A Hybrid Control Approach to the Next-Best-View Problem using Stereo Vision

Charles Freundlich, Philippos Mordohai and Michael M. Zavlanos

Abstract— **In this paper, we consider the problem of precisely localizing a group of stationary targets using a single stereo camera mounted on a mobile robot. In particular, assuming that at least one pair of stereo images of the targets is available, we seek to determine where to move the stereo camera so that the localization uncertainty of the targets is minimized. We call this problem the Next-Best-View problem. The advantage of using a stereo camera is that, using triangulation, the two simultaneous images can yield range and bearing measurements of the targets, as well as their uncertainty. We use a Kalman filter to fuse location and uncertainty estimates as more measurements are acquired. Our solution to the Next-Best-View problem is to iteratively minimize the fused uncertainty of the targets' locations subject to field-of-view constraints. We capture these objectives by appropriate artificial potentials on the camera's relative frame and the global frame, respectively. In particular, with every new observation, the mobile stereo camera computes the new next best view on the relative frame and subsequently realizes this view in the global frame via gradient descent on the space of robot positions and orientations, until a new observation is made. Integration of next best view with motion planning results in a hybrid system, which we illustrate in computer simulations.**

I. Introduction

The increasing capabilities of mobile robots illuminate the need for robotic systems that are able to operate outside the controlled infrastructure of lab environments. Such environments, equipped with e.g., Vicon systems, provide robots with continuous and precise position and orientation information [1]. This information is not available outside the lab, where the robots should be able to self localize. In such settings, allowing one or several sensors to be mobile has been shown to be advantageous in terms of its effect on localization accuracy [2], [3].

The novelty of this work lies in the use of stereo vision for target localization. We consider a single robot equipped with a stereo camera overlooking a group of stationary targets. The advantage of stereo vision, compared to the use of monocular camera systems, is that it provides both depth and bearing measurements of a target from a single pair of simultaneous images. Differentiation of these measurements provides an estimate for the uncertainty of the target's location, which has been shown to grow quadratically with depth [4]–[6]. As such, we leverage the inherent uncertainty of stereo vision to define the Next-Best-View (NBV) as the position and orientation of a stereo camera that, given a sequence of observations of a collection of targets, minimizes their localization uncertainty.

The NBV problem has often been formulated as the selection of the next image from a given finite set using sampling or grid-based methods [7]–[13]. While these methods can apply multi-view constraints and obtain uncertainty estimates that depend on factors such as viewing distance and camera resolution, they can only select among the input set of images. Furthermore, they have to satisfy constraints related to maintaining consistency between images [9], require *a priori* models of the environment [10], or use heuristic estimates of the covariance [10], [11]. Our approach guides the sensor to the NBV based on gradient descent of an analytical representation of the uncertainty.

Approaches capable of computing the next best viewing position have been proposed in the cooperative localization literature [2], [14]–[19]. They typically employ abstract sensor models and approximations of the uncertainty in the range and bearing measurements, which are often treated independently with respect to range and to each other. On the other hand, assuming noise is dominated by quantization of pixel coordinates and propagating uncertainty from pixel to target coordinates, we obtain more accurate estimates of the structure of the covariance matrix, which captures uncertainty. This is true for both the instantaneous uncertainty of one measurement and for filtered uncertainty of the full sequence of measurements. As a result, our objective function is a tighter approximation of the true uncertainty. Gradient descent of this objective guides the robot to more effective viewing positions, e.g., the NBV. To realize the NBV in the global coordinate frame, we use gradient descent in the space of robot positions and orientations. While respecting field of view constraints, the robot moves until a next observation is made, which, in turn, determines a new next best viewing position to be realized. Integration of NBV with continuous motion planning gives rise to a hybrid system that drives a robot in the direction that minimizes localization uncertainty of the targets.

The paper is organized as follows. Section II outlines the system model, our assumptions, and the Kalman filter (KF), which fuses the observation sequence in real time. Section III determines the NBV in the camera coordinate system. Section IV realizes the NBV in the global coordinate frame. Sections V and VI show simulations of our approach and conclude the paper.

This work is supported in part by the National Science Foundation under Grant No. DGE-0742462 and IIS-1217797

Charles Freundlich is with the Dept. of Mechanical Engineering, Stevens Institute of Technology, Hoboken, NJ 07030, USA cfreundl@stevens.edu. Philippos Mordohai is with the Dept. of Computer Science, Stevens Institute of Technology, Hoboken, NJ 07030, USA mordohai@stevens.edu. Michael M. Zavlanos is with the Dept. of Mechanical Engineering and Materials Science, Duke University, Durham, NC 27708, USA michael.zavlanos@duke.edu.

