Industrial robot trajectory measurement by target tracking and 3D coordinate measurement in large-scale assembly cell

Mingda Harvey Yang [1], Mohammed A Isa [1], Zi Wang [2], David T Branson III [3], Samanta Piano [1]

[1] Manufacturing Metrology Team, Faculty of Engineering, Advanced Manufacturing Building, University of Nottingham Jubilee Campus, Nottingham, NG8 1BB, UK
[2] Centre for Aerospace Manufacturing, Faculty of Engineering, Advanced Manufacturing Building, University of Nottingham Jubilee Campus, Nottingham, NG8 1BB, UK
[3] Nottingham Advanced Robotics Laboratory, Faculty of Engineering, Advanced Manufacturing Building, University of Nottingham Jubilee Campus, Nottingham, NG8 1BB, UK
* Corresponding Author: mingda.yang@nottingham.ac.uk

Keywords: photogrammetry, robot trajectory measurement, machine vision, spherical target detection, coordinate measurement

Abstract

Ensuring precise and accurate six-degree-of-freedom (6 DoF) positioning of industrial robots is crucial in large-scale automotive and aerospace assembly cells, where even slight positional errors can compromise product quality and production efficiency. Traditional laser tracker–based methods commonly require multiple trackers to acquire complete 6 DoF measurements, leading to high costs and constraints such as limited measurement range, line-of-sight restrictions, and declining accuracy in large workspaces. This study presents a low-cost approach that combines pre-characterized target tracking with photogrammetry to overcome these limitations. By mounting a specially designed target artifact on a robot’s end effector and capturing images from multiple high-resolution cameras, the proposed system performs camera calibration and multi-view triangulation to calculate the artifact’s three-dimensional (3D) position and orientation. Preliminary results validate the method’s efficacy, showing that the edges of spherical targets can be accurately fitted in three dimensions. In the next phase of research, the corrected pose data from the vision system will be transmitted back to the robot, enabling a closed-loop feedback control strategy. Comprehensive validation trials will further quantify the system’s accuracy and robustness, laying the groundwork for broader adoption in large-scale industrial assembly applications.

1. Introduction

In the automobile and aerospace large-scale assembly cells, the precision and accuracy of industrial robot movements across all six degrees of freedom (6 DoF) are paramount for guaranteeing high product quality and operational efficiency. In the aerospace sector, where large airplane components such as fuselage panels or wing sections must be assembled within expansive manufacturing cells, the position and orientation of robots must be controlled with an exceptional degree of accuracy to ensure proper fit-up (Muelaner and Paul, 2008). Traditional trajectory measurement methods often use laser tracker to measure the tool centre point (TCP) of the robot to get the position and orientation relative to a reference, but the disadvantage of this approach is that a single tracker and a single reflector can only measure three DoF translation information, achieving full six DoF measurement typically requires multiple laser trackers and reflectors, thereby increasing both complexity and cost (Khanesar et al. 2023). Moreover, for large-scale production lines, the laser-based tracking approach can suffer from limited measurement ranges, line-of-sight constraints, extensive setup time, and the possibility of degraded accuracy over large distances. Therefore, this study introduces a low-cost approach that integrates pre-characterized target tracking and 3D photogrammetry using machine vision to capture the edges of a target artefact and thus perform 6 DoF pose measurements of robot TCP. The central concept is to attach a specially designed target artefact onto the robot’s end-effector, then track this target using multiple cameras positioned around the assembly environment. By capturing the edges of each unique visual markers from the target artefact, the system is able to reliably compute both translation and rotation of the robot’s TCP. This approach not only avoids the high costs associated with multiple laser trackers but also promises a more flexible and scalable configuration—especially beneficial when dealing with extremely large assembly fixtures found in aircraft manufacturing.

At the core of industrial photogrammetry lies the use of high-resolution cameras to capture images of objects and the use of computer vision and optical measurement principles to analyse these images (Sims-Waterhouse et al. 2017). In particular, this study introduces a trinocular photogrammetry setup: three high-resolution cameras stationed around the assembly cell, each capturing images of the robot-mounted target from different perspectives. Through rigorous calibration procedures that factor in intrinsic and extrinsic parameters, the system ensures that the triangulation computations remain accurate and robust across the workspace. By measuring the spatial relationship among the spherical features, it becomes possible not only to ascertain the TCP’s position but also to derive the target’s 3D orientation—thereby enabling a full 6 DoF measurement of the robot’s end-effector. Furthermore, the use of multiple targets on a single artefact or a set of artefacts can enhance measurement redundancy, improving the overall accuracy and reliability of the pose estimation.

