Controlling a Robotic Stereo Camera Under Image Quantization Noise

The International Journal of Robotics Research
XX(X):1–17

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Charles Freundlich[1], Yan Zhang[1], Alex Zihao Zhu[2], Philippos Mordohai[3], and Michael M. Zavlanos[1]

Abstract

In this paper, we address the problem of controlling a mobile stereo camera under image quantization noise. Assuming that a pair of images of a set of targets is available, the camera moves through a sequence of Next-Best-Views (NBVs), i.e., a sequence of views that minimize the trace of the targets' cumulative state covariance, constructed using a realistic model of the stereo rig that captures image quantization noise and a Kalman Filter (KF) that fuses the observation history with new information. The proposed algorithm decomposes control into two stages: first the NBV is computed in the camera relative coordinates, and then the camera moves to realize this view in the fixed global coordinate frame. This decomposition allows the camera to drive to a new pose that effectively realizes the NBV in camera coordinates while satisfying Field-of-View constraints in global coordinates, a task that is particularly challenging using complex sensing models. We provide simulations and real experiments that illustrate the ability of the proposed mobile camera system to accurately localize sets of targets. We also propose a novel data-driven technique to characterize unmodeled uncertainty, such as calibration errors, at the pixel level and show that this method ensures stability of the KF.

Keywords

Range Sensing, Motion Control, Mapping

1 Introduction

Active robotic sensors are rapidly gaining viability in environmental, defense, and commercial applications. As a result, developing information-driven sensor strategies has been the focus of intense and growing research in artificial intelligence, control theory, and signal processing. We focus on stereoscopic camera rigs, that is, two rigidly connected cameras in a pair. Specifically, we address the problem of determining the trajectory of a mobile robotic sensor equipped with a stereo camera rig so that it localizes a collection of possibly mobile targets as accurately as possible under image quantization noise.

The advantage of binocular vision, compared to the use of monocular camera systems, is that it provides both depth and bearing measurements of a target from a pair of simultaneous images. Assuming that noise is dominated by quantization of pixel coordinates (Blostein and Huang 1987; Matthies and Shafer 1987; Chang et al. 1994) we use the measurement Jacobian to propagate the error from the pixel coordinates to the target coordinates relative to the stereo rig. In particular, we approximate the pixel error as Gaussian and propagate the noise to the target locations, giving rise to fully correlated second order error statistics, or measurement error covariance matrices, which capture target location uncertainty. The resulting second order statistic is an accurate representation of not only the eigenvalues but also the eigenvectors of the measurement error covariance matrices, which play a critical role in active sensing as they determine viewing directions from where localization uncertainty can be further decreased.

Assuming that a pair of images of the targets is available, in this paper, we iteratively move the stereo rig through a sequence of configurations that minimize the trace of the targets' cumulative covariance. This cumulative covariance is constructed using a Kalman Filter (KF) that fuses the observation history with the predicted instantaneous measurement covariance obtained from the proposed stereoscopic sensor model. Differentiating this objective with respect to the new instantaneous measurement in the relative camera frame provides the Next Best View (NBV), i.e., the new relative distance and direction from where a new measurement should be obtained. Then, the stereo rig moves to realize this NBV using a gradient descent approach in the joint space of camera rotations and translations. Once the NBV is realized in the global frame, the camera takes a new pair of images of the targets that are fused with the history using the KF to update the prior cumulative covariance of the targets, and the process repeats with the determination of a new NBV. The sequence of observations and resulting NBVs, generated by the proposed iterative scheme, constitutes a switching signal in the continuous motion control. During motion, appropriate

[1]Duke University, Dept. of Mechanical Engineering and Materials Science
[2]University of Pennsylvania, Dept. of Computer and Information Science
[3]Stevens Institute of Technology, Dept. of Computer Science

Corresponding author:
Charles Freundlich, Dept. of Mechanical Engineering and Materials Science, Duke University, Durham, NC, 27708
Email: charles.freundlich@duke.edu

barrier potentials prevent targets from exiting the camera's geometric Field-of-View (FoV). As we illustrate in computer simulations and real-world experiments, the resulting sensor trajectory balances between reducing range and diversifying viewpoints, a result of the eigenvector information contained in the posterior error covariances. This behavior of our controller is notable when compared to existing sensor guidance approaches that adopt approximations to the error covariance that are not direct functions of the stereo rig calibration and the pixel observations themselves (Le Cadre and Gauvrit 1996; Passerieux and Van Cappel 1998; Logothetis et al. 1997; Stroupe and Balch 2005; Zhou and Roumeliotis 2011; Olfati-Saber 2007; Chung et al. 2006).

1.1 Related Work

Our work is relevant to a growing body of literature that addresses control for one or several mobile sensors for the purpose of target localization or tracking (Fox et al. 2000; Roumeliotis and Bekey 2002; Spletzer and Taylor 2003; Stroupe and Balch 2005; Chung et al. 2006; Yang et al. 2007; Morbidi and Mariottini 2013; Zhou and Roumeliotis 2011). These methods use sensor models that are based only on range and viewing angle. These models, if used for stereo triangulation, can not accurately capture the covariance among errors in measurement coordinates, nor can they capture dependence on range and viewing angle. It is also common to ignore directional field of view constraints by assuming omnidirectional sensing. In this paper, we derive the covariance specifically for triangulation with a calibrated stereo rig. The derived measurement covariance, when fused with a prior distribution, provides our controller with critical directional information that enables the mobile robot to find the NBV, defined as the vantage point from where new information will reduce the posterior variance of the targets' distribution by the maximum amount.

Recent work (Ponda and Frazzoli 2009; Adurthi et al. 2013; Ding et al. 2012) brought about by developments in fixed-wing UAV control, addresses the autonomous visual tracking and localization problem using optimal control over information-based objectives using monocular vision. Ponda and Frazzoli (2009) define an objective function based on the trace of the covariance matrix of the target location and determine the next best view by a numerical gradient descent scheme. Ding et al. (2012) also minimize the trace of the fused covariance by guiding multiple non-holonomic Unmanned Aerial Vehicles (UAVs) that use the Dubins car model. Adurthi et al. (2013) use receding horizon control to maximize mutual information. Their method avoids replanning at every step by doing so only when the Kullback-Leibler divergence between the most recent target location probability density function (pdf) and the pdf that was used in the planning phase differ by a user-specified threshold. The Dynamic Programming (DP) approaches of Ponda and Frazzoli (2009); Adurthi et al. (2013); Ding et al. (2012) have complexity that grows exponentially in the horizon length. In this paper, our proposed analytical and closed-form expression for the gradient guides image collection in all position and orientation directions continuously. Although it does not plan multiple steps into the future, our controller is adaptive due to its feedback nature; each decision predicts new sensor locations from where new measurements optimize the estimated target locations based on the full, fused observation history.

As far back as Bajcsy (1988), computer vision researchers have recognized that sensing decisions should be based on an exact sensor model, and that robotic vision, like human vision, can benefit from mobility. Relevant prior work on active vision controls the image collection process for digital cameras through a discretized pose space by optimizing a scalar function of the covariance of feature-points on an object that is to be reconstructed. Specifically, Trummer et al. (2010) focus on the maximum eigenvalue of the posterior covariance, Wenhardt et al. (2007b) the entropy, and Dunn et al. (2009) the expected quality of the next view. While these works do obtain uncertainty estimates that depend on factors such as viewing distance and camera resolution, which improves accuracy in 3D reconstruction, they do not operate continuously in 3D or consider dynamic environments. Morbidi and Mariottini (2013) optimize an objective function that depends on the covariance matrix of the KF, rather than the measurement error covariance matrix. The authors derive upper and lower bounds on the covariance matrix at steady-state and validate their method in simulation. Stereoscopic vision sensors in continuous pose space are employed by Shade and Newman (2010), similar to the work proposed here. However, this work (Shade and Newman 2010) is concerned with exploration of an indoor environment and not with refining localization estimates.

When the target configurations can be collectively modeled by a coarse mesh in space, the NBV problem becomes similar to active inspection. Several researchers have addressed this problem using approximate dynamic programming, by formulating it as coverage path planning. Galceran and Carreras (2013) provide an overview of coverage problems in mobile robotics, where the goal is to plan sensor paths that "see" every point on the surface mesh. Similarly, Wang et al. (2007) propose solving the traveling view path planning problem using approximate integer programming on a network flow model. Papadopoulos et al. (2013) enforce differential constraints for this problem. Hollinger et al. (2013) invoke adaptive submodularity, which argues that greedy approaches to measurement acquisition may outperform dynamic programming approaches that do not replan as measurements are acquired. While relevant to this work from a sensor planning perspective, active inspection methods do not address target localization. Moreover, dynamic and integer programming methods tend to be computationally expensive, especially for high dimensional spaces as those resulting from the presence of mobile targets. In this paper, we assume that there are no occlusions and, therefore, coverage (or detection) can be obtained if FoV constraints are met. Moreover, we assume no correspondence errors between images. These assumptions allow us to develop a control systems approach to the target localization problem that is based on computationally efficient, analytic, expressions for the camera motion and image collection process, as well as on precise sensor models that can result in more accurate localization.

We note briefly that this paper is based on preliminary results contained in our prior publications (Freundlich et al. 2013a,b). These early works used simplified versions of the

noise model and global controller and lacked experimental validation.

1.2 Contributions

Our proposed control decomposition and resulting hybrid scheme possess a number of advantages compared to other methods that control directly the full non-linear system or resort to dynamic programming for nonmyopic planning. While these methods can have their own benefits, they also suffer from drawbacks. In particular, controlling directly the full non-linear system can be subject to multiple local minima that might be difficult to handle. On the other hand, dynamic programming formulations suffer from computational complexity due the size of the resulting state-spaces and often resort to abstract sensor models to help reduce complexity (Logothetis et al. 1997; Stroupe and Balch 2005; Singh et al. 2007; Logothetis et al. 1998; Frew 2003; Adurthi et al. 2013; Ding et al. 2012). Additionally, these approaches use discrete methods, e.g., the exhaustive search of Frew (2003) and the gradient approximations of Singh et al. (2007), to achieve the desired control task. Instead, decomposing control in the global and relative frames allows us to consider separately high-level planning, defined by the image collection/sensing process, and low-level planning, i.e., motion control of the camera. An advantage of this decomposition is that, given an NBV in the relative frame, there are infinite ways that the camera can realize this NBV in the global frame. This provides choices to the motion controller that otherwise could be subject to local stationary points due to the nonlinear coupling between sensing and planning. We provide a stability proof of the motion controller, while extensive computer simulations and experimental results have shown that even when FoV constraints are considered, local minima are not an issue and can be avoided by simple tuning of a gain parameter. The control decomposition also allows us to introduce Field-of-View (FoV) constraints that have not been previously used in the NBV context due to the complexity of their implementations. Most authors have used omnidirectional sensor models to circumvent these difficulties. The FoV constraints naturally enter the motion controllers when control is decomposed in the global and relative frames. Finally, to the best of our knowledge, the approaches by (Stroupe and Balch 2005; Zhou and Roumeliotis 2011; Olfati-Saber 2007; Adurthi et al. 2013; Ding et al. 2012) rely on having large numbers of sensors, e.g., 20 or 60, and consider a single target, while our method enables one sensor to track multiple targets as long as they satisfy the FoV constraints.

In summary, the contribution of this work is that we address the multi-target, single-sensor problem employing the most realistic sensor model among continuous-space approaches in the literature that rely on the gradient of an optimality metric of the error covariance for planning. Additionally, to the best of our knowledge, this work is the first to include FoV constraints within the NBV setting. We also model image quantization noise directly. This allows us to accurately model the second order error statistics of the target location uncertainty based on the actual pixel error distribution. While other sources of error, such as association (matching) errors or occlusion, contribute to target localization error, a simultaneous and exact treatment of all error sources for the purposes of active sensing is an open problem. In this work we have addressed an essential contributor, that is, quantization. We have also proposed a novel data-driven technique to account for unmodeled uncertainty, such as system calibration errors, that is necessary to transition the proposed theoretical results to practice. In particular, for long range stereo vision, calibration errors are unique to the particular stereo rig used. They can cause severe bias and ill conditioned covariance matrices that may be completely different from one stereo rig to another. As our method heavily relies on the KF, ensuring that measurements are unbiased and that we have a reliable estimate for their covariance is crucial to both convergence of the estimator and for generating sensible closed-loop robot trajectories. Our proposed data-driven technique corrects measurements at the pixel level and empirically calculates their predictive error covariances. Specifically, using a sufficiently large training set of stereo image pairs, we determine the empirical error covariance, which we propagate to world coordinates and use for both path planning and state estimation. To the best of our knowledge, our data-driven approach to estimating the error statistics in the pixel coordinates is novel. Most relevant literature in stereo vision assumes arbitrarily large such statistics, so that the estimation process is stable. In practice, we found this step to be crucial for accurate triangulation and fusion of multiple measurements, even in our controlled lab environment.

