Probabilistic Camera-to-Kinematic Model Calibration for Long-Reach Robotic Manipulators in Unknown Environments

1st Petri Mäkinen
Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
petri.makinen@tuni.fi

2nd Pauli Mustalahti
Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
pauli.mustalahti@tuni.fi

3rd Sirpa Launis
Rock Technologies and Drilling
Sandvik Mining and Construction Oy
Tampere, Finland
sirpa.launis@sandvik.com

4th Jouni Mattila
Automation Technology and Mechanical Engineering
Tampere University
Tampere, Finland
jouni.mattila@tuni.fi

***Abstract*—In this paper, we present a methodology for extrinsic calibration of a camera attached to a long-reach manipulator in an unknown environment. The methodology comprises coarse frame alignment and fine matching based on probabilistic point set registration. The coarse frame alignment is based on the known initial pose and assists in the fine matching step, which is based on robust generalized point set registration that utilizes position and orientation data. Comparison with other methods utilizing only position data is provided. The first 6 DOF point set is obtained using a SLAM algorithm running on a camera attached near the tip of a manipulator, whereas the second point set is obtained using a kinematic model and joint encoders. Real-time experiments and a use case are presented. The results demonstrate that the proposed methodology is suited for the application, and that it can be useful in operations requiring precise visual measurements obtained near the tip of the manipulator.**

***Index Terms*—robot vision systems, simultaneous localization and mapping, iterative methods**

I. Introduction

Visual sensors, such as different types of cameras and laser scanners, have seen some significant technological advances in hardware and software in the past decades. These types of sensors are able to provide large amounts of information related to the surroundings, which, due to more affordable and increasing processing power, have been widely adopted in numerous applications, especially in the manufacturing industry in controlled factory environments. However, visual sensors are challenging to utilize in harsh working environments, as the sensors often lack the robustness and reliability required in, for example, mobile work machines that do not operate in strictly controlled environments. Despite this problem, the current direction in the heavy-duty mobile machine industry is toward autonomous systems, where an affordable perception system is essential. This calls for new technologies for perception systems that perform in a robust manner under uncertainties arising from inconsistent working environments and the characteristics of robotic manipulators used in mobile machinery, such as structural flexibility and actuator backlash.

The extrinsic camera calibration problem arises when information measured in a camera's coordinate system needs to be expressed with respect to another sensor's coordinate system. For example, mounting a visual sensor on a robotic manipulator and using the sensor data for control purposes requires determining the sensor's position and orientation in relation to the manipulator's coordinate system, typically defined by its set kinematic model and joint encoders. This is known as the eye-in-hand calibration problem. Finding this extrinsic calibration has been examined in, for example, [1] and [2], where three separate methods were presented. However, each method relied on a visible reference object or point, which is problematic to realize outside controlled environments, such as factories employing stationary industrial robots. Few studies exist for large-scale manipulators working in unstructured or unknown environments, where predefined objects for extrinsic calibration are not practical or available.

Point sets, or point clouds, are a common method of processing and visualizing 3D vision data. These point sets can be used for several applications, such as map building, searching for and tracking known objects, or extrinsic camera calibration by utilizing point sets obtained from two sources. In point set registration, the goal is to find the correspondence between a measured point set and a reference point set. The correspondence between the two point sets is described by a transformation comprising rotation and translation components. Many methods exist for point set registration; the most well-known is the iterative closest point (ICP) algorithm [3] and its numerous variants. In [4], an ICP-based method for extrinsic calibration of an eye-in-hand 2D LiDAR sensor in unstructured environments was presented. A small-scale industrial robot was used in the experiments. Other types of more sophisticated algorithms utilizing 3 degrees of freedom (DOF) position data have also been proposed, such as coherent point drift (CPD) [5] that adopts a probabilistic approach using a Gaussian mixture model (GMM). However, in robotic applications, pose data (3 DOF position and 3 DOF orientation) are readily available. Until recently, point set registration methods utilized only 3 DOF position data, which may not be optimal for robotic applications, as half of the available data is not utilized in the registration process. However, a robust generalized point set registration method was proposed in [6], which builds on the CPD algorithm by incorporating orientation data via the von Mises-Fisher mixture model (FMM) [7]. The resulting hybrid mixture model (HMM) comprises a GMM for position data and an FMM for orientation data, which is perceived as useful in robotic applications especially due to the availability of 6 DOF pose data.

This paper is a continuation of our previous research [8], [9], in which new visual sensor system solutions were investigated for long-reach robotic manipulators in unknown environments, especially underground. In this paper, we focus on the development of a robust, generalized methodology for on-site extrinsic camera-to-kinematic model calibration in such applications. Specifically, 1) an outline for optimal extrinsic camera calibration for long-reach robotic manipulators in unknown environments is presented, with 2) comparison to other similar methods, and 3) real-time experiments with a visual servo use case is discussed.