A key advantage of this vision-based approach is its adaptability to large-scale environments. Unlike laser trackers, which must maintain uninterrupted line-of-sight, multiple cameras arranged around an assembly cell can provide overlapping fields of view, reducing occlusion risks. Additionally, such a photogrammetry system can scale to even larger workspaces simply by adding more cameras or repositioning them without incurring the steep costs associated with multiple trackers. The system’s reliance on standard machine vision hardware also presents an opportunity for cost reduction, particularly when compared to specialized metrology equipment used in precision assembly contexts. The machine-vision based system has many applications in object identification, human detection and component inspection (e.g. robot motion capture from OptiTrack, 3D Arena from Hexagon and IONA from Insphere). The proposed research expands the

capability of machine-vision cameras towards measurement guided assemblies. Building upon the philosophy of reconfigurable manufacturing systems, these types of multi-functional assets are much encouraged than traditional single-purposed assets.

The proposed 3D vision and photogrammetry method addresses several challenges associated with traditional laser-based measurement systems for industrial robotics in large-scale assembly cells. By integrating advanced image processing, robust calibration techniques, and redundant visual data acquisition, this method offers an accurate, flexible, and cost-effective means of capturing full 6 DoF robot trajectories. Consequently, it has significant potential for deployment in both automotive and aerospace assembly lines, where maintaining high accuracy across expansive work envelopes is critical for achieving reliable, repeatable, and efficient production outcomes.

2. Methodology

2.1 Camera characterisation

For coordinate measurements using a vision system, the accurate camera characterisation is fundamental for establishing precise relationships between 3D world coordinates and their corresponding 2D projections in an image plane (Ma et al. 2001). In a monocular setup, a single camera is characterised by its intrinsic parameters (K) including focal length $f$and principal point (c_x, c_y), radial and tangential distortion parameters, and extrinsic parameters (rotation R and translation T with respect to a world coordinate system). Under the pinhole camera model, the homogeneous image coordinates $p = [u\ v\ 1]^T$ of a 3D point $P = [X^w\ Y^w\ Z^w]^T$ are given by:

$$s\begin{bmatrix}u\\v\\1\end{bmatrix} = \begin{bmatrix}f_x & 0 & c_x\\0 & f_y & c_y\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}R_{11} & R_{12} & R_{13} & T_1\\R_{21} & R_{22} & R_{23} & T_2\\R_{31} & R_{32} & R_{33} & T_3\end{bmatrix}\begin{bmatrix}X_w\\Y_w\\Z_w\\1\end{bmatrix}. \quad (1)$$

To account for lens distortion, radial distortion parameters (k_1, k_2, k_3) and tangential distortion parameters (p_1, p_2) need to be employed. The image pixels $(u\ ,\ v)$ are first converted to normalised coordinates $(x\ ,\ y)$:

$$x = \frac{u-c_x}{f_x},\ y = \frac{u-c_y}{f_y}. \quad (2)$$

The distance r from the current point to the optical axis can then be calculated by:

$$r^2 = u^2 + v^2 \quad (3)$$

If the ideal coordinates (without distortion) are (x, y), the actual camera captures are (x', y'), which further adjust the pinhole projection:

$$x' = x(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_1xy + p_2(r^2 + 2x^2) \quad (4)$$

$$y' = y(1 + k_1r^2 + k_2r^4 + k_3r^6) + 2p_2xy + p_1(r^2 + 2y^2) \quad (5)$$

The distorted pixel coordinates (u', v') are obtained by remapping (x', y') from:

$$\begin{bmatrix}u'\\v'\\1\end{bmatrix} = \begin{bmatrix}f_x & 0 & c_x\\0 & f_y & c_y\\0 & 0 & 1\end{bmatrix}\begin{bmatrix}x'\\y'\\1\end{bmatrix} \quad (6)$$

In a binocular camera system, the caracterisation not only involves the intrinsic parameters of both left and right cameras ($K_l\ and\ K_r$) but also the extrinsic relationship (relative position and posture) between them.

After completing the single camera characterisation of the two cameras respectively, binocular camera characterisation is mainly focus on determine the $R|T$ between the two cameras based on epipolar geometry shows in figure 1. The left camera is set as the reference, and the extrinsic parameters of the right camera can be expressed via rotation R_{lr} and translation T_{lr}. When a point in the scene is projected onto two images, the two projected points are not arbitrary, but are subject to the mathematical relationship constraint of 'epipolar geometry'. For corresponding feature points x_l and x_r in images coordinate from left and right camera, the epipolar geometry imposes the constraint:

$$x_r^T F x_l = 0 \quad (7)$$

Where F is the 3×3 fundamental matrix, the relationship between F and the camera's intrinsic parameters is:

$$F = K_r^{-T}\ E\ K_l^{-1} \quad (8)$$

Where E is the essential matrix, the pixel coordinates of the two cameras can be converted to their respective normalised coordinates through their intrinsic parameters:

$$\tilde{x}_l = K_l^{-1}\ x_l\ ,and\ \ \tilde{x}_r = K_r^{-1}\ x_r \quad (9)$$

The epipolar constraint now becomes:

$$\tilde{x}_r^T\ E\ \tilde{x}_l = 0 \quad (10)$$

Here the essential matrix E is defined as:

$$E = [T_{lr}]_\times\ R_{lr} \quad (11)$$

Where R_{lr} is the 3×3 rotation matrix of right camera relative to the left camera, T_{lr} is the 1×3 translation matrix from the optic centers of the two cameras, $[T_{lr}]_\times$ represents the skew-symmetric matrix generated by T_{lr},

$$[T_{lr}]_\times = \begin{bmatrix}0 & -T_z & T_y\\T_z & 0 & -T_x\\-T_y & T_x & 0\end{bmatrix} \quad (12)$$

Then using a system of equations consisting of many sets of corresponding feature points, E is solved for by minimising the reprojection error, and then the candidate values for R_{lr} and T_{lr} are obtained by performing the singular value decomposition (SVD) (Stewart, 1993) of E.

$$E = U\ D\ V^T \quad (13)$$

where U, V are orthogonal matrix and D is a diagonal matrix, all with dimensions of 3×3

Perform the triangulation to the feature points deteacted from the left and right cameras, the resulting 3D points should appear in front of the camera (positive depth values), thus removing the erroneous solution (negative depth value) and determining the pose of right camera with respect to the left camera. Finally, bundle adjustment is performed on all parameters (K_l, K_r, R_{lr}, T_{lr}) and the 3D coordinates of all extracted feature points to improve accuracy and minimize the reprojection error on all images (Triggs et al. 2002).

In this study, experiments were conducted using a trinocular vision system consisting of three machine vision cameras $C_1\ C_2\ C_3$, which can be considered as three stereo pairs, $(C_1, C_2), (C_1, C_3),$ and (C_2, C_3) respectively. After the binocular camera calibration for each stereo pair, the relative spatial $R|T$ relationship between the two cameras can be obtained as(R_{12}, T_{12}), (R_{13}, T_{13}) with the optic center of C_1 as the origin and (R_{23}, T_{23}) with the optic center of C_2 as the origin. By enforcing all three stereo camera pairs on a single coordinate system, one reduces redundant degrees of freedom and improves robustness. Fix C_1 as the origin of a world coordinate system, the extrinsic parameters of C_2 and C_3 relative to C_1 can then be deduced by transformation matrix.

If H denotes a transformation of a point from coordinate system A to coordinate system B, then for this rigid homogeneous transformation matrix:

$$H = \begin{bmatrix}R & T\\0^T & 1\end{bmatrix} \quad (14)$$

Then its inverse transformation H^{-1} represent the rigid transformation from B to A, which is given by:

$$H^{-1} = \begin{bmatrix} R^T & -R^T T \\ 0^T & 1 \end{bmatrix} \quad (15)$$

In this three-camera system, for the two sub-camera coordinate systems C_2 and C_3, we know their respective 4×4 transformation matrices relative to the world coordinate system C_1:

$$H_{C_2}^{C_1}: \text{from } C_2 \text{ to } C_1$$

$$H_{C_3}^{C_1}: \text{from } C_3 \text{ to } C_1$$

The transformation matrix from C_2 to C_3 can be obtained through the following relationship:

$$H_{C_2}^{C_3} = (H_{C_3}^{C_1})^{-1} \; H_{C_2}^{C_1} \quad (16)$$

After the above calculations, the transformation relationship between the three cameras in the C_1 coordinate system can be found, and finally the consistency of the multi-camera calibration can be verified by mapping the 3D points between different camera coordinate systems. The known transformation are:

$$H_{C_2}^{C_1} \; (C_2 \text{ to } C_1) \; and \; H_{C_3}^{C_2} \; (C_3 \text{ to } C_2)$$

For any 3D point, the result of mapping this point from the camera coordinate system C_3 to C_2 and then from C_2 to C_1 should be very similar to the result of mapping directly from the C_3 to C_1, there:

$$H_{C_3}^{C_1} \approx H_{C_3}^{C_2} \cdot H_{C_2}^{C_1} \quad (17)$$

Based on the relative translation between each stereo pair, the baseline length (B, distance between the optical centers of each two cameras) can also be calculated.