We note that this paper is based on preliminary results that can be found in (Freundlich et al. 2013a,b). The main differences between our preliminary work in (Freundlich et al. 2013a,b) and the work proposed here are the following. First, here we present thorough experimental results that validate our approach; the first of their kind for stereo vision. Second, the controller proposed in (Freundlich et al. 2013a,b) realizes a NBV that does not place the targets on the positive z-axis (viewing direction) of the stereo rig. Looking straight at the targets results in more accurate observations. The controller proposed here has this property. As a result, the correctness proofs in this paper are different compared to (Freundlich et al. 2013a,b). Finally, the noise model in this paper is based on an empirical model of quantization noise in stereo vision, as opposed to constant pixel noise covariance in (Freundlich et al. 2013a,b).

The paper is organized as follows. Section 2 formulates the visual target tracking problem. Section 3 discusses the NBV in the camera-relative coordinate system. Section 4 presents the gradient flow in the global coordinate frame. Section 5 illustrates the proposed integrated hybrid system via computer simulations for static and mobile target localization and discusses ways to integrate FoV constraints in the proposed controllers. Section 6 gives experimental validation of our claims and describes the data-driven noise modeling strategy. Section 7 concludes the work.

2 System Model

Consider a group of n mobile targets, indexed by $i \in \mathcal{N} = \{1 \dots n\}$, with initially unknown positions $\mathbf{x}_i$. Consider also a mobile stereo camera located at $\mathbf{r}(t) \in \mathbb{R}^3$ and with

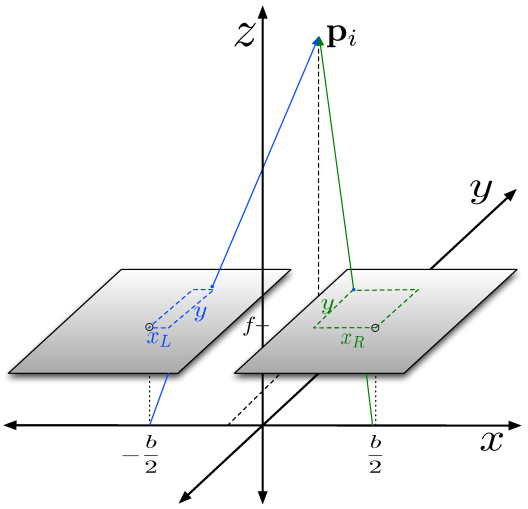

Figure 1. Stereo geometry in 3D. Two rays from the camera centers to a target located at $\mathbf{p}_i$ creates a pair of image coordinates, (x_L, y) and (x_R, y).

orientation $R(t) \in SO(3)$, where $SO(3)$ denotes the special orthogonal group of dimension three, with respect to a fixed global reference frame at time $t \geq 0$. A coordinate frame anchored to the stereo camera, hereafter referred to as the relative coordinates, is oriented such that, without loss of generality, the x-axis joins the centers of two monocular cameras and the positive z-axis measures range. We denote the two cameras by Left (L) and Right (R). The (L) and (R) camera centers are thus located at $(-b/2, 0, 0)$ and $(b/2, 0, 0)$ in the relative coordinates, where b denotes the stereo baseline (see Fig. 1).

The position of target i with respect to the relative camera frame can be expressed as

$$\mathbf{p}_i \triangleq \mathbf{p}(x_{Li}, x_{Ri}, y_i) = \frac{b}{x_{Li} - x_{Ri}} \begin{bmatrix} \frac{1}{2}(x_{Li} + x_{Ri}) \\ y_i \\ f \end{bmatrix}, \quad (1)$$

where f denotes the focal length of the camera lens, measured in pixels, and x_{Li}, x_{Ri}, and y_i denote the pixel coordinates of target i on the left and right camera images, as in Fig. 1, where we note that y_i is equal in the left and right image by the epipolar constraint. Given the orientation and position of the mobile camera, it is useful to consider the location of target i in global coordinates

$$\mathbf{x}_i \triangleq R(t)\mathbf{p}_i + \mathbf{r}(t). \quad (2)$$

In practice, we can only observe quantized versions of the image coordinate tuples (x_{Li}, x_{Ri}, y_i) once they are rounded to the nearest pixel centers, which we hereafter denote by $\breve{x}_{Li}$, $\breve{x}_{Ri}$, and $\breve{y}_i$. In view of (1), the corrupted observation $(\breve{x}_{Li}, \breve{x}_{Ri}, \breve{y}_i)$ carries its quantization error into the observed coordinates $\mathbf{p}_i$ of target i, causing non-Gaussian error distributions (Blostein and Huang 1987; Chang et al. 1994). For convenience, we follow Matthies and Shafer (1987); Förstner (2005) and approximate the quantized error in the pixels as Gaussian to allow uncertainty propagation from image to world coordinates. The noise propagation takes place under a linearization of the measurement equation, so that the localization error of the target in the relative camera frame will also be Gaussian with mean $\breve{\mathbf{p}}_i = \mathbf{p}(\breve{x}_{Li}, \breve{x}_{Ri}, \breve{y}_i)$. It follows from (2) that the global location estimate $\breve{\mathbf{x}}_i$ is also subject to Gaussian noise.

Targets may be mobile, so we denote the ground truth full state of target i by $\mathbf{z}_i = [\mathbf{x}_i^\top \ \dot{\mathbf{x}}_i^\top \ \ddot{\mathbf{x}}_i^\top]^\top$. Then, $\mathbf{x}_i$ and $\mathbf{z}_i$ are related by $\mathbf{x}_i = H\mathbf{z}_i$, where $H = [1\ 0\ 0] \otimes I_3$, where $\otimes$ represents the Kronecker product. Thus, we can think of $\breve{\mathbf{x}}_i$ as a noisy copy of the zero-th order terms of $\mathbf{z}_i$,

$$\breve{\mathbf{x}}_i = H\mathbf{z}_i + \mathbf{v}_i, \quad (3)$$

where $\mathbf{v}_i$ is a white noise vector. We hereafter denote the covariance of $\mathbf{v}_i$ by $\Sigma_i \in \mathbb{S}_+^3$, where $\mathbb{S}_+^3$ denotes the set of 3×3 symmetric positive definite matrices. In Section 3, we discuss an explicit form of Σ_i that depends on the measurement itself.

2.1 Kalman Filtering to Fuse the Target Observations

Assume that the stereo camera has made a sequence of observations of the targets. Introduce an index $k \geq 0$ associated with every observation to obtain $\breve{\mathbf{x}}_{i,k}$ and associated covariances $\Sigma_{i,k}$ from (3). Our goal is to create accurate state information for a group of targets based on a sequence of such observations. For this, we use a Kalman filter (KF), which is an efficient filter that can incorporate a sequence of noisy measurements within a system model to create accurate state estimates.

We model the continuous time evolution of target i's motion with the discrete time linear equation

$$\mathbf{z}_{i,k} = \Phi\mathbf{z}_{i,k-1}. \quad (4)$$

In (4), Φ is the state transition matrix, which is unknown to the observer. Adaptive procedures for determining Φ are well studied in the literature on mobile target tracking (Singer 1970; Rong Li and Jilkov 2003). Zero velocity and constant acceleration models of the target trajectory, which we discuss in Section 5, are modeled over a short time interval Dt by

$$\Phi_{\dot{\mathbf{x}}_i=\mathbf{0}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \otimes I_3 \text{ and} \quad (5a)$$

$$\Phi_{\ddot{\mathbf{x}}_i=\mathbf{0}} = \begin{bmatrix} 1 & Dt & \frac{Dt^2}{2} \\ 0 & 1 & Dt \\ 0 & 0 & 1 \end{bmatrix} \otimes I_3, \quad (5b)$$

The KF recursively creates state estimates, which we denote by $\hat{\mathbf{z}}_i$, and their associated covariances, which we denote by U_i. In particular, given prior estimates $\hat{\mathbf{z}}_{i,k-1|k-1}$ and $U_{i,k-1|k-1}$, we update the state estimates and fuse the covariance matrices according to the following KF:

$$\hat{\mathbf{z}}_{i,k|k-1} = \Phi\hat{\mathbf{z}}_{i,k-1|k-1} \quad (6a)$$

$$U_{i,k|k-1} = \Phi U_{i,k-1|k-1}\Phi^\top + W \quad (6b)$$

$$K_k = U_{i,k|k-1}H^\top \left[HU_{i,k|k-1}H^\top + \Sigma_{i,k}\right]^{-1} \quad (6c)$$

$$\hat{\mathbf{z}}_{i,k|k} = \hat{\mathbf{z}}_{i,k|k-1} + K_k\left[\breve{\mathbf{x}}_{i,k} - H\hat{\mathbf{z}}_{i,k|k-1}\right] \quad (6d)$$

$$U_{i,k|k} = U_{i,k|k-1} - K_k H U_{i,k|k-1} \quad (6e)$$

where W is the process noise covariance matrix and accounts for the approximate nature of Φ. Singer (1970) gives a closed

form for this matrix,

$$W = \begin{bmatrix} Dt^5/20 & Dt^4/8 & Dt^3/6 \\ Dt^4/8 & Dt^3/3 & Dt^2/2 \\ Dt^3/6 & Dt^2/2 & Dt \end{bmatrix} \otimes I_3. \quad (7)$$

From equation (6e) and the results of Bishop and Welch (2001), a closed form expression for the fused covariance estimate follows in the form of a Lemma.

Lemma 1. *Let $U_{i,k|k-1}$ denote the predicted covariance of all prior observations and $\Sigma_{i,k}$ denote the covariance of the most recent measurement. Then, the location estimate of target i, $H\hat{\mathbf{z}}_{i,k|k}$, has a covariance matrix, which we hereafter denote by $\Xi_{i,k}$, given by*

$$\Xi_{i,k} \triangleq HU_{i,k|k}H^\top = \left[\left(HU_{i,k|k-1}H^\top\right)^{-1} + \Sigma_{i,k}^{-1}\right]^{-1}. \quad (8)$$

Proof. From the definition of $\Xi_{i,k}$ we have that,

$$U_{i,k|k} = U_{i,k|k-1} - K_k HU_{i,k|k-1}$$
$$HU_{i,k|k}H^\top = H\left(U_{i,k|k-1} - K_k HU_{i,k|k-1}\right)H^\top = \Xi_{i,k}.$$

To simplify the analysis, let $U = U_{i,k|k-1}$ and $\Sigma = \Sigma_{i,k}$. Substituting the Kalman gain K_k from (6c), we have

$$\begin{aligned}
\Xi_{i,k} &= HUH^\top\left[I - \left(HUH^\top + \Sigma\right)^{-1}HUH^\top\right] \\
&= HUH^\top\Big[\left(HUH^\top + \Sigma\right)^{-1}\left(HUH^\top + \Sigma\right) - \\
&\qquad\qquad \left(HUH^\top + \Sigma\right)^{-1}HUH^\top\Big] \\
&= HUH^\top\left(HUH^\top + \Sigma\right)^{-1}\left(HUH^\top + \Sigma - HUH^\top\right) \\
&= HUH^\top\left(HUH^\top + \Sigma\right)^{-1}\Sigma \\
&= \Sigma\left(HUH^\top + \Sigma\right)^{-1}HUH^\top \\
&= \left(\left(HUH^\top\right)^{-1}\left(HUH^\top + \Sigma\right)\Sigma^{-1}\right)^{-1} \\
\Xi_{i,k} &= \left[\Sigma^{-1} + \left(HUH^\top\right)^{-1}\right]^{-1}.
\end{aligned}$$

These manipulations are legal because covariance matrices are positive definite and therefore symmetric and invertible.

2.2 The Next Best View Problem

Suppose there are $k-1$ available observations of the group of targets in $\mathcal{N}$, and let

$$HU_{s,k|k-1}H^\top \text{ with } s = \underset{j\in\mathcal{N}}{\operatorname{argmax}}\left\{\mathbf{tr}\left[HU_{j,k|k-1}H^\top\right]\right\} \quad (9)$$

denote the predicted covariance of the worst localized target and

$$HU_{c,k|k-1}H^\top = \frac{1}{n}\sum_{i\in\mathcal{N}} HU_{i,k|k-1}H^\top. \quad (10)$$

denote the average of all predicted target covariances at iteration k. The problem that we address in this paper is as follows.

Problem 1. Next-Best-View. *Given the predicted covariance of the worst localized target $HU_{s,k|k-1}H^\top$ (respectively, the average of the targets' predicted covariances $HU_{c,k|k-1}H^\top$) and the predicted next location $\hat{\mathbf{z}}_{s,k|k-1}$ of target s (respectively, the average of the targets' predicted locations $\hat{\mathbf{z}}_{c,k|k-1}$), determine the next pose of the stereo rig $(\mathbf{r}(t_k), R(t_k))$ so that $\mathbf{tr}[\Xi_{s,k}]$ (respectively, $\mathbf{tr}[\Xi_{c,k}]$) is minimized.*

To solve Problem 1 we make the following assumptions:

(A1) noise is dominated by quantization of pixel coordinates;
(A2) correct correspondence of the targets between the images in the stereo rig exists;
(A3) if targets are in the field of view of the cameras, then they are not occluded by any obstacle in space.

