the pixel points on the checkerboard have equivalent characteristics, so that there is a lot of redundant information learned by CNN.

Based on the above analysis, we conclude that CNN has two main defects in 3D attitude angle estimation. Firstly, the shallow layers of CNN are sensitive to the local regions, and the pixel points in the different areas of the checkerboard represent the same attitude angle. However, the illumination, lens distortion, and noise have regional differences that cause different effects on the same attitude angle. So that the features learned by CNN are unstable, resulting in an unreliable estimation. Secondly, CNN extracts a large amount of redundant information, which increases the parameters in the network and slows down the convergence speed.

Therefore, fully considering the characteristics of attitude angle estimation, a fully connected network is more suitable. It implicitly constructs the mapping relationship between the pixel coordinate input of the 2D image and the output of 3D attitude angle. The avoidance of redundant information guarantees a significant performance improvement of FADN.

D. Analysis of the Number of Feature Point

At the end of the first stage of our method, we integrate the coordinates of the feature points into an array as the input of network. This involves the Perspective N Points (PNP) problem in the field of computer vision, which is first proposed by Fishler *et al.* [32]. Through the image coordinate of N feature points, we can calculate the pose of a camera.

F_o represents object coordinate system, while F_c represents camera coordinate system. $\mathcal{R} = \{P_1^o, P_2^o, \dots, P_n^o\}$ is a set of n feature points on an object, where $P_i^o = (X_i^o, Y_i^o, Z_i^o)^T$. And $\mathcal{R} = \{P_1^c, P_2^c, \dots, P_n^c\}$ is the set of these points in the camera coordinate system, where $P_i^c = (X_i^c, Y_i^c, Z_i^c)^T$. And $\mathcal{S} = \{q_1, q_3, \dots, q_n\}$ is the set of corresponding points on the imaging plane, and $q_i = (x_i, y_i, f)^T$.

The attitude of F_o relative to F_c can be expressed by rotation matrix R and translation vector t .

$$P_i^c = Z_i^c q_i = R P_i^o + t, \quad i \in \{1, 2, \dots, n\} \quad (12)$$

For simplicity, we use the following equation:

$$S P_i = \lambda_i q_i \quad (13)$$

We define $P_i = (X_i^o, Y_i^o, Z_i^o, 1)^T$, $q_i = (x_i, y_i, 1)^T$, $\lambda_i = Z_i^c$, and redefine $R = [e_1, e_2, e_3]$, where $e_i \in \mathbb{R}^3, i = 1, 2, 3$. Thus, we can get $S = [e_1, e_2, e_3, t]$, which is a matrix of four times three.

After matrix transposition, we rewrite it as $S^* = [t_1, t_2, t_3]^T$, where $t_i \in \mathbb{R}^4, i = 1, 2, 3$. Taking S^* into (13), we can get the following three equation relationships.

$$t_1^T P_i = \lambda_i x_i, \quad t_2^T P_i = \lambda_i y_i, \quad t_3^T P_i = \lambda_i \quad (14)$$

Take the third equation into the first two.

$$\begin{cases} t_1^T P_i = t_3^T P_i x_i \\ t_2^T P_i = t_3^T P_i y_i \end{cases} \quad (15)$$

Each feature point provides two equations. Matrix S includes 12 parameters, which describe a rotation and translation. In an ideal situation, three points can solve the 2D/3D position problem.

However, factors including illumination, lens distortion, and noise cause interference. So that it requires more feature points to enhance robustness. Whereas, increasing the number of feature points has negative impacts, such as increase of training time, instability of test results and so on.