The cameras used in this work are Basler acA5472-17um camera with the Ricoh FL-BC1220-9M 1.0/1.1 fixed focal lenses, we use 450 ms as the exposure time for the image acquisition of both checkerboard and target artefact. An A0 checkerboard (Non-centrosymmetric) is used for the camera calibration, which contains each square grid with a side length of 54.5 mm, and 19×12 corner points. The three cameras were placed in suitable locations outside of the robot's working area (1m away from the closest position where the robot can move after loading the target) to ensure that they have overlapped field of views (FoV), and a total of 40 different checkerboard positions were placed and 120 photos were taken by three cameras. Figure 2 shows the view from each camera at the same checkerboard position. The visualization of the spatial camera positions for the final experiment is shown in figure 3, baseline length B(C_1, C_2) = 528.7865 mm; B(C_1, C_3) = 391.4946 mm; B(C_2, C_3) = 507.9919 mm.

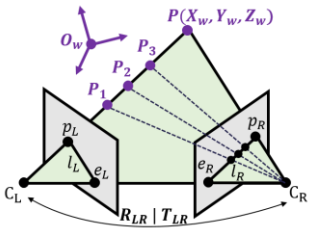


Figure 1. Epipolar geometry in stereo camera pair
C_l, C_r: Optical center of left and right camera
e_l, e_r: Epipole of left and right camera
l_l, l_r: Epipolar line of left and right camera
$P_1 to \; P_3$: Entity points from world coordinate system
O_w: Origin of the world coordinate system

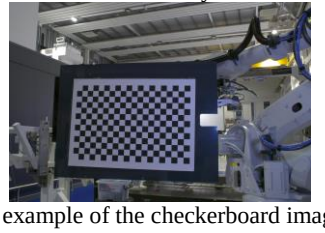


Figure 2. One example of the checkerboard image acquisition by this trinocular camera system

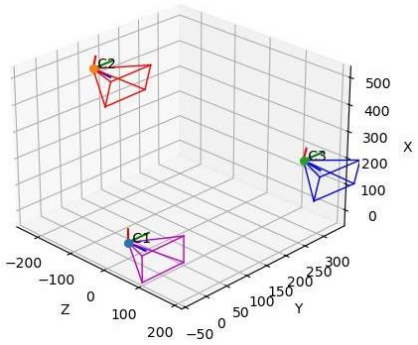


Figure 3. Characterised camera pose visualisation in fused world coordinate system (C_1 as origin).

2.2 Spatial encoded target design

The artefact used in this work is made up of eleven white Nylon plastic spheres that each one are attached onto a dark plates, the diameters (D) of the spheres were selected from the discrete set [35, 40, 60] mm. This ensures that the pixel value occupied by each sphere in each image is at least 50 × 50 pixel in the subsequent target artefact image acquisition step. By mixing diameters on a single plate, the system can reliably distinguish among multiple spheres, even when partial occlusions or differing viewing angles occur (Isa et al. 2024). Neighbouring spheres on the artefact are separated by distances (L) chosen from [225, 240, 255, 270, 285, 300] mm. Each group of three adjacent spheres adopts a unique combinations of D-L so that no two triplets sphere sets share the same arrangement of sphere diameter and spacing (figure 4). This uniqueness provides a critical geometric cue that allows the machine vision algorithm to recognize exactly which region of the artefact is in view. Hence, even when some spheres are temporarily obscured, the system can use the D-L pattern to infer unambiguously which portion of the target is being observed. Each set of three adjacent spheres has distinct diameters and spacing, every set of spheres in the artefact images taken from various camera positions and orientations can be accurately identified. This design guarantees global uniqueness: no identified set spheres is mismatched.

Figure 5 shows the CAD model of the target artefact and the assembled hardware prototype, respectively. To verify whether each sphere met the design requirements after assembly, we used a Mitutoyo Crysta Apex S7106 contact coordinate measuring machine (CMM) with a stylus of 2 mm diameter to measure both the diameter and the position of each sphere (in the X,Y,Z directions). After two repetitions of CMM measurements, the measured diameter of each sphere and the 3D coordinates of each sphere centre was used as the reference for the following vision system feature extraction.