Assumption (A1) allows us to isolate, analyze, and control the effect of pixelation noise on the target localization process. While other sources of noise do exist, pixelation noise does in fact dominate for small disparities, e.g., when the camera is far away from the targets. The effect of other noise sources can be critical for the stability of the KF, and we discuss a novel data-driven approach to obtain empirical models of these uncertainties in Section 6. Assumptions (A2) and (A3) allow us to simplify the problem formulation. It is well known that correspondence and occlusion are both important problems and, being such, have received significant attention in the computer vision literature, see e.g., Scharstein and Szeliski (2002) and the references therein. In this paper, assumptions (A2) and (A3) allow us to obtain analytic and computationally efficient solutions to the Next-Best-View and target localization problem using exact models of stereo vision sensors. In situations where correspondence errors and occlusions do not raise significant challenges, e.g., for sparse target configurations, our approach can have significant practical applicability.

In problem 1, we have chosen the trace as a measure of uncertainty among other choices, such as the determinant or the maximum eigenvalue. (A similar choice was made by Ponda and Frazzoli (2009).) Wenhardt et al. (2007a) shows that all such criteria behave similarly in practice. Since minimization of $\mathbf{tr}[\Xi_{s,k}]$ is associated with improving localization of the worst localized target, we call it the *supremum objective*. We call minimization of $\mathbf{tr}[\Xi_{c,k}]$ the *centroid objective*. $\Xi_{s,k}$ will depend only on the predicted next position of the worst localized target, which we denote by $\mathbf{p}_{s,k}$, but $\Xi_{c,k}$ will depend on the predicted next positions $\mathbf{p}_{i,k}$ of all $i \in \mathcal{N}$.

Attempting to solve Problem 1 by simultaneously controlling the covariances of all targets requires a nonconvex constraint to maintain consistency between images. We note that, when we employ the supremum or centroid objective, the decision process comprises two nonlinear procedures: triangulation and Kalman Filtering.

3 Controlling the Relative Frame

Assume that $k-1$ observations are already available. Our goal in this section is to determine the next best target locations $\mathbf{p}_{s,k}$ or $\mathbf{p}_{c,k}$ on the relative camera frame so that if a new observation is made with the targets at these new relative locations, the fused localization uncertainty, which is captured by $\Xi_{s,k}$ or $\Xi_{c,k}$, is optimized. For this, we need to express the instantaneous covariance $\Sigma_{i,k}$ of target i as a

function of the relative position $\mathbf{p}_{i,k}$. To simplify notation, in this section we drop the subscripts s, c, and o. We will also drop the subscript k when no confusion can occur.

From (1), we know that $\mathbf{p}$ depends on the noisy vector $(\check{x}_L, \check{x}_R, \check{y})$, which we assume has some known or easily estimated covariance Q. * In the experiments of Section 6, we propose a new data-driven linear model to estimate Q. Let J be the Jacobian of $\mathbf{p} \triangleq \mathbf{p}(x_L, x_R, y)$ evaluated at the point $(\check{x}_L, \check{x}_R, \check{y})$, given by

$$J = \frac{b}{(\check{x}_L - \check{x}_R)^2} \begin{bmatrix} -\check{x}_R & \check{x}_L & 0 \\ -\check{y} & \check{y} & \check{x}_L - \check{x}_R \\ -f & f & 0 \end{bmatrix}. \tag{11}$$

Then, the first order (linear) approximation of $\mathbf{p}$ about the point $(\check{x}_L, \check{x}_R, \check{y})$ is

$$\mathbf{p}(x_L, x_R, y) \approx \mathbf{p}(\check{x}_L, \check{x}_R, \check{y}) + J \begin{bmatrix} \check{x}_L - x_L \\ \check{x}_R - x_R \\ \check{y} - y \end{bmatrix}. \tag{12}$$

Since $\mathbf{p}(\check{x}_L, \check{x}_R, \check{y})$ corresponds to the current mean estimate of target coordinates, it is constant in (12). Therefore, the covariance of $\mathbf{p}$ in the relative camera frame is $JQJ^\top$. Fusing covariance matrices as in Lemma 1 requires that they are represented in the same coordinate system. To represent the covariance $JQJ^\top$ in global coordinates, we need to rotate it by an amount corresponding to the camera's orientation at the time this covariance is evaluated. Assuming that consecutive observations are close in space, so that the camera makes a small motion during the time interval $[t_{k-1}, t_k]$, we may approximate the camera's rotation $R(t)$ at time $t \in [t_{k-1}, t_k]$ by its initial rotation $R(t_{k-1})$. We note that this approximation will be inaccurate if the robot moves long distances between consecutive observations. For the use case discussed in Section 6, the robot takes multiple observations per second, so this approximation is not an issue. Denoting $R(t_{k-1})$ by R, the instantaneous covariance of $\mathbf{p}$ can be approximated by

$$\Sigma = \mathrm{cov}[\mathbf{p}(\check{x}_L, \check{x}_R, \check{y})] \approx RJQJ^\top R^\top. \tag{13}$$

In view of (1) and (11), the covariance in (13) is clearly a function of the target coordinates on the relative image frame. Using this model of measurement covariance, we define the uncertainty potential

$$h(\mathbf{p}) = \mathbf{tr}\{\Xi\}, \tag{14}$$

Then, the next best view vector that minimizes h can be obtained using the gradient descent update

$$\mathbf{p}_k = \mathbf{p}_{k-1} - K \int_0^T \nabla_{\mathbf{p}} h\left(\mathbf{p}(\tau)\right) d\tau, \tag{15}$$

where K is a gain matrix. The length $T > 0$ of the integration interval is chosen so that the distance between $\mathbf{p}_k$ and $\mathbf{p}_{k+1}$ is less than the maximum distance the robot can travel before another NBV is calculated at time t_k. The following result provides an analytical expression for the gradient of the potential h in (15).

Proposition 2. *The j-th coordinate of the gradient of h with respect to $\mathbf{p}$ is given by*

$$\frac{\partial h}{\partial [\mathbf{p}]_j} = \mathbf{tr}\left\{ \Sigma^{-1} \Xi^2 \Sigma^{-1} \frac{\partial \Sigma}{\partial [\mathbf{p}]_j} \right\}, \tag{16}$$

where $\frac{\partial \Sigma}{\partial [\mathbf{p}]_j}$ is the partial derivative of Σ with respect to the j-th coordinate of $\mathbf{p}$, and $j = x, y, z$, corresponds to the three dimensions of $\mathbf{p}$.

It is not difficult to show that

$$\frac{\partial h}{\partial [\mathbf{p}]_j} = -\mathbf{tr}\left\{ \Xi^2 \frac{\partial \left[(HUH^\top)^{-1} + \Sigma^{-1} \right]}{\partial [\mathbf{p}]_j} \right\}. \tag{17}$$

Note that the covariance of all prior fused measurements $HUH^\top$ is a constant with respect to the next best view $\mathbf{p}$ and, therefore, its derivative with respect to $\mathbf{p}$ is zero, i.e., $\partial \left(HUH^\top\right)^{-1} / \partial [\mathbf{p}]_j = 0$. The derivative of Σ^{-1} with respect to $[\mathbf{p}]_j$ leads to an expression for the derivative in the right-hand-side of (17) that retreives (16).

In what follows, we apply the chain rule to calculate $\partial \Sigma / \partial [\mathbf{p}]_j$ in (16). In particular, since we hold R constant during the relative update, we have that the partial derivatives of Σ in the directions $[\mathbf{p}]_j$ for $j = x, y, z$ are taken only with respect to the entries of $JQJ^\top$, i.e.,

$$\frac{\partial \Sigma}{\partial [\mathbf{p}]_j} = R \left(\frac{\partial J}{\partial [\mathbf{p}]_j} QJ^\top + JQ \frac{\partial J^\top}{\partial [\mathbf{p}]_j} \right) R^\top. \tag{18}$$

Then, using the chain rule,

$$\frac{\partial J}{\partial [\mathbf{p}]_j} = \frac{\partial J}{\partial x_L} \frac{\partial x_L}{[\mathbf{p}]_j} + \frac{\partial J}{\partial x_R} \frac{\partial x_R}{[\mathbf{p}]_j} + \frac{\partial J}{\partial y} \frac{\partial y}{[\mathbf{p}]_j}. \tag{19}$$

The need arises to express the pixel coordinate tuple (x_L, x_R, y) as a function of the location of target in relative coordinates $\mathbf{p}$. This is available via the inverse of (1), given by

$$\begin{bmatrix} x_L \\ x_R \\ y \end{bmatrix} = \frac{f}{[\mathbf{p}]_z} \begin{bmatrix} [\mathbf{p}]_x + \frac{b}{2} \\ [\mathbf{p}]_x - \frac{b}{2} \\ [\mathbf{p}]_y \end{bmatrix}. \tag{20}$$

Then, (19) can be evaluated by finding the partial derivative of J with respect to (x_L, x_R, y) and the partial derivatives of the entries of (20) with respect to each coordinate of $\mathbf{p}$. Using these derivatives, all terms in (18) are accounted for, which completes the proof of Proposition 2.

4 Controlling the Global Frame

The update in (15) provides the desired change in relative target coordinates $\mathbf{p}_{o,k} - \mathbf{p}_{o,k-1}$ of target o in the camera frame, where o stands for 'objective' and can be either s or c, depending on the objective defined in Problem 1. Our goal in this section is to determine a new camera position $\mathbf{r}(t_k)$ and orientation $R(t_k)$ in space that realizes the change in view, effectively arriving at the Next Best View of the target located at $\hat{\mathbf{x}}_o$. Transforming the change of view into global

*Recall that we approximate the uniform pixelation noise as Gaussian, hence the approximate nature of Q.

coordinates, the goal position $\mathbf{r}^*$ is defined as

$$\mathbf{r}^* \triangleq \mathbf{r}(t_{k-1}) + R(t_{k-1})(\mathbf{p}_{o,k} - \mathbf{p}_{o,k-1}). \quad (21)$$

The ability to rotate the camera in addition to translating it means that there are infinitely many poses in the global frame that realize the NBV in relative coordinates. The goal orientation is defined to be any pose such that the point $\hat{\mathbf{x}}_o$ lies on the z-axis of the camera relative coordinate system, i.e., the camera is looking straight at the centroid (or supremum) target location. To achieve this new desired camera position and orientation, we define the following potential which we seek to minimize:

$$\psi(\mathbf{r}, R) = \overbrace{\|\mathbf{r} - \mathbf{r}^*\|^2}^{\text{position}} + \overbrace{\left\|R^\top \hat{\mathbf{z}} - \mathbf{e}_3\right\|^2}^{\text{orientation}}, \quad (22)$$

where

$$\hat{\mathbf{z}} = \frac{\hat{\mathbf{x}}_o - \mathbf{r}^*}{\|\hat{\mathbf{x}}_o - \mathbf{r}^*\|} \text{ and } \mathbf{e}_3 = [0\ 0\ 1]^\top. \quad (23)$$

In (23), $\hat{\mathbf{z}}$ is the direction in global coordinates from the desired robot position $\mathbf{r}^*$ to the estimated target-objective location $\hat{\mathbf{x}}_o$, and $\mathbf{e}_3$ is the unit vector in the direction of the robot's view in relative coordinates, defined to be the z-axis. Note that the robot and target cannot be located at the same point, because this would violate field of view constraints.

To minimize ψ, we define the following gradient flow for all time $t \in [t_{k-1}, t_k]$

$$\dot{\mathbf{r}} = -\nabla_{\mathbf{r}}\psi(\mathbf{r}, R) \quad (24a)$$

$$\dot{R} = -R\nabla_R\psi(\mathbf{r}, R), \quad (24b)$$

in the joint space of camera positions $\mathbb{R}^3$ and orientations $SO(3)$, where $\nabla_{\mathbf{r}}\psi$ and $\nabla_R\psi$ are the gradients of ψ with respect to $\mathbf{r}$ and R. After initializing the gradient flow (24) at $\big(\mathbf{r}(t_{k-1}), R(t_{k-1})\big)$, the following lemma shows that if $R(t_{k-1}) \in SO(3)$ and $R(t)$ evolves as in (24b) and $\nabla_R\psi(\mathbf{r}, R)$ is skew-symmetric, then $R(t) \in SO(3)$ for all time $t \in [t_{k-1}, t_k]$; see Zavlanos and Pappas (2008).

