Calibration of a Dynamic Camera Cluster for Multi-Camera Visual SLAM

Arun Das* and Steven L. Waslander†

***Abstract*— Multi-camera clusters used for visual SLAM assume a fixed calibration between the cameras, which places many limitations on its performance, and directly excludes all configurations where a camera in the cluster is mounted to a moving component. In this work, we present a calibration method for *dynamic* multi-camera clusters, where one or more of the cluster cameras is mounted to an actuated mechanism, such as a gimbal or robotic manipulator. Our calibration approach parametrizes the actuated mechanism using the Denavit-Hartenberg convention, then determines the calibration parameters which allow for the estimation of the time varying extrinsic transformations between camera frames. We validate our calibration approach using a dynamic camera cluster consisting of a static camera and a camera mounted to a pan-tilt unit, and demonstrate that the dynamic camera cluster can be provide accurate tracking when used to perform SLAM.**

I. Introduction

Vision based navigation systems are instrumental in providing precise localization information in situations where GPS quality is poor or intermittent, or greater accuracy is required than what GPS alone can provide. A recent trend is the use of omnidirectional vision and multi camera clusters (MCCs), which help to increase localization accuracy and solution integrity by observing features over a large section of the environment [1], [2]. Although effective for precise visual navigation, current state-of-the-art MCCs assume a fixed calibration between the cameras, which excludes the incorporation of cameras mounted to actuated mechanisms.

Many different configurations of multi-camera systems have been successfully used for localization and mapping. Stereo cameras are a widely used approach, where two forward facing cameras are calibrated and rectified so that epipoloar search can be easily performed between the two images [3]. The large overlapping field-of-view (FOV) allows for the calculation of corresponding feature point depths in every pair of collected images, which has led to the development of many stereo based approaches for visual odometry and SLAM in applications such as planetary exploration [4], autonomous driving [5], and aerial robots [6].

As an alternative, more general MCCs provide advantages to stereo based configurations, as the addition of an arbitrary number of cameras with multiple viewpoints allows for more robust tracking and mapping operations in three main ways. First, the ability of the MCC to take measurements over a wide FOV helps with camera localization robustness by better constraining the motion solution, and preventing feature starvation by consistently tracking features over longer durations and over large viewpoint changes. Second, a wider FOV allows for robust map generation and point triangulation by collecting more feature measurements across the whole environment. Finally, so long as the extrinsic calibration is known, multi-camera systems do not require overlap in the FOV to resolve the scale of the solution [7].

Although capable of performing accurate localization in a variety of environments, a major disadvantage of all multi camera systems to date is that they require a fixed calibration between cameras to provide the solution at the correct scale. The fixed extrinsic calibration of the cluster places many limitations on MCC performance. First, any camera cluster must be re-calibrated if a new configuration is required, which is especially tedious and time consuming when the vehicle is deployed in the field. Second, since the MCC is fixed to the vehicle frame, the observation viewpoints of the cameras are highly dependent on the vehicle motion. The coupling of the vehicle motion and camera observation viewpoints is especially problematic if the vehicle undergoes motions which make the vision solution degenerate, or if the vehicle motion results in the camera cluster observing areas of low texture where only poor feature measurements are possible. Finally, many systems, such as UAVs, cannot use the existing gimballed camera payload to assist with the visual navigation. Since current state of the art multi-camera solutions require fixed calibrations between cameras, the gimballed camera is generally only employed for data collection purposes.

In this work, we propose a calibration procedure which will enable the estimation of time varying extrinsic calibrations for multi-camera clusters, which can then easily be used in existing vision based tracking and SLAM systems. We formulate the calibration process for a *dynamic* multi-camera cluster, in which some or all of the cameras in the cluster are mounted to mechanisms which allow them to move independently of each other. Our calibration procedure then determines the parameters of the system, such that the transformation between cameras can be computed using only the control inputs to the mechanisms. Using the Denavit-Hartenberg (DH) convention, we formulate the kinematic transform chain which describes the transformation between a static camera and an actuated camera, then use visual data from a fiducial target to estimate the DH parameters of the mechanism. We experimentally demonstrate our approach on a camera mounted to a pan-tilt unit, and show that a high quality calibration is achievable in the case where there is sufficient motion of both the actuated camera and the fiducial target, and also in the case where the target is kept stationary and only the camera mechanism undergoes motion. Finally, the calibrated pan/tilt dynamic camera cluster is used in

* Ph.D. Student, Mechanical and Mechatronics Engineering, University of Waterloo; adas@uwaterloo.ca.

† Assistant Professor, Mechanical and Mechatronics Engineering, University of Waterloo; stevenw@uwaterloo.ca

a multi-camera visual SLAM system, and is able to track the camera cluster motion in an indoor space approximately 15m×10m×15m in size, with an average accuracy of 2.3cm, which is comparable to the performance of a fixed cluster with the same number of cameras. This is, to the best of our knowledge, the first dynamic camera cluster system used for visual SLAM.

II. Related Works

Much work has been done on multi-sensor calibration problems for robotics applications. Existing approaches have been able to perform high quality extrinsic calibrations between camera and IMU sensors [8], as well as perform observability analysis to determine when the calibration fails [9]. Precise extrinsic calibration between cameras and 3D LIDAR sensors have also been achieved using both gradient based methods [10], and information theoretic approaches [11]. The camera-to-camera calibration problem is also well studied, as it essential for MCC based slam systems.

Current camera-to-camera calibration approaches typically use fiducal markers to generate common observations between cameras [7], [12], [13], though unsupervised methods which use natural features in the environment from pre-existing maps or online SLAM solutions have also provided good results [14]. Although there has been significant work done in the area of camera to camera calibration, we have not found any existing results for camera to camera calibration through an actuated mechanism.

The hand-eye calibration problem, from the field of robotic manipulators, consists of computing the relative position and orientation between the motion frame of a mechanism, and a sensor which is rigidly mounted to the mechanism. The main focus for the hand-eye problem is simultaneously estimating the relative translation between a camera mounted to a robotic manipulator and the manipulator's end effector frame, as well as the transformation between the manipulator's base frame and the camera's motion base frame [15]. Originally developed for camera to manipulator calibration, the hand-eye problem also describes other calibration tasks, such as camera to odometry calibration [16], and the calibration between a camera cluster and a motion tracking system [7]. Although the hand-eye problem is similar to the dynamic MCC calibration problem, the hand-eye calibration assumes that the parameters of the mechanism's forward kinemetics (such as the DH parameters), are known, whereas our dynamic MCC calibration requires estimation of these parameters.