E. Analysis of Joint Output and Independent Output

FADN better overcomes the deficiencies of CNN and achieves a desired performance. But there is still room for improvement. Without loss of generality, the rotation of the cooperative object can be decomposed into pitch angle θ , rotation angle ψ , and roll angle φ . From (4) to (11), we know the impacts of θ and ψ . And the object rotates around the Z -axis, which affects Δx and Δy , as Fig. 2(d).

$$\begin{aligned} \Delta x &= |x - x_0| = ||(x_0, y_0)||_2 \times \cos \text{Rot}(Z, \varphi) - x_0| \\ &= |\xi_{(x_0, y_0)} \cos(\varphi + \varphi_0) - \xi_{(x_0, y_0)} \cos \varphi_0| \end{aligned} \quad (16)$$

$$\begin{aligned} \Delta y &= |y - y_0| = ||(x_0, y_0)||_2 \times \sin \text{Rot}(Z, \varphi) - y_0| \\ &= |\xi_{(x_0, y_0)} \sin(\varphi + \varphi_0) - \xi_{(x_0, y_0)} \sin \varphi_0| \end{aligned} \quad (17)$$

The changes of pixel coordinates caused by these three angles are not equivalent. Based on the analysis, we can prove that the θ and ψ are equivalent and different from φ . The φ has more characteristics, so that it converges easier and more quickly. As a result, this inequality will lead to mutual interference and degradation of network performance. Therefore, we creatively adjust the network architecture to output independently, as shown in Fig. 3. The model of the network should be consistent with the application characteristics. In terms of network architecture, we output a three-dimensional vector in the joint mode and three one-dimensional vectors in the independent output mode. The output of the three vectors and the independent loss values during the training process ensure that the attitude angles do not interfere with each other.

IV. EXPERIMENTS

To evaluate the performance of FADN, we build a dataset using the rendering software 3d Max. And a series of experiments are all carried out on this basis. We implement our

TABLE I: Comparison between Different Methods

Method	Roll Error (°)	Pitch Error (°)	Rotation Error (°)
DARPA-OE	0.200	0.200	0.130
NASA-AVGS	0.200	0.200	0.130
NASA-NGAVGS	0.150	0.150	0.340
ETS-VII	0.130	0.130	0.130
FADN	0.036	0.009	0.010

–The **Orbital Express (OE)** [33] undertakes the engineering tasks of automatic meeting and docking of space. Test ASTRO and NEXT at docking on track and a variety of service operations.
 –The **AVGS** [34] is a sensor mounted on the OE to provide the relative position and attitude between ASTRO and NEXTSet.
 –The **NGAVGS** [35] further improves the technical indicators based on AVGS.
 –The **ETS-VII** [36] is a two-way remote operation test platform built by the Japan Space Agency and has been subjected to engineering rendezvous and docking test.

Fig. 3: The network architecture with Joint/Independent Output mode

method using the python deep learning library Keras which is based on the backend of TensorFlow and run it on Intel Xeon E5-2620 v4 @2.10G and two NVIDIA TITAN XP GPUs.

A. Dataset

At present, deep learning based methods are not widely utilized in industrial applications. And the existing datasets are not suitable for industrial applications in real scenarios. Therefore, for further industrial implementation, we build a dataset that is closer to the application background. The cooperative object is a regular cylinder rendered by 3d Max, of which the diameter is 100 mm and the height is 200 mm, as shown in Fig. 4. The marker is attached to this cooperative object. And we use the target camera in 3d Max to capture the images of the marker, which are exactly the same as the ones taken in the real scenarios. The resolution of the image is 640×480 . The focal length of the target camera is 228.712 mm, the viewing angle is 9 degrees, and other parameters are defaults. The cooperative object is initially placed 600 mm from the target camera and on the optical axis, and it can

TABLE II: Comparison between FADN and CNN on Training and Test

Method	Training indicator	Indicator value	Roll Error (°)	Pitch Error (°)	Rotation Error (°)	Global MAE (°)
CNN	MSE	0.03210000	0.043410082	0.019008463	0.010299280	0.024239275
FADN	MSE	0.00087737	0.036935433	0.009307614	0.009920013	0.018721020

Fig. 4: The cooperative object from 3d Max perspective

Fig. 5: The test result of FADN on three attitude angles. **angx/y/z** represents the Roll/Pitch/Rotation Angle.

change the distance within a certain range.