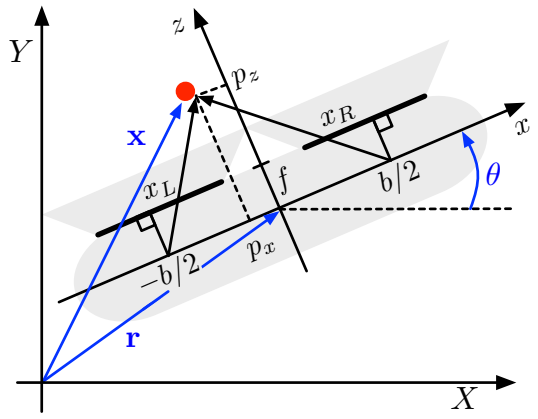

Fig. 1. A diagram of a single target (red dot) in both relative (black) and global (blue) coordinate frames. The camera is enlarged for clarity.

II. System Model & Problem Formulation

Consider a group of N point targets, indexed by $i \in I = \{1 \dots N\}$. We assume that the targets are fixed and we denote their, initially unknown, global coordinates by $\mathbf{x}_i \in \mathbb{R}^2$. Consider also a mobile, stereo, camera and let $\mathbf{r}(t) \in \mathbb{R}^2$ denote its position on the plane and $R(t) \in SO(2)$ its orientation at time $t \geq 0$, where $SO(2)$ the special orthogonal group of dimension 2. Since the camera an targets lie in a plane, the y coordinates are omitted, and the camera only measures x and z. The stereo camera consists of a pair of two monocular cameras, referred to as the left (L) and right (R) cameras, located at coordinates $-b/2$ and $b/2$, respectively, with respect to the origin of the binocular camera system, where b denotes the baseline measured in meters; see Fig. 1. Let x_{Li} and x_{Ri} denote the x-axis coordinates of target i, measured in pixels, on the left and right camera images, respectively. Then, the position of target i with respect to the relative, camera, frame is

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

where f denotes the focal length of the camera lens, measured in pixels. Note that x_{Li} and x_{Ri} are measured in pixels and can only take integer values. Since the actual coordinates of target i on the two images can be anywhere within these pixels, we may assume that they are uniformly distributed around the pixel centers. We denote the pixel centers by $\hat{x}_{Li}$ and $\hat{x}_{Ri}$, which now take values in $\mathbb{Z}$. In view of (1), the above pixelation errors on the images work their way in the coordinates $\mathbf{p}_i$ of target i in space causing non-Gaussian error distributions [4], [6]. For convenience, we follow [5], [20] and approximate the uniform pixelation errors as Gaussian to allow uncertainty propagation from image to world coordinates. Under this assumption, the localization error of the target in the relative camera frame will also be Gaussian with mean $\hat{\mathbf{p}}_i = \mathbf{p}(\hat{x}_{Li}, \hat{x}_{Ri})$ and covariance $U_i \in \mathbb{S}^2_+$ in global coordinates, where $\mathbb{S}^2_+$ denotes the set of 2×2 symmetric positive definite matrices.

Now, assume that the stereo camera has made a sequence of observations of the targets and introduce an index $k \geq 0$ associated with every observation. Moreover, let $\hat{\mathbf{x}}_{i,k} \in \mathbb{R}^2$ denote the mean and $U_{i,k} \in \mathbb{S}^2_+$ the covariance of the location of target i on the global frame associated with observation k. Similar to [21]–[23], these observations can be fused using a linear Kalman filter (KF) to significantly increase localization accuracy of the targets. In particular, let $\hat{\mathbf{x}}_{i,k} = \mathbf{x}_i + \mathbf{v}_{i,k}$ denote the measured, noisy, coordinates of target i in the global frame, where $\mathbf{x}_i$ are the actual coordinates of target i (which do not change with k for fixed targets) and $\mathbf{v}_{i,k} \in \mathbb{R}^2$ is the realization of zero-mean Gaussian noise with instantaneous covariance matrix $U_{i,k}$. Then, $\hat{\mathbf{x}}_{i,k}$ can be related to the target coordinates $\hat{\mathbf{p}}_{i,k}$ in the relative camera frame as

$$\begin{bmatrix} \hat{\mathbf{x}}_{i,k} \\ 1 \end{bmatrix} = \begin{bmatrix} R(t_k) & \mathbf{r}(t_k) \\ \mathbf{0_{1\times 2}} & 1 \end{bmatrix} \begin{bmatrix} \hat{\mathbf{p}}_{i,k} \\ 1 \end{bmatrix}. \tag{2}$$