A two-step methodology is proposed, in which the first step is coarse alignment of the camera frame (or coordinate system) by utilizing a kinematic model of the manipulator and the known initial pose. The second step is fine matching of pose data sequences using robust generalized point set registration [6], a method that benefits not only from position data but also from orientation data and is robust against noise and outliers that can be an encumbrance in visual measurements. For comparison, the fine matching step is also realized with the CPD algorithm [5] and a least-squares-based estimation method [10] that only utilize 3 DOF position data. It is assumed that the intrinsic parameters of the camera are pre-calibrated. Real-time experiments are presented using a laboratory-installed hydraulic crane with 5 m reach. For fine matching, the pose trajectory data are obtained using a camera located near the tip of the manipulator, with a simultaneous localization and mapping (SLAM) algorithm providing the pose estimates. A kinematic model with joint encoders is used to obtain the second set of pose trajectory data. After computing the optimal extrinsic calibration matrix, we apply it to a use case of driving the manipulator to a specific feature detected with the camera. This type of operation is very common in mining and is relatable to bolting, for example, in which supportive rods are inserted into drill holes. In this paper, only a planar case was examined, and ArUco markers [11] were used as the specific features to detect.

The paper is organized as follows: In Section II, we describe the methodology of 6 DOF pose trajectory registration; in Section III, we present the experimental setup, which was used in the real-time experiments; in Section IV, we present the measurements and results; and finally, in Section V, we conclude the paper.

II. Methodology

A. Coarse Frame Alignment

A coarse frame alignment between the camera frame and the encoder-based tool center point (TCP) frame is required for initialization. This alignment reduces the number of iterations in the fine matching step, while also reducing the possibility of the registration algorithm converging to local minima that do not produce correct matching results.

The coarse frame alignment is performed based on the known initial pose of the encoder-based TCP and applying a rigid transformation to the camera frame to roughly align the axes with the encoder-based frame axes. This step must be carefully performed to avoid issues when employing Euler angles.

B. Robust Generalized Point Set Registration

The fine matching of the 6 DOF point sets is based on a probabilistic hybrid mixture model (HMM) [6], [12] that utilizes position and orientation data. Specifically, a GMM is used to model positional uncertainties, whereas an FMM is used to model the orientation uncertainties. The optimal (rigid) transformation between two point sets is solved iteratively using the expectation-maximization (EM) algorithm [13]. The notations used in the HMM formulation are as follows:

- M – Number of points in the encoder-based point set,
- N – Number of points in the SLAM-based point set,
- $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_M] \in \mathbb{R}^{3\times M}$ – encoder-based TCP position vector set,
- $\hat{\mathbf{Y}} = [\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M] \in \mathbb{R}^{3\times M}$ – encoder-based TCP orientation unit vector set,
- $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_N] \in \mathbb{R}^{3\times N}$ – SLAM-based position vector set,
- $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_N] \in \mathbb{R}^{3\times N}$ – SLAM-based orientation unit vector set.

The encoder-based points in $\mathbf{Y}$ are considered the GMM centroids, and the respective unit orientation vectors in $\hat{\mathbf{Y}}$ are considered the mean directions of the FMM. The SLAM-based points in $\mathbf{X}$ are generated by the GMM, and the respective orientation unit vectors in $\hat{\mathbf{X}}$ are generated by the FMM. The goal is to find the optimal rigid transformation (rotation and translation) between the two pose trajectory data sequences $(\mathbf{X}, \hat{\mathbf{X}})$ and $(\mathbf{Y}, \hat{\mathbf{Y}})$. The probability density function of the HMM is expressed as follows:

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n) = \sum_{m=1}^{M+1} P(m) p(\mathbf{x}_n, \hat{\mathbf{x}}_n | m), \tag{1}$$

where

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m) = \frac{\kappa}{(2\pi\sigma^2)^{\frac{3}{2}} 2\pi(e^{\kappa} - e^{-\kappa})} e^{\kappa(\mathbf{R}\hat{\mathbf{y}}_m)^{\mathrm{T}}\hat{\mathbf{x}}_n - \frac{1}{2\sigma^2}||\mathbf{x}_n - (\mathbf{R}\mathbf{y}_m + \mathbf{t})||^2}. \quad (2)$$

The variance parameter of the GMM is denoted by $\sigma^2 \in \mathbb{R}$, the concentration parameter of the FMM is denoted by κ, $(\mathbf{x}_n, \hat{\mathbf{x}}_n)$, $(\mathbf{y}_m, \hat{\mathbf{y}}_m)$ denote arbitrary data points in the point sets, and $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$ denote the rotation and translation transformations applied to $(\mathbf{Y}, \hat{\mathbf{Y}})$, respectively. The assumption is made that the position and orientation data are independent.

To account for noise and outliers in the SLAM-based pose data, an additional uniform distribution is added to the model:

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n|M+1) = \frac{1}{N} \quad (3)$$

with equal membership probabilities $P(m) = \frac{1}{M}$ assumed for the GMM components. The complete HMM is now as follows:

$$p(\mathbf{x}_n, \hat{\mathbf{x}}_n) = w\frac{1}{N} + (1-w)\sum_{m=1}^{M} \frac{1}{M} p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m), \quad (4)$$

where $w \in [0, 1]$ denotes the weight of the uniform distribution. To find the optimal set of parameter estimates $\mathbf{R}, \mathbf{t}, \kappa$, and σ^2, the following negative log-likelihood function is to be minimized:

$$E(\mathbf{R}, \mathbf{t}, \kappa, \sigma^2) = -\sum_{n=1}^{N} \log \sum_{m=1}^{M+1} P(m) p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m). \quad (5)$$

The EM algorithm is used to obtain the parameter estimates in an iterative manner. New parameters are found by minimizing the complete negative log-likelihood function:

$$Q = -\sum_{n=1}^{N}\sum_{m=1}^{M+1} P^{old}(m|\mathbf{x}_n, \hat{\mathbf{x}}_n) \log(P^{new}(m) p^{new}(\mathbf{x}_n, \hat{\mathbf{x}}_n|m)). \quad (6)$$

