Relative pose estimation of cooperative target based on monocular vision

Weiru Wang¹, Jiasheng Liu*, Peizhang Wu¹, Jianpo Guo¹
¹Ninth department, 713th Research Institute of China Shipbuilding Industry Corporation, Zhengzhou 450015, China
*Corresponding author’s e-mail: Qingw713@outlook.com

Abstract: With the vigorous development of the global logistics industry, warehousing automation has become a popular applied research direction in computer vision. In this paper, the fork and the fork hole’s relative posture during the forklift picking process were taken as the research object. The relative position of the two is solved based on the Aruco cooperative logo PnP algorithm. In this paper, the initial values of the relative poses of the two are addressed by the linear algorithm DLT, and the relative poses are further optimized by Bundle Adjustment (BA). The iterative error observation function is constructed based on the visual pose estimation model, and the Gauss-Newton method is adopted. The least-square fitting iterative process is carried out many times, and the relative position and posture of the fork and the fork hole with high precision are finally obtained.

1. Introduction
The forklift is an industrial handling vehicle, which is often used to load, unload, and stack commodities. It has a significant role in the warehousing and logistics industry. However, in actual operation, it is difficult to align the fork with the fork hole, and it takes a long time. This situation causes a significant amount of time, labor, and material costs to sink. Therefore, for automated warehousing, the forklift's autonomous fork picking of goods needs to be solved [2]. The high-precision relative position and attitude estimation of the fork and the fork hole is the critical link in the forklift's autonomous fork picking process [3]. This paper presents a method for estimating the relative pose of the target based on monocular vision, which is used to obtain the fork's relative posture and the fork hole.

Pose estimation based on vision has vast advantages such as low cost, high precision, non-contact, and a large amount of information[4]. The traditional visual pose estimation methods are divided into two types based on cooperative goals and non-cooperative goals. The pose estimation method based on the collective target needs to artificially set some collaborative signs with known geometric structure feature information on the destination, which can significantly improve the detection accuracy of the goal[5][6]. Compared with other cooperative symptoms, the Aruco code can provide enough corresponding information to obtain the camera pose. At the same time, its internal binary system makes the sign maintain absolute stability in error checking and correction, and the corners are easy to extract. Therefore, this article chooses the Aruco code as the cooperation mark, which is placed near the fork hole. The positional relationship is shown in Figure 1.
In most existing vision-based pose estimation methods, there are two necessary steps: feature extraction and pose parameter solution. The extracted features are used to construct 2D-3D correspondence, and the pose estimation is implemented according to the composed communication. Because the feature points of Aruco codes are rich in information, the features used by the solution method in this paper mainly feature points.

In the case of known camera internal parameters, it is a classic PnP problem (Perspective-N-Point) to solve the relative pose relationship between the target space coordinate system and the camera coordinate system from the corresponding relationship between the feature points and their image points[7]. PnP problem-solving methods are roughly divided into two types: linear algorithms and nonlinear iterative algorithms. P3P has at most four solutions, and Gao et al. gave the complete solution classification by Wu method[8] Quan and Lan gave linear SVD algorithms for P4P and P5P problems[9]. They transformed PnP problems into multiple P3P issues, but the algorithm's computational complexity increased. Moreno and Lepetit proposed an efficient and high-precision linear algorithm EPnP[10] with computational complexity. Still, after experimental verification, the EPnP algorithm is more sensitive to the depth information of the picture[12], that is, when the camera is too close to the target At this time, the algorithm has poor stability and low accuracy. Another typical linear algorithm is the direct linear transformation (DLT immediate linear transformation) algorithm[11], which ignores the orthogonal constraints of the rotation matrix. Through the matrix transformation, linear constraint equations for 12 independent variables can be obtained. The computational complexity of the DLT algorithm is only, but it is greatly affected by noise, and the accuracy is not high. The classic nonlinear iterative algorithm is to minimize a particular cost function as the goal, transform the pose solution problem into a nonlinear least-squares problem, and iteratively solve it using the nonlinear optimization algorithms Gauss-Newton method or Levenberg-Marquardt method. The nonlinear iterative algorithm has high precision and has high anti-noise performance. Still, the calculation amount of this method is much more significant than that of the linear algorithm. The solution process is very time-consuming, and it depends on the initial value of the iteration.

This paper proposes a pose estimation algorithm based on feature point positioning technology, which can effectively and quickly solve the relative pose of the fork and the fork hole, the algorithm obtains the low-precision estimated initial value through DLT solution and further performs several nonlinear iterative operations on the initial cost to get a high-precision six-dimensional posture of the relative position of the fork and the fork hole.