The class of calibration methods related to our problem is known as *kinematic calibration*, and seeks to refine the forward kinematic parameters of robotic manipulators in order to improve overall end effector positioning performance. Generally, the kinematic parameters are optimized by comparing the motion of the end effector to the predicted motion of the mechanism given the forward kinematic parameters and the joint angles. External measurement of the end effector can be performed using a variety of methods, such as using co-ordinate measurement machines (CMM)[17] and externally mounted theodolites [18], however the cost of such measurement equipment is typically quite high, which has motivated the use of low-cost camera based solutions for kinematic calibration.

Camera based kinematic calibration for manipulators consists of taking relative measurements between a camera mounted on the manipulator and a fiducial target in the environment, or mounting the target on the manipulator and placing a static camera in the environment [19]. The work most similar to ours uses a laser pointer as a target and estimates the DH-parameters of two pan-tilt units with attached cameras, mounted to a manipulator end effector [20]. Although such approaches use the camera and fiducal marker to perform the calibration, the estimated parameters only describe the forward kinematics of the manipulator with respect to the *robot base*, whereas calibration of the dynamic MCC, for use in a SLAM problem, requires knowledge of the *camera to camera* calibration, which only include the mechanism's kinematic parameters as part of the total transformation between camera co-ordinate frames.

III. Background

General Rigid Body Transformation: Let a point in 3D be denoted as $\mathbf{p} \in \mathbb{R}^3$. In order to transform points between co-ordinate frames, we will define the rigid body transformation between frames a and b as $\mathbf{T}_\tau^{a:b} \in \mathbb{SE}(3)$, where $\mathbf{T}_\tau^{a:b} : \mathbb{R}^3 \mapsto \mathbb{R}^3$ and $\tau \in \mathbb{R}^6$ is a parameter vector which is used to construct $\mathbf{T}_\tau^{a:b}$. The parameter vector, $\tau = [r_x\, r_y\, r_z\, t_x\, t_y\, t_z]^T$ is composed of three rotation parameters, r_x, r_y, r_z, which denote the 3-2-1 Euler angle rotations, and three translation parameters, t_x, t_y, t_z, which denote the translations along the x, y, and z axes of frame a, respectively. We shall also define the function $\nu(\mathbf{T}) : \mathbb{SE}(3) \mapsto \mathbb{R}^6$, which computes the transformation parameters from the transformation matrix.

Denavit-Hartenburg Parameterization:

The Denavit-Hartenburg (DH) convention is a widely used method to assign co-ordinate frames to the links of a robotic manipulator. Here, we will provide a brief overview of the DH approach for a serial manipulator with rotational joints. For more detailed information, we refer the reader to some of the popular references for manipulator modelling and control [21], [22].

Suppose co-ordinate frame, $\mathcal{F}_i$, is attached to the i^{th} link of a robotic manipulator. The DH convention uses four independent parameters to define the transformation between adjacent links, as depicted in Figure 1.

Consider the two adjacent co-ordinate frames $\mathcal{F}_{i-1}$ and $\mathcal{F}_i$ from Figure 1. In order to construct co-ordinate frame i using the DH convention, the z axis of the frame is placed co-incident with the joint angle. Then, a common normal direction between z_{i-1} and z_i can be determined as $n_i = {}^{z_{i-1}\times z_i}/_{\|z_{i-1}\times z_i\|}$.

Using the common normal, the x_i axis is placed along n_i and points from z_{i-1} to z_i, and the intersection of the x_i and z_i axes define the origin, $\mathcal{O}_i$, of frame $\mathcal{F}_i$. With the x_i and z_i axes defined, the y_i axis is constructed on the frame according to the right-hand rule. Typically,

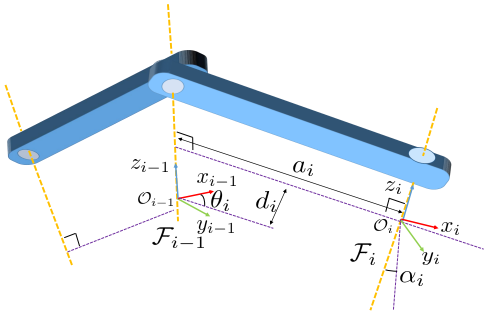


Fig. 1. Example of DH convention between two rotational joints

frames are assigned in a sequential fashion, starting from the end effector frame and ending at the base frame of the mechanism.

With the i^{th} frame constructed, the transformation between $i-1^{\text{th}}$ and i^{th} frames can be defined using the DH parameters. First, frame $\mathcal{F}_{i-1}$ is rotated about axis z_{i-1} by the joint rotation parameter θ_i. Second, frame $\mathcal{F}_{i-1}$ is translated along the z_{i-1} axis by the link offset parameter d_i. Third, $\mathcal{F}_{i-1}$ is translated along the direction of the x_i axis by the link length parameter, a_i. Finally, the $\mathcal{F}_{i-1}$ frame is rotated about the x_i axis by the twist angle parameter, α_i. After applying the transformations with the four parameters, frames $\mathcal{F}_{i-1}$ and $\mathcal{F}_i$ are co-incident.

In this work, we shall denote the DH parameters which describe the transformation between frames $i-1$ and i on an actuated mechanism as $\omega_i = [\theta_i, d_i, a_i, \alpha_i]^T \in \mathbb{R}^4$. Using the DH parameters, a homogeneous rigid body transformation, $\mathbf{T}_{\omega_i}^{i-1:i} \in \mathbb{SE}(3)$, can be computed as

$$\mathbf{T}_{\omega_i}^{i-1:i} = \begin{bmatrix} c(\theta_i) & -s(\theta_i)c(\alpha_i) & s(\theta_i)s(\alpha_i) & a_i c(\theta_i) \\ s(\theta_i) & c(\theta_i)c(\alpha_1) & -c(\theta_i)s(\alpha_i) & a_i s(\theta_i) \\ 0 & s(\alpha_i) & c(\alpha_i) & d_i \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

where $c(\cdot)$ and $s(\cdot)$ denote the $cos(\cdot)$ and $sin(\cdot)$ of an angle, respectively.