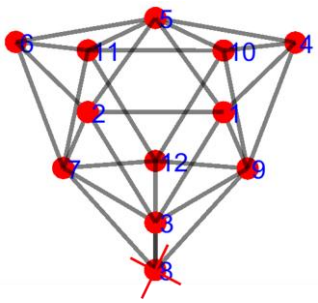


Figure 4. Spherical feature region of interest (RoI, No. 8 are used as the attachment to the robot TCP)

Figure 5. CAD model of the spatial encoded target artefact prototype

2.3 Robot motion and target artefact image acquisition

In the current experimental setup, we focus on assessing translational accuracy by keeping the robot’s end-effector orientation fixed while moving the TCP along the x-axis. Specifically, within the robot’s own coordinate system, the TCP travels from an initial position [1228,955,1890] on the left to a final position [108,955,1890] on the right (unit mm). To collect image data, the robot pauses at every 70 mm step along the x-direction. At each stop, the robot dwells briefly while all three cameras capture images of the robot-mounted artefact from their respective viewpoints, and in total 16 intermediate positions were used to capture images (figure 6 illustrates the straight-line path followed by the robot TCP, and the three-camera position in the world coordinate system). The TCP position read in the robot's coordinate system also needs to be recorded at the same time (shown in the jogging section from the teach pendent), the position information records the translation (T_x, T_y, T_z) and rotation quaternion $(q = (w,x,y,z))$ of the TCP in the robot coordinate system, which needs to be converted to a 3×3 rotation matrix using equation (18) for subsequent operations:

$$R = \begin{bmatrix} 1-2(y^2+z^2) & 2(xy-wz) & 2(xz+wy) \\ 2(xy+wz) & 1-2(x^2+z^2) & 2(yz-wx) \\ 2(xz-wy) & 2(yz+wx) & 1-2(x^2+y^2) \end{bmatrix} \quad (18)$$

Figure 7 a) and b) show the variation of xyz and quaternions q1-q4 of the robot TCP throughout the image acquisition process, respectively. This process ensures the vision system can operate the robust triangulation and pose estimation at each robot stop point. Figure 8 shows the captured images of the target artefact from three cameras perspective.

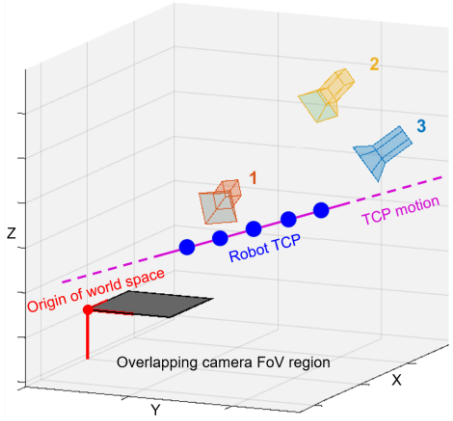
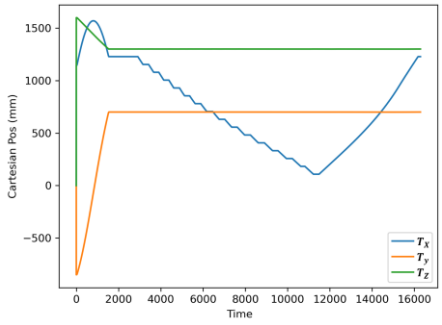
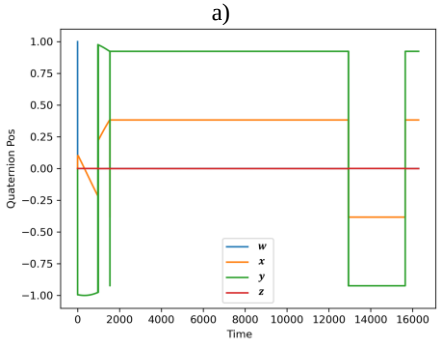


Figure 6. Target artefact images acquisition by this trinocular camera system, red line shows the TCP motion trajectory and the blue dots shows the TCP position for each image acquisition

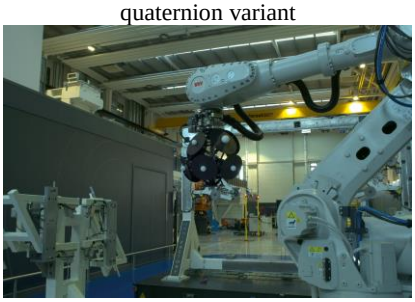


a)

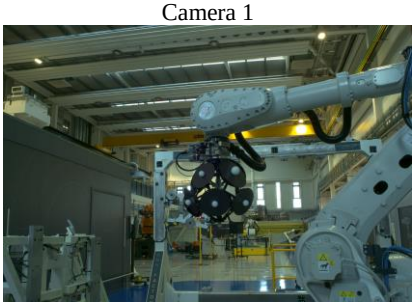


b)

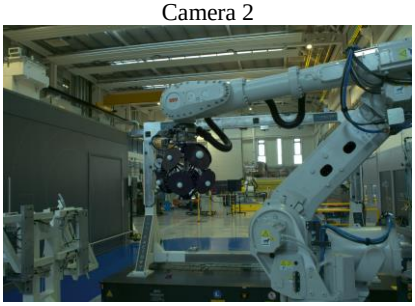
Figure 7. a) Robot TCP translation motion, b) TCP rotation quaternion variant



