Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

A Monocular Target Pose Estimation System based on An Infrared Camera

Jiahui Lin*, Han Ma*, Jiyu Cheng, Peng Xu and Max Q.-H. Meng, *Fellow*, *IEEE*

***Abstract*— We present an accurate and robust pose estimation system based on infrared reflective markers and a monocular camera with an infrared filter. The infrared reflective markers are pasted on target object and the camera is mounted on the mobile robot. In the system initialization phase, the correspondence between the markers and image observations are calculated by our correspondence search algorithm. After the initialization, correspondence between markers and observations can be predicted with pose computed in last frame. Thereafter, P2P algorithm based on LevenbergMarquardt is applied to optimize the pose of the current frame. The experiment result shows that our system has larger positioning range than ArUco markers. In addition, this method can estimate object pose in most perspective and is robust to occlusion.**

***Index Terms*— Robotics, Perception, Pose estimation, image processing, PnP**

I. Introduction

Localization and navigation are two fundamental capabilities for mobile robots. A robot needs to know its position by estimating the positions of the landmarks with respect to itself in order to navigate around the environment. In this work, we focus on vision-based target pose estimation. In an indoor scenario, a mobile robot is used to collect the target object. This task can be divided into 3 stages, stage *I* searching objects in a global map, stage *II* estimating the pose of the target object and navigating to the target object, stage *III* approaching and catching target object properly. In this paper, we proposed a method which can be used in the target pose estimation of stage *II*. The mobile robot and the target object is shown in Fig. 1. To achieve this task, firstly, the target object should be distinguished in a cluttered environment. Secondly, the features acquired are used to estimate the pose of target. Thirdly, the pose of the target is sent to the robot navigation system which can guide the robot to approach the target. The last step is to get robot closer to the target by visual servoing based on infrared reflective markers and here we don't provide a detailed description in this paper.

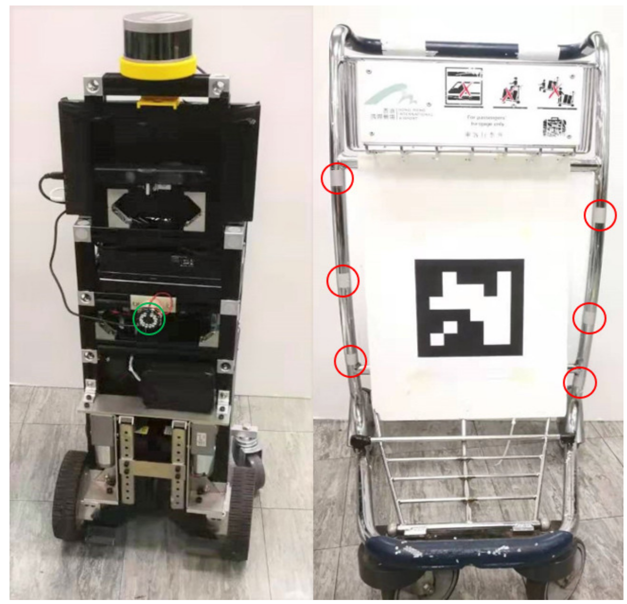

Fig. 1. The left part of this figure is the mobile robot used and the right part is a trolley which is the target object for our experiment. The green circle on the left is an infrared camera and the red circles on the right are infrared reflective markers. There is also an ArUco Marker on the trolley for the contrast experiment.

We propose a target pose estimation system that consists of a camera with an infrared-pass filter and multiple infrared reflective markers. The markers are attached to the target object, while the camera is mounted on the mobile robot. Since this system is based on infrared images, the process of marker detection is much easier than that in RGB images. Besides, this system is applied in indoor environment, thus, infrared noise from sunlight can be excluded. The noise from reflective smooth surface and some infrared light source also can be filtered by our method. To address the occlusion issue, we use multiple markers and choose several of them arbitrarily to estimate the pose of the target. Thus, we can also estimate target pose correctly even when some markers are invisible.

In our experiments, we compared the performance of our system with those using ArUco marker [4] and OptiTrack motion capture system.