Image Projections of 3D Points: The projection function, which maps a point, $\mathbf{p}_i$, to a pixel location on the 2D image plane is defined as

$$\begin{aligned} \Psi(\mathbf{p}_i) &: \mathbb{R}^3 \mapsto \mathbb{P}^2 \\ \Psi(\mathbf{p}_i) &= [u_i \, v_i]^T, \end{aligned} \tag{1}$$

where u_i and v_i are the pixel co-ordinates of the projected point along the u and v image directions, respectively. In this work, we do not assume a specific type of camera model, however, it is important to consider the decrease in sensitivity of the measurement Jacobian for wide angle camera models such as the Taylor model [2]. In such cases, image measurements of the points seen near the boundary of the lens' field of view are less sensitive to perturbations of the point position in 3D, thus degrading the information quality required for precise localization of the camera [23]. In this work, the sensitivity issue can be mitigated by ensuring sufficient point measurements are collected over the entire image plane.

IV. Problem Formulation

This section will describe the calibration process for a dynamic MCC where some or all of the cameras are non-static. First, we will formulate the calibration process between a single static camera and a camera mounted to an actuated mechanism, which will be referred to as the *actuated camera*. Our calibration process requires a region of overlapping FOV between the static and actuated camera, but only over a subset of all possible configurations of the actuated camera. Second, we will describe an extension of the static-to-actuated camera calibration case which will allow for calibration of the actuated-to-actuated camera case. Using the static-to-actuated and actuated-to-actuated camera calibration techniques, the extrinsics of any arbitrary dynamic camera cluster can be calibrated in a pair-wise fashion, provided each calibration pair has sufficient field-of-view overlap.

A. Static-to-Actuated Camera Calibration

The aim of the calibration process is to determine the rigid body transformation between the static camera frame, and the actuated camera frame, $\mathbf{T}_{\Theta,\lambda}^{s:m}$, where $\Theta \in \mathbb{R}^n$ is the set of *estimated* parameters which is used to build the rigid body transform, and $\lambda \in \mathbb{R}^d$ is the set of *measured* parameters which are used to build the transformation. Generally speaking, the measured parameters are available from either known inputs to the mechanism, or can be measured using sensor feedback. For the calibration, the transformation between cameras has the form $\mathbf{T}_{\Theta,\lambda}^{s:m} = \mathbf{T}_{\tau_m}^{e:m}\mathbf{T}_{\omega,\lambda}^{b:e}\mathbf{T}_{\tau_s}^{s:b}$, where $\mathbf{T}_{\tau_s}^{s:b}$ defines the transformation between the static camera and mechanism base frame, $\mathbf{T}_{\omega,\lambda}^{b:e}$ defines the transformation from the base frame of the mechanism to the end effector frame, and $\mathbf{T}_{\tau_m}^{e:m}$ defines the transformation from the end effector frame to the actuated camera frame. Note that $\mathbf{T}_{\omega,\lambda}^{b:e}$ is a chain of transforms through the mechanism's links computed using its forward kinematics, and is a function of its DH parameters and control inputs.

In order to perform the calibration between the static camera and one of the actuated cameras, a fiducial marker is used to collect feature measurements in both cameras. Note that any marker, such as an AprilTag [24] or chess board is suitable, so long as the scale of the points can be determined using a target of known dimension. Measurements of the marker are taken from both cameras simultaneously, which requires that the two cameras share an overlapping field of view. Although it is possible to calibrate a multi camera cluster with completely disjoint or non overlapping fields of view, such a calibration requires motion of the camera cluster and tracking of natural feature points from non-fiducial sources [1], [7], which is left as direction of future work for the dynamic MCC case.

Using the measurements and known scale of the fiducial marker, it is possible for the observing camera to compute its 3D pose relative to the marker frame using well studied techniques such as the perspective-n-point algorithm [25] or a simple bundle adjustment approach [3]. Given the pose of the observing camera relative to the marker frame, we can

determine the position of marker points relative to the camera frame.

For each instance where both the actuated and static camera capture measurements to the fiducial marker, we can define the i$^{\text{th}}$ *measurement set* as $Z_i = \{P_i^s, P_i^m, Q_i^s, Q_i^m, \lambda_i\}$, where P_i^s and P_i^m is the set of marker points defined in the frames of the static and actuated cameras, respectively, Q_i^s and Q_i^m is the set of measurements to the marker points, as observed by the static and actuated cameras, respectively, and λ_i is the set of joint inputs for the mechanism at snapshot i. Note that the measurement sets only include corresponding points visible in both cameras, so consequently $|P_i^s| = |P_i^m| = |Q_i^s| = |Q_i^m|$. In order to produce a high quality calibration, multiple measurement sets need to be collected, while ensuring sufficient excitation of the joint inputs by collecting measurements from many different configurations of the actuated camera.

Using the measurement set and the transformation between camera frames, we can now define the reprojection error between the marker point j in the static camera frame and the corresponding measured point in the actuated camera frame, for measurement set i, as

$$e_j^m(\Theta, \lambda_i) = z_j^m - \Psi^m(\mathbf{T}_{\Theta,\lambda_i}^{s:m} \mathbf{p}_j^s) \tag{2}$$

where $z_j^m \in Q_i^m$ is the measurement of point j, from measurement set Q_i^m, observed in the actuated camera, and $\mathbf{p}_j^s \in P_i^s$ is the 3D position of point j, from the point set P_i^s, as observed from the static camera. Since both the actuated and static camera observe the same marker at each snapshot, we can similarly compute the error for points observed in the actuated frame and projected into the static frame,

$$e_j^s(\Theta, \lambda_i) = z_j^s - \Psi^s((\mathbf{T}_{\Theta,\lambda_i}^{s:m})^{-1} \mathbf{p}_j^m) \tag{3}$$

where $z_j^s \in Q_i^s$ is the measurement of point j, from measurement set Q_i^s, observed in the static camera, and $\mathbf{p}_j^m \in P_i^m$ is the 3D position of point j, from the point set P_i^m, as observed from the actuated camera. The total squared reprojection error as a function of the estimation parameters, $\Lambda(\Theta) : \mathbb{R}^n \mapsto \mathbb{R}$ over all of the collected measurement sets, $\Gamma = \{Z_1, Z_2, \ldots, Z_k\}$, is defined as

$$\begin{aligned} \Lambda(\Theta) = \sum_{Z_i \in \Gamma} \sum_{j=1}^{|P_i^s|} & e_j^m(\Theta, \lambda_i)^T e_j^m(\Theta, \lambda_i) \\ & + e_j^s(\Theta, \lambda_i)^T e_j^s(\Theta, \lambda_i) \end{aligned} \tag{4}$$