Obviously, the smaller the range of angle changes, the smaller the difference between images, and the more difficult the estimation is. To reflect the superiority of FADN and lay a foundation for subsequent high-precision industrial practices, our own dataset is in a small range with a high accuracy. The range of rotation for each attitude angle is from positive to negative 1 degree at the accuracy of 0.001 degree. Thus, traversing all combinations of these three angles will produce 2001^3 images, which places very high demands on hardware devices. So we randomly generate a million images using 3d Max as the training set. The corresponding 3D attitude angle values are set as labels for supervised training. In the same way, 1000 images are generated as the test set.

B. Evaluation Metrics

To evaluate the performance of the network comprehensively, we adopt two types of evaluation metrics, mean square error (MSE) and mean absolute error (MAE). In the process of training, MSE is used as a loss function to characterize the degree of network fitting. K is a batch in the network training; M is a single image in the batch; $\hat{y}_{(\theta, \psi, \varphi)}$ is the predicted attitude angle value of M , and $y_{(\theta, \psi, \varphi)}$ is the corresponding label.

$$MSE = \frac{1}{|K|} \sum_{M \in K} \|\hat{y}_{(\theta, \psi, \varphi)} - y_{(\theta, \psi, \varphi)}\|_2^2 \quad (18)$$

During testing, MAE is the absolute difference between the output result and the corresponding label, including three single attitude angle errors and Global MAE. V is the test set; N is an image from it; $\hat{y}_{(*)}$ is the predictive value, and $y_{(*)}$ is the corresponding label value. $*$ represents three attitude angles θ , ψ , and φ . Global MAE is an average of these three

errors.

$$MAE(*) = \frac{1}{|V|} \sum_{N \in V} |\hat{y}_{(*)} - y_{(*)}| \quad (19)$$

C. Experiments of End-to-end Estimation

In Section III, we conduct a set of complete and systematic visual theory analyses. The related theory lays a solid foundation for using FADN to estimate the cooperative object's 3D attitude angle. Here, we implement FADN to achieve a complete end-to-end training and test on our dataset.

During the training, we design a fully connected neural network with 15 hidden layers. At the same time, the output is designed as a 3-dimensional port to correspond to the dimension of attitude angle. We use stochastic gradient descent (SGD) to train our network, and set the learning rate to 0.0003 and the momentum to 0.9. And 100 images from the test set are randomly selected to evaluate the performance of FADN. Fig. 5 shows the experimental results.

It is well known that in the field of aeronautics and astronautics, the estimation of attitude angle has higher accuracy requirements. In order to intuitively reflect their requirements, we list a few of the technical indicators and accuracy in some aerospace organizations and engineering projects in Table I [37]. In actual scenarios, their accuracy reaches minute. In our simulation experiments, the accuracy of FADN can reach second. Despite the background gap, it initially illustrates the feasibility and high performance of FADN and lays the foundation for subsequent research, which will be discussed in Section V.

Further, we also conduct a comparative experiment between FADN and CNN. Table II shows the details. During training, FADN has a smaller MSE, easier convergence, and better

TABLE III: Comparison of Network Depth

Method	MSE	Roll Error (°)	Pitch Error (°)	Rotation Error (°)	Global MAE (°)
FADN-D-15	0.00087737	0.0369354334063	0.00930761391297	0.00992001306884	0.018721020
FADN-D-25	0.00083823	0.0367685680891	0.0088886652974	0.00989393024564	0.018517122
FADN-D-33	0.00082239	0.0629554123477	0.01047703092720	0.01000872308280	0.027813722

FADN-D-X represents network models with different depths, and X corresponds to the network depth.

Fig. 6: The effect of the number of feature points on MSE and Global MAE

network fitting. The same conclusion can be obtained during testing, and the estimations of FADN are more accurate.