The remainder of the paper is organized as follows. Section

* The authors contribute equally to this paper
The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China. email: jiahuilin@cuhk.edu.hk, hanma@link.cuhk.edu.hk, jycheng@ee.cuhk.edu.hk, peterxu@link.cuhk.edu.hk, max.meng@cuhk.edu.hk

II reviews the related work on Monocular pose estimation systems. Section III describes the hardware components of our system and present the structure overview of the system. Section IV describes the overall algorithm in detail. Section V evaluates the performance of the system by experiments.

II. Related Work

Because of the convenience of monocular pose estimation system, it has been widely studied in recent years. [1] proposed a system based on visible spectrum, so cluttered environment and low light environment will result in performance decreasing. Their system use only four markers to perform pose estimation, which can not work properly encountering occlusion issue. Our proposed system can handle occlusion, as long as the number of visible markers is larger than four. ARTag [2] is widely used for pose estimation, which distinguish different markers by different IDs. Because the IDs is generated by different patterns, ARTag have to attach to a large and flag area of target object. In practice, target object may do not have any available large flat areas and only some small curve surface can be used. Besides, the pattern of ARTag has to be large enough to be distinguished, which restrict the effective distance of this kind of marker. ARTag is planar marker, so it's only visible in a limited range. ApirlTag [3] and ArUco [4] has the same aforementioned problems.

The pose-tracking method presented in [5] is based on Active LED Markers. Infrared LEDs blinking at known frequencies are used. This approach works well with low latency. Nonetheless, its precision is limited by the low spatial resolution (128x128) of the DVS. [6] use visible light communitcation to broadcast self-identiy LEDs as global landmarks, while the proposed method use serval identical markers without any communication system. [7] proposed a monocular pose estimation system based on infrared LEDs. In this approach, infrared LEDs and infrared camera are mounted on different robots for mutual localization propose. However, all methods using active markers are not suitable for the pose estimation of no power supply target object.

Nowadays, motion capture systems like OptiTrack are very popular for pose estimation, because of good performance in precision. However, these systems are all expensive and inconvenient, because multiple cameras need to be installed in the fixed position, which make them not suitable for large-scale environment.

III. System Prerequisites

Our system consists of a mobile robot with a 3 dimension Lidar for navigation and mapping, an infrared camera for object pose estimation, serval infrared reflective markers at known positions on the target object. We have already built a global map with a Lidar in the experimental scene and set up a navigation system. The infrared camera is embedded with infrared LEDs and an infrared-pass filter. As the height, roll angle, and yaw angle between camera and trolley are fixed which can be measured in the motion capture system, the pose of the trolley has only 3 DoF. The intrinsic parameters of the camera are obtained with the camera calibration tools of ROS[1]. With at least two markers on the target object captured by the camera, the 6 DOF pose of the target in the camera frame can be estimated. The placement of the infrared reflective markers must avoid symmetric and coplanar to reduce ambiguities of the pose estimation. To increase the robustness of the target object pose estimation, markers on the target object should be visible from as many perspectives as possible. Accuracy can be increased by enlarging the distance between markers on the target objects. To measure the configuration of the makers, the target object is placed in the motion capture system. The object frame can be created by aligning the origin of the motion capture system to one of the markers. To measure the roll angle, yaw angle, and height between the infrared camera and marker on the target object, we can created rigid body for camera and markers and get the transform between camera and camera rigid body with hand-eye calibration tools [8].

IV. Methodology

A. Overview

The flowchart of the overall process is presented in Fig. 2. In the initialization phase, markers configuration and the current camera image serve as the input. Firstly, the marker observations on the image are detected and we search the correspondences between observations and markers for system initialization. The correspondence search algorithm is introduced in section *D* in details. In pose updated phase, we use prediction method in section *F* to determine the correspondences between the observation and markers with previous pose. The correspondences are used by P2P algorithm in section *E* to estimate the pose of current frame.