Finally, to perform the calibration and determine the optimal parameters, Θ^*, Equation (4) is optimized in order to find the parameters which minimize the total reprojection error,

$$\Theta^* = \underset{\Theta \in \mathbb{R}^n}{\operatorname{argmin}} \, \Lambda(\Theta). \tag{5}$$

Note that (5) describes an unconstrained nonlinear optimization which can be solved using a variety of methods such as Gauss-Newton or Levenberg-Marquardt.

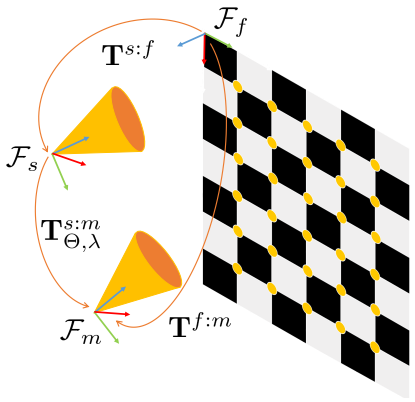


Fig. 2. Example transformation loop for a measurement set. Since both the moving and static camera make observations to the same marker, it is possible to travel from the fiducial frame $\mathcal{F}_f$, through both camera frames, $\mathcal{F}_s$ and $\mathcal{F}_m$, and back to the fiducial frame $\mathcal{F}_f$.

Since both the static and actuated cameras view the same fiducial marker, a constraint on estimated transformation, $\mathbf{T}_{\Theta,\lambda_i}^{s:m}$, is present as a result of a closed loop transformation chain. As illustrated in Figure 2, for each measurement set, we have that the transform, $\mathbf{T}_{\Theta,\lambda_i}^{f:f}$, which travels from the fiducial frame $\mathcal{F}_f$, through both camera frames, $\mathcal{F}_s$ and $\mathcal{F}_m$, and back to the fiducial frame, can be written as $\mathbf{T}_{\Theta,\lambda_i}^{f:f} = (\mathbf{T}^{f:m})^{-1} \mathbf{T}_{\Theta,\lambda_i}^{s:m} \mathbf{T}^{f:s}$. As the transformation is closed loop, we also require that $\mathbf{T}_{\Theta,\lambda_i}^{f:f} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Since the loop constraint is present for all collected measurement sets, we can now reformulate the problem as a constrained optimization over all measurement sets,

$$\begin{aligned} \min \quad & \Lambda(\Theta) \\ \text{subject to} \quad & \mathbf{T}_{\Theta,\lambda_i}^{f:f} = \mathbf{I} \quad \text{for } i = 1, \ldots |\Gamma|. \end{aligned} \tag{6}$$

The decision of using the constrained or unconstrained optimizations is dependent on the accuracy with which the camera-to-fiducial transforms, $\mathbf{T}^{f:s}$ and $\mathbf{T}^{f:m}$, can be computed, as the presence of noise in these transformations make it difficult to exactly enforce the constraint when solving (6).

B. Actuated-to-Actuated Camera Calibration

Compared to the static-to-actuated calibration, the calibration of an actuated-to-actuated camera pair requires the calculation of an additional transform between the base frames of the each mechanism, $\mathbf{T}^{b_1:b_2}$, as depicted in Figure 3. Suppose the camera pair consists of two cameras, Camera 1 and Camera 2. To determine the unknown transform, the camera pair is calibrated by first holding Camera 1 stationary using the static control input, $\bar{\lambda}_1$, and performing the static-to-actuated calibration by moving Camera 2, which results in the estimation of calibration parameters $\Theta_2 = [\tau_2, \gamma_2]^T$, where τ_2 are the parameters describing the transform from $\mathcal{F}_{c_1}$ to $\mathcal{F}_{b_2}$, and γ_2 are the parameters which describe the transformation from $\mathcal{F}_{b_2}$ to $\mathcal{F}_{c_2}$. The static-to-actuated calibration is then performed again, except now holding Camera 2 stationary using the static control input, $\bar{\lambda}_2$, and performing the calibration by moving Camera 1, which results in the estimation of calibration parameters $\Theta_1 = [\tau_1, \gamma_1]^T$, where

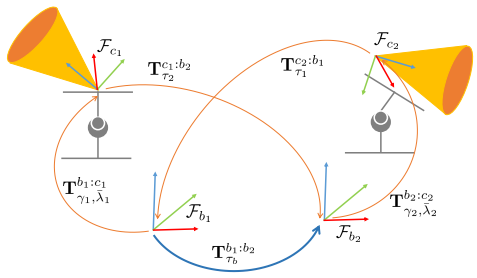


Fig. 3. Frame transformations for actuated-to-actuated camera calibration case. Note the unknown transformation, $\mathbf{T}_{\tau_b}^{b_1:b_2}$, depicted with the blue arrow.

τ_1 are the parameters describing the transform from $\mathcal{F}_{c_2}$ to the $\mathcal{F}_{b_1}$, and γ_1 are the parameters which describe the transformation from $\mathcal{F}_{b_1}$ to $\mathcal{F}_{c_1}$.