Lemma 3. *Let $\Omega(t)$ be skew-symmetric $\forall\, t \geq 0$ and define the matrix differential equation $\dot{R}(t) = R(t)\Omega(t)$. Then, $R(t) \in SO(n)\, \forall\, t \geq 0$ if $R(0) \in SO(n)$.*

In other words, the gradient flow (24b) is implicitly constrained to the set of Special Euclidean transformations during the minimization of ψ.

4.1 Closed Form Motion Controllers

In the remainder of this section we provide analytic expressions for the gradients in (24). We also use these expressions to show that the closed loop system (24) minimizes ψ. The first proposition identifies the gradient of ψ with respect to R. To prove it, we use the matrix inner product $\langle A, B\rangle = \mathbf{tr}(A^\top B)$, which has the following property.

Lemma 4. *For any square matrix A and skew-symmetric matrix Ω of appropriate size, $2\langle A, \Omega\rangle = \langle A - A^\top, \Omega\rangle$.*

Proof. We have that $2\langle A, \Omega\rangle = \langle A, \Omega\rangle + \langle \Omega, A\rangle = \mathbf{tr}(A^\top\Omega + \Omega^\top A) = \mathbf{tr}((A^\top - A)\Omega) = \langle A - A^\top, \Omega\rangle$.

Proposition 5. *The gradient of ψ with respect to R is given by the skew-symmetric matrix*

$$\nabla_R\psi = R^\top\hat{\mathbf{z}}(R^\top\hat{\mathbf{z}} - \mathbf{e}_3)^\top - (R^\top\hat{\mathbf{z}} - \mathbf{e}_3)\hat{\mathbf{z}}^\top R. \quad (25)$$

Proof. For any skew symmetric matrix Ω,

$$\left\|(R(I+\Omega))^\top\hat{\mathbf{z}} - \mathbf{e}_3\right\|^2 = \left\|R^\top\hat{\mathbf{z}} - \mathbf{e}_3 - \Omega R^\top\hat{\mathbf{z}}\right\|^2.$$

Let $\mathbf{v} = R^\top\hat{\mathbf{z}} - \mathbf{e}_3$ to simplify notation. Using the first order approximation of the neighborhood of the rotation matrix $R(\Omega) \approx I + \Omega$, where Ω is skew-symmetric, and using Lemma 4 along with the basic properties of inner products, we have that

$$\begin{aligned}\psi(\mathbf{r}, R(I+\Omega)) &= \|\mathbf{r} - \mathbf{r}^*\|^2 + \left\|\mathbf{v} - \Omega R^\top\hat{\mathbf{z}}\right\|^2 \\ &= \|\mathbf{r} - \mathbf{r}^*\|^2 + \|\mathbf{v}\|^2 - 2\left\langle\mathbf{v}, \Omega R^\top\hat{\mathbf{z}}\right\rangle + o(\|\Omega\|) \\ &= \psi(\mathbf{r}, R) - 2\left\langle\mathbf{v}\hat{\mathbf{z}}^\top R, \Omega\right\rangle + o(\|\Omega\|) \\ &= \psi(\mathbf{r}, R) + \left\langle R^\top\hat{\mathbf{z}}\mathbf{v}^\top - \mathbf{v}\hat{\mathbf{z}}^\top R, \Omega\right\rangle + o(\|\Omega\|),\end{aligned}$$

from which we identify $R^\top\hat{\mathbf{z}}\mathbf{v}^\top - \mathbf{v}\hat{\mathbf{z}}^\top R$ as $\nabla_R\psi(\mathbf{r}, R)$, and the result follows immediately.

Additionally, we have from elementary calculus that

$$\begin{aligned}\nabla_{\mathbf{r}}\psi(\mathbf{r}, R) &= 2(\mathbf{r} - \mathbf{r}^*) \quad (26)\\ &= 2\left(\mathbf{r} - \mathbf{r}(t_{k-1}) - R(t_{k-1})(\mathbf{p}_{o,k} - \mathbf{p}_{o,k-1})\right).\end{aligned}$$

The following result shows that the closed loop system (24) is globally asymptotically stable about the minimizers of ψ.

Theorem 6. *The trajectories of the closed loop system (24) globally converge to the set of minimizers of the function ψ.*

Proof. By inspection of (22), $\psi(\mathbf{r}, R) \geq 0$, with equality if and only if $R^\top\hat{\mathbf{z}} = \mathbf{e}_3$ and $\mathbf{r} = \mathbf{r}^*$. In the remainder of the proof, we show that ψ is a suitable Lyapunov function for the closed loop system (24), and the set of equilibrium points is exactly the set of minimizers of ψ. To begin, let $\mathbf{v}$ be defined as above, so that

$$\begin{aligned}\dot{\psi}(\mathbf{r}, R) &= 2\langle\mathbf{r} - \mathbf{r}^*, \dot{\mathbf{r}}\rangle + 2\left\langle R^\top\hat{\mathbf{z}} - \mathbf{e}_3, \dot{R}^\top\hat{\mathbf{z}}\right\rangle \\ &= 2\langle\mathbf{r} - \mathbf{r}^*, -\nabla_{\mathbf{r}}\psi(\mathbf{r}, R)\rangle + 2\left\langle\mathbf{v}, (-R\nabla_R\psi(\mathbf{r}, R))^\top\hat{\mathbf{z}}\right\rangle \\ &= 2\langle\mathbf{r} - \mathbf{r}^*, -2(\mathbf{r} - \mathbf{r}^*)\rangle + 2\left\langle\mathbf{v}, (R^\top\hat{\mathbf{z}}\mathbf{v}^\top - \mathbf{v}\hat{\mathbf{z}}^\top R)R^\top\hat{\mathbf{z}}\right\rangle \\ &= -4\|\mathbf{r} - \mathbf{r}^*\|^2 + 2\left(\left\langle\mathbf{v}, R^\top\hat{\mathbf{z}}\mathbf{v}^\top R^\top\hat{\mathbf{z}}\right\rangle - \left\langle\mathbf{v}, \mathbf{v}\hat{\mathbf{z}}^\top RR^\top\hat{\mathbf{z}}\right\rangle\right) \\ &= -4\|\mathbf{r} - \mathbf{r}^*\|^2 + 2\left(\left(\mathbf{v}^\top R^\top\hat{\mathbf{z}}\right)^2 - \|\mathbf{v}\|^2\right). \quad (27)\end{aligned}$$

The Cauchy-Schwartz inequality implies that

$$\left(\mathbf{v}^\top R^\top\hat{\mathbf{z}}\right)^2 \leq \|\mathbf{v}\|^2\left\|R^\top\hat{\mathbf{z}}\right\|^2 = \|\mathbf{v}\|^2,$$

so that (27) is the sum of two nonpositive terms. Thus, $\dot{\psi} \leq 0$, with equality if and only if both of the nonpositive terms are zero. In particular, $\dot{\psi}(\mathbf{r}, R) = 0$ if and only if $\mathbf{r} = \mathbf{r}^*$ and

$$\begin{aligned}\|\mathbf{v}\|^2 &= (\hat{\mathbf{z}}^\top R\mathbf{v})^2 \\ (\hat{\mathbf{z}}^\top R - \mathbf{e}_3^\top)(R^\top\hat{\mathbf{z}} - \mathbf{e}_3) &= (\hat{\mathbf{z}}^\top R(R^\top\hat{\mathbf{z}} - \mathbf{e}_3))^2 \\ 2(1 - \hat{\mathbf{z}}^\top R\mathbf{e}_3) &= (1 - \hat{\mathbf{z}}^\top R\mathbf{e}_3)^2 \\ 1 - (\hat{\mathbf{z}}^\top R\mathbf{e}_3)^2 &= 0,\end{aligned}$$

which implies $\hat{\mathbf{z}}^\top R = \mathbf{e}_3^\top$ for all critical points. Invoking the Lyapunov Stability Theorem, the result follows.

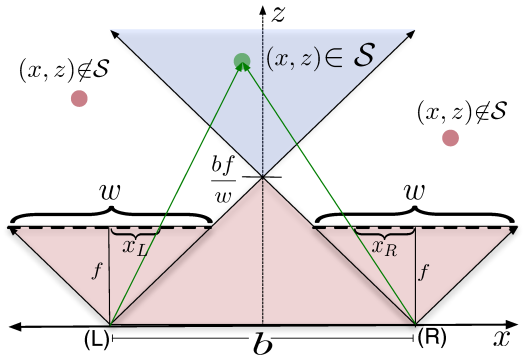

Figure 2. The field of view for a stereo camera in the xz plane. The field of view in the yz plane is similar.

Note that the system (24) evolves during the time interval $[t_{k-1}, t_k]$, until a new observation of the targets is made at time t_k. This time interval might not be sufficient for the camera to realize exactly the NBV. Nevertheless, Theorem 6 implies that at time t_k, the position and orientation of the camera is closer to desired NBV than it was at time t_{k-1}. By appropriately choosing the length of the time interval $[t_{k-1}, t_k]$, we may ensure that for practical purposes the camera almost realizes the NBV.

5 Performance of the Integrated Hybrid System

In this section, we illustrate our approach in computer simulations. We begin by discussing a practical method of how to incorporate field of view constraints in the hybrid system, which is used in our simulations and experimental results.

5.1 Incorporating Field of View Constraints

For a 3D point to appear in a given image, that point must lie within the field of view of both cameras in the stereo pair as the robot rotates and translates in an effort to minimize (22). We assume that the (L) and (R) cameras have identical square sensors with a 70° field of view, which, when combined with the image width w, determines the focal length f. Let

$$\mathcal{S} = \left\{ [x, y, z]^\top \in \mathbb{R}^3 \colon\ |x| \leq \frac{wz - bf}{2f},\ |y| \leq \frac{zw}{2f},\ z > \frac{bf}{w} \right\}$$

denote the set of points in relative coordinates that are visible to both cameras in the pair. This set is the intersection of two pyramids facing the positive z direction with vertices located at the two camera centers. Figure 2 visualizes the set $\mathcal{S}$ in two dimensions (blue shaded region). Note that the intersection of the two pyramids is located at $z = \frac{bf}{w}$, and therefore any point with $z < \frac{bf}{w}$ can not be in view of both cameras.

Maintaining all targets within the FoV $\mathcal{S}$ requires that the camera positions and orientations evolve in the set

$$\mathcal{D} = \left\{ (\mathbf{r}, R) \in \mathbb{R}^3 \times SO(3) \ :\ \{\mathbf{p}_i\}_{i\in\mathcal{N}} \in \mathcal{S} \right\} \tag{28}$$

for all time. To ensure invariance of the set $\mathcal{D}$, we define the potential functions

$$\phi_{i1}(\mathbf{r}, R) = \left(\frac{w[\mathbf{p}_i]_z - bf}{2f} \right)^2 - [\mathbf{p}_i]_x^2, \tag{29a}$$

$$\phi_{i2}(\mathbf{r}, R) = \left(\frac{w[\mathbf{p}_i]_z}{2f} \right)^2 - [\mathbf{p}_i]_y^2, \tag{29b}$$

$$\phi_{i3}(\mathbf{r}, R) = [\mathbf{p}_i]_z^2 - \left(\frac{bf}{w} \right)^2, \tag{29c}$$

that are positive if $(\mathbf{r}, R) \in \mathcal{D}$, where $[\mathbf{p}_i]_x$, $[\mathbf{p}_i]_y$, and $[\mathbf{p}_i]_z$ are the x, y, and z coordinates of target i in the relative camera frame, that can be expressed in terms of the camera position and orientation as

$$[\mathbf{p}_i]_x = \left\langle \mathbf{e}_1, R^\top(\hat{\mathbf{x}}_i - \mathbf{r}) \right\rangle, \tag{30a}$$

$$[\mathbf{p}_i]_y = \left\langle \mathbf{e}_2, R^\top(\hat{\mathbf{x}}_i - \mathbf{r}) \right\rangle, \tag{30b}$$

$$[\mathbf{p}_i]_z = \left\langle \mathbf{e}_3, R^\top(\hat{\mathbf{x}}_i - \mathbf{r}) \right\rangle, \tag{30c}$$

where $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are the unit vectors in the standard basis. Then, given an estimate of target locations $\hat{\mathbf{x}}_i$ for $i = 1, \ldots, n$, we augment the potential ψ from (22) by adding barrier functions $1/\phi_{ij}$ that will grow without bound anytime a target is close to the boundary of the feasible set $\mathcal{D}$. The repulsive force supplied by ϕ_i is regulated by a user defined penalty parameter $\rho > 0$. The artificial potential function, incorporating the desired FoV constraints, is given by

$$\hat{\psi}(\mathbf{r}, R) = \psi(\mathbf{r}, R) + \frac{\rho}{n} \sum_{i\in\mathcal{N}} \sum_{j=1}^{3} g(\phi_{ij}), \tag{31}$$

where $g \colon \mathbb{R} \to \mathbb{R}$ is a barrier potential, and multiplication by $1/n$ ensures that the number of targets does not affect the strength of the penalty. The penalty parameter ρ is set sufficiently small so that $\hat{\psi}$ approximates ψ when $(\mathbf{r}, R)$ is in the interior of $\mathcal{D}$ while maintaining that $\hat{\psi}$ becomes extremely large for $(\mathbf{r}, R)$ that approach the boundary of $\mathcal{D}$. Replacing ψ with $\hat{\psi}$ in the gradient flow in Algorithm 1 provides the desired potential that realizes the NBV and respects FoV constraints. In the simulations, we set $g(a) = \frac{1}{a}$.