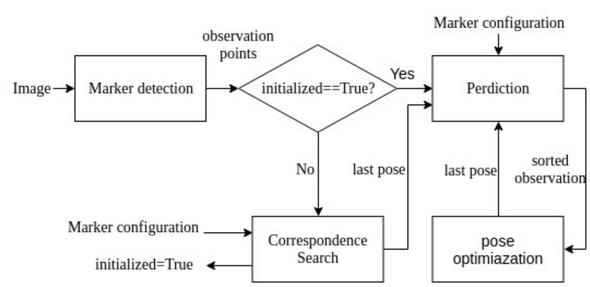

Fig. 2. Algorithm flowchart

B. Notation

We denote the marker positions on the target object as $m_i \in \mathbb{R}^3$, the number of marker on the target object as n_M,

[1]https://www.ros.org

and the marker configuration as $M = \{m_1, m_2, ..., , m_{nM}\}$. The observations of the marker in the image are denoted as $o_j \in \mathbb{R}^2$ in normalized pixel coordinates. The number of observations is n_O and the set of observations is $O = \{o_1, o_2, ..., , o_{nO}\}$. A matching between marker m_i and observation o_j is called a correspondence $c_k = (m_i, o_j) \in \mathcal{C}$. In this application, we split the object state vector to two part since some parameters in the object state vector are considered as constant. As we use Euler angle for orientation representation, the object pose vector is denoted as $X = [x, z, \beta] \in \mathcal{X}$ where β is the pitch angle. The roll angle, yaw angle and height between the trolley and camera is denoted as $Y = [y, \alpha, \gamma] \in \mathcal{Y}$, where α, γ is the roll yaw angle respectively. Although using Euler angle will suffer from data instabilities, we still apply the Euler angle parameterization to make the rotation parameters completely independent.

C. Marker Detection

As we are using a camera with an infrared-pass filter, the camera is only sensitive to 840nm infrared light. With the infrared fill lights embedded in the front of the camera, the infrared-reflective markers will appear very bright in the image compared to their environment on the indoor scene. Thus, a fixed threshold is sufficient to detect the infrared reflective markers. (For the outdoor scene, the light intensity of infrared fill light from infrared LEDs should be large compared to the sunlight.) As for some irregular noise from the reflection of smooth surface, a filter based on the shape of marker is used. In our marker configuration, the projection of markers in the image always shows as rectangles with fixed length-width ratio, which is a good condition for shape based filter.

D. Correspondence Search

After marker detection, we get the observations $\mathcal{O}$ of markers $\mathcal{M}$ on the image. However we cannot distinguish corresponding matches. Thus, we need another step to determine the correspondence $\mathcal{C}$ between observations $\mathcal{O}$ and markers $\mathcal{M}$, so that we can utilize P3P algorithm to compute poses. P3P can get one solution from each combination of four observations in $\mathcal{O}_4$ and each permutation of four markers in $\mathcal{M}_4$. Then, we use this solution to reproject markers $m_l \in \mathcal{M} \setminus \mathcal{M}_4$ which were not used in P3P back to image. Each observation is matched with a unique marker with minimum reprojection error. For each correspondences, the sum of reprojection error can be computed after each observation is matched with a marker. We iteratively search correspondences and find a correspondence with minimum reprojection error which means the estimation initial pose P_{init} is the desire solution. This procedure is summarized in **Algorithm 1**.

For $n_\mathcal{O}$ observations on image and $n_\mathcal{M}$ markers on target object, we have N cases.

$$N = \binom{n_\mathcal{O}}{4} \cdot \frac{n_\mathcal{M}!}{(n_\mathcal{M} - 4)!} \tag{1}$$

The value of N increases rapidly as $n_\mathcal{O}$ and $n_\mathcal{M}$ increase. In practice, we use only a few markers and we only search correspondences when robot initialize the system. Thus this is not an issue.

Algorithm 1 Correspondence Search

$e_{min} \leftarrow -\infty$
$P_{init} \leftarrow 0$
for each $\mathcal{O}_4 \in Combinations(\mathcal{O}, 4)$ **do**
    for each $\mathcal{M}_4 \in Permutations(\mathcal{M}, 4)$ **do**
        $\mathcal{M}_l \leftarrow \mathcal{M} \setminus \mathcal{M}_4$
        $\mathcal{O}_l \leftarrow \mathcal{O} \setminus \mathcal{O}_4$
        $P \leftarrow P3P(\mathcal{O}_4, \mathcal{M}_4)$
        $found \leftarrow 0$
        for each $m \in M_l$ **do**
            $p_m \leftarrow reproject(m, P)$
            for each $o \in \mathcal{O}_l$ **do**
                $e_i = \|p_m - o\|^2$
            end for
            $e_m = min(e_i)$
        end for
        if $e_{min} > \sum_{i=o}^{m} e_i$ **then**
            $e_m = \sum_{i=o}^{m} e_i$
            $P_{init} \leftarrow P$
        end if
    end for
end for
return P_{init}