At $k = 0$, $\mathbf{x}_i$ is unknown, and we take a measurement $\hat{\mathbf{x}}_{i,0}$ and calculate its covariance, $U_{i,0}$ by methods discussed in section III. Because the targets are assumed to be fixed, we predict that at $k = 1$, these values will not change. In other words, based on the events at $k = 0$, we predict that at $k = 1$, $\hat{\mathbf{x}}_{i,1|0} = \hat{\mathbf{x}}_{i,0}$ and $U_{i,1|0} = U_{i,0}$. If at $k = 1$ a new observation is made, then new instantaneous measurements $\hat{\mathbf{x}}_{i,1}$ and $U_{i,1}$ are obtained that do not depend on any prior event. The purpose of the KF is to fuse the new measurement with the history of measurements to create an estimate $\hat{\mathbf{x}}_{i,1|1}$ and a covariance $U_{i,1|1}$. All covariance matrices are defined in the global coordinate system in this paper.

In general, at time $k+1$, we have access to (i) the stand alone measurements $\hat{\mathbf{x}}_{i,k+1}$ and $U_{i,k+1}$ and (ii) the prediction $\hat{\mathbf{x}}_{i,k+1|k}$ and its predicted covariance $U_{i,k+1|k}$, which are based on the entire measurement history. Let

$$e_{i,k+1} = \hat{\mathbf{x}}_{i,k+1} - \hat{\mathbf{x}}_{i,k+1|k} \tag{3}$$

be the discrepancy between the prediction and the measurement or the location of target i at time $k+1$. Also let the innovation covariance matrix for that target be

$$S_i = U_{i,k+1|k} + U_{i,k+1}. \tag{4}$$

We fuse the prior and current measurements according to the linear Kalman Filter equation

$$\hat{\mathbf{x}}_{i,k+1|k+1} = \hat{\mathbf{x}}_{i,k+1|k} + W_i e_{i,k+1}, \tag{5}$$

$$U_{i,k+1|k+1} = U_{i,k+1|k} - W_i S_i W_i^T, \tag{6}$$

where the gain matrix is $W_i = U_{i,k+1|k} S_i^{-1}$. Using equation (6) we obtain a closed form expression for the fused covariance $U_{i,k+1|k+1}$. In particular, we have the following lemma, which follows from [24] using the simple form of S_i and W_i as defined above.

Lemma 2.1: Let $U_{i,k+1|k}$ denote the fused covariance of all prior measurements and $U_{i,k+1}$ denote the covariance of the most recent measurement. Then, the updated covariance $U_{i,k+1|k+1}$ obtained by the linear KF is given by $U_{i,k+1|k+1} = (U^{-1}_{i,k+1|k} + U^{-1}_{i,k+1})^{-1}$.

We can now state the problem that we address in this paper. For this, let $U_{s,k+1|k} = U_{i,k+1|k}$ with $i =$

$\text{argmax}_{j\in I}\left\{\mathbf{tr}\left[U_{j,k+1|k}\right]\right\}$ denote the covariance of the worst localized target up to observation k, and $U_{c,k+1|k} = \frac{1}{N}\sum_{i\in I} U_{i,k+1|k}$ denote the average of all target covariances up to observation k. Then we have:

Problem 1 (Next Best View): Given the covariance of the worst localized target $U_{s,k+1|k}$ (respectively, the average of the targets' covariances $U_{c,k+1|k}$), determine $U_{s,k+1}$ (respectively, $U_{c,k+1}$) so that $\mathbf{tr}[U_{s,k+1|k+1}]$ (respectively, $\mathbf{tr}[U_{c,k+1|k+1}]$) is minimized.