Then, the encoder-based TCP data $(\mathbf{Y}, \hat{\mathbf{Y}})$ are transformed by applying $\mathbf{R}$ and $\mathbf{t}$. Ignoring constants independent of $\mathbf{R}, \mathbf{t}, \kappa$, and σ^2, (6) is reformulated as follows:

$$Q(\mathbf{R}, \mathbf{t}, \kappa, \sigma^2) = \sum_{n=1}^{N}\sum_{m=1}^{M} p_{mn}\left(\frac{1}{2\sigma^2}||\mathbf{x}_n - (\mathbf{R}\mathbf{y}_m + \mathbf{t})||^2 - \kappa((\mathbf{R}\hat{\mathbf{y}}_m)^{\mathrm{T}}\hat{\mathbf{x}}_n)\right) + \frac{3}{2}N_{\mathbf{P}} \log\sigma^2 + N_{\mathbf{P}} \log(e^{\kappa} - e^{-\kappa}) - N_{\mathbf{P}} \log\kappa, \quad (7)$$

where $p_{mn} = P^{old}(m|\mathbf{x}_n, \hat{\mathbf{x}}_n)$, $N_{\mathbf{P}} = \sum_{n=1}^{N}\sum_{m=1}^{M} p_{mn}$. The Bayes theorem is used to compute the posterior probabilities p_{mn} as follows:

$$P^{old}(m|\mathbf{x}_n, \hat{\mathbf{x}}_n) = \frac{P(m)p(\mathbf{x}_n, \hat{\mathbf{x}}_n|m)}{p(\mathbf{x}_n, \hat{\mathbf{x}}_n)}. \quad (8)$$

According to the EM algorithm, the parameters $\mathbf{R}, \mathbf{t}, \kappa$ and σ^2 are updated in an iterative manner until convergence. The optimal translation $\mathbf{t}^*$ is obtained by minimizing (7) with respect to $\mathbf{t}$, whereas the optimal rotation matrix $\mathbf{R}^*$ is obtained by minimizing (7) with respect to $\mathbf{R}$, respectively. The resulting solutions are as follows:

$$\mathbf{R}^* = \mathbf{V}\ \mathrm{diag}([1, 1, \det(\mathbf{V}\mathbf{U}^{\mathrm{T}})])\ \mathbf{U}^{\mathrm{T}} \quad (9)$$

$$\mathbf{t}^* = \boldsymbol{\mu}_x - \mathbf{R}^*\boldsymbol{\mu}_y, \quad (10)$$

where the mean positional vectors for each point set are defined as follows:

$$\boldsymbol{\mu}_x = \frac{1}{N_{\mathbf{P}}}\mathbf{X}\mathbf{P}^{\mathrm{T}}\mathbf{1}, \quad \boldsymbol{\mu}_y = \frac{1}{N_{\mathbf{P}}}\mathbf{Y}\mathbf{P}\mathbf{1}, \quad (11)$$

$\mathbf{P} \in \mathbb{R}^{M\times N}$ has elements p_{mn} in (8), and $\mathbf{1}$ is a vector of ones. The singular value decomposition (SVD) of $\mathbf{H} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}}$ is used to obtain $\mathbf{V}$ and $\mathbf{U}$, where $\mathbf{H} = \mathbf{H}_1 + \mathbf{H}_2$, $\mathbf{H} \in \mathbb{R}^{3\times 3}$ and

$$\mathbf{H}_1 = \mathbf{Y}'\mathbf{P}\mathbf{X}', \quad \mathbf{H}_2 = \hat{\mathbf{Y}}\mathbf{P}\hat{\mathbf{X}}^{\mathrm{T}}. \quad (12)$$

The matrices $\mathbf{Y}'$ and $\mathbf{X}'$ contain de-meaned positional data $\mathbf{y}'_m = \mathbf{y}_m - \boldsymbol{\mu}_y$ and $\mathbf{x}'_n = \mathbf{x}_n - \boldsymbol{\mu}_x$.

The variance parameter of the GMM is updated by minimizing (7) with respect to σ^2:

$$\sigma^2 = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} p_{mn}(||\mathbf{x}_n - (\mathbf{R}\mathbf{y}_m + \mathbf{t})||^2)}{3N_{\mathbf{P}}}. \quad (13)$$

The concentration parameter (κ) of the FMM is updated using two parts [7]. The first part r_1 results from orientation error and is computed as follows:

$$r_1 = \frac{1}{N_{\mathbf{P}}}\sum_{n=1}^{N}\sum_{m=1}^{M} p_{mn}(\mathbf{R}\hat{\mathbf{y}}_m)^{\mathrm{T}}\hat{\mathbf{x}}_n. \quad (14)$$

The second part r_2 is caused by positional error and is computed as follows:

$$r_2 = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M} p_{mn}\mathbf{x}'_n{}^{\mathrm{T}}\mathbf{R}\mathbf{y}'_m}{\sum_{n=1}^{N}\sum_{m=1}^{M} p_{mn}||\mathbf{R}\mathbf{y}'_m||\ ||\mathbf{x}'_n||}. \quad (15)$$

Then, κ is updated with $\kappa = r(3 - r^2)/(1 - r^2)$, where $r = vr_1 + (1 - v)r_2$, in which $v = 0.5$.