Camera 1



Camera 2



Camera 3

Figure 8. Example of captured images for the same TCP pose with different camera perspective (C_1 C_2 C_3 from top to bottom)

2.4 Spherical target detection method and preliminary result

A sphere based detection strategy is presented to facilitate pose determination and tracking by utilising spatially encoded sphere identities. The process begins by converting input images from RGB to single-channel grayscale to minimise processing complexity and eliminate distortions caused by uneven colour contrasts. Next, through gradient detection, line spread function (LSF), and edge spread function (ESF) analysis, a refined set of edge points defining the spherical target's region of interest (RoI) is extracted. When sampling the edge profile, a cross-sectional line is drawn across the suspected sphere boundary, and gradient values along this line are recorded. Edge points that fail to meet the edge-response criteria or exceed noise model thresholds are discarded to eliminate false edges caused by extreme brightness or darkness. Based on the estimated noise level, a 5σ standard deviation threshold is applied to ensure the retention of only reliable edge points. Figure 9 illustrates a cross-sectional line of a sampled edge profile. ESF (Masaoka et al. 2014) describes the intensity transition from one side of the edge to the other. In this work ESF are modelled using logistic function thereby better characterizing the transition from dark to bright (or vice versa), and the exact location of the edge can be derived more accurately from the fitted curve (figure 10). By applying Gaussian fitting to the LSF (Rossmann, 1969), the sub-pixel edge location (at the point of maximum gradient) can be determined (figure 11).

The projection of the sphere in the image plane is inferred from the candidate RoI (edge points) of the sphere. When the camera view is nearly orthogonal to the sphere, the projected shape is close to a circle. In this case a Circular Hough Transform (CHT) (Yuen et al. 1990) can be used to detect the centre and radius of that circle, estimated and compensated the eccentricity for all detected image features. When the tilt angle of the sphere is large, the projection of the sphere in the 2D image approximates an ellipse and requires an ellipse fitting approach, given a set of edge points $[(x_i, y_i)]$, the general ellipse equation can be formulated as:

$$\alpha x^2 + \beta xy + \gamma y^2 + \delta x + \epsilon y + 1 = 0 \quad (19)$$

And solve for $(\alpha, \beta, \gamma, \delta, \epsilon)$ in a least-squares sense. With the known camera intrinsic, extrinsic (R|T) and distortion parameters, the ellipse parameters can be back-projected into 3D space to estimate the true center and radius of the corresponding sphere. Figure 12 a) shows the detected ellipse and b) shows the corresponding incidence relation by epipolar geometry constrain and the four epipolar lines.

Once a sphere's 3D position and approximate diameter are known, the system assigns an identity based on the CMM-measured reference of the entire artefact. Figure 13 shows the detection result, from a) the corresponding sphere indices [5, 10, 11, 12] and from b) the corresponding sphere indices [4, 9, 10, 11, 12]. By determining each sphere's identity, the system can compute the artefact's 6 DoF pose relative to the camera(s). This is crucial for downstream robot TCP position tracking task (Isa et al. 2024).

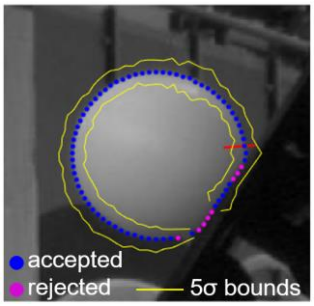


Figure 9. Spherical RoI feature detection

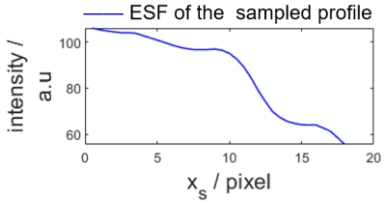


Figure 10. ESF of the sampled edge profile

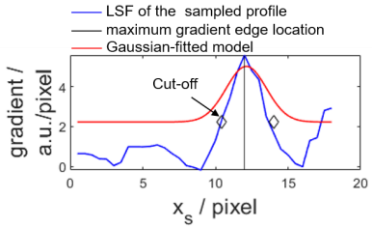
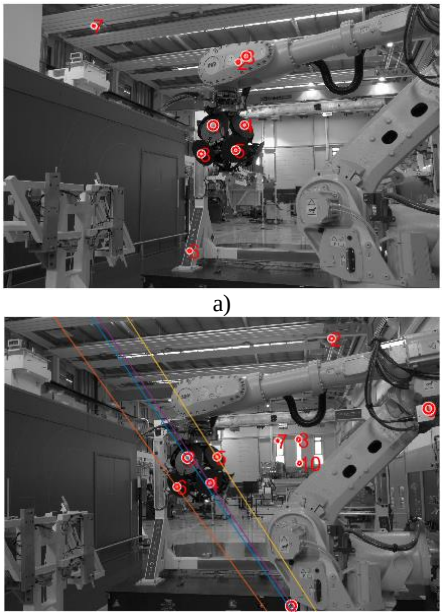