2. Method

2.1 Pose transformation theory
Transforming from one coordinate system to another requires two processes: posture transformation and parallel translation. That is a rotation through three posture angles and adaptation in the direction of three coordinate axes. This article refers to three coordinate systems, namely the camera coordinate system $O_c - X_cY_cZ_c$, fork coordinate system (reference coordinate system) $O - XYZ$, fork hole coordinate system (world coordinate system) $O_w - X_wY_wZ_w$, the three The positional relationship is shown in Figure 1.
In Figure 1, transforming from the coordinate system $O_0 – X_0 \ Y_0 \ Z_0$ to the coordinate system $O – XYZ$ requires rotating the $O_0 – X_0 \ Y_0 \ Z_0$ coordinate system around the $X_0$, $Y_0$, and $Z_0$ axes by angles $A_1$, $A_2$, and $A_3$, respectively, and the corresponding rotation matrices are $R_x(A_1)$, $R_y(A_2)$, $R_z(A_3)$, the translation coordinates are $T_1$, $T_2$, $T_3$. Let the rotation matrix $R = R_x(A_1)R_y(A_2)R_z(A_3)$, translation matrix $T = [T_1 \ T_2 \ T_3]^T$, then the coordinates of the Q point in Figure 1 are converted from $O_0 – X_0 \ Y_0 \ Z_0$ to $O – XYZ$ homogeneous coordinate expression:

$$
\begin{bmatrix}
X \\
Y \\
Z \\
1
\end{bmatrix} =
\begin{bmatrix}
R & T \\
0 & 1
\end{bmatrix}
\begin{bmatrix}
X_0 \\
Y_0 \\
Z_0 \\
1
\end{bmatrix}
$$

(1)

2.2 Vision estimation model

The estimation model is established according to the camera perspective projection imaging model $O_c – X_c \ Y_c \ Z_c$ is the camera coordinate system, $O_0 – X_0 \ Y_0 \ Z_0$ is the world coordinate system, $o – uv$ is the image coordinate system, and $o_{xy} – xy$ is the intersection point of the $Z_c$ axis and the image(Image center point). The conversion relationship of the pixel homogeneous coordinate vector $u = [u, v, 1]^T$ at point Q and the camera similar coordinate vector $X_{cam} = [X_c, Y_c, Z_c, 1]^T$:

$$
Z_c \begin{bmatrix}
u \\
v \\
1
\end{bmatrix} = K[I|0] \begin{bmatrix}
X_c \\
Y_c \\
Z_c \\
1
\end{bmatrix} =
\begin{bmatrix}
f_x & 0 & c_x & 0 \\
0 & f_y & c_y & 0 \\
0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
X_c \\
Y_c \\
Z_c \\
1
\end{bmatrix}
$$

(2)

Among them, $K$ represents the internal parameter matrix of the camera, $f_x, f_y$ represent the equivalent focal length of the camera, $c_x, c_y$ are the pixel coordinates of the image center point. The internal camera parameters can be obtained by camera calibration. In this paper, the calibration toolbox of MATLAB is used to calibrate the camera.

Combining formula (2), the corresponding relationship between homogeneous pixel coordinates and world seamless coordinate vector $X_w = [X_w, Y_w, Z_w, 1]^T$ is:

$$
Z_c u = K[R_{cw}|T_{cw}]X_w
$$

(3)
Where $R_{CW}$ is a $3 \times 3$ rotation matrix, and $T_{CW}$ is a translation matrix between the camera coordinate system and the world coordinate system.

2.3 Direct Linear Transformation

Noting $U = Z_C[u, v, 1]^T = [U_1, U_2, U_3]^T$, according to equation (3), we get the homogeneous coordinate vector of the 2D point to the 3D aspect of the pinhole camera machine Projection matrix $H$:

$$H = K[R_{CW}|T_{CW}]$$

(4)

We record the jth row of matrix $H$ as $h^j$ and simplify to get three equations about $H$:

$$U \times HX_W = \begin{bmatrix} U_2h^3TX_W - U_3h^2TX_W \\ U_3h^1TX_W - U_1h^3TX_W \\ U_1h^2TX_W - U_2h^1TX_W \end{bmatrix} = \begin{bmatrix} 0^T \\ -U_3X_W^T \\ 0^T \end{bmatrix} = 0$$