After successful convergence, the optimal calibration matrix for fine matching is written as follows:

$$\mathbf{T}_{fm} = \begin{bmatrix} & \mathbf{R}^* & & \mathbf{t}^* \\ 0 & 0 & 0 & 1 \end{bmatrix}. \quad (16)$$

During iteration, the algorithm was stopped if one of the following conditions was met: $\sigma^2 < 10^{-6}$, $|\sigma_{i+1}^2 - \sigma_i^2| < 10^{-6}$, or 100 iterations were reached. The maximum concentration parameter was also set as $\kappa_{max} = 100$ to avoid computational issues.

The initial iteration parameters were set as follows: $\mathbf{R} = \mathbf{I} \in \mathbb{R}^{3\times 3}$, $\mathbf{t} = \mathbf{0}$, $\sigma_0^2 = \sum_{n+1}^{N}\sum_{m+1}^{M}||\mathbf{x}_n - \mathbf{y}_m||^2/(3MN)$, and $\kappa = 1$.

Finally, the extrinsic camera-to-kinematic model calibration matrix is formulated as follows:

$$\mathbf{T} = \mathbf{T}_{fm}^{-1}\mathbf{T}_{cfa}\mathbf{T}_{slam}, \quad (17)$$

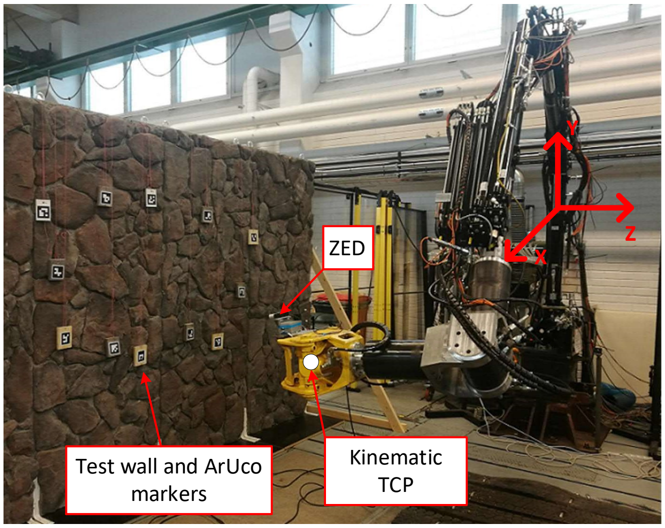

Fig. 1. The experimental setup showing the manipulator, the ZED attached to the claw, a test wall, and 12 ArUco markers placed in the workspace. The base frame of the manipulator is also (roughly) shown.

where $\mathbf{T}_{cfa} \in \mathbb{R}^{4\times4}$ denotes the coarse frame alignment homogeneous transformation matrix, and $\mathbf{T}_{slam} \in \mathbb{R}^{4\times4}$ denotes a single SLAM pose expressed with a homogeneous transformation matrix.

C. Orientation Magnitude Correction

As the FMM employs orientation *unit* vectors, the computed transformation matrix (16) cannot directly produce transformed orientations with true magnitudes. This is resolved by using the encoder measured magnitudes as references. The mathematical expression is as follows:

$$\theta^{slam}_{corr} = \begin{cases} \theta^{slam} - |\theta^{slam}_f - \theta^{enc}_f|, & \text{if } \theta^{slam}_f > \theta^{enc}_f \\ \theta^{slam} + |\theta^{slam}_f - \theta^{enc}_f|, & \text{else} \end{cases}, \tag{18}$$

where θ represents a current Euler angle, and θ_f denotes the final value of the respective variable in a calibration data sequence.

III. Experimental Setup

The experimental setup is shown in Fig. 1. The main components and systems as follows:

- HIAB033 hydraulic crane with an additional 3 DOF wrist, and each joint was equipped with an incremental encoder,
- ZED stereo camera running a SLAM algorithm,
- A dSPACE real-time control platform,
- A test wall comprising decorative stones to simulate a mine and provide visual features,
- Markers attached to the wall acting as specific features.

A dSPACE DS1005 PPC controller board served as the real-time control system, and a 2 ms sampling period was used in the experiments.

TABLE I
DH Parameters of HIAB033 With a 3 DOF Wrist

Joint	α_i	a_i	θ_i	d_i
Rotation	$\pi/2$	a_1	θ_1	d_1
Lift	0	a_2	θ_2	0
Tilt	$\pi/2$	a_3	$\theta_3+\pi/2$	d_3
Wrist 1	$\pi/2$	0	θ_4	d_4
Wrist 2	$-\pi/2$	0	θ_5	0
Wrist 3	0	0	θ_6	d_6

A. HIAB033 Hydraulic Crane With 3 DOF Wrist

A forward kinematic representation of the manipulator is formulated using the Denavit-Hartenberg (DH) parameters, which are presented in Table I in symbolic form. The rigid transform from the base frame to the TCP frame, $\mathbf{T}_{enc}$, is formulated as follows:

$$\mathbf{T}_{enc} = \mathbf{T}_{j1}\mathbf{T}_{j2}\mathbf{T}_{j3}\mathbf{T}_{j4}\mathbf{T}_{j5}\mathbf{T}_{j6}, \tag{19}$$

where joint specific transforms $\mathbf{T}_{ji},\ i \in \{1, ..., 6\}$ are computed with

$$\mathbf{T}_i = \begin{bmatrix} c\theta_i & -s\theta_i c\alpha_i & s\theta_i s\alpha_i & a_i c\theta_i \\ s\theta_i & c\theta_i c\alpha_i & -c\theta_i s\alpha_i & a_i s\theta_i \\ 0 & s\alpha_i & c\alpha_i & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{20}$$

while applying the respective DH parameters for each joint. Additionally, $s = sin$ and $c = cos$.