E. Pose Optimization

The algorithm named as P2P in this part is similar to a Levenberg-Marquardt based PnP algorithm, except that P2P only requires two point pairs to get the optimize solution. As the infrared camera is fixed on the mobile robot and the trolley performs pitch angle rotation only, the roll angle ,yaw angle, and height between infrared camera and trolley remained unchanged. Therefore, we can optimize the pose with respect to x, y, γ, considering roll, yaw and height as observations. Only two pairs of points with four constants is needed to get the optimized solution while accuracy and robustness can be improved by using more correspondences in P2P optimization. The optimized pose is the one that gives the minimum reprojection error with respect to all correspondences in $\mathbb{C}$. This optimization process can be replaced by others PnP algorithm in order to achieve 6 DoF pose estimation, such as method in [10] [11]. For 3 DoF pose estimation, the cost function is

$$f(X) = \sum_{<m,o> \in \mathcal{C}} (p(X) - o) \tag{2}$$

where $p : \mathbb{R}^3 \times \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}^2$ projects a marker into the normalized image panel. To get the optimized trolley pose, X^*, we minimize the reprojection error with the initial pose X generated by P3P algorithm and computed the residual, ΔX, that is

$$\Delta X = \arg\min_{\Delta X} \frac{1}{2} \|J(X)\Delta X + f(X)\|^2 \tag{3}$$

where $J : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ is the Jacobian matrix of the cost function with respect to X. To compute J, let o be the image marker observation, P^w be the marker 3D position in trolley frame, P^c be the marker 3D position in camera frame, p be the marker marker reprojection point, and X, Y is the state vector of the trolley in camera frame. Their relation is given by

$$P^c = R_{roll}(Y_1)R^{pitch}(X_2)R^{yaw}(Y_2)P^w + \begin{bmatrix} X_0 \\ Y_0 \\ X_1 \end{bmatrix} \tag{4}$$

$$p = \begin{bmatrix} P_0^c / P_2^c \\ P_1^c / P_2^c \end{bmatrix} \tag{5}$$

Where $R_{roll}, R_{pitch}, R_{yaw}$ is the rotation matrix in x, y, z axis respectively. We use yaw, pitch, roll order from global to local and vice versa. Therefore, J can be computed with the chain rule, that is

$$J = \frac{\partial f(X)}{\partial X} = \frac{\partial f(X)}{\partial p} \frac{\partial p}{\partial P^c} \frac{\partial P^c}{\partial X} \tag{6}$$

Using automatic differentiation in Ceres[2] can handle the derivatives without computing the closed-form derivatives. The solution of residual is given by

$$J = QR \tag{7}$$

$$\Delta X = -R^{-1}Q^T f \tag{8}$$

where Q is an orthonormal matrix and R is an upper triangular matrix, according to the method of Bjorck [12].

F. Prediction

Considering that if we search correspondences in every frame, the algorithm will be time consuming. Therefore, we use pose computed from last frame to predict pose in the current frame. The detail of the prediction algorithm is described as follows. The pose of last frame is defined as P_{t-1} and the pose of current frame is defined as P_t. The subscript t is the time stamp of each frame. Then, the markers $\mathcal{M}$ are reprojected back to image according to P_{t-1}. The reprojections of $\mathcal{M}$ can be noted as $\mathcal{M}_r$. Because the time interval between two frames is very short, every points in $\mathcal{M}_r$ is very close to one of the obversations in $\mathcal{O}$ in current frame. This kink of information is used to determine the correspondences between $\mathcal{M}_r$ and $\mathcal{O}$. For every points in $\mathcal{O}$, we can search $\mathcal{M}_r$ for a point which is the closest to it and then put the index of this point in a index set $\mathcal{I}_{matched}$. After that, we can get the correspondences between $\mathcal{M}$ and $\mathcal{O}$ regarding to $\mathcal{I}_{matched}$, which are used to optimise the pose of current time by P2P described in *E*. To judge whether the prediction succeeds or not, the distance threshold λ_d is defined. If every points o_i in $\mathcal{O}$ have a reprojection point whose distance to o_i is less than λ_d, the prediction succeeds, otherwise the prediction fails. The process of the prediction and pose optimization is summarized in **Algorithm 2**.