As illustrated in Figure 3, we can now define two equivalent transformation loops using the estimated parameters from the static-to-dynamic calibrations and the static control inputs $\bar{\lambda}_1$ and $\bar{\lambda}_2$,

$$\begin{aligned}\mathbf{T}_{\tau_b}^{c_2:c_2} &= \mathbf{T}_{\gamma_2,\bar{\lambda}_2}^{b_2:c_2}\mathbf{T}_{\tau_b}^{b_1:b_2}\mathbf{T}_{\tau_1}^{c_2:b_1}\\ \mathbf{T}_{\tau_b}^{c_1:c_1} &= \mathbf{T}_{\gamma_1,\bar{\lambda}_1}^{b_1:c_1}(\mathbf{T}_{\tau_b}^{b_1:b_2})^{-1}\mathbf{T}_{\tau_2}^{c_1:b_2},\end{aligned} \tag{7}$$

where $\tau_b \in \mathbb{R}^6$ are the parameters describing the transformation between frames $\mathcal{F}_{b_1}$ and $\mathcal{F}_{b_2}$. Using the transformation loops from (7), we can now estimate the parameters of the unknown base to base transformation, τ_b, such that $\mathbf{T}_{\tau_b}^{c_2:c_2} = \mathbf{T}_{\tau_b}^{c_1:c_1} = \mathbf{I}$. Let us define the error vector for the transformation loop for Camera 1 and Camera 2 as $e_{c_1}(\tau_b) = \nu(\mathbf{T}_{\tau_b}^{c_1:c_1})$ and $e_{c_2}(\tau_b) = \nu(\mathbf{T}_{\tau_b}^{c_2:c_2})$, respectively. Then, the cost function, $\Omega(\tau_b) : \mathbb{R}^6 \mapsto \mathbb{R}$, which penalizes the error in the loop transformations from (7), is given as,

$$\Omega(\tau_b) = \begin{bmatrix} e_{c_1}(\tau_b) \\ e_{c_2}(\tau_b) \end{bmatrix}^T \begin{bmatrix} e_{c_1}(\tau_b) \\ e_{c_2}(\tau_b) \end{bmatrix}, \tag{8}$$

which can be optimized using any unconstrained nonlinear method in order to find the optimal parameters for the unknown transformation, $\mathbf{T}_{\tau_b}^{b_1:b_2}$. Once the base mechanism to base mechanism transformation is determined, the calibration is complete, as the forward kinematics between both actuated cameras are fully defined.

C. Estimated Parameter Reduction

Given that the joint angle, θ_i, for each joint in the mechanism can be estimated, or in many cases, directly measured using rotary encoder feedback, the number of estimated DH parameters required for the calibration of each link is 3. Thus, for the general camera-to-camera calibration problem through a mechanism with N joints, we will have N measured parameters and $12+3N$ estimated parameters. It should be noted, however, that some of the degrees of freedom present in the mechanism's kinematic chain are contained within the six degree of freedom transformations from the static camera to the mechanism base, and from the mechanism's end effector to the moving camera frame. Recall that in the DH convention, the z axis of each joint's frame is along the axis of revolution for the joint. Since there is flexibility in placing the base frame of the mechanism through the calibration process, we can assume without loss of generality, that the six degree of freedom transform between the static camera and the base frame can be determined such that the base frame's z axis is aligned with the axis of rotation of the first joint, and the origin of the base frame is placed on the common normal defined by the frame of the first link. Such an approach means the link offset parameter, d_1, between the base frame and first link is zero, and thus no longer needs to be estimated. Also, for a mechanism with N joints, the N^{th} set, or the last set, of DH parameters will define the transformation between the N^{th} joint and the end effector frame. However, the physical placement of the end effector frame is not important for the camera-to-camera calibration process, as we estimate an additional six degree of freedom transformation between the end effector frame and the camera frame. Thus, without loss of generality, the parameters d_N, a_N, and α_N can be set to zero and removed from the set of estimated parameters. Therefore, for a general static-to-actuated camera calibration, the number of estimated parameters required is actually $8+3N$. Note that in the case of a single link there is only one link offset parameter which can be removed ($d_1 = d_N$), thus when $N = 1$, the number of estimation parameters required for the calibration is 12.

V. Calibration of a Pan-Tilt Unit

We shall now apply the general static-to-actuated formulation described in Section IV-A to a dynamic MCC consisting of one static camera and one camera mounted to a pan-tilt unit with two degrees of freedom, with direct measurements of the joint angles available. Figure 4 depicts the co-ordinate frames and associated transforms for the dynamic MCC.

The total transformation between $\mathcal{F}_s$ and $\mathcal{F}_m$, $\mathbf{T}_{\Theta,\lambda}^{s:m}$, is given as $\mathbf{T}_{\Theta,\lambda}^{s:m} = \mathbf{T}_{\tau_2}^{t:m}\mathbf{T}_{\omega_2}^{p:t}\mathbf{T}_{\omega_1}^{b:p}\mathbf{T}_{\tau_1}^{s:b}$, where $\mathbf{T}_{\tau_1}^{s:b}$ is the transformation from the static camera frame to the mechanism base frame, $\mathbf{T}_{\omega_1}^{b:p}$ is the transformation from the base frame to the pan frame, $\mathbf{T}_{\omega_2}^{p:t}$ is the transformation between the pan frame and tilt frame, and $\mathbf{T}_{\tau_2}^{t:m}$ is the transformation between the tilt frame and the actuated camera frame. Since measurements are available to the pan and tilt angle DH parameters of the mechanism, we shall define the measured parameters as $\lambda = [\theta_1, \theta_2]^T$. Using the parameter reduction method described in Section IV-C, the estimated parameters can be defined as $\Theta = [\tau_1, \hat{\omega}_1, \tau_2]^T \in \mathbb{R}^{14}$, where $\tau_1 \in \mathbb{R}^6$ are the six parameters for the transformation between Camera 1 and the pan-tilt base frame, $\hat{\omega}_1 = [a_1, \alpha_1]^T$ are the DH parameters for the pan-tilt unit, and $\tau_2 \in \mathbb{R}^6$ are the six parameters for the transformation between the pan-tilt end-effector frame and the frame of Camera 2.