Figure 11. LSF of the sampled edge profile


a)

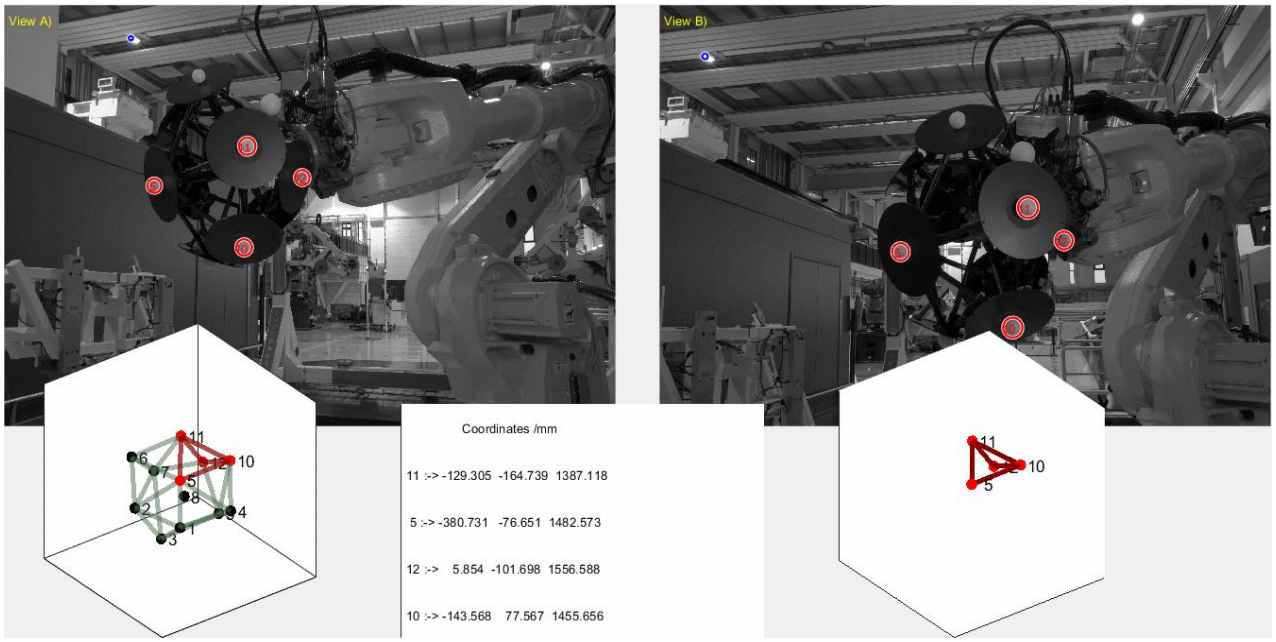
b)
Figure 12. Corresponding ellipse from two images