In problem 1, we have chosen the trace as a measure of total uncertainty among other choices, such as the determinant or the maximum eigenvalue. It is shown in [25] that all such criteria behave similarly in practice. Since minimization of $\mathbf{tr}[U_{s,k+1|k+1}]$ is associated with improving localization of the worst localized target, we call it the *supremum objective*. Accordingly, we call minimization of $\mathbf{tr}[U_{c,k+1|k+1}]$ the *centroid objective*. Clearly, $U_{s,k+1|k+1}$ depends only on the position of the worst localized target, which we denote by $\mathbf{p}_{s,k+1}$, but $U_{c,k+1|k+1}$ depends on the positions $\mathbf{p}_{i,k+1}$ of all targets. Attempting to find a $U_{c,k+1}$ that solves Problem 1 by controlling the relative coordinates $\mathbf{p}_i$ of all targets simultaneously requires a nonconvex constraint to maintain consistency between images. Instead, we can think of the array of fixed targets as a mobile rigid body in the relative coordinates and place a virtual target at the centroid, $\mathbf{p}_{c,k+1|k} = \frac{1}{N}\sum_{i\in I}\mathbf{p}_{i,k+1|k}$. The centroid serves as a proxy for all targets.

As we discuss in the following sections, instantaneous covariances depend on $\mathbf{p}_{s,k+1|k}$ or $\mathbf{p}_{c,k+1|k}$, which in turn are functions of the stereo camera position $\mathbf{r}$ and orientation R. Since expressing the covariances directly in terms of the camera translation and rotation results in highly nonlinear expressions that are difficult to control, we propose an alternative approach. In particular, we decompose optimization of the above objectives in the relative camera frame and the global frame. During the former stage, we find the vector that solves Problem 1. We denote this vector by $\mathbf{p}_{o,k+1}$, where o stands for s or c depending on the objective for which we solve [cf. the *supremum* or the *centroid objective*].. This vector is then realized by an appropriate rotation and translation of the camera in the global space. Integration of the two stages results in a hybrid control scheme, where the next best views obtained by every new observation correspond to the switching signal in the continuous motion of the camera.

III. Controlling the Relative Frame

Assume that k observations are already available and let t_k denote the time instant corresponding to the k-th observation. Our goal in this section is to determine the next best target locations $\mathbf{p}_{s,k+1}$ or $\mathbf{p}_{c,k+1}$ on the relative camera frame so that if a new observation is made at time t_{k+1} with the targets at these new relative locations, it will optimize the fused localization uncertainty. In particular, let $Q = \text{cov}([\hat{x}_{Li} \quad \hat{x}_{Ri}]) \approx \text{diag}\left[\sigma_L^2 \quad \sigma_R^2\right]$ denote the approximate covariance of the target coordinates x_{Li} and x_{Ri} on the left and right image frames, respectively, where σ_L^2 and σ_R^2 denote the associated variances.[1] Let also J_i be the Jacobian of $\mathbf{p}_i \triangleq \mathbf{p}(x_{Li}, x_{Ri})$ evaluated at the point $(\hat{x}_{Li}, \hat{x}_{Ri})$. Then, the first order (linear) approximation of $\mathbf{p}_i = \mathbf{p}(x_{Li}, x_{Ri})$ about the point $(\hat{x}_{Li}, \hat{x}_{Ri})$ is $\mathbf{p}(x_{Li}, x_{Ri}) \approx \mathbf{p}(\hat{x}_{Li}, \hat{x}_{Ri}) + J_i[x_{Li} \quad x_{Ri}]^T$. Since $\mathbf{p}_i(\hat{x}_{Li}, \hat{x}_{Ri})$ corresponds to the current mean estimate of target coordinates, the covariance of $\mathbf{p}_i$ in the relative camera frame is nothing but $J_i Q J_i^T$. This is the standard way to propagate error from one set of variables to another when there is linear dependence, e.g., when the first set of variables can be written as a linear combination of the second. However, fusing covariance matrices as in Lemma 2.1 requires that they are represented in the global frame. To represent the covariance $J_i Q J_i^T$ 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 $[t_k, t_{k+1}]$, we may approximate the camera's rotation $R(t)$ at time $t \in [t_k, t_{k+1}]$ by its initial rotation $R(t_k)$ at time t_k. Then the covariance of $\mathbf{p}_i$, rotated to global coordinates, at any time instant $t \in [t_k, t_{k+1}]$, can be approximated by

$$U_i = \text{cov}[\mathbf{p}(\hat{x}_{Li}, \hat{x}_{Ri})] \approx R(t_k) J_i Q J_i^T R^T(t_k), \qquad (7)$$