Algorithm 2 Prediction

$\mathcal{I}_{matched} \leftarrow [\ ]$
$\mathcal{M}_r \leftarrow reproject(\mathcal{M}, P_{t-1})$
for $o_i \in \mathcal{O}$ **do**
    for $m_j \in \mathcal{M}_r$ **do**
        if $distance(o_i, m_j) \leq \lambda_d$ **then**
            $\mathcal{I}_{matched}.append(j)$
        else
            $\mathcal{I}_{matched}.append(-1)$
        end if
    end for
end for
if $find - 1 in \mathcal{I}_{matched}$ **then**
    Do correspondence search
else
    $P_t \leftarrow P2P(\mathcal{O}, \mathcal{M}, \mathcal{I}_{matched})$
end if

V. Experiment

A. Benchmark

To evaluate our system, we compare it with ArUco marker [4] and OptiTrack motion capture system. A RER-USBFHDO camera with an infrared-pass filter, a resolution of 640X480 pixels, and a field of view of 90° was used for the experiment. Because the infrared reflective markers can be observed by OptiTrack motion capture system, and the relative position of markers can be acquired directly. To evaluate rotation property and translation property respectively, we conduct two experiments. Firstly, we compare the translation in x direction and z direction and β angle with ArUco maker and OptiTrack. We choose ArUco marker with edge length of 20.0cm to enlarge the working distance. Because the target object in our experiments is a trolley which has no large flat area to paste ArUco maker of such size, we first stick ArUco marker to a flat slab and then attach the flat slab to the trolley. Secondly, since the effective β angle range is limited to $(-90°, 90°)$, we evaluate the rotation property of our system by OptiTrack motion capture system in β range $[0°, 360°]$. The β angle aforementioned is with respect to camera coordinate system. All the algorithms are implemented by C++.

[2]http://ceres-solver.org

In the experiments, we stick 6 infrared reflective markers to the trolley as shown in Fig. 1. In the comparison with ArUco marker and OptiTrack motion capture system, we turn off the infrared LEDs on the mobile robot using infrared from OptiTrack motion capture system. We use ASUS Xtion to capture ArUco marker. The extrinsic parameters between the camera and its end-effector in OptiTrack calibrated with the hand-eye calibration tool [8]. The relative position between ASUS Xtion and infrared camera is also acquired by OptiTrack motion capture system. We align the data from ASUS Xtion and the data from infrared camera by time stamp.

B. Experiment I

In this experiment, the mobile robot kept stationary and the trolley was moved in the visible area of the infrared camera. During the experiment, the trolley was moved from near to far and then from far to near with respect to the infrared camera and we also rotate the trolley during this process. The result of x translation, z translation and β angle are shown in Fig. 3, Fig. 4 and Fig. 5 respectively.