D. The Number of Feature Points

The number of feature points determines the input array dimension of the network, thus affecting the estimation accuracy. To determine the most efficient form of input array, we explore the effect of the number of feature points.

Fig. 6 shows the effect of the number of feature points on the MSE and Global MAE. In terms of MSE, the network has a higher error when solving 3 points, and the error is significantly reduced during the solution of 4 points. When the number of points is more than 3, the MSE value tends to be stable at a low value, which indicates that the network achieves an active fitting. At the same time, the Global MAE value tends to be smaller and more durable after the number of feature points is more than 6, which indicates that FADN achieves higher estimation precision.

Fig. 7 shows the relationship between the number of feature points and the MAE of three attitude angles in detail. The MAEs of the three attitude angles are significantly different, and the MAE of the roll angle is much more larger than the other two. But as the number of feature points increases, their trends are the same. When the feature points exceed 3, the MAE drops drastically. When the feature points are more than 6, the MAE tends to be stable at a low value.

According to the previous theoretical analysis, we know that FADN can fit 2D/3D mapping relationship when the number

Fig. 7: The effect of the number of feature points on MAE of the three attitude angles

Fig. 8: The effect of the network width on MAE of the three attitude angles

of feature points is more than 3 and less than 6, but this is a multi-solution case. And the experimental results also prove that it is not stable. When the number of feature points reaches 6, FADN can fully achieve a stable solution. Considering the interference of nonlinear factors such as illumination and camera distortion, we should have more than 6 points to increase robustness. However, some of the negative effects follow. After the trade-off, we finally choose the 8-point solution which can better enable FADN to adequately learn the 2D/3D mapping relationship and achieve higher accuracy.

TABLE IV: Comparison of Network Width

Method	MSE	Time/epoch (s)	Roll Error (°)	Pitch Error (°)	Rotation Error (°)	Global MAE (°)
FADN-W-128	0.00097208	8.05	0.03721	0.00885	0.00980	0.01862
FADN-W-256	0.00088498	8.20	0.03516	0.00959	0.00994	0.01823
FADN-W-512	0.00089990	11.70	0.03668	0.00879	0.00998	0.01849
FADN-W-1024	0.00084695	32.51	0.03694	0.00931	0.00992	0.01872
FADN-W-2048	0.00087737	113.47	0.03822	0.00839	0.01021	0.01894
FADN-W-4096	0.00084407	485.42	0.03927	0.00901	0.00980	0.01936

FADN-W-X represents network models with different widths, and X corresponds to the width of the bottom part.

E. Parameter Setting of FADN

1) *Network Depth*: The depth of the network affects the complexity and expressiveness of FADN. We construct three fully connected networks with different hidden layer depths 15, 25 and 33 respectively. The relevant results are shown in Table III. As the depth of the network increases, the MAE of the three angles and the global attitude decreases first and then increases. According to the theoretical analysis and experimental results, the shallow network expression ability is not enough to fit mapping relationship, while the deep network is too complex to represent. Considering the MSE and MAE, the performance of FADN-D-25 is better than the other two depth models.

2) *Network Width*: The width of the network is critical to the dimension of the weight matrix, which affects the number of the parameters and the training time. We compare six fully connected network models with different network width distributions.

The width of the network is designed according to the following principles. All network layers are divided into three parts: the bottom part, the middle part, and the upper part. The width of the bottom part is n , the middle part $2n$, and the upper part is $n/2$. For example, in a 15-layer network, FADN-W-2048 represents that the width of the bottom 5 layers is 2048, the middle 5 layer 4096, and the upper 5 layer is 1024.

In Table IV, as the network width increases, the training time increases dramatically. Because the increase of network nodes in each layer leads to a surge in the total network parameters. Fig. 8 shows the results on MAE of three attitude angles and Global MAE. Under the condition of accuracy of 0.001 degree, the width of the network has a certain influence on the estimation results. Weighing the accuracy and the time, the FADN-W-256 achieves higher performance.