In what follows we derive analytical expressions for the gradients of $\hat{\psi}$. In particular, we have that

$$\nabla_{\mathbf{r}}\hat{\psi}(\mathbf{r}, R) = \nabla_{\mathbf{r}}\psi + \frac{\rho}{n} \sum_{i\in\mathcal{N}} \sum_{j=1}^{3} g'(\phi_{ij}) \nabla_{\mathbf{r}}\phi_{ij}, \tag{32a}$$

$$\nabla_{R}\hat{\psi}(\mathbf{r}, R) = \nabla_{R}\psi + \frac{\rho}{n} \sum_{i\in\mathcal{N}} \sum_{j=1}^{3} g'(\phi_{ij}) \nabla_{R}\phi_{ij}. \tag{32b}$$

The derivative of the barrier function, g', is available from elementary calculus. The gradients in (32) with respect to $\mathbf{r}$ and R can be obtained by application of the chain rule as

$$\begin{aligned} \nabla_{\mathbf{r}}\phi_{ij} &= \frac{\partial\phi_{ij}}{\partial[\mathbf{p}_i]_x}\nabla_{\mathbf{r}}[\mathbf{p}_i]_x + \frac{\partial\phi_{ij}}{\partial[\mathbf{p}_i]_y}\nabla_{\mathbf{r}}[\mathbf{p}_i]_y + \frac{\partial\phi_{ij}}{\partial[\mathbf{p}_i]_z}\nabla_{\mathbf{r}}[\mathbf{p}_i]_z \\ \nabla_{R}\phi_{ij} &= \frac{\partial\phi_{ij}}{\partial[\mathbf{p}_i]_x}\nabla_{R}[\mathbf{p}_i]_x + \frac{\partial\phi_{ij}}{\partial[\mathbf{p}_i]_y}\nabla_{R}[\mathbf{p}_i]_y + \frac{\partial\phi_{ij}}{\partial[\mathbf{p}_i]_z}\nabla_{R}[\mathbf{p}_i]_z. \end{aligned} \tag{33}$$

The coefficients in (33) can be obtained by differentiating (29). The following propositions provide the gradients of $[\mathbf{p}_i]_x, [\mathbf{p}_i]_y$, and $[\mathbf{p}_i]_z$ with respect to R and $\mathbf{r}$.

Proposition 7. *The gradients of $[\mathbf{p}_i]_x, [\mathbf{p}_i]_y$, and $[\mathbf{p}_i]_z$ with respect to R are given by the skew symmetric matrices*

$$\nabla_R[\mathbf{p}_i]_x = \frac{1}{2}\left[R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\mathbf{e}_1^\top - \mathbf{e}_1(\hat{\mathbf{x}}_i - \mathbf{r})^\top R\right], \quad (34a)$$

$$\nabla_R[\mathbf{p}_i]_y = \frac{1}{2}\left[R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\mathbf{e}_2^\top - \mathbf{e}_2(\hat{\mathbf{x}}_i - \mathbf{r})^\top R\right], \quad (34b)$$

$$\nabla_R[\mathbf{p}_i]_z = \frac{1}{2}\left[R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\mathbf{e}_3^\top - \mathbf{e}_3(\hat{\mathbf{x}}_i - \mathbf{r})^\top R\right]. \quad (34c)$$

Proof. The procedure here is nearly identical to the method in Proposition 5. Specifically,

$$\begin{aligned}[\mathbf{p}_i]_x(\mathbf{r}, R(I+\Omega)) &= \left\langle \mathbf{e}_1, (R(I+\Omega))^\top(\hat{\mathbf{x}}_i - \mathbf{r})\right\rangle \\ &= \left\langle \mathbf{e}_1, (I+\Omega)^\top R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\right\rangle \\ &= \left\langle \mathbf{e}_1, R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\right\rangle - \left\langle \mathbf{e}_1, \Omega R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\right\rangle \\ &= [\mathbf{p}_i]_x(\mathbf{r}, R) - \left\langle \mathbf{e}_1(\hat{\mathbf{x}}_i - \mathbf{r})^\top R, \Omega\right\rangle.\end{aligned}$$

Again we can use Lemma 4 to obtain that

$$\left\langle \mathbf{e}_1(\hat{\mathbf{x}}_i - \mathbf{r})^\top R, \Omega\right\rangle = \frac{1}{2}\left\langle \mathbf{e}_1(\hat{\mathbf{x}}_i - \mathbf{r})^\top R - R^\top(\hat{\mathbf{x}}_i - \mathbf{r})\mathbf{e}_1^\top, \Omega\right\rangle,$$

from which we can identify the gradient as the term linear in Ω, and the proof follows. The gradients of the other two coordinates are found analogously.

Note that the gradients of the functions $[\mathbf{p}_i]_x, [\mathbf{p}_i]_y$, and $[\mathbf{p}_i]_z$ with respect to R are skew-symmetric, as required for (24b) to ensure that $R \in SO(3)$ for all time; see Lemma 3. From elementary calculus, the gradients of $[\mathbf{p}_i]_x, [\mathbf{p}_i]_y$, and $[\mathbf{p}_i]_z$ with respect to $\mathbf{r}$ are

$$\nabla_{\mathbf{r}}[\mathbf{p}_i]_x = -R\mathbf{e}_1,\ \nabla_{\mathbf{r}}[\mathbf{p}_i]_y = -R\mathbf{e}_2,\ \nabla_{\mathbf{r}}[\mathbf{p}_i]_z = -R\mathbf{e}_3. \quad (35)$$

5.2 Outline of Controller

Algorithm 1 outlines the hybrid controller developed in Sections 3 and 4. After initialization, Step 1 determines the NBV according to either the *supremum objective* or the *centroid objective*. Given a frame rate and sensor speed, we set the integration interval T so that the distance between $\mathbf{p}_{o,k-1}$ and $\mathbf{p}_{o,k}$ is the maximum distance the camera can travel before making a new observation. Each time a new observation is made, Step 1 returns a new NBV $\mathbf{p}_{o,k}$, which constitutes a discrete switch in the potential $\hat{\psi}$ in Step 2. This switch results in a new motion plan that guides the robot to a position and orientation that realizes the new NBV. The camera moves according to Step 2 until a new measurement is taken, at which point we set $k := k+1$ and return to Step 1.

5.3 Static Target Localization

We begin this section by illustrating our approach for a simple scenario involving an array of five stationary targets in two dimensions. In this case, the mobile stereo camera effectively has only two motion primitives available: "reduce depth" and "diversify the viewing angle." Thus, the optimal controller will be a state-dependent combination of these two primitives, which should emerge naturally by minimizing the objective function we have described herein. For comparison, we present a "straight baseline" and a "circle baseline," which exclusively utilize one of these two motion primitives. Specifically, the circle baseline moves the robot in a circle about the cluster of targets, and the straight baseline drives the robot closer to the targets. We require that all methods travel the same distance in each iteration, except for the straight baseline that stops moving once FoV constraints tend to become violated. To test the validity of our assumption that pixel-noise due to quantization is uniform on the image plane, all simulated observations in this section have been quantized at the pixel level.

Algorithm 1 Hybrid control in the relative and global frames.

Require: A position $\mathbf{r}(t_{k-1})$ and orientation $R(t_{k-1})$ of the camera and estimated positions $\hat{\mathbf{x}}_{i,k-1}$ of the targets.

1: Find the next best view associated with objective "o" according to equation (15):

$$\mathbf{p}_{o,k} = \mathbf{p}_{o,k-1} - K\int_0^T \nabla h\left(\mathbf{p}_o(\tau)\right) d\tau.$$

2: Move the camera according to the system (24):

$$\dot{\mathbf{r}} = -\nabla_{\mathbf{r}}\hat{\psi}(\mathbf{r}, R),$$
$$\dot{R} = -R\nabla_R\hat{\psi}(\mathbf{r}, R),$$

for a time interval of length $t_k - t_{k-1}$ in order to realize the next best view $\mathbf{p}_{o,k}$ obtained from step 1.

3: At time t_k observe targets and incorporate new estimates and covariances into KF as in (6a) and (6e). Increase the observation index k by 1 and return to step 1.

Figure 3 shows robot trajectories for this simple example. It can be seen that reducing range results in slightly better short-term performance for this situation, but once FoV constraints are nearly violated, repetitive observations from the same spot have correlated noise, leading to divergence of the KF; a similar behavior can be observed in Figure 5, which refers to the 3D example bellow. On the other hand, the circle baseline and the supremum and centroid objectives continuously move the camera, so the i.i.d. noise assumption is not violated and the localization error throughout the simulation keeps being reduced, as can be seen in the shrinking confidence ellipses in Figure 4. By combining the two motion primitives automatically, the supremum and centroid objectives reduce noise by an order of magnitude compared to the straight baseline (before its KF diverges) after around 23 observations.

The following set of simulations considers target localization in three dimensions. The goal is to evaluate our algorithm against the baseline methods. We use an image resolution of 1024×1024 pixels. The unit of measure is the distance between the two cameras in the stereo rig, or the baseline, which is the characteristic length in stereo vision. It is depicted as b from Figure 2. The stereo rig moves 10% of its baseline between successive images, which corresponds to a Dt in the simulations of 0.1. The matrix Q

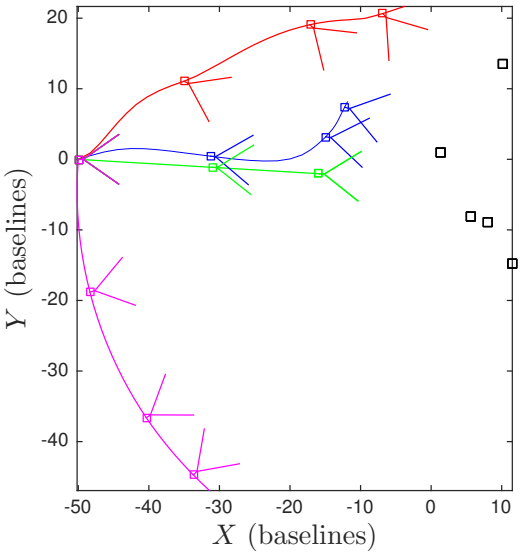

Figure 3. An example of the trajectories generated by our algorithm and the baseline methods, shown in 2D for readability. Red denotes the trajectory of the supremum, blue denotes the centroid, magenta denotes the circle baseline method, and green the straight baseline method. The □ symbols show the ground truth target locations. The triangles emanating from the trajectories represents the orientation and field of view for each objective (see fig. 2).

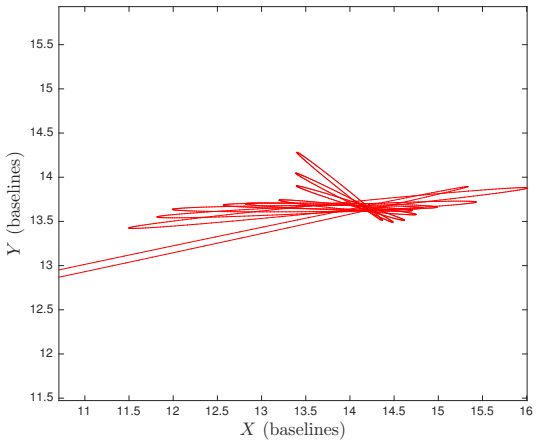

Figure 4. Ten of the 3σ confidence intervals produced by the supremum objective for one of the targets in Figure 3.