To obtain the next best estimate of the targets' locations on the relative camera frame that optimizes localization uncertainty, we define the uncertainty potential $h : \mathbb{R}^2 \to \mathbb{R}_+$ such that

$$\begin{aligned} h(\mathbf{p}_{o,k+1}) &= \mathbf{tr}\left[U_{o,k+1|k+1}\right] \\ &= \mathbf{tr}\left[\left(U_{o,k+1|k}^{-1} + U_o^{-1}\right)^{-1}\right], \end{aligned} \qquad (8)$$

where we have used the result in Lemma 2.1. Then the target coordinates $\mathbf{p}_{o,k+1}$ that locally minimize (8) can be determined by

$$\mathbf{p}_{o,k+1} = \mathbf{p}_{o,k} - \int_0^T \frac{\partial h(\mathbf{p}_{o,k+1}(\tau))}{\partial \mathbf{p}_{o,k+1}} d\tau. \qquad (9)$$

The length $T > 0$ of the integration interval is chosen sufficiently small so that our assumption that $R(t_k)$ remains approximately constant during the update holds. The following result provides an analytical expression for the gradient of the potential h in (9).

Proposition 3.1: The gradient of h with respect to $\mathbf{p}_o$ is given by

$$\frac{\partial h}{\partial [\mathbf{p}_o]_v} = \mathbf{tr}\left[U_o^{-1}\left(U_{o,k+1|k}^{-1} + U_o^{-1}\right)^{-2} U_o^{-1} \frac{\partial U_o}{\partial [\mathbf{p}_o]_v}\right], \qquad (10)$$

where v can be either x or z, depending on which coordinate of $\mathbf{p}_o$ we differentiate.

The proof of Proposition 3.1 depends on the following Lemmas, which we present without proof due to space limitations.

Lemma 3.2: Let M be a nonsingular matrix and $f(M) = \mathbf{tr} M^{-1}$. Then, $\nabla_M f(M) = -M^{-2}$.

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

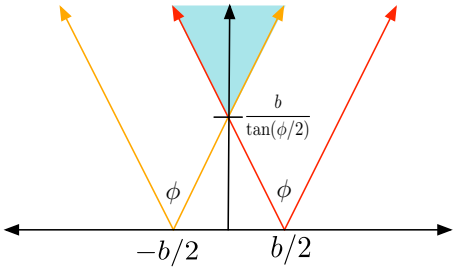

Fig. 2. The field of view for a 2D stereo camera.

Lemma 3.3: Let $C(x)$ be a nonsingular matrix and let x be a scalar. Then, $\frac{\partial C^{-1}(x)}{\partial x} = -C^{-1}(x)\frac{\partial C(x)}{\partial x}C^{-1}(x)$.

Applying Lemmas 3.2 and 3.3 to (8) gives (10), which completes the proof of Proposition 3.1. The term $\partial U_o/\partial[\mathbf{p}_o]_v$ in (10) can be found from (7) and (1) using elementary calculus.

IV. Controlling the Global Frame

The update in (9) provides the relative target coordinates $\mathbf{p}_{o,k+1}$ on the camera frame from where, if a new observation $k+1$ is taken, the localization uncertainty associated with objective "o" is minimized. Our goal in this section is to drive the camera to a new position $\mathbf{r}$ and orientation R in space that realizes the next best view $\mathbf{p}_{o,k+1}$. For this, let

$$\psi(\mathbf{r}, R) = \|R\mathbf{p}_o - \hat{\mathbf{x}}_o + \mathbf{r}\|_F^2, \tag{11}$$

denote a positive semidefinite function that becomes zero only if the next best view is realized in the global frame, where $\|\cdot\|_F$ is the Frobenius norm and we have dropped dependence of $\hat{\mathbf{x}}_{o,k+1}$ and $\mathbf{p}_{o,k+1}$ on the observation index to simplify notation. To capture field of view constraints, let

$$\mathbf{g}_i(\mathbf{r}, R) = \alpha^2\left(\mathbf{e}_2^T R^T(\hat{\mathbf{x}}_i - \mathbf{r}) - c\right)^2 - \left(\mathbf{e}_1^T R^T(\hat{\mathbf{x}}_i - \mathbf{r})\right)^2, \tag{12}$$

where ϕ is the field of view of the camera, $\alpha = \tan(\phi/2)$, $c = b/(2\alpha)$, $\mathbf{e}_1 = [1\ 0]^T$, and $\mathbf{e}_2 = [0\ 1]^T$. $\mathbf{g}_i(\mathbf{r}, R) > 0$ if and only if the position $\hat{\mathbf{x}}_i$ of target i lies in the camera's field-of-view; see Fig. 2. Then, to realize the next best view $\mathbf{p}_o$ while maintaining all targets in the camera's field of view we equivalently need to minimize ψ while ensuring $\mathbf{g}_i > 0$ for all targets $i \in I$. For this, we combine (11) and (12) in the artificial potential function $\hat{\psi} : \mathbb{R}^2 \times SO(2) \to \mathbb{R}_+$ with

$$\hat{\psi}(\mathbf{r}, R) = \psi(\mathbf{r}, R) + \rho \sum_{i \in I} \frac{1}{\mathbf{g}_i(\mathbf{r}, R)}, \tag{13}$$

where $\rho > 0$ is a penalty parameter. The terms $1/\mathbf{g}_i$ in (13) serve as barrier potentials, since $\hat{\psi} \to \infty$ whenever there exists a target $i \in I$ for which $\mathbf{g}_i \to 0$.