3) *Network Output Structure*: According to the analysis, the three attitude angles are not equivalent in the mapping relationship. The experimental results in Fig. 5, Fig. 7, Fig. 8, Table II and Table III all verify this analysis. The MAE of roll angle is higher than the other two angles. To effectively solve this problem and optimize the performance of FADN, we design a mode of independent output. First and foremost, it is necessary to ensure the dimensional matching of the output. So in the joint output mode, we output a 3-dimensional vector, and three 1-dimensional vectors in the independent output mode. In the network construction level, three dense nodes are set up for output, each with their own separate loss function to ensure training independence.

TABLE V: The Results of Network Output Structure

Method	Roll Error (°)	Pitch Error (°)	Rotation Error (°)
FADN-Jo	0.036935433	0.009307614	0.009920013
FADN-In	0.035938685	0.008320475	0.010075036

FADN-Jo/In represents the join/independent network output mode.

Table V shows the experimental results, which are consistent with the theoretical analysis. The independent output mode performs better, effectively reducing the MAE of these three attitude angles. The reason is that the independent output mode helps to avoid mutual interference caused by the three attitude angles in the neural network fitting process.

V. CONCLUSION AND DISCUSSION

The prosperity of deep learning has inspired us to rethink the attitude angle estimation. This technique overcomes the tedious manual workload of traditional methods. Therefore, it is of great significance to apply the method of deep learning to practical industrial scenarios. We are taking the first step, constructing a fast and high-precision neural network FADN. It is a whole new network based on dense layer and fully considers the characteristics of attitude angle estimation. We propose an adaptive feature points extraction algorithm in the Marker Process stage to get the precise pixel coordinates as the input of FADN. Furthermore, we make innovative adjustments to the network output. The independent output mode effectively avoids mutual interference between the three attitude angles and achieves higher precision.

We construct our own dataset, which is close to the actual application background. The experimental results fully illustrate the feasibility and superiority of FADN, which is the premise of industrialization. Although our simulation experiment settings are as close to the real scenarios as possible, there is still a background gap. The key to this problem is to ensure the consistency of the rendered images and reality, so as to transform our method into actual applications. Therefore, we will discuss several possible strategies and our future work.

For consistency, FADN is trained on the photos taken by real camera. The problem that ensues is the generation and storage of large-scale training samples. Therefore, reducing the required samples is the key. A convolution-deconvolution network helps to solve this problem. Firstly, benchmark photos of each step are taken in a single dimension. For example, cooperative object only rotates around the $X - axis$. As

a result, the real camera captures 2001 photos. Similarly, another 2001*2 images of the other two attitude angles can also be generated. Then, the convolutional-deconvolutional networks can synthesize training samples using benchmark images. Finally, during the test, the image needs to be decoded accordingly to ensure consistency.

If FADN is trained with rendered images, the problem that follows is the inconsistency of parameters between the target camera and the real camera. Therefore, the point is to enhance the adaptability of FADN and the robustness to camera parameters. More types of images will be added to the training dataset, such as taken by different target cameras, having different noises and lighting conditions.

In addition, there is another way to train FADN using the dataset which proportionally blends the rendered and real images. Thus, the diversity of training samples helps to improve the adaptability of the network. At the same time, more specific issues need to be further determined, such as the fusion ratio and network parameters.

In summary, there are many possibilities to apply FADN to actual scenarios, and the corresponding problems will be encountered. In future work, we will further compare the feasibility of these above strategies and pursue exclusively to industrialize FADN.

REFERENCES

- [1] W. Lin, A. Anwar, Z. Li, M. Tong, J. Qiu, and H. Gao, "Recognition and pose estimation of auto parts for an autonomous spray painting robot," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1709–1719, March 2019.
- [2] Z. He, Z. Jiang, X. Zhao, S. Zhang, and C. Wu, "Sparse template-based 6D pose estimation of metal parts using a monocular camera," *IEEE Transactions on Industrial Electronics*, pp. 1–1, 2019.
- [3] J. Park, S. Heo, K. Lee, H. Song, and S. Lee, "Robust facial pose estimation using landmark selection method for binocular stereo vision," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, October 2018, pp. 186–190.
- [4] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *2017 International Conference on 3D Vision (3DV)*, October 2017, pp. 506–516.
- [5] J. Peng, W. Xu, B. Liang, and A. Wu, "Pose measurement and motion estimation of space non-cooperative targets based on laser radar and stereo-vision fusion," *IEEE Sensors Journal*, vol. 19, no. 8, pp. 3008–3019, April 2019.
- [6] S. Mahendran, H. Ali, and R. Vidal, "3D pose regression using convolutional neural networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, July 2017, pp. 494–495.
- [7] C. Xu, L. Zhang, L. Cheng, and R. Koch, "Pose estimation from line correspondences: A complete analysis and a series of solutions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1209–1222, June 2017.
- [8] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, *Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes*, 2012.
- [9] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, *Learning 6D Object Pose Estimation Using 3D Object Coordinates*, 2014.
- [10] T. Hodan, P. Haluza, S. Obdrzalek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects," in *Applications of Computer Vision*, 2017.
- [11] N. Haoachine, M. Berger, and S. Cotin, "Simultaneous pose estimation and augmentation of elastic surfaces from a moving monocular camera," in *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, Sep. 2016, pp. 199–202.
- [12] D. Chwa, A. P. Dani, and W. E. Dixon, "Range and motion estimation of a monocular camera using static and moving objects," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 4, pp. 1174–1183, July 2016.
- [13] Y. Furukawa and J. Ponce, "Accurate camera calibration from multi-view stereo and bundle adjustment," *International Journal of Computer Vision*, vol. 84, no. 3, pp. 257–268, 2008.
- [14] F. Tian, W. Feng, Q. Zhang, X. Wang, J. Sun, V. Loia, and Z. Liu, "Active camera relocation from a single reference image without hand-eye calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.
- [15] F. Konigseder, W. Kemmetmuller, and A. Kugi, "Attitude estimation using redundant inertial measurement units for the control of a camera stabilization platform," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 5, pp. 1837–1844, 2016.
- [16] B. Jiang, J. Yang, Z. Lv, and H. Song, "Wearable vision assistance system based on binocular sensors for visually impaired users," *IEEE Internet of Things Journal*, vol. 6, no. 2, pp. 1375–1383, 2018.
- [17] R. Sakurai, S. Yamane, and J. Lee, "Restoring aspect ratio distortion of natural images with convolutional neural network," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 1, pp. 563–571, January 2019.
- [18] C. Hong, J. Yu, J. Zhang, X. Jin, and K. Lee, "Multi-modal face pose estimation with multi-task manifold deep learning," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2018.
- [19] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [22] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.
- [23] M. Bojarski, A. Choromanska, K. Choromanski, B. Firner, L. Jackel, U. Muller, and K. Zieba, "VisualBackProp: efficient visualization of CNNs," 2016.
- [24] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 3364–3372.
- [25] B. Tekin, S. N. Sinha, and P. Fua, "Real-time seamless single shot 6D object pose prediction," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018, pp. 292–301.
- [26] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [27] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *2017 IEEE International Conference on Computer Vision (ICCV)*, October 2017, pp. 1530–1538.
- [28] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 1019–1034, May 2015.
- [29] C. Chen, B. Jiang, and X. Jin, "Parameter transfer extreme learning machine based on projective model," in *2018 International Joint Conference on Neural Networks (IJCNN)*, July 2018, pp. 1–8.
- [30] R. Hartley and A. Zisserman, "Multiple view geometry in computer vision," *Kybernetes*, vol. 30, no. 9/10, pp. 1865 – 1872, 2008.
- [31] J. J. Craig, *Introduction to Robotics: Mechanics and Control*, 1955.
- [32] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [33] R. T. Howard, A. F. Heaton, R. M. Pinson, C. L. Carrington, and J. E. Johnson, "The advanced video guidance sensor: Orbital express and the next generation," 2008.
- [34] R. T. Howard and T. C. Bryan, "DART AVGS flight results," *Proc Spie*, vol. 6555, 2007.
- [35] J. Lee, C. Carrington, S. Spencer, T. Bryan, R. Howard, and J. Johnson, "Next generation advanced video guidance sensor: Low risk rendezvous

and docking sensor,” in *AIAA SPACE 2008 Conference & Exposition*, 2008, p. 7838.

- [36] T. Imaida, Y. Yokokohji, T. Doi, M. Oda, and T. Yoshikawa, “Ground-space bilateral teleoperation of ETS-VII robot arm by direct bilateral coupling under 7-s time delay condition,” *IEEE Transactions on Robotics and Automation*, vol. 20, no. 3, pp. 499–511, 2004.
- [37] L. Zhao, “Research on mono-vision pose measurement based on features of target,” Ph.D. dissertation, Institute of Optics and Electronics of Chinese Academy of Sciences, 2014.

Jiachen Yang (M’16) received the BSc, MSc and PhD degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2002, 2005 and 2009, respectively. He is currently a professor at School of Electronical and Information Engineering, Tianjin University. From 2014 to 2015, he was a visiting scholar with the Department of Computer Science, School of Science, Loughborough University, U.K. in 2019. He is the leader of Lab of Stereo Visual Information Processing in Tianjin University. His research interests include

image processing, artificial intelligence and information security. In these areas, he has published more than 100 technical articles in refereed journals and proceedings. He is also on the editorial boards of IEEE Access, Sensors, Multimedia Tools and Applications, etc. and held special issue on IEEE Transactions on Industrial Informatics, IEEE Access and Sensors as Lead Guest Editor.

Meng Xi received the B.S. degree in the internet of things engineering from Tianjin University, Tianjin, China, in 2018, where she is currently pursuing the Ph.D. degree at the school of electrical and information engineering. Her research interests include image processing and artificial intelligence.

Bin Jiang received the B.S. and M.S. degree in communication and information engineering from Tianjin University, Tianjin, China, in 2013 and 2016, where he is currently pursuing the Ph.D. degree. He is also a visiting scholar in Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, US, where he is a member of Security and Optimization for Networked Globe Laboratory. He has published more than 30 technical articles in refereed journals and proceedings including IEEE

Transactions on Cybernetics, IEEE Transactions on Industrial Informatics, IEEE Internet of Things Journal et al. His research interests lie in Multimedia Quality and Information Security.

Jiabao Man received the B.S degree in communication and information engineering from Hebei University of Technology, Tianjin, China, in 2017. He is currently pursuing the M.S. degree at school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include multimedia quality evaluation, stereo vision research and pattern recognition.

Qinggang Meng (M06-SM18) received the B.S. and M.S. degrees from the School of Electronic Information Engineering, Tianjin University, China, and the Ph.D. degree in computer science from Aberystwyth University, U.K. He is a Professor with the Department of Computer Science, Loughborough University, UK. He is a fellow of the Higher Education Academy, UK. His research interests include biologically inspired learning algorithms and developmental robotics, service robotics, robot learning and adaptation, multi-UAV cooperation, human motion analysis and activity recognition, activity pattern detection, pattern recognition, artificial intelligence, and computer vision.

Baihua Li received her B.S. and M.S. degrees in Electronic Engineering from Tianjin University and her Ph.D. degree in Computer Science from Aberystwyth University. She has worked at Tianjin University and Manchester Metropolitan University before she joined the Department of Computer Science at Loughborough University. Her research emphasizes innovations and novel applications of internet of things, computer vision and pattern recognition techniques in various fields. More than 50 papers have been published in high impact journals

and conferences of international standard, such as Pattern Recognition, IEEE Trans Syst Man Cybern and IEEE Trans Biomed Eng.