B. Visual Measurements

A Stereolabs ZED stereo camera was used in the experiments. It was installed near the tip of the manipulator, and the ROS node provided by the manufacturer was used to publish 720p images.

For SLAM, the open-source version of ORB-SLAM2 Stereo [14] was utilized. The algorithm ran in real time and the pose data were transmitted to the dSPACE controller board via UDP. A 2.5×4 m textured wall served as the main feature extraction area for the SLAM algorithm, because the main focus of this research is underground applications.

For detecting specific markers on the wall, the OpenCV ArUco detection library was used. Twelve ArUco markers were placed around the workspace of the manipulator, as in Fig. 1.

C. Robot Control

Quintic polynomial path planning [15] was used to generate trajectories, and a P-controller with a first-order time delay (PT1 control) was used on the actuator level. The controller's transfer function is described as follows:

$$G(s) = \frac{K_P}{\tau s + 1}. \tag{21}$$

The time delay term (τ) enables larger proportional gain (K_P) values, which reduces static positioning errors when driving to a specific point.

Furthermore, the manipulator was constrained so that only the first three joints (rotation, lift, and tilt) were used for

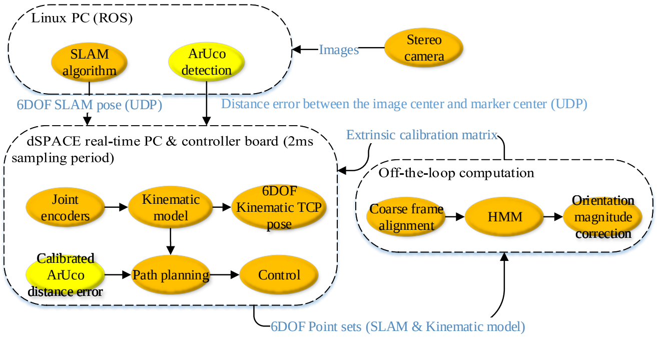
Linux PC (ROS)
SLAM algorithm
ArUco detection
Images
Stereo camera
6DOF SLAM pose (UDP)
Distance error between the image center and marker center (UDP)
dSPACE real-time PC & controller board (2ms sampling period)
Extrinsic calibration matrix
Off-the-loop computation
Joint encoders
Kinematic model
6DOF Kinematic TCP pose
Coarse frame alignment
HMM
Orientation magnitude correction
Calibrated ArUco distance error
Path planning
Control
6DOF Point sets (SLAM & Kinematic model)

Fig. 2. A simplified diagram of the experimental setup: The camera algorithms were processed on a dedicated Linux PC running ROS and the desired camera measurements were sent to the dSPACE real-time control PC via UDP. The camera-to-kinematic model calibration was processed outside the 2 ms control loop. The resulting extrinsic calibration matrix, computed using the two point sets, was then updated in the main control system.

motion, whereas the wrist joints moved only to keep the orientation of the wrist constant.

A simplified diagram of the experimental setup is shown in Fig. 2, in which the orange blocks are related to the overall system, whereas the yellow blocks are related to the use case regarding the ArUco markers.

IV. Measurements and Results

First, a calibration measurement was conducted, in which the manipulator was arbitrarily moved around the workspace to obtain pose data sequences using SLAM and encoder measurements. The recorded data were used to compute the optimal calibration matrix by first applying coarse alignment transform to the SLAM-based pose data by using (17). Then, the coarse frame aligned pose data were used for fine matching, i.e., point set registration with the encoder-based pose data by using the robust generalized point set registration algorithm (4)–(16). The three point sets (encoder-based, SLAM-based with coarse frame alignment, and SLAM-based after fine matching) are shown in Fig. 3. The black points represent the encoder-based TCP position data, whereas the red point sets represent the SLAM-based position data before and after fine matching. The individual pose variables are presented in Fig. 4, where the black lines denote the encoder-based pose variables, and the red lines represent the calibrated SLAM-based pose variables. As illustrated, the algorithm was able to accurately match the pose trajectories resulting from arbitrary motions. Two additional separate calibrations were performed, using the same coarse frame alignment transform, for which the results are shown in Fig. 5 and Fig. 6. The number of iterations required for the fine matching varied between 20 and 25.

After each calibration, the manipulator was driven to 12 different ArUco markers that were placed around the workspace. Monocular detection was employed by using the left lens of ZED, and the middle of the image was treated as the TCP that was to be driven to a marker center. An example image of the left camera view is shown in Fig. 7. The metric distance between the camera center and a marker center was computed based on the known marker size from the image. Then, the point distance was calibrated with the camera-to-kinematic model calibration. Only the rotation part of (16) was required to transform the camera reference to the kinematic frame, meaning this use case does not suffer from the larger position errors in the calibration. The Euclidean distance errors for each marker, for each of the three calibrations, are documented in Table II. The errors were measured from the images. The average positioning error in each measurement was less than 1.0 cm, which is acceptable for this type of application. Only planar results are presented, as the ZED camera was not able to provide reliable depth measurements.