VI. Experimental Validation

The proposed calibration approach is experimentally validated using a dynamic camera cluster consisting of two Ximea xIQ cameras which operate at 60fps and 900×600 resolution. To build the dynamic MCC, one camera is

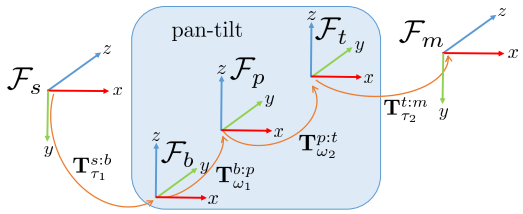


Fig. 4. Frame transformations for a general pan-tilt unit with two degrees of freedom.

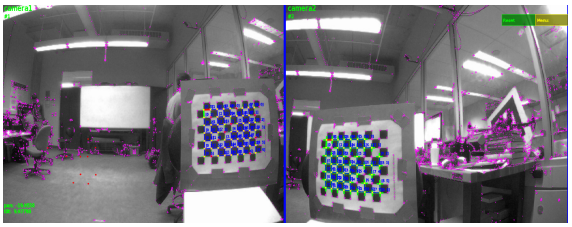


Fig. 5. Example measurement set acquisition with the pan and tilt angles set to -20 and 10 degrees, respectively. For data collection, the pan and tilt axes are actuated to the minimum and maximum angles which allow for overlap in the field of views of the cameras (approximately $\pm$ 30 degrees for pan, and $\pm$ 20 degrees for tilt.

statically mounted and the other is mounted to a FLIR-D46-17 pan-tilt unit with two degrees of freedom, with both cameras forward facing at zero pan and tilt angle. The Taylor camera model [2] is used to accommodate the 120 degree FOV wide angle lenses fitted to the cameras. The pan-tilt unit uses stepper motors to determine the joint angles, which are reported at a 30Hz update rate. Note that for this particular pan-tilt unit, the pan and tilt joint axes intersect at right angles to each other, which results in a concentric configuration where the origins of both joint frames are coincident and offset to each other by a 90 degree angle. This results in a further reduction in the number of estimated parameters, as the concentric configuration implies $a_1 = 0$ and $\alpha_1 = \pi/2$, thus leaving 12 parameters to estimate.

Using the pan-tilt based dynamic MCC, we perform two sets of experiments. First, we collect 3D point data using a chessboard marker and show that it is possible to achieve a high quality calibration. Second, we integrate the calibrated cluster into the multi-camera parallel tracking and mapping (MCPTAM) [1] pipeline [1], [7], and demonstrate the effectiveness of the dynamic MCC in a SLAM scenario.

A. Calibration of the Pan-Tilt Unit

Similar to existing system ID and calibration methods, sufficiently rich input data is required to ensure the estimated parameters can be accurately determined. Existing marker based MCC calibration relies on relative motion between the marker and camera rig to collect 3D point and image measurement information from multiple viewpoints, whereas a dynamic MCC is able to observe the marker from different viewpoints by actuating the camera. To that end, we perform and compare two calibrations. In the first, excitation of the image input is provided by moving both the marker and the actuated camera through various configurations (dual-excitation), and in the second, excitation of the image input is accomplished by only moving the camera (single-excitation) and keeping the target stationary. In both cases, approximately 130 measurement sets are collected with a chessboard target consisting of 35 points. An example of the measurement set collection is presented in Figure 5. Furthermore, we test both the dual-excitation and single-excitation cases using the unconstrained and constrained methods for calibration described in section IV-A.

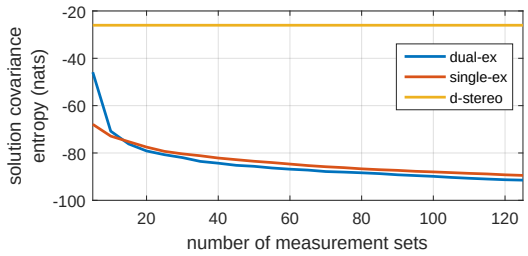
Verification of the calibrations are performed with respect to an independently collected *validation set*, which consists of an additional 130 measurement sets, with excitation of the marker and pan and tilt angles. Summary statistics of the pixel reprojection errors are provided in Table I.

In general, all methods, with the exception of the constrained dual-excitation approach, produced good calibrations, as the mean pixel reprojection error was less than 1. However, the unconstrained dual-excitation method provided a calibration with the lowest error mean and covariance, likely due to the richness of the input images collected by moving both the marker and the actuated camera. Note that the constrained dual-excitation approach produced a less accurate calibration compared to the unconstrained dual-excitation method, which can be attributed to noise or error in the fiducial to camera transforms which could be present from observing the marker with large skew angles, or near edges of the image where pixel sensitivity is reduced. In the constrained single-excitation case, the static camera observed the marker roughly straight-on and near the centre of the image, which would reduce noise in the marker to camera transform and help with achieving an accurate calibration. Interestingly, the calibration provided by the single-excitation methods are of comparable accuracy to the unconstrained dual-excitation approach, demonstrating that a high quality calibration is possible by leaving both the marker and camera cluster static, and providing image excitation by only moving the actuated camera.

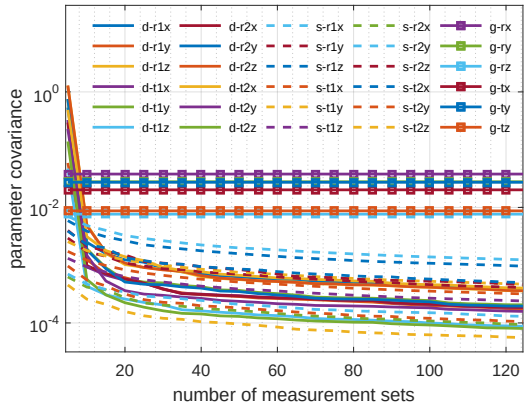
The effectiveness of the single-excitation approach is further verified in Figure 6, which captures the uncertainty in the calibration parameters as a function of the measurement sets. We compare the unconstrained dual-excitation and constrained single-excitation cases, as they provided the best calibration for the dual-excitation and single-excitation cases, respectively, and a *degraded stereo* (d-stereo) configuration. For the degraded stereo, the dynamic MCC was held static in a forward facing configuration, and a six degree of freedom transformation between the camera frames was estimated using a single image of the fiducial target. For the single-excitation case, the degraded stereo provides a baseline to illustrate the improvement in calibration quality that is possible when one of the cameras is actuated.