To minimize the potential $\hat{\psi}$, let $t_k > 0$ denote the time instant associated with observation k and for all time $t \in [t_k, t_{k+1}]$ we define the gradient flow

$$\dot{\mathbf{r}} = -\nabla_{\mathbf{r}}\hat{\psi}(\mathbf{r}, R), \tag{14a}$$

$$\dot{R} = R\nabla_R\hat{\psi}(\mathbf{r}, R), \tag{14b}$$

on the joint space of camera positions $\mathbb{R}^2$ and orientations $SO(2)$. It is well known that if $R(t_k) \in SO(2)$, the gradient $\nabla_R\hat{\psi}(\mathbf{r}, R)$ is a skew-symmetric matrix, and $R(t)$ evolves as in (14b), then $R(t) \in SO(2)$ for all time $t \in [t_k, t_{k+1}]$; see, e.g., [26].

The benefit of the gradient flow (14b) is that it implicitly ensures the nonconvex constraint that $R(t)$ must be a rotation matrix during the minimization of $\hat{\psi}$. In the remainder of this section we provide analytic expressions for the gradients in (14) and show that the closed loop system minimizes $\hat{\psi}$. In particular, we have the following results.

Lemma 4.1: The negative gradient of ψ with respect to R is given by the skew-symmetric matrix

$$\nabla_R\psi(\mathbf{r}, R) = \mathbf{p}_o(\mathbf{r} - \hat{\mathbf{x}}_o)^T R - R^T(\mathbf{r} - \hat{\mathbf{x}}_o)\mathbf{p}_o^T. \tag{15}$$

Proof: Using the first order approximation of the neighborhood of the rotation matrix R, $R(\Omega) \approx R(I + \Omega)$, where Ω is skew-symmetric, we have that

$$\begin{aligned}
&\psi(\mathbf{r}, R(I+\Omega)) = \\
&= \mathbf{tr}\Big[\big(R(I+\Omega)\mathbf{p}_o + \mathbf{r} - \hat{\mathbf{x}}_o\big)\big(R(I+\Omega)\mathbf{p}_o + \mathbf{r} - \hat{\mathbf{x}}_o\big)^T\Big] \\
&\approx \psi(\mathbf{r}, R) + \mathbf{tr}\Big[(R\mathbf{p}_o + \mathbf{r} - \hat{\mathbf{x}}_o)(R\Omega\mathbf{p}_o)^T \\
&\qquad\qquad + (R\Omega\mathbf{p}_o)(R\mathbf{p}_o + \mathbf{r} - \hat{\mathbf{x}}_o)^T\Big] \\
&= \psi(\mathbf{r}, R) + \mathbf{tr}\Big\{\Big[\mathbf{p}_o(\mathbf{r} - \hat{\mathbf{x}}_o)^T R - R^T(\mathbf{r} - \hat{\mathbf{x}}_o)\mathbf{p}_o^T\Big]\Omega\Big\}
\end{aligned}$$

where we have ignored terms of the order of Ω^2. Defining the matrix inner product as $\langle A, B\rangle = \mathbf{tr}(A^T B)$ (on $SO(n)$ this is proportional to the Killing form), we can identify the negative gradient of the function ψ at R by $\nabla_R\psi(\mathbf{r}, R) = \mathbf{p}_o(\mathbf{r} - \hat{\mathbf{x}}_o)^T R - R^T(\mathbf{r} - \hat{\mathbf{x}}_o)\mathbf{p}_o^T$. ∎