To compare the HMM-based 6 DOF point set registration method with methods utilizing only 3 DOF position data, offline data analysis was conducted for the three measurements. Namely, the very similar CPD algorithm [5] employing a GMM and a simple pairwise least-squares-based estimation algorithm [10] were chosen for comparison purposes. The coarse frame alignment and the orientation magnitude correction steps were performed identically in each case, with only the fine matching step changing. Furthermore, to test the robustness of the three algorithms, Gaussian noise was injected to the X-axis position with varying signal-to-noise ratios (SNR). The SNRs tested were 10, 20, and 30 dB. The effect of added noise to the signal is illustrated in Fig. 8. The root mean square errors (RMSE) for 3 DOF position and 3 DOF orientation in each measured case are presented in Tables III-V. As shown, despite the arbitrary motions in each measurement, the resulting errors are very similar. The position errors are on the centimeter range, whereas the orientation errors are less than 2°. The errors follow from kinematic inaccuracies, for example, due to flexibility, which makes perfect pose trajectory matching practically impossible. The bending of the manipulator is witnessed especially in the X-axis orientations estimated with the SLAM algorithm. Visual measurements are also susceptible to outliers and errors, however, they performed well in the experiments.

As seen from Tables III-V, minimal differences can be found between the fine matching algorithms. When Gaussian noise is added to the X-axis position signal, however, the utilization of orientation data in the HMM appears to slightly improve the matching result compared to the CPD, which only incorporates position data via a GMM. We also experimented by adding similar noise to the other signals, including orientations, which showed uniform results with the presented case. However, to obtain the best result, the weight of the uniform distribution in (4) should be tuned. Same values were used for both the HMM-based method and the CPD. In the cases where noise was added, the least-squares-based algorithm provided the least accurate results. It is worth noting the least-squares-based method required pairwise point sets, whereas the HMM-based method and the CPD are able to process point sets that do not match in length.

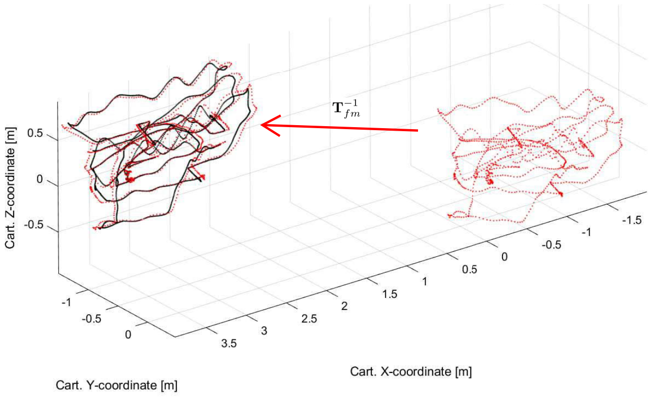

Fig. 3. For computing the calibration matrix, the manipulator was moved arbitrarily around the workspace, while the pose trajectories were recorded for point set registration. The black points represent the encoder-based point set. The right side red points represent the coarse frame aligned SLAM-based point set, whereas the left side red points represent the same SLAM-based point set after fine matching using the HMM.

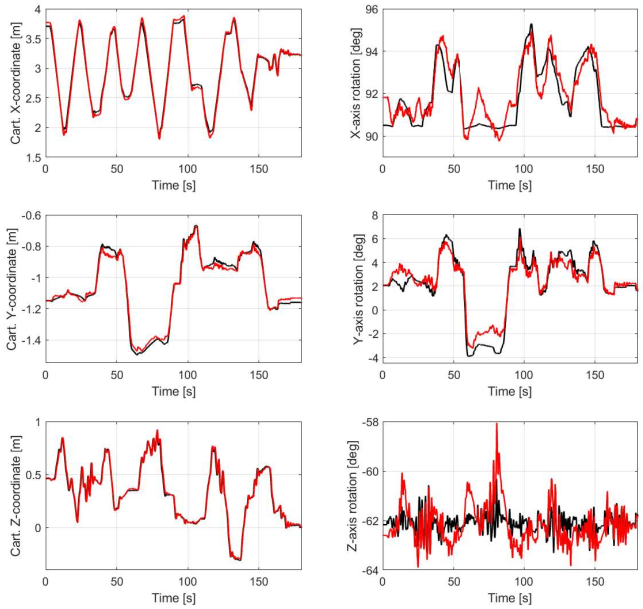

Fig. 4. The first calibration data sequence pose variables: The black lines denote the encoder-based pose variables, whereas the red lines denote the calibrated SLAM-based pose variables.

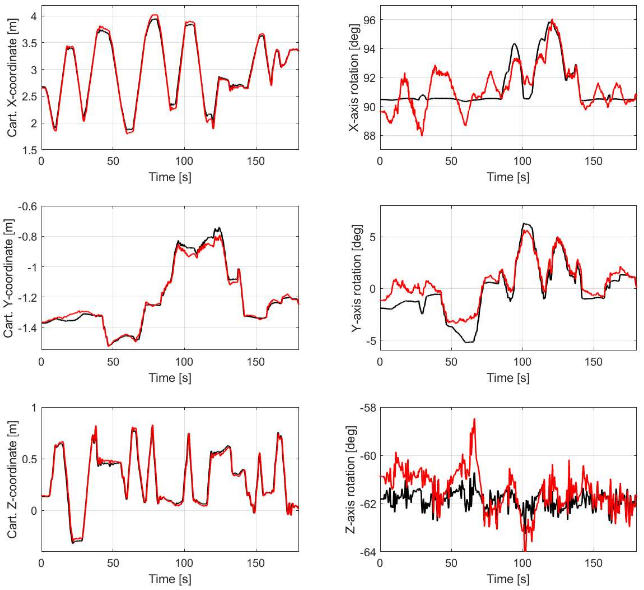

Fig. 5. The second calibration data sequence pose variables: The black lines denote the encoder-based pose variables, whereas the red lines denote the calibrated SLAM-based pose variables.