To study how the calibration quality is affected by the number of measurement sets, the calibration method is performed repeatedly, first using 5 randomly selected measurement sets, then adding an additional 5 unique and randomly selected measurement sets to every subsequent optimization

[1]The MCPTAM code is available on GitHub: https://github.com/ wavelab/mcptam



(a)



(b)

Fig. 6. Parameter uncertainty plotted as a function of the number of measurement sets used to determine the calibration. (a) displays the entropy of the covariance matrix, while (b) shows the variances of individual parameters. Entries with the prefixes 'd-', 's-', and'g-'denote the dual-excitation, single-excitation, and degraded-stereo cases, respectively. Note that the large uncertainty for the dual-excitation case at 5 measurement sets is likely a result of those particular randomly selected measurement sets causing a degenerate solution due to lack of excitation of the pan and tilt angles.

epoch. For each epoch, we estimate the solution covariance with $\Sigma = (J_e^T J_e + \sigma \operatorname{diag}(J_e^T J_e))^{-1}$, where J_e is the Jacobian matrix of the stacked error equations defined in (2), σ is a regularization factor, and $\operatorname{diag}(J_e^T J_e)$ defines a regulation term which allows for the inversion of $J_e^T J_e$ in degenerate situations. Figure 6(a) displays the entropy of the covariance matrix, while figure 6(b) displays the individual parameter covariances (the diagonal elements of Σ), with respect to the number of measurement sets used in the calibration. As seen in figure 6(a), both the single- and dual-excitation cases have comparable improvements in the solution confidence as the number of measurements sets used increases, with both cases settling at roughly -90 nats after 130 measurement sets, and the single-excitation case performing significantly better than the degraded stereo case. Since the entropy of the covariance matrix provides a single measure on the overall estimation uncertainty, it is also important to inspect the individual parameter uncertainty, which is provided in Figure 6(b). Although some of the individual parameters for the single-excitation case are not as well estimated when compared to the dual-excitation case, in practice the calibration may still perform well when implemented in the SLAM system.

B. Use of a Dynamic MCC in MCPTAM

At the time of each image acquisition, the extrinsic camera transformation is updated using the calibrated cluster

TABLE I

SUMMARY STATISTICS FOR PIXEL REPROJECTION OF VALIDATION SET

	mean (cam 1)	**mean (cam 2)**	**covariance (cam 1)**	**covariance (cam 2)**
uncon. dual	0.9713	0.9318	0.2076	0.2584
ucon. single	0.9818	1.0601	0.3028	0.2089
con. dual	1.5866	1.6418	0.3103	0.5069
con. single	0.9938	0.9622	0.2446	0.2605

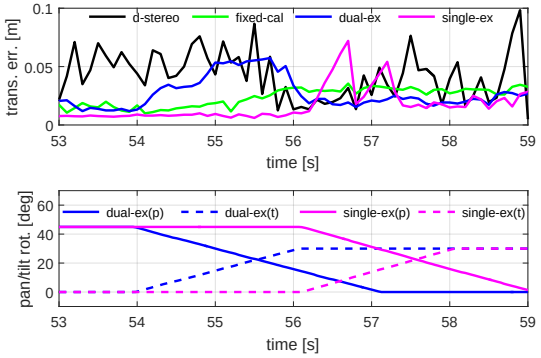


Fig. 7. Translation error for the tested calibration types, as well as pan (p) and tilt (t) angle trajectories for the dual-excitation and single-excitation cases, for a segment of the motion trajectory. The error spikes which coincide with non-zero velocity of the tilt axis are a result poor tracking caused by image vibration.

parameters and the reported pan and tilt angles. Since the camera image and joint angle acquisition rates are 60Hz and 30Hz, respectively, linear interpolation of the angles is performed, based on the acquisition timestamps, to determine the pan and tilt angles for each collected image. Although we demonstrate SLAM results using the MCPTAM method, the dynamic MCC can be integrated into any multi-camera SLAM system using a similar approach.

For the SLAM experiment, we move the pan-tilt MCC through an indoor environment approximately 15m×10m×15m in size, with the cluster starting on the ground, traversing multiple loops through the room, and returning to its starting location. We perform four separate motion trials in order to compare the dual-excitation case, the single-excitation case, the d-stereo case, and a baseline fixed calibration (fixed-cal) performed using the extrinsic calibrator included with MCPTAM, with the pan-tilt cluster held static at zero pan and tilt angles. For the dual-excitation, and single-excitation cases, the pan and tilt joints are actuated in a periodic trajectory ($\pm$ 45 degrees from the zero angle for pan, and to +40 degrees for tilt). Note that we did not produce joint trajectories with negative tilt angles, as the resulting camera viewpoint of the textureless floor would pose a difficulty in tracking not present in the fixed-cal or d-stereo cases. Ground truth of the motion is captured by an OptiTrack indoor positioning system (IPS), which is capable of providing tracking at 100Hz with millimeter translation accuracy and sub-degree rotation accuracy. To compare the motion tracks, the SLAM and IPS motion solutions are aligned using an off-line calibration method [7].

Table II presents the median translation and rotation errors

TABLE II
AVERAGE ERROR STATISTICS FOR SLAM TRAJECTORY

	trans. err. [cm]	rot. err. [deg]
fixed calibration	2.341	0.6850
dual-excitation	2.503	0.7831
multi-excitation	2.419	0.7606
degraded-stereo	4.303	1.1748

for the four tested cases. It is evident that the dual-excitation and single-excitation approaches achieve comparable performance to the fixed calibration case, while the degraded stereo case performs the worst. The slight increase in the error of the dynamic MCC cases compared to the fixed-cal case is due to high frequency vibrations from the tilt axis stepper motor causing image distortion of the actuated camera. Vibration issues can be mitigated through the use of brushless or direct drive servo motors in the actuated mechanism, as is commonly used on most UAV gimbals where vibration mitigation is important. The increase in error caused by joint axis vibration is further demonstrated in Figure 7, which plots the translation error and the pan and tilt angles over a segment of the SLAM trajectory where the vibration issue is prominent. From Figure 7, it is clear that the spikes in the error for the dynamic cases coincide with non-zero velocity of the pan and tilt joints. Once the joint velocity is zero and the cluster settles in its new configuration, the dynamic cluster is able to track with comparable accuracy to the fixed calibration case. Also note the high frequency error signal exhibited by the d-stereo case, which is a result of poor tracking performance caused by the imprecise calibration.