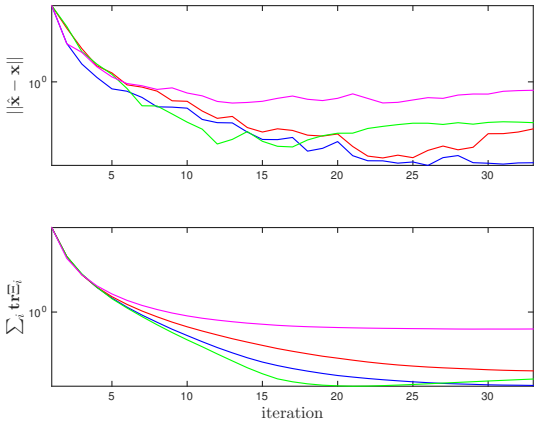

Figure 5. Average localization error (top panel) and trace (bottom panel) of the position covaraince of the all targets versus iteration, averaged over 50 simulations. Red denotes the trajectory of the supremum, blue denotes the centroid, magenta denotes the circle baseline method, and green the straight baseline method.

was set to the identity matrix. In every simulation, the robot begins 50 baselines west of a cluster of targets, which are placed according to a uniform random distribution in the unit cube centered at the origin. The penalty parameter $\rho = 100$ ensures that all targets remain within the camera's 70° field of view throughout. The length of the time interval $t_{k+1} - t_k$ between two consecutive observations is chosen so that the robot either realizes the NBV, i.e., achieves $\psi = 0$ in (24), or the robot travels the maximum allowed distance between observations. The gain parameter, K from (15), is set to $K = \mathrm{diag}(1, 1, 7)$. The observers that follow the circle baseline method and straight baseline method at each iteration travel a distance equal to the maximum of the distances that the supremum and centroid traveled in that iteration. All motion plans make the same amount of total observations. All use identical camera parameters. All observations suffer from quantization noise after pixel coordinates are rounded to the nearest integer.

Figure 5 shows the average total error and the trace of the target location covariance matrices for 50 simulations. In the bottom panel, evidently the straight baseline method outperforms the supremum and centroid objective in terms of the trace of the posterior covariance matrices, up to the point when it stops being able to move. This is because the centroid and supremum objectives also obtain control inputs from a penalty function, which repels the robot from views that allow targets near the field of view boundary. The straight baseline method, on the other hand, can go to the point when one of the targets is on the outer edge of the image, allowing it to get closer. The centroid and supremum objectives still outperform the straight baseline method in terms of localization error. Note also that once the stereo rig following the straight baseline method stops moving, it suffers from the same quantized noise in every observation, which is biased, and causes the KF to diverge. The KFs from the rigs following the circle baseline and the centroid and supremum objectives do not diverge because the individual measurement bias changes when the relative vector changes, effectively de-correlating the errors.

5.4 Mobile Target Localization

In these simulations, the mobile stereo camera localizes a group of mobile targets that move in the Olympic ring pattern. The observers, two cameras implementing the supremum and centroid objectives, use the constant acceleration model from (5). As a simple example, Figures 6 and 7 show an example of the mobile target simulation in two dimensions. We present the results of 50 simulations for the mobile target scenario, again subject to quantized noise from pixelation and again in three dimensions. All constants used in the mobile simulations are the same as the static simulations.

The top panel of Figure 8 shows the average error during the 50 simulations with mobile targets in three dimensions. Because none of the targets stray far from the rest, the *centroid objective* has a slight advantage over the supremum

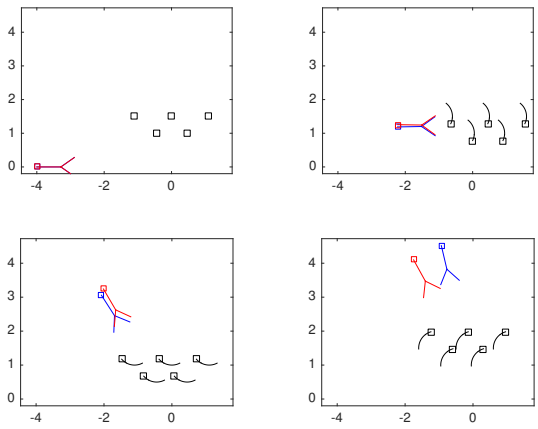

Figure 6. An example of the trajectories generated by our algorithm when the targets are mobile in two dimensions, with time processing in clockwise order. Red denotes the supremum and blue denotes the centroid. The □ symbols show the ground truth target locations, with tails to show their motion. The triangles emanating from the trajectories represents the orientation and field of view for each objective (see Figure 2). All units are in baselines.

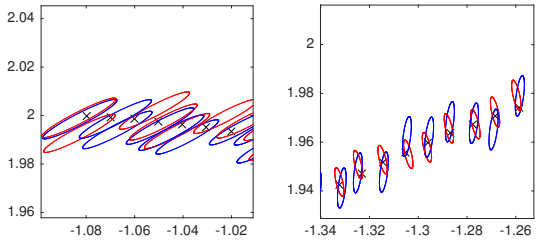

Figure 7. A closeup of the beginning (left panel) and end (right panel) of the left-most target trajectory in Figure 6. 95% Confidence ellipses associated with each objective are plotted. Red denotes the supremum's confidence ellipses, and blue denotes the centroid.

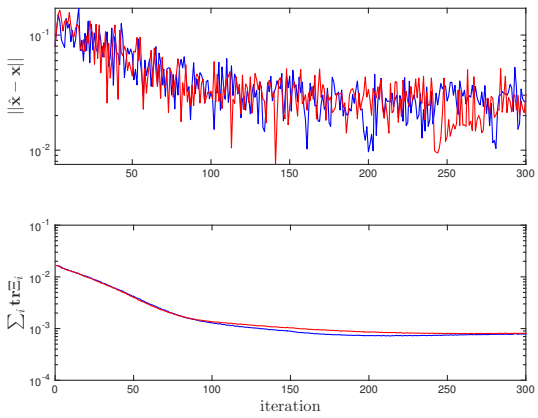

Figure 8. Localization error (top panel) and trace (bottom panel) of the position covariance of the all targets versus iteration for the mobile simulation shown in Figure 6.

objective. We also performed simulations with asymmetric data sets and outliers, which favored the supremum objective. Any nondecreasing properties in the top panel of Figure 8 are due to quantized observations. The correlation coefficient between the time series representing the target error (top panel of Figure 8) and that representing the traces of the covariance matrix sequence (bottom panel of Figure 8) is

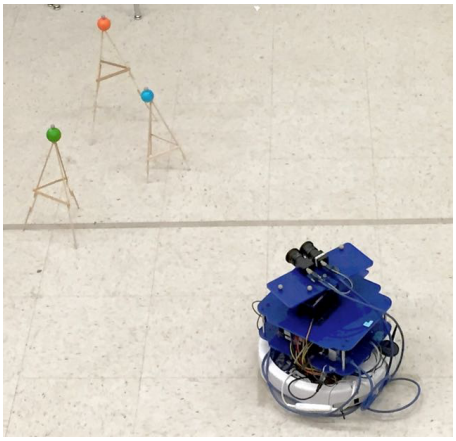

Figure 9. Overhead photograph of the experimental setup

0.84 for the centroid objective and 0.87 for the supremum objective, showing that these proxies are reasonable for localization accuracy. We also note that the flattening out of the objective function value, plotted in the bottom panel of Figure 8, is due to the process noise covariance (7) preventing the KF-updated covariance from converging to the zero matrix. This term prevents the covariance from converging to zero in the mobile target case, and instead holds it near the heuristic value given in (7). Overall localization accuracy could be further improved by *a priori* knowledge of motion model, the on-line adaptive modeling of Rong Li and Jilkov (2003), and using multiple observers, as in Roumeliotis and Bekey (2002).

6 Experiments

In this section, we present experiments using a single ground robot (iRobot Create™), pictured in Figure 9, to localize a set of stationary targets, for which we used colored ping pong balls. The robot carries a stereo rig with 4 cm baseline mounted atop a servo that can rotate the rig $\pm 180°$. The rig uses two Point Grey Flea3™ cameras with resolution 1280×1024. To simulate long distance localization, all images are downsampled by a factor of 24 so that the effective resolution is 54×43, allowing us to operate with disparities at or below ten in our laboratory environment. The robot is equipped with an on-board computer with 8GB RAM and an Intel Core i5-3450S processor. All image processing and triangulation is done on-board using C++ and run on Robot Operating System (ROS). We used the Eigen library for mathematical operations and the OpenCV library for HSV color detection in our controller.

We use the Bouguet (2004) toolbox to calibrate the intrinsic and extrinsic parameters of the stereo rig offline. For self-localization, our laboratory is equipped with an OptiTrack™ array of infrared cameras that tracks reflective markers that are rigidly attached to the robot. The robot is equipped with an 802.11n wireless network card, which it uses to retrieve its position and orientation by reading a ROS topic that is broadcast over wifi. To evaluate the localization accuracy of our algorithm, in addition to saving the robot trajectories, we fix markers to the targets, and the motion capture system records their ground truth locations as well. Finally, note that estimation takes place in three

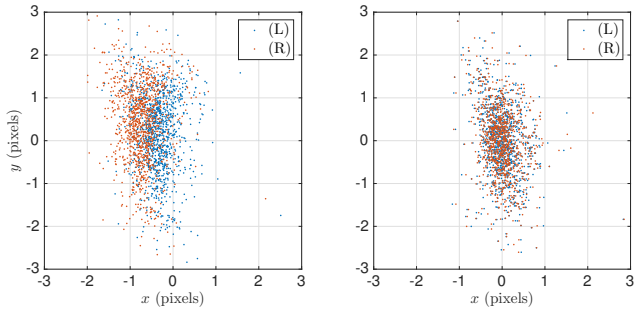

Figure 10. Scatter plots of the residual errors $\epsilon_\ell^{\text{uc}}$ (left panel) and ϵ_ℓ (right panel) for the training data.

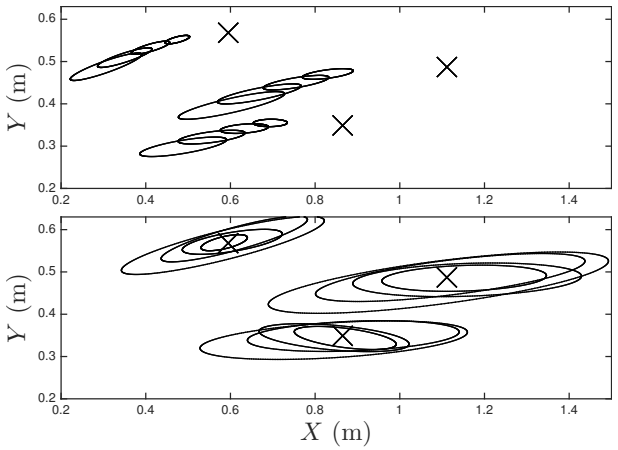

Figure 11. Projecting the Kalman Filtered 95% confidence ellipses onto the X-Y (ground) plane using the raw/uncorrected (top panel) and corrected (bottom panel) pixel observations. The $\times$'s denote the true target locations. The data used to generate these plots were obtained during experimental trials on unseen data. Projections onto the X-Z and Y-Z planes gave similar results.

dimensions, whereas the experimental platform is a ground robot confined to the plane. All navigation and waypoint tracking relies on a PID controller using the next waypoint, defined by the differential flow in (24), as the set point. In the experiment, robots generally came within 2 cm of their target waypoints. The servo is capable of orienting the stereo camera with accuracy of $\pm 1^\circ$ compared to the global controller. No collision avoidance, aside from the implicit collision avoidance from the FoV constraints presented in Section 5.1, is used in the implementation.

6.1 Noise Modeling

In this paper, we have assumed that pixel measurement errors are subject to a known zero mean Normal noise distribution with covariance Q. The goal of this subsection is to ensure that this assumption is satisfied in practice. In particular, we use training data to remove average bias in the pixel estimates and estimate Q for our experimental setup. This is critical for a variety of reasons:

- If the mean of the pixel measurements is biased, then the KF will not converge to the ground truth.
- If Q is an under-approximation to the actual covariance of random errors at the pixel level, then the KF will become inconsistent and will not converge to the ground truth, if it converges at all.
- If our choice of Q is too conservative or heuristic, it may not be informative enough to be useful in the decision process at the core of the controller.

We also want to test the system in relatively extreme conditions, particularly at long ranges (small disparities), where Freundlich et al. (2015) shows that triangulation error distributions are heavy tailed, biased away from zero, and highly asymmetric, which can exacerbate problems caused by calibration errors.

To address these challenges, we adopt a data-driven approach using linear regression in the pixel domain. Using a set of $n = 600$ pairs of training images for the robot at various ranges and viewing angles, we obtain a regression that maps raw pixel observations (x_L, x_R, y) to their best linear unbiased estimate (x_L^c, x_R^c, y^c), hereafter referred to as the *corrected* measurement. To acquire training data for the regression, we project the motion capture target locations, i.e., the ground truth, onto the camera image sensors, using the mapping in (20). This yields n individual output vectors Y_ℓ for $\ell = 1, \ldots, n$, which we stack into an $n \times 3$ matrix of outputs Y. We also use a color detector (the same detector that is used in the experiments) to obtain n raw pixel observations. We then compute five features and, because the data are not centered, include one constant, for each raw pixel tuple according to the model

$$X_\ell = \left[1,\, y_\ell,\, d_\ell,\, x_{L,\ell} + x_{R,\ell},\, yd_\ell,\, \frac{x_{L,\ell} + x_{R,\ell}}{d_\ell}\right], \quad (36)$$

where $d_\ell = x_{L,\ell} - x_{R,\ell}$. Stacking the X_ℓ into an $n \times 6$ matrix, we have a linear model $Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta}$ is a 6×3 matrix of coefficients and $\boldsymbol{\epsilon}$ is an $n \times 3$ matrix of errors. We refer to the raw pixels as *uncorrected*. The associated error vectors (computed with respect to the uncorrected pixels and the projected ground truth) $\boldsymbol{\epsilon}_\ell^{\text{uc}}$ for $\ell = 1, \ldots, n$ are plotted in the left panel of Figure 10. In the scatter plot it can be seen that the mean error is nonzero, contributing average bias to individual measurements. Also note the apparent skew of the error distribution in the vertical (y) direction.