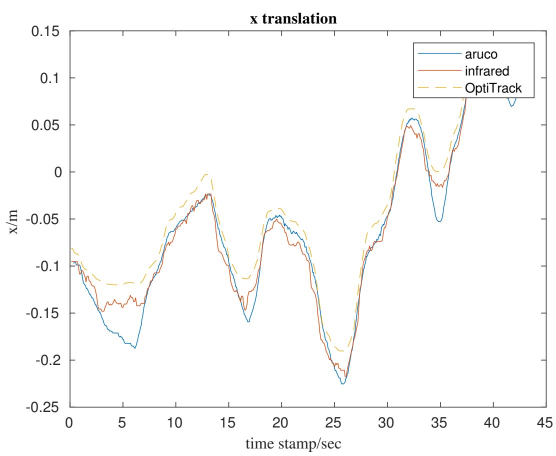

Fig. 3. x translation compared with ArUco marker

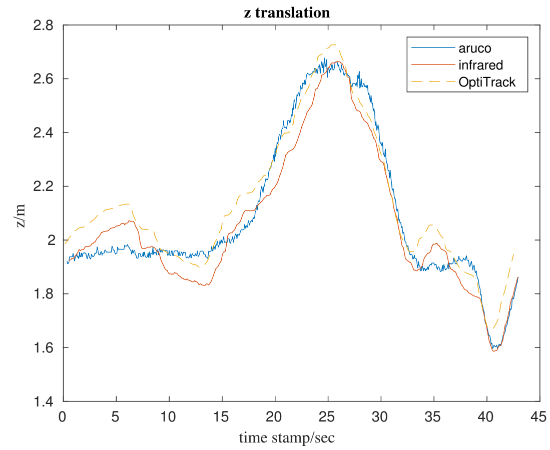

Fig. 4. z translation compared with ArUco marker

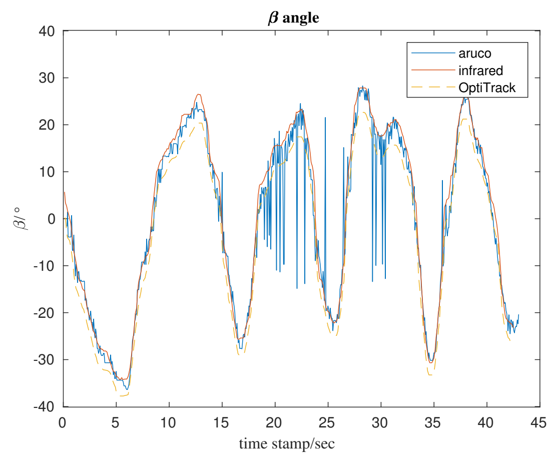

Fig. 5. β angle values compared with ArUco marker

From Fig. 3 and Fig. 4, we know that the translation property of our system is comparative to ArUco marker. Fig. 5 shows that the pose of the ArUco is unstable in motion, but the estimated pose of our system keeps smooth. The reason is that in the middle part the trolley is farthest to the mobile robot and the detection of ArUco marker is based on edge detection, which is sensitive to light condition and distance between the object and the camera. When the distance goes farther, the contour of ArUco marker becomes more blurry. Thus, the result generated by Aruco marker degrades unstable. However, our method is based on the infrared image which is hardly interferred by the indoor lighting condition. Besides, the discrete markers can span a large volume easily which can increase the accuracy of the PnP algorithm. Therefore, the proposed algorithm can perform precise and accuracy pose estimation in a long range compared to the approach of the ArUco marker. To address the occlusion problem, one or two infrared reflective makers is blocked after pose initialization during the experiment. Our system still performs robust pose estimation with occlusions. Fig. 6 shows the detection result of 6 markers and one of the 6 markers is blocked. All the result of this experiment are summarized in TABLE I.

C. Experiment II

Considering that the effective rotation estimation range of ArUco is limited, we also use OptiTrack motion capture system for generating the ground truth to evaluate the rotation property of our system. In this experiment, we rotate the trolley for $0°$ to $360°$ and compare the result from our system with that from OptiTrack motion capture system. The result is shown in Fig. 7. The OptiTrack motion capture system we used comprises 8 cameras in eight different directions, so

Evaluation	$RMSE_x/m$	$\lvert meanError_x \rvert /m$	$RMSE_z/m$	$\lvert meanError_z \rvert /m$	$RMSE_\beta/\circ$	$\lvert meanError_\beta \rvert /\circ$
infrared	0.0893	0.0383	0.0722	0.0717	8.4193	4.3252
ArUco	0.1708	0.1696	0.1027	0.0381	43.5430	24.1061

TABLE I

Experiment I result: $RMSE_x$,$RMSE_z$ and $RMSE_x$ are the RMSE of error of x translation, z translation and β respectively. $|meanError_x|$, $|meanError_z|$ and $|meanError_\beta|$ are the mean of error of x translation, z translation and β respectively.

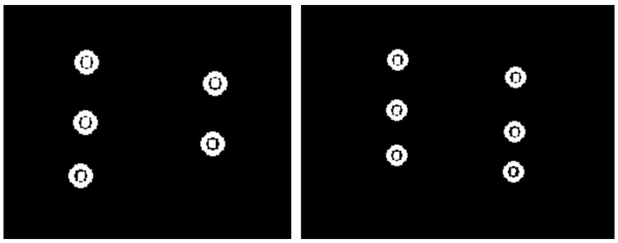

Fig. 6. The white circles in the figure are observations of markers. The left figure is the detection result when one of the six markers is occluded. The right figure is the detection result when none of the markers is blocked.

it can always generate high precision estimation result. Our system which using only a single camera generate very close results to OptiTrack, beacasue our configured markers are visible in most perspective. According to Fig. 7, the estimated rotation of the proposed method is close to the ground truth, with 5.14 degree RMSE.