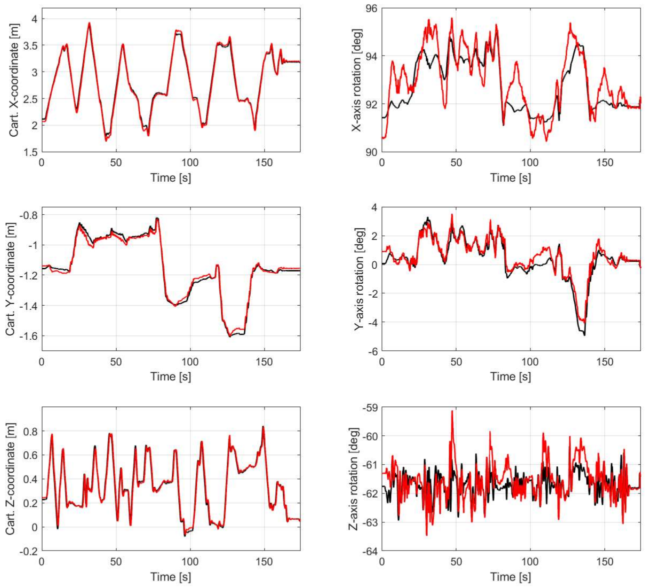

Fig. 6. The third calibration data sequence pose variables: The black lines denote the encoder-based pose variables, whereas the red lines denote the calibrated SLAM-based pose variables.

V. Discussion and Conclusion

In this paper, a methodology for camera-to-kinematic model calibration was proposed, with camera-aided operations for long-reach manipulators in unknown environments as motivation. The goal of this method is to be able to perform fast extrinsic camera calibration easily on the worksite, with arbitrary manipulator motions and in unknown environments. The methodology comprised coarse frame alignment based on the known initial pose of the manipulator and fine matching based on the robust generalized point set registration that benefits not only from position data but also from orientation data, which is perceived as optimal for robotic applications that have complete pose data. Comparison with two other methods utilizing only position data was conducted in offline data analysis, with the results suggesting that utilizing both the orientation and position data is most efficient. As the FMM resolves orientation using unit vectors, a simple solution for correcting the transformed orientation magnitudes using the joint sensors present in the system was shown.

Real time experiments were conducted using a hydraulic

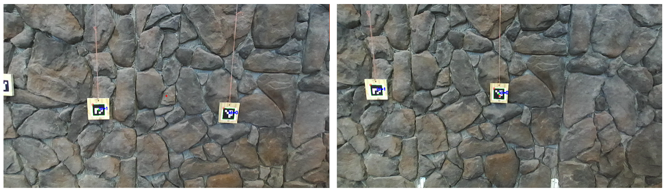

Fig. 7. The left image shows the initial pose of the camera, whereas the right image shows a control result of driving the TCP (image center) to a specific marker.

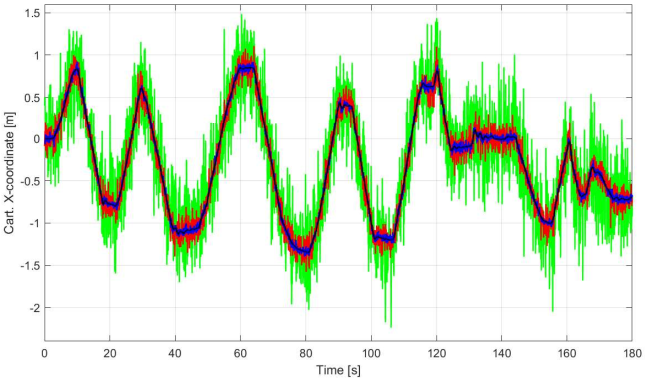

Fig. 8. Gaussian noise added to the X-axis position signal. The black line denotes the raw signal, the green line denotes SNR 10 dB, the red line denotes SNR 20 dB, and the blue line denotes SNR 30 dB.

TABLE II
EUCLIDEAN DISTANCE ERRORS BETWEEN THE IMAGE CENTER AND MARKER CENTERS

	Meas. 1 [m]	Meas. 2	Meas. 3
ArUco#1	0.0085	0.0073	0.0082
ArUco#2	0.0075	0.0066	0.0095
ArUco#3	0.0047	0.0077	0.0031
ArUco#4	0.0052	0.0079	0.0072
ArUco#5	0.0063	0.0097	0.0095
ArUco#6	0.0077	0.0100	0.0083
ArUco#7	0.0103	0.0127	0.0096
ArUco#8	0.0108	0.0100	0.0110
ArUco#9	0.0104	0.0088	0.0104
ArUco#10	0.0105	0.0114	0.0102
ArUco#11	0.0114	0.0104	0.0112
ArUco#12	0.0065	0.0092	0.0083
Avg.	0.0083	0.0093	0.0089