Lemma 4.2: The negative gradient of $\mathbf{g}_i$ with respect to R is given by the skew-symmetric matrix

$$\begin{aligned}
&\nabla_R\mathbf{g}_i(\mathbf{r}, R) = \qquad\qquad (16)\\
&\alpha^2\Big(\mathbf{e}_2^T R^T(\hat{\mathbf{x}}_i - \mathbf{r}) - c\Big)\Big(\mathbf{e}_2(\hat{\mathbf{x}}_i - \mathbf{r})^T R - R^T(\hat{\mathbf{x}}_i - \mathbf{r})\mathbf{e}_2^T\Big) - \\
&- \Big(\mathbf{e}_1^T R^T(\hat{\mathbf{x}}_i - \mathbf{r})\Big)\Big(\mathbf{e}_1(\hat{\mathbf{x}}_i - \mathbf{r})^T R - R^T(\hat{\mathbf{x}}_i - \mathbf{r})\mathbf{e}_1^T\Big).
\end{aligned}$$

Proof: Omitted; analogous to that of Lemma 4.1. ∎

Note that the negative gradients of the functions ψ and $\mathbf{g}_i$ with respect to R are both skew-symmetric matrices, as required for (14b) to ensure that $R \in SO(2)$ for all time $t \in [t_k, t_{k+1}]$. The gradients of ψ and $\mathbf{g}_i$ with respect to $\mathbf{r}$ are

$$\nabla_{\mathbf{r}}\psi(\mathbf{r}, R) = 2(R\mathbf{p}_o - \hat{\mathbf{x}}_o + \mathbf{r}) \tag{17}$$

and

$$\begin{aligned}
\nabla_{\mathbf{r}}\mathbf{g}_i(\mathbf{r}, R) = {}& 2\left(\mathbf{e}_1^T R^T(\hat{\mathbf{x}}_i - \mathbf{r})\right)R\mathbf{e}_1 \\
& - 2\alpha^2\left(\mathbf{e}_2^T R^T(\hat{\mathbf{x}}_i - \mathbf{r}) - c\right)R\mathbf{e}_2.
\end{aligned} \tag{18}$$

Then, the gradients of $\hat{\psi}$ required in (14) are

$$\nabla_{\mathbf{r}}\hat{\psi}(\mathbf{r}, R) = \nabla_{\mathbf{r}}\psi(\mathbf{r}, R) - \rho\sum_{i \in I}\frac{\nabla_{\mathbf{r}}\mathbf{g}_i(\mathbf{r}, R)}{\mathbf{g}_i^2(\mathbf{r}, R)} \tag{19a}$$

$$\nabla_R\hat{\psi}(\mathbf{r}, R) = \nabla_R\psi(\mathbf{r}, R) - \rho\sum_{i \in I}\frac{\nabla_R\mathbf{g}_i(\mathbf{r}, R)}{\mathbf{g}_i^2(\mathbf{r}, R)}, \tag{19b}$$

with $\nabla_R\psi(\mathbf{r}, R)$, $\nabla_R\mathbf{g}_i(\mathbf{r}, R)$, $\nabla_{\mathbf{r}}\psi(\mathbf{r}, R)$, and $\nabla_{\mathbf{r}}\mathbf{g}_i(\mathbf{r}, R)$ obtained from (15), (16), (17) and (18), respectively. Note

Algorithm 1 Hybrid control in the relative and global frames.

Require: A position $\mathbf{r}(t_k)$ and orientation $R(t_k)$ of the camera and estimated positions $\hat{\mathbf{x}}_{i,k}$ of the targets, so that $\mathbf{g}_i(\mathbf{r}(t_k), R(t_k)) > 0$ for all targets $i \in I$.

1: Find the next best view associated with objective "o" according to equation (9):

$$\mathbf{p}_{o,k+1} = \mathbf{p}_{o,k} - \int_0^T \frac{\partial h(\mathbf{p}_{o,k+1}(\tau))}{\partial \mathbf{p}_{o,k+1}} d\tau.$$

2: Move the camera according to the system (14):

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

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

3: At time t_{k+1} observe targets and incorporate new estimates and covariances into KF as in (5) and (6). Increase the observation index k by 1 and return to step 1.

that $\nabla_{\mathbf{r}} \hat{\psi}(\mathbf{r}, R)$ is a positive gradient, while $\nabla_R \hat{\psi}(\mathbf{r}, R)$ is a negative gradient.

V. Simulation Results

In this section we illustrate our approach in computer simulations. Subject to pixelated images (quantized noise), we compare the localization performance of the proposed two motion objectives, namely the *supremum objective* and the *centroid objective*, to two heuristic motion plans that are subject to the same parameters. The first of the heuristic motion plans is the *circle baseline*, which guides the robot on a circle with center at the estimated centroid of the targets. The second heuristic, the *straight baseline*, guides the robot as close to the targets as possible without allowing any to leave the field of view. Once this limit is reached, the robot stops and continues to take measurements. In both heuristic motion plans, the robot is always oriented toward its current estimate of the centroid of the targets.