Using the model with the feature vector described in (36) and applying the ordinary least squares estimator, the maximum likelihood estimate of the coefficient matrix is $\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top Y$. Using $\hat{\boldsymbol{\beta}}$, the residual covariance in the pixel measurements we obtained is

$$Q = \begin{bmatrix} 0.1297 & 0.1267 & -0.0882 \\ 0.1267 & 0.1355 & -0.0819 \\ -0.0882 & -0.0819 & 0.6988 \end{bmatrix}.$$

Note that the standard deviation of the y pixel value, corresponding to the variances in the lower right entry of the above matrix, corresponds to errors in the height of the ping pong ball center in vertical world-coordinates. The right panel of Fig. 10 shows the residual errors in the training set $\boldsymbol{\epsilon}_\ell$ for $\ell = 1, \ldots, n$ for the corrected vector $X\hat{\boldsymbol{\beta}}$.

To use the learned model online, new raw observations (x_L, x_R, y) are converted to corrected pixels (x_L^c, x_R^c, y^c) based on the associated new feature vector and $\hat{\boldsymbol{\beta}}$. Then, the

robot triangulates the relative location of the target via (1) using the corrected pixels, propagates Q via the Jacobian, rotates the covariances, and finally translates the estimates to global coordinates. Fig. 11 compares the projection of Kalman Filtered 95% confidence ellipses onto the X-Y (ground) plane using the raw/uncorrected and corrected pixel observations on data that was acquired during the experimental trials. To generate the plot in the top panel, which corresponds to the result if the raw pixels are used, we computed the empirical covariance of the raw residual errors $\epsilon_\ell^{\text{uc}}$ for $\ell = 1, \ldots, n$.

6.2 Results

We conducted sixty total static localization experiments – thirty using the supremum objective and thirty using the centroid objective. Figures 12 and 13 (a) show paths followed by the robot during the experimental trials using the setup shown in Figure 9.

Figures 12 and 13 (b) and show scatter plots of the errors from all thirty experiments using each control objective. Each point in these plots represents the Euclidean distance between filtered estimates and ground truth locations of the ping pong balls provided by the motion capture system. In each experiment, we collected fifteen images, and in each iteration three targets were observed. Accordingly, the plots have fifteen bands, each with thirty total points, representing the filtered error in a particular target for a particular experiment. The mean error for each target across all thirty experiments is drawn on the plot to guide the eye though the scatter plots.

Figures 12 and 13 (b) reveal the presence of outliers in the localization. Note from the figure that the KF still converges to ground truth. We can also see that the overall spread of the bands in the scatter plots is decreasing, reflecting that the control objective is indeed minimized. On average, the error in each target was reduced by about half, which is less of a reduction than what was observed in simulation. One reason for this, aside from the presence of unmodeled noise, is the fact that our lab has only about four square meters of usable area, so the diversity of viewpoints is not as rich as in the simulations.

Figures 12 and 13 (c) show the trace of the filtered error covariance for the same data that was used to plot Figures 12 and 13 (b). The points in the scatter plots reflect the posterior variance of each target for each simulation, and again the mean over the thirty experiments using each control objective is drawn on the plot to guide the eye.

6.3 Comparison to existing heuristic methods that use discrete pose space

We conducted static target localization experiments to compare the localization performance of our NBV method to a heuristic method that employs a discretization of the pose space, similar to the approaches discussed in (Dunn et al. 2009; Wenhardt et al. 2007a,b). Specifically, the heuristic we implement is based on discretizing the stereo rig's pose space, calculating the objective value in Problem 1 at all possible next poses and choosing the one having minimum objective value. This approach is in line with every stereo camera-based approach that we are aware of, in that they all

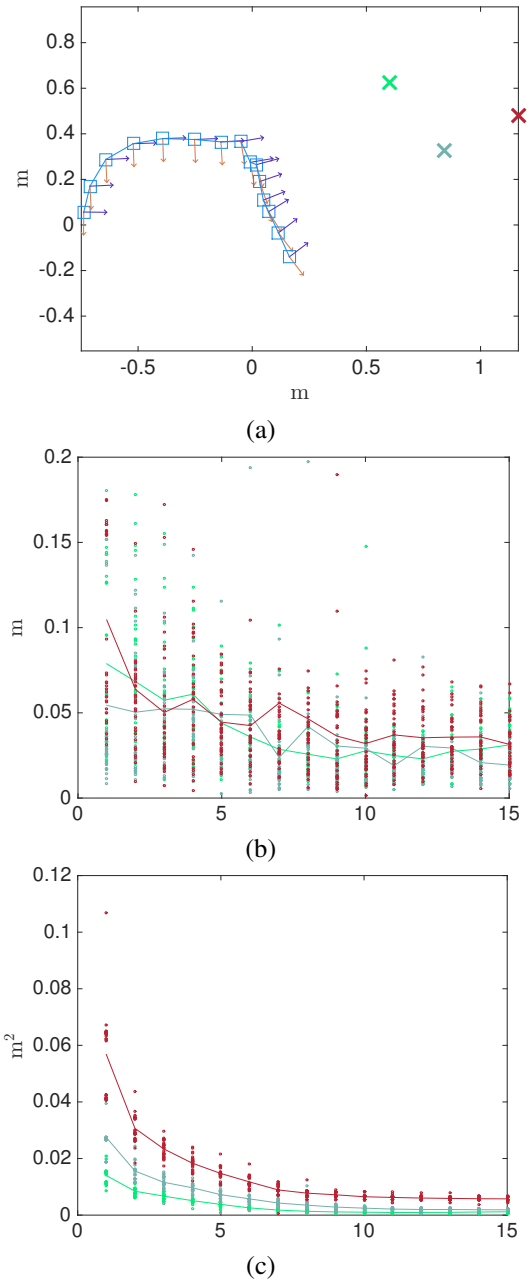

Figure 12. Plotting the trajectory of the robot along with the locations of the three static targets for one of the thirty experiments using the supremum objectives. The blue line represents the position and the □'s are the locations from where an image was taken. The orientation of the robot (projected onto the plane) is represented at each imaging location by a set of orthogonal axes. Scatter plots of the filtered error and the trace of the filtered error covariance in all targets for all thirty experiments are also shown. Each target is plotted using a unique color corresponding to the $\times$. In the scatter plots, colors correspond to (a), and a line is drawn to guide the eye through the means for each target in the experiment. The horizontal axes in these plots are the number of images taken.

(except Ponda and Frazzoli (2009), which we have discussed in the introduction) select from discrete next view sets. Note that we cannot fairly compare the localization accuracy of a single camera to a stereo rig (the rig always wins), and we can not compare a stereo rig to a LiDAR system (the LiDAR always wins, assuming that data associations can be established).

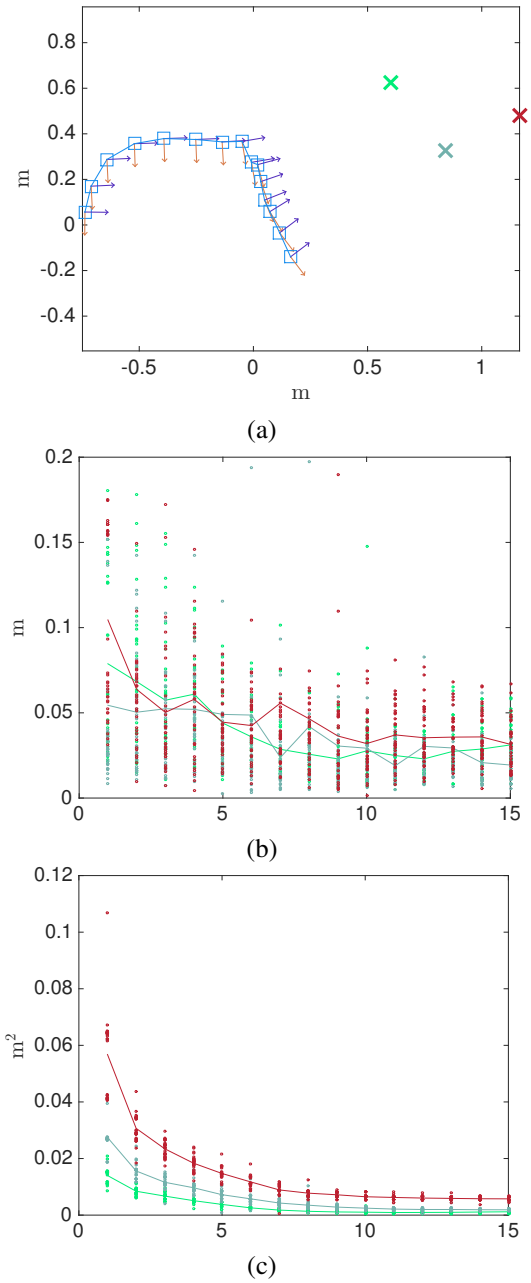

Figure 13. Identical plots to Fig. 12, however the data reported correspond to the centroid objective experiments.

In our experiments we focus on the supremum objective, so that the objective value calculated by the heuristic method is the trace of the filtered covariance matrix of the worst localized target. We tested two different ways of discretizing the pose space, namely, a square grid and a triangular grid, as shown in Figures 14 (a)-(b). In all experiments, the robot started from the same pose. Moreover, we required that both our NBV method and the heuristic travel approximately the same amount of distance and take the same number of images. In this way, different trajectories can be compared in terms of their ability to localize the targets. This requirement also specifies the edge length for the square and triangular grids. In this experiment, we set the total number of images that each method can take equal to ten and the edges of the square and the equilateral triangle cells were both set to be 0.25m. At each node of the grid, the stereo rig is oriented towards the estimated position of the worst localized target. This is also the behavior achieved by the NBV method with the supremum objective. We ran each method twenty times and below present our results.

Figure 14 (a)-(c) shows sample paths followed by the robot using the heuristic method for the two different grids and the NBV approach, respectively, for one of the twenty trials. To take ten images, the heuristic method that uses the square grid travels an average distance of 2.3058m, the heuristic method that uses the triangular grid travels 2.2198m on average, and our proposed method travels 2.1949m on average. We note that the heuristic method is highly sensitive to the grid size and error covariance matrices during the measuring process, for both types of grids. Specifically, during twenty trials of experiments with grid size of 0.25m, the heuristic generated trajectories that contain small cycles (back and forth motion among a few cells) for both grid types; see, e.g., Figure 14 (a) and (d). In fact, we were unable to find a grid size that does not generate such motion artifacts for the square grid heuristic; in every single trial and for every grid size we observed a behavior similar to the one shown in Figure 14 (a). On the other hand, after much trial and error we found that, for the particular set up of targets in our lab, a 0.25m grid size can produce reasonable trajectories for the triangular grid heuristic, as shown in Figure 14 (b). Nevertheless, this behavior was not consistent, as seen in Figure 14 (d) for the same grid size. Our continuous space NBV controller, shown in Figure 14 (c), selects the next pose in a continuous pose space and automatically balances the strategies between varying viewing angle and approaching targets. Figures 15 (a)-(b) demonstrate the localization performance of the NBV method compared to the heuristic method. Figure 15 (a) shows the filtered localization error during each one of the ten iterations (after each image was taken), averaged over the twenty trials. The localization error of the heuristic method for the square grid (magenta line) eventually diverges due to measurement bias that causes the KF to diverge. This is the result of observing the targets from the same position. A similar behavior was also observed for the straight baseline in the simulations; see Section 5.3. When the grid edge length is chosen as 0.25m, the heuristic method for the triangular grid (green line) achieves similar localization error as our NBV method (red line). Figure 15 (b) shows the trace of filtered error covariance for the heuristic and the NBV method, averaged over the twenty trials. In this case, the NBV method outperforms the heuristic for both grid types. While the heuristic confined to the square performs extremely poorly because it does not approach the targets, the heuristic method on the triangular grid, while slightly better, still does not perform as well as the proposed continuous space method. Finally, note that the grid size of 0.25m was selected after laborious tuning to remove such artifacts, suggesting that our continuous method will perform better in general situations than the discrete pose space alternatives.