VII. CONCLUSION

This work presents the calibration of a dynamic MCC, which allows for the use of time varying camera-to-camera extrinsic transformations in multi-camera visual SLAM. The unknown parameters of the actuated mechanism are parameterized using the DH convention, and calibrated using a fiducial marker of known scale. We experimentally demonstrate the validity of the calibration on a pan-tilt based dynamic cluster using the MCPTAM method. Our future work includes performing degeneracy analysis for the general dynamic camera cluster, testing our approach on a wider class of actuated gimbals and mechanisms, integrating dynamic MCCs into other existing SLAM methods, and performing active gaze selection on the actuated camera to help improve localization accuracy.

REFERENCES

[1] A. Harmat, M. Trentini, and I. Sharf, "Multi-camera tracking and mapping for unmanned aerial vehicles in unstructured environments," *Journal of Intelligent & Robotic Systems*, vol. 78, no. 2, pp. 291–317, 2014.

[2] D. Scaramuzza, A. Martinelli, and R. Siegwart, "A flexible technique for accurate omnidirectional camera calibration and structure from motion," in *IEEE International Conference on Computer Vision Systems (ICVS)*. New York, NY: IEEE, Jan. 2006, pp. 45–45.

[3] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge University Press, 2003.

[4] T. Howard, A. Morfopoulos, J. Morrison, Y. Kuwata, C. Villalpando, L. Matthies, and M. McHenry, "Enabling continuous planetary rover navigation through FPGA stereo and visual odometry," in *IEEE Aerospace Conference*, Big Sky, MT, 2012, pp. 1–9.

[5] I. Cvisic and I. Petrovic, "Stereo odometry based on careful feature selection and tracking," in *European Conference on Mobile Robots (ECMR)*, Lincoln, UK, Sept 2015, pp. 1–6.

[6] S. Shen, Y. Mulgaonkar, N. Michael, and V. Kumar, "Vision-based state estimation for autonomous rotorcraft mavs in complex environments," in *IEEE International Conference on Robotics and Automation (ICRA)*, Karlsruhe, Germany, 2013, pp. 1758–1764.

[7] A. Tribou, Michael J. and. Harmat, D. Wang, I. Sharf, and S. L. Waslander, "Multi-camera parallel tracking and mapping with non-overlapping fields of view," *International Journal of Robotics Research*, vol. 34, no. 12, pp. 1480–1500, December 2015.

[8] P. Furgale, J. Rehder, and R. Siegwart, "Unified temporal and spatial calibration for multi-sensor systems," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, Nov 2013, pp. 1280–1286.

[9] J. Kelly and G. S. Sukhatme, "Visual-inertial sensor fusion: Localization, mapping and sensor-to-sensor self-calibration," *The International Journal of Robotics Research*, vol. 30, no. 1, pp. 56–79, 2011.

[10] J. Levinson and S. Thrun, "Unsupervised calibration for multi-beam lasers," in *International Symposium on Experimental Robotics*, Delhi, India, 2014, pp. 179–193.

[11] G. Pandey, J. R. McBride, S. Savarese, and R. M. Eustice, "Automatic targetless extrinsic calibration of a 3D lidar and camera by maximizing mutual information." in *AAAI National Conference on Artificial Intelligence*, Toronto, Canada, July 2012, pp. 2053–2059.

[12] P. Lbraly, E. Royer, O. Ait-Aider, C. Deymier, and M. Dhome, "Fast calibration of embedded non-overlapping cameras," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011, pp. 221–227.

[13] B. Li, L. Heng, K. Koser, and M. Pollefeys, "A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, Nov 2013, pp. 1301–1307.

[14] L. Heng, M. Burki, G. H. Lee, P. Furgale, R. Siegwart, and M. Pollefeys, "Infrastructure-based calibration of a multi-camera rig," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, 2014, pp. 4912–4919.

[15] K. Daniilidis, "Hand-eye calibration using dual quaternions," *The International Journal of Robotics Research*, vol. 18, no. 3, pp. 286–298, 1999.

[16] A. Censi, A. Franchi, L. Marchionni, and G. Oriolo, "Simultaneous calibration of odometry and sensor parameters for mobile robots," *IEEE Transactions on Robotics*, vol. 29, no. 2, pp. 475–492, 2013.

[17] M.-S. Kim, H.-S. Yoo, S.-W. Cho, H.-S. Chang, and G. Spur, "A new calibration method," *CIRP Annals-Manufacturing Technology*, vol. 39, no. 1, pp. 421–424, 1990.

[18] D. Whitney, C. Lozinski, and J. M. Rourke, "Industrial robot forward calibration method and results," *Journal of dynamic systems, measurement, and control*, vol. 108, no. 1, pp. 1–8, 1986.

[19] D. C.-C. Lu, "Kinematic calibration of serial manipulators using relative measurements," Master's thesis, Ottawa-Carleton Institute for Mechanical and Aerospace Engineering Department, Jan 2014.

[20] C. S. Gatla, R. Lumia, J. Wood, and G. Starr, "Calibrating pan-tilt cameras in robot hand-eye systems using a single point," in *IEEE International Conference on Robotics and Automation (ICRA)*, Roma, Italy, April 2007, pp. 3186–3191.

[21] M. W. Spong and M. Vidyasagar, *Robot dynamics and control*. John Wiley & Sons, 2008.

[22] R. S. Hartenberg and J. Denavit, *Kinematic Synthesis of Linkages*. McGraw-Hill, 1964.

[23] A. Das and S. L. Waslander, "An entropy based approach to keyframe selection for multi-camera parallel tracking and mapping," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Hamburg, Germany, October 2015.

[24] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 2011, pp. 3400–3407.

[25] V. Lepetit, F. Moreno-Noguer, and P. Fua, "EPNP: An accurate O(n) solution to the PNP problem," *International Journal of Computer Vision*, vol. 81, no. 2, pp. 155–166, 2009.