All simulations were performed using image width equal to 1024 pixels and a baseline (b from Fig. 1) of 5 cm. The standard deviation of the Gaussian approximation to quantization noise was set equal to 0.25 pixels. In every simulation, the robot begins 1.5 m west of a cluster of targets, which are placed according to a uniform random distribution in the unit circle. The penalty parameter, $\rho = 1\text{e}{-5}$, ensured that all targets remained within the camera's 70° field of view throughout. The *circle baseline* and *straight baseline* traveled a distance equal to whichever proposed gradient method went further. All motion plans made the same amount of total observations.

Implementation of the *supremum objective* and the *centroid objective* are outlined in Algorithm 1. In step 1, we set the integration time interval T so that the distance between $\mathbf{p}_{o,k+1}$ and $\mathbf{p}_{o,k}$ is at most 0.1 mm, the maximum allowed distance the camera is allowed to travel before taking a new measurement. Once the new next best view $\mathbf{p}_{o,k+1}$ has been determined in the relative frame, step 2 of Algorithm 1 drives the camera in the global frame to realize $\mathbf{p}_{o,k+1}$. The camera moves until one of two events occurs. Either the next best view is successfully realized, or the robot moved the maximum distance.

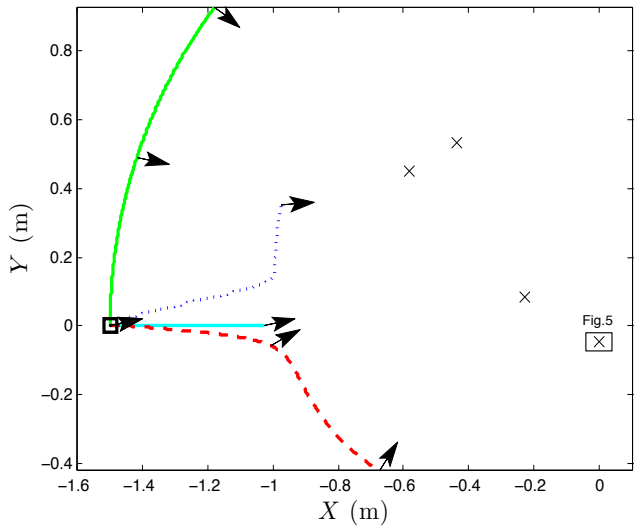

Fig. 3. Camera trajectories. Colors and line styles correspond to Fig. 4. The four exes are ground truth target locations.

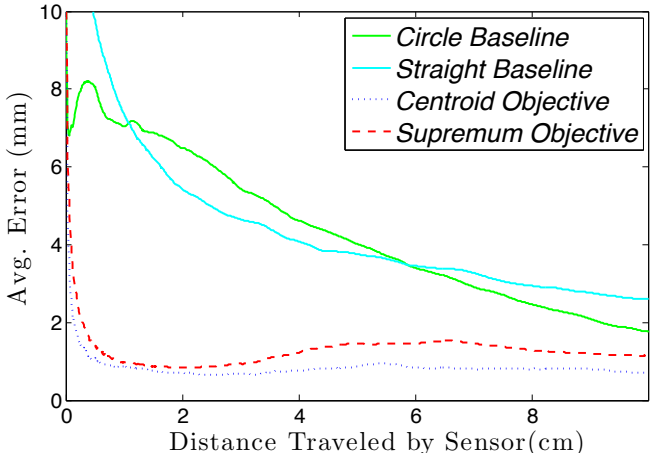

Fig. 4. Per target error between ground truth and KF outputs for the four motion plans.

We performed 100 simulations of the proposed gradient-based motion plans and the heuristic baselines. Each simulation began by generating a random cluster of four targets and running 10,000 iterations of Algorithm 1. All used identical parameters. All observations were faced with quantization noise after pixel coordinates are rounded to the nearest integer. Figure 3 shows an example of camera trajectories in one of the simulations. Figure 4 shows the average localization error per target over all simulations for the first 15 cm traveled by each sensor. After 15 cm, which is equivalent to 1500 observations, all methods performed similarly, except the *straight baseline*, which accumulates error once it stops moving, when it suffers from the same quantized noise in every observation. Figure 5 is a close up on the target marked in Figure 3. It shows the true location