TABLE III
ROOT MEAN SQUARE ERRORS FOR 3 DOF POSITION AND 3 DOF ORIENTATION IN THE FIRST MEASUREMENT

Fig. 4 data	HMM	CPD	Least-Squares
Raw signals [m]	0.0331	0.0331	0.0332
Raw signals [deg]	1.1864	1.1864	1.1849
SNR 10 dB [m]	0.0337	0.0339	0.0340
SNR 10 dB [deg]	1.1894	1.1998	1.1867
SNR 20 dB [m]	0.0272	0.0273	0.0274
SNR 20 dB [deg]	1.1876	1.1882	1.1844
SNR 30 dB [m]	0.0259	0.0259	0.0260
SNR 30 dB [deg]	1.1858	1.1863	1.1848

TABLE IV
ROOT MEAN SQUARE ERRORS FOR 3 DOF POSITION AND 3 DOF ORIENTATION IN THE SECOND MEASUREMENT

Fig. 5 data	HMM	CPD	Least-Squares
Raw signals [m]	0.0375	0.0375	0.0393
Raw signals [deg]	1.5895	1.5895	1.5926
SNR 10 dB [m]	0.0335	0.0351	0.0335
SNR 10 dB [deg]	1.5896	1.5896	1.5931
SNR 20 dB [m]	0.0286	0.0287	0.0289
SNR 20 dB [deg]	1.5925	1.5925	1.5932
SNR 30 dB [m]	0.0292	0.0292	0.0302
SNR 30 dB [deg]	1.5897	1.5897	1.5924

TABLE V
ROOT MEAN SQUARE ERRORS FOR 3 DOF POSITION AND 3 DOF ORIENTATION IN THE THIRD MEASUREMENT

Fig. 6 data	HMM	CPD	Least-Squares
Raw signals [m]	0.0287	0.0287	0.0296
Raw signals [deg]	0.9665	0.9665	0.9679
SNR 10 dB [m]	0.0239	0.0240	0.0248
SNR 10 dB [deg]	0.9663	0.9665	0.9678
SNR 20 dB [m]	0.0258	0.0259	0.0267
SNR 20 dB [deg]	0.9677	0.9680	0.9680
SNR 30 dB [m]	0.0239	0.0240	0.0248
SNR 30 dB [deg]	0.9663	0.9665	0.9678

manipulator with three moving joints. The results showed that the proposed methodology was able to match the pose variables sufficiently in each measured case. Inaccuracies in the matching result were caused by, for example, the rigidity assumption in the kinematic formulation. Furthermore, in the use case, the results were promising for visually assisted operations in applications involving long-reach manipulators with uncertainties, as an acceptable average positioning error was achieved. Some challenges include reliance on the performance of the SLAM algorithm in the sense that the variables may drift during the calibration sequence, for example. Another challenge is that if the camera and the kinematic TCP are on different rotation axes, the two point sets cannot be matched with good accuracy due to the camera's offset.

ACKNOWLEDGMENT

This work was supported by the Doctoral School of Industry Innovations (DSII) of Tampere University.

REFERENCES

[1] R. Y. Tsai, R. K. Lenz *et al.*, "A new technique for fully autonomous and efficient 3 d robotics hand/eye calibration," *IEEE Transactions on robotics and automation*, vol. 5, no. 3, pp. 345–358, 1989.

[2] C.-C. Wang, "Extrinsic calibration of a vision sensor mounted on a robot," *IEEE Transactions on Robotics and Automation*, vol. 8, no. 2, pp. 161–175, 1992.

[3] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," in *Sensor Fusion IV: Control Paradigms and Data Structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–607.

[4] A. Peters, A. Schmidt, and A. C. Knoll, "Extrinsic calibration of an eye-in-hand 2d lidar sensor in unstructured environments using icp," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 929–936, 2020.

[5] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[6] Z. Min, J. Wang, and M. Q.-H. Meng, "Robust generalized point cloud registration using hybrid mixture model," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4812–4818.

[7] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, and G. Ridgeway, "Clustering on the unit hypersphere using von Mises-Fisher distributions." *Journal of Machine Learning Research*, vol. 6, no. 9, 2005.

[8] P. Mäkinen, M. M. Aref, J. Mattila, and S. Launis, "Application of simultaneous localization and mapping for large-scale manipulators in unknown environments," in *Cybernetics and Intelligent Systems (CIS) and IEEE Conf. Robotics, Automation and Mechatronics (RAM), 2019 IEEE 9th Inter. Conf.* IEEE, 2019.

[9] P. Mäkinen, P. Mustalahti, S. Launis, and J. Mattila, "Redundancy-based visual tool center point pose estimation for long-reach manipulators," in *2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2020, pp. 1387–1393.

[10] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 13, no. 04, pp. 376–380, 1991.

[11] S. Garrido-Jurado, R. Muñoz-Salinas, F. J. Madrid-Cuevas, and M. J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280–2292, 2014.

[12] Z. Min, J. Wang, and M. Q.-H. Meng, "Robust generalized point cloud registration with orientational data based on expectation maximization," *IEEE Transactions on Automation Science and Engineering*, vol. 17, no. 1, pp. 207–221, 2019.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[14] R. Mur-Artal and J. D. Tardós, "ORB-SLAM2: an open-source SLAM system for monocular, stereo and RGB-D cameras," *IEEE Trans. Robot.*, vol. 33, no. 5, pp. 1255–1262, 2017.

[15] R. N. Jazar, *Theory of Applied Robotics - Kinematics, Dynamics, and Control*. Dordrecht, the Netherlands: Springer, 2010.