(5)

If the orthogonal constraint of the rotation matrix $R_{CW}$ is ignored and each element of $H$ is used as an independent variable, the following two independent equations can be derived:

$$\begin{bmatrix} X_W^T \\ 0^T \end{bmatrix} = \begin{bmatrix} h^1 \\ h^2 \end{bmatrix}$$

(6)

Each feature point can provide two independent equations of equation (6). The projection matrix $H$ has 12 unknown quantities. When the number of feature points $N \geq 6$, we can obtain the unique solution of the mapping matrix $H$. Through formulas (1) and (4), we can get the six-dimensional pose of the world coordinate system 3D point relative to the camera coordinate system. However, the pose accuracy achieved by the DLT solution is low, and we can improve efficiency by minimizing the reprojection error.

2.4 Bundle Adjustment

The essence of Bundle Adjustment is an optimization model to minimize reprojection errors. The reprojection error is the projection of real 3D space points on the image plane $u_i$ (the pixels on the image) and reprojection $v_i$ (the virtual pixels obtained by performing the second projection calculation using the 3D points according to the currently calculated projection matrix).

![Figure 2 Schematic diagram of reprojection error](image)

The reprojection error of a single point is:

$$e_i = \|u_i - v_i\|$$

(7)

The BA model minimizes the sum of all reprojection errors:

$$\min_{R,T,X} \sum_i \|u_i - v_i\|$$

(8)
This article uses the Gauss-Newton method to optimize the six-dimensional pose obtained by the DLT algorithm iteratively. The quadratic term of the Taylor expansion of the function is removed:

\[ f(x + \delta x) \approx f(x) + J(x)\delta x \]

In the formula, J is the Jacobi matrix obtained by obtaining the first-order partial derivative of the matrix function. The ideal value of the reprojection error, \( e_i = \|u_i - v_i\| \), is 0, so if we directly set the function to 0, the iterative optimization model of BA can be obtained:

\[ J\delta x = -f(x) \]

\[ x = x + \delta x \]

In summary, this paper constructs two iterative error observation functions \( F_{ui}, F_{vi} \):

\[
\begin{bmatrix}
F_{ui} \\
F_{vi}
\end{bmatrix} = 
\begin{bmatrix}
X_{wi}^T h_1 - u_i X_{wi}^T h_3 \\
X_{wi}^T h_2 - v_i X_{wi}^T h_3
\end{bmatrix}
\]

The variables of the iterative function are six-dimensional pose vector \( p_e = [A_1, A_2, A_3, T_1, T_2, T_3]^T \), then the first-order partial derivatives of and are respectively obtained:

\[
\begin{bmatrix}
J_{ui} \\
J_{vi}
\end{bmatrix} = 
\begin{bmatrix}
\frac{\partial F_{ui}}{\partial A_1}, \frac{\partial F_{ui}}{\partial A_2}, \frac{\partial F_{ui}}{\partial A_3}, \frac{\partial F_{ui}}{\partial T_1}, \frac{\partial F_{ui}}{\partial T_2}, \frac{\partial F_{ui}}{\partial T_3} \\
\frac{\partial F_{vi}}{\partial A_1}, \frac{\partial F_{vi}}{\partial A_2}, \frac{\partial F_{vi}}{\partial A_3}, \frac{\partial F_{vi}}{\partial T_1}, \frac{\partial F_{vi}}{\partial T_2}, \frac{\partial F_{vi}}{\partial T_3}
\end{bmatrix}
\]

The BA optimization iteration model in this paper:

\[
\begin{bmatrix}
J_{ui} \\
J_{vi}
\end{bmatrix} \delta p_e = - \begin{bmatrix} F_{ui} \\
F_{vi}\end{bmatrix}
\]

\[ p_e = p_e + \delta p_e \]

3. Experiment

Based on the above algorithm principle, the overall process design of the pose estimation method in this paper is shown in Figure 3:

Figure 3 Pose estimation algorithm flow chart, image acquisition

Wherein like reference calibration outside the machine, when adjusting the forks coincides with the coordinates of the world coordinate system, when the feature points \( N \geq 6 \), According to formulas (1) and
The rotation translation matrix of the camera relative to the fork coordinate system can be obtained, so the estimation model in this paper is:

\[ U = K[R_c|T_c]X = K[R_c|T_c] \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} X_w \]  

(16)