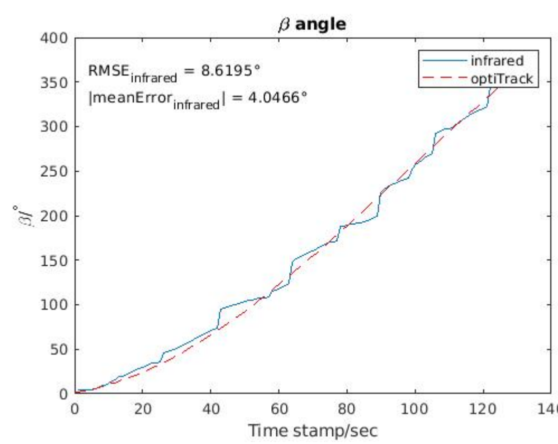

Fig. 7. β angle values compared with OptiTrack motion capture system

VI. Conclusion

Compared to the ArUco marker, the system proposed has better performance in accuracy and precision, since it has a lower mean error and root-mean-square error (RMSE) than the ArUco marker approach. The proposed method chooses markers adaptively which makes the system robust to deal with the occlusion issue. Because of passive infrared reflective markers, the proposed method can hardly be interfered by lighting conditions and cluttering background while the ArUco marker will be affected by motion blur and lighting conditions. Corresponding search is only performed in the initialization phase of the system and the prediction algorithm is time efficient. Thus, our system can be used in real time applications.

References

[1] A. Breitenmoser, L. Kneip and R. Siegwart, "A monocular vision-based system for 6D relative robot localization," 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, 2011, pp. 79-85. doi: 10.1109/IROS.2011.6094851

[2] M. Fiala, "ARTag, a fiducial marker system using digital techniques," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 2005, pp. 590-596 vol. 2. doi: 10.1109/CVPR.2005.74

[3] E. Olson, "AprilTag: A robust and flexible visual fiducial system," 2011 IEEE International Conference on Robotics and Automation, Shanghai, 2011, pp. 3400-3407. doi: 10.1109/ICRA.2011.5979561

[4] Garrido-Jurado, Sergio, Rafael Muoz-Salinas, Francisco Jos Madrid-Cuevas, and Manuel Jess Marn-Jimnez. "Automatic generation and detection of highly reliable fiducial markers under occlusion." Pattern Recognition 47, no. 6 (2014): 2280-2292.

[5] A. Censi, J. Strubel, C. Brandli, T. Delbruck and D. Scaramuzza, "Low-latency localization by active LED markers tracking using a dynamic vision sensor," 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, Tokyo, 2013, pp. 891-898. doi: 10.1109/IROS.2013.6696456

[6] Liang, Qing, Jiahui Lin, and Ming Liu. "Towards Robust Visible Light Positioning Under LED Shortage by Visual-inertial Fusion." 2019 International Conference on Indoor Position and Indoor Navigation, Pisa, Italy, 2019

[7] M. Faessler, E. Mueggler, K. Schwabe and D. Scaramuzza, "A monocular pose estimation system based on infrared LEDs," 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, 2014, pp. 907-913. doi: 10.1109/ICRA.2014.6906962

[8] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3D robotics hand/eye calibration," in IEEE Transactions on Robotics and Automation, vol. 5, no. 3, pp. 345-358, June 1989. doi: 10.1109/70.34770

[9] Xiao-Shan Gao, Xiao-Rong Hou, Jianliang Tang and Hang-Fei Cheng, "Complete solution classification for the perspective-three-point problem," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 8, pp. 930-943, Aug. 2003. doi: 10.1109/TPAMI.2003.1217599

[10] C. -. Lu, G. D. Hager and E. Mjolsness, "Fast and globally convergent pose estimation from video images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 6, pp. 610-622, June 2000. doi: 10.1109/34.862199

[11] F. Moreno-Noguer, V. Lepetit and P. Fua, "Accurate Non-Iterative O(n) Solution to the PnP Problem," 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1-8. doi: 10.1109/ICCV.2007.4409116

[12] A. Bjrck, Numerical Methods for Least Squares Problems, SIAM, Philadelphia, 1996.