7 Conclusions

In this paper, we addressed the multi-target, single-sensor problem employing the most realistic sensor model in the literature. Our approach relies on a novel control decomposition in the relative camera frame and the global

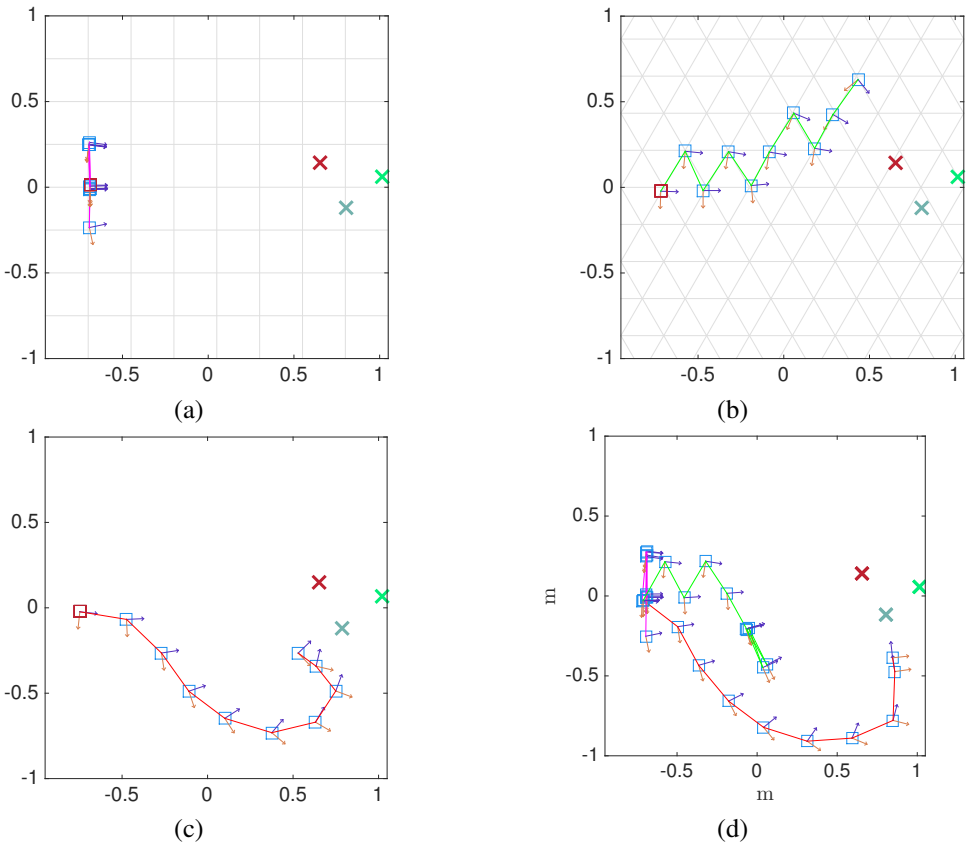

Figure 14. Exaples of trajectories generated from heuristic comparisons on a 0.25m grid size to our method. From the top: the heuristic method for the square grid (a), triangular grid (b), and the proposed supremum objective (c). Red squares are the initial poses of the stereo rig during the experiments. Selected trajectories of the heuristic and the proposed method that contain interesting motion artifacts are shown in (d).

frame. In the relative frame, we modeled quantization noise and did not operate under a Gaussian noise assumption at the pixel level (as range/bearing models assume). Our approach avoids setting covariance by deriving Σ from the uniform distribution. This allows us to obtain the Next Best View from where the targets can be observed in order to minimize their localization uncertainty. We obtain this NBV using gradient descent on appropriately defined potentials, without sampling the pose space or having to select from a set of previously recorded image pairs. Compared to previous gradient-based approaches, our integrated hybrid system is more precise since it derives Gaussian parameters from the quantization noise in the images. Furthermore, our approach does not assume omnidirectional sensors, but instead imposes field of view constraints.

8 Funding

This work is supported in part by the National Science Foundation under awards No. CNS-1302284, IIS-1217797, and IIS-1637761

References

Adurthi N et al. (2013) Optimal information collection for nonlinear systems- an application to multiple target tracking and localization. In: *IEEE American Control Conf. (ACC)*. pp. 3864–3869.

Bajcsy R (1988) Active perception. *Proceedings of the IEEE* 76(8): 966–1005.

Bishop G and Welch G (2001) An introduction to the kalman filter. *Proc. of SIGGRAPH, Course* 8(27599-23175): 41.

Blostein SD and Huang TS (1987) Error analysis in stereo determination of 3-d point positions. *IEEE Trans. on Pattern Analysis and Machine Intell.* 9(6): 752–766.

Bouguet J (2004) Camera calibration toolbox for matlab (software). *California Institute of Technology, Pasadena, http://www. vision. caltech. edu/bouguetj/calib_doc* .

Chang CC, Chatterjee S and Kube PR (1994) A quantization error analysis for convergent stereo. In: *Int. Conf. on Image Proc. (ICIP)*. Austin, Texas: IEEE, pp. II: 735–739.

Chung TH, Burdick JW and Murray RM (2006) A decentralized motion coordination strategy for dynamic target tracking. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Orlando, FL: IEEE, pp. 2416 –2422.

Ding C, Morye A, Farrell J and Roy-Chowdhury A (2012) Coordinated sensing and tracking for mobile camera platforms. In: *IEEE American Control Conf. (ACC)*. Montreal, Canada: IEEE, pp. 5114–5119.

Dunn E, Van Den Berg J and Frahm JM (2009) Developing visual sensing strategies through next best view planning. In: *IEEE Int. Conf. on Intell. Robots and Systems (IROS)*. St. Louis, USA: IEEE, pp. 4001 –4008. DOI:10.1109/IROS.2009.5354179.

Förstner W (2005) Uncertainty and projective geometry. In: Bayro-Corrochano E (ed.) *Handbook of Geometric Computing*.

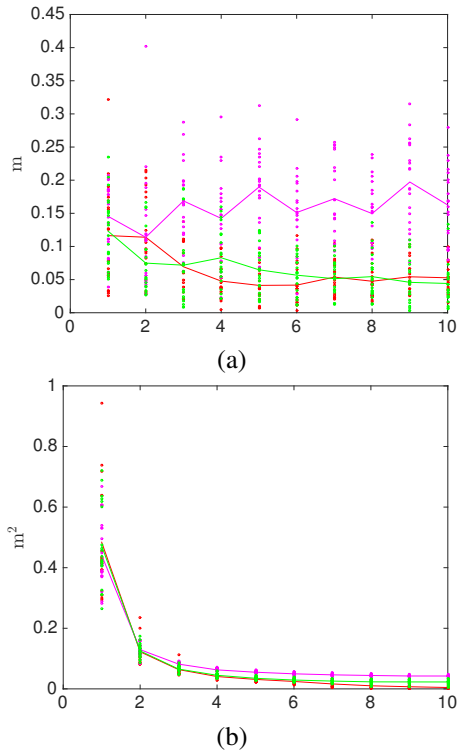

Figure 15. Filtered localization error (a) and the trace of the error covariance (b) of the green target from Fig. 14, averaged over twenty trials for each one of the three methods. Red corresponds to the proposed NBV method, magenta to the heuristic method for the square grid, and green to the triangular grid heuristic.

Springer, pp. 493–534.

Fox D, Burgard W, Kruppa H and Thrun S (2000) A probabilistic approach to collaborative multi-robot localization. *Autonomous Robots* 8(3): 325–344.

Freundlich C, Mordohai P and Zavlanos MM (2013a) A hybrid control approach to the next-best-view problem using stereo vision. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Karlsruhe, DE, pp. 4478–4483. DOI:10.1109/MRA.2012.2201600.

Freundlich C, Mordohai P and Zavlanos MM (2013b) Hybrid control for mobile target localization with stereo vision. In: *IEEE Conf. on Decision and Control (CDC)*. Firenze, Italy, pp. 2635–2640.

Freundlich C, Mordohai P and Zavlanos MM (2015) Exact bias correction and covariance estimation for stereo vision. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, pp. 3296–3304.

Frew EW (2003) *Trajectory design for target motion estimation using monocular vision*. PhD Thesis, Ph. D. dissertation, Stanford University, Stanford, CA.

Galceran E and Carreras M (2013) A survey on coverage path planning for robotics. *Robotics and Autonomous Systems* 61(12): 1258–1276.

Hollinger GA, Englot B, Hover FS, Mitra U and Sukhatme GS (2013) Active planning for underwater inspection and the benefit of adaptivity. *Int. J. Robotics Research* 32(1): 3–18. DOI:10.1177/0278364912467485.

Le Cadre JP and Gauvrit H (1996) Optimization of the observer motion for bearings-only target motion analysis. In: *Data Fusion Symposium, 1996. ADFS'96., First Australian*. IEEE, pp. 190–195.

Logothetis A, Isaksson A and Evans RJ (1998) Comparison of suboptimal strategies for optimal own-ship maneuvers in bearings-only tracking. In: *American Control Conference, 1998. Proceedings of the 1998*, volume 6. Philadelphia, PA: IEEE, pp. 3334–3338.

Logothetis A et al. (1997) An information theoretic approach to observer path design for bearings-only tracking. In: *IEEE Conf. on Decision and Control (CDC)*, volume 4. San Diego, CA, pp. 3132–3137.

Matthies LH and Shafer SA (1987) Error modelling in stereo navigation. *IEEE Journal of Robotics and Automation* 3(3): 239–250.

Morbidi F and Mariottini GL (2013) Active target tracking and cooperative localization for teams of aerial vehicles. *IEEE Transactions on Control Systems Technology* 21(5): 1694–1707.

Olfati-Saber R (2007) Distributed tracking for mobile sensor networks with information-driven mobility. In: *IEEE American Control Conf. (ACC)*. New York, NY, pp. 4606–4612.

Papadopoulos G, Kurniawati H and Patrikalakis NM (2013) Asymptotically optimal inspection planning using systems with differential constraints. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Karlsruhe, DE: IEEE, pp. 4126–4133.

Passerieux JM and Van Cappel D (1998) Optimal observer maneuver for bearings-only tracking. *IEEE Trans. on Aerospace and Elec. Sys.* 34(3): 777–788.

Ponda S and Frazzoli E (2009) Trajectory optimization for target localization using small unmanned aerial vehicles. In: *AIAA Conf. on Guidance, Navigation, and Control*. Chicago, IL: IEEE.

Rong Li X and Jilkov V (2003) Survey of maneuvering target tracking. part I. dynamic models. *IEEE Trans. on Aerospace and Elec. Sys.* 39(4): 1333 – 1364. DOI:10.1109/TAES.2003.1261132.

Roumeliotis S and Bekey G (2002) Distributed multirobot localization. *IEEE Trans. on Robotics and Automation* 18(5): 781 – 795.

Scharstein D and Szeliski R (2002) A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *Int. J. Computer Vision* 47(1-3): 7–42.

Shade R and Newman P (2010) Discovering and mapping complete surfaces with stereo. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Anchorage, AK: IEEE, pp. 3910–3915.

Singer R (1970) Estimating optimal tracking filter performance for manned maneuvering targets. *IEEE Trans. on Aerospace and Elec. Sys.* AES-6(4): 473 –483. DOI:10.1109/TAES.1970.310128.

Singh SS, Kantas N, Vo BN, Doucet A and Evans RJ (2007) Simulation-based optimal sensor scheduling with application to observer trajectory planning. *Automatica* 43(5): 817–830.

Spletzer JR and Taylor CJ (2003) Dynamic sensor planning and control for optimally tracking targets. *Int. J. Robotics Research* 22(1): 7–20.

Stroupe AW and Balch T (2005) Value-based action selection for observation with robot teams using probabilistic techniques. *Robotics and Autonomous Systems* 50(2-3): 85 – 97.

Trummer M et al. (2010) Online next-best-view planning for accuracy optimization using an extended e-criterion. In: *Int. Conf. on Pattern Recognition*. Istanbul, Turkey: IEEE, pp. 1642–1645.

Wang P, Krishnamurti R and Gupta K (2007) View planning problem with combined view and traveling cost. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Roma, Italy: IEEE, pp. 711–716.

Wenhardt S, Deutsch B, Angelopoulou E and Niemann H (2007a) Active visual object reconstruction using d-, e-, and t-optimal next best views. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. Minneapolis, MN.

Wenhardt S et al. (2007b) An information theoretic approach for next best view planning in 3-d reconstruction. In: *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1. pp. 103 –106. DOI:10.1109/ICPR.2006.253.

Yang P, Freeman R and Lynch K (2007) Distributed cooperative active sensing using consensus filters. In: *IEEE Int. Conf. on Robotics and Automation (ICRA)*. Roma, Italy: IEEE, pp. 405 –410.

Zavlanos MM and Pappas GJ (2008) A dynamical systems approach to weighted graph matching. *Automatica* 44(11): 2817–2824.

Zhou K and Roumeliotis S (2011) Multirobot active target tracking with combinations of relative observations. *IEEE Trans. on Robotics* 27(4): 678 –695.