3. Achieved accuracy and validation

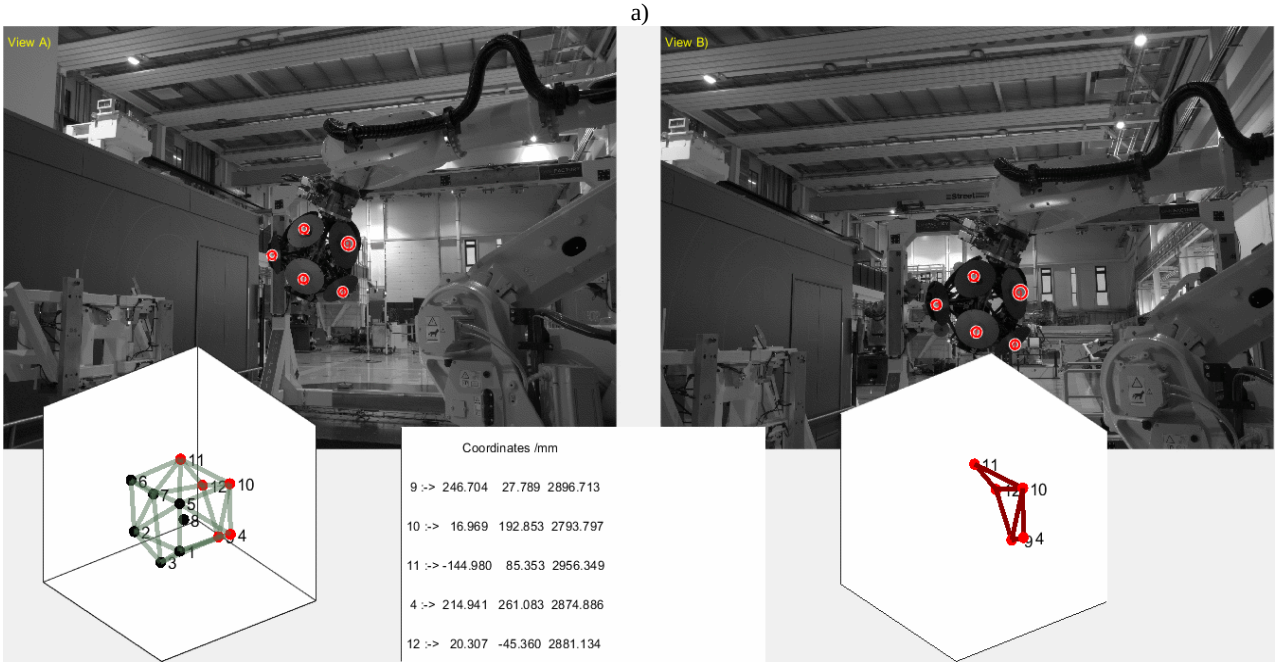
Future work will investigate if the results of this study meet existing measurement verification standards, including VDI/VDE 2634 or ISO 10360 for the 3D coordinate measurement (vision part) and ISO 9283 for the robot TCP positioning accuracy. Finally, the vision system should attain uncertainty within 0.4 mm (+/- 0.02 mm/3 mrad) in the working area of 3 m × 2 m × 1.5 m roughly based on robot pick and place bracket storage position (Figure 8 bottom left). As a validation of the 6 DoF position and orientation accuracy for the vision system, we will use two commercial instruments to implement the validation: Geodetic Systems V-STARS and Leica T-Mac30-I with laser tracker.

Acknowledgements

The authors would like to thank the UKRI Research England Development (RED) Fund for funding this work via the Midlands Centre for Data-Driven Metrology and acknowledge Made Smarter Innovation - Research Centre for Connected Factories (EP/V062123/1) for funding this study.



a)



b)

Figure 13. Detected sphere identification with CMM data

References

Muelaner, J.E., and Paul, M., 2008. Large scale metrology in aerospace assembly. 5th International Conference on Digital Enterprise Technology.

Khanesar, M.A., Yan, M., Isa, M., Piano, S. and Branson, D.T., 2023. Precision denavit–hartenberg parameter calibration for industrial robots using a laser tracker system and intelligent optimization approaches. Sensors, 23(12), p.5368.

Sims-Waterhouse, D., Piano, S., and Leach, R., 2017. Verification of micro-scale photogrammetry for smooth three-dimensional object measurement. Measurement Science and Technology, 28(5), p.055010.

Ma, Y., Soatto, S., Kosecka, J., and Sastry, S., 2003. An invitation to 3d vision: From images to models springer verlag.

Stewart, G., W., 1993. On the early history of the singular value decomposition. *SIAM review*, *35*(4), 551-566.

Triggs, B., McLauchlan, P., F., Hartley, R., I., and Fitzgibbon, A., W., 2000. Bundle adjustment—a modern synthesis. In Vision Algorithms: Theory and Practice: International Workshop on Vision Algorithms Corfu, Greece, September 21–22, 1999 Proceedings (pp. 298-372). Springer Berlin Heidelberg.

Isa, M.A., Leach, R., Branson, D., and Piano, S., 2024. Vision-based detection and coordinate metrology of a spatially encoded multi-sphere artefact. Optics and Lasers in Engineering, 172, p.107885.

Masaoka, K., Yamashita, T., Nishida, Y., and Sugawara, M., 2014. Modified slanted-edge method and multidirectional modulation transfer function estimation. Optics express, 22(5), 6040-6046.

Rossmann, K., 1969. Point spread-function, line spread-function, and modulation transfer function: tools for the study of imaging systems. Radiology, 93(2), 257-272.

Yuen, H. K., Princen, J., Illingworth, J., and Kittler, J., 1990. Comparative study of Hough transform methods for circle finding. Image and vision computing, 8(1), 71-77.

Duda, R.O., and Hart, P.E., 1972. Use of the Hough transformation to detect lines and curves in pictures. Communications of the ACM, 15(1), pp.11-15.

Hartley, R., Zisserman, A., 2003. Multiple view geometry in computer vision. Cambridge University Press.