In this paper, an industrial-grade infrared camera was used. When the fork coordinate system coincided with the fork hole coordinate system (both relative 6-dimensional poses are 0), images were collected, and the poses are solved. During the experiment, the camera signed cooperation from 0.97m. We kept the fork, fork hole camera relatively static. The data collected in the experiment was 500 groups, and the results were as follows:

Figure 4 showed the actual and predicted values of the X-axis translation distance \( T_X \) and the rotation angle \( A_X \), the Y-axis translation distance \( T_Y \), and the rotation angle \( A_Y \). After a comparative analysis, we could obtain that the X-axis translation distance \( T_X \) static error was about 2mm, the dynamic error was in the range of [0mm, 4mm]; the X-axis rotation angle \( A_X \) static error was about 0.05°, the dynamic error was in the field of [-0.05°, 0.1°]; the Y-axis translation distance \( T_Y \) static error was about -1.5mm, the active error is [0.5mm, -3mm] interval; Y-axis rotation angle \( A_Y \) static error was around 0.05°, the dynamic error was [0.15°, 0°] interval.
Figure 5 showed the actual and predicted values of the Z-axis translation distance $T_z$ and the rotation angle $A_z$. After a comparative analysis, we could get that the Z-axis translation distance $T_z$ had a static error of about 0mm and a dynamic error of [-1.5mm, 2mm]; The static error of Z-axis rotation angle $A_z$ was about -0.05°, and the dynamic error was in the range of [-0.15°, 0.05°].

The accuracy of the translation distance of the pose estimation algorithm in this paper reached a millimeter level through experimental verification. The accuracy of the rotation angle was within 0.2°.

4. Conclusions
In this paper, a pose estimation algorithm is studied based on at least six coplanar feature points. Based on the pinhole camera perspective projection model, it is proposed to use the DLT linear algorithm to obtain the initial value of the relative pose of the fork and fork hole, and the Gauss-Newton nonlinear iterative algorithm is used to optimize it, which improves the accuracy of pose estimation. Experimental results show that under ideal conditions, the translation distance error does not exceed 4mm, and the rotation angle error does not exceed 0.2°. The algorithm has an extensive fluctuation range of pose estimation value, and further optimization is needed.

References
[1] Wang Shulin. Structural strength and manufacturing of forklift and fork[J]. Technology and Enterprise, 2016(09): 215+217.
[2] Wang Peng. Robot motion planning for warehouse automation [D]. Zhejiang University, 2017.
[3] Kouskouridas, R., Charalambous, K, Gasteratos, A. 2012. 6DoF object pose measurement by a monocular manifold based pattern recognition technique. Meas. Sci. Technol. 23, 11 (Oct. 2012), 286-286. DOI = 10.1088/09570233/23/11/114005.
[4] Ma, Y., Zhao, H., Gu, F., Bu, P. 2017. Center extraction deviation correction of SMD-LEDs in the target-based vision measurement system. Meas. Sci. Technol. 28, 4 (Jan. 2017). DOI = 10.1088/1361-6501/aa5bfa.
[5] Eberli D, Scaramuzza D, Weiss S, et al. Vision-Based Position Control for MAVs Using One Single Circular Landmark[J]. Journal of Intelligent & Robotic Systems, 2011, 61(1-4):495-512
[6] Wang Tianyu, Dong Wenbo, Wang Zhenyu. Pose measurement system based on monocural vision and fixed target [J]. Infrared and Laser Engineering, 2017, 46(3): 146-153.
[7] Tang Jianliang. Perspective n-point positioning (PnP) research [D]. Institute of Mathematics and Systems Science, Chinese Academy of Sciences, 2003
[8] Wu WJ. Basic Principles of Mechanical Theorem Proving in Geometries, vol. I: Part of Elementary Geometries. Beijing: Science Press, 1984 (in Chinese).

[9] Quan L, Lan ZD. Linear N-point camera pose determination. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1999,1(8): 774–780. [doi: 10.1109/34.784291].

[10] Lepetit V, Moreno-Noguer F, Fua P. EPnP: An accurate O(n) solution to the PnP problem. Int'1 Journal of Computer Vision, 2009, 81(2):155–166. [doi: 10.1007/s11263-008-0152-6]

[11] Abdel-Aziz YI, Karara HM. The direct linear transformation from the comparator coordinates into object space coordinates in close-range photogrammetry. In: Proc. of the ASP/UI Symp. on Close-Range Photogrammetry. 1971. 1–18.

[12] Wohler C. 3D ComputerVision Efficient Methods and Application[M]. Springer, 2009: 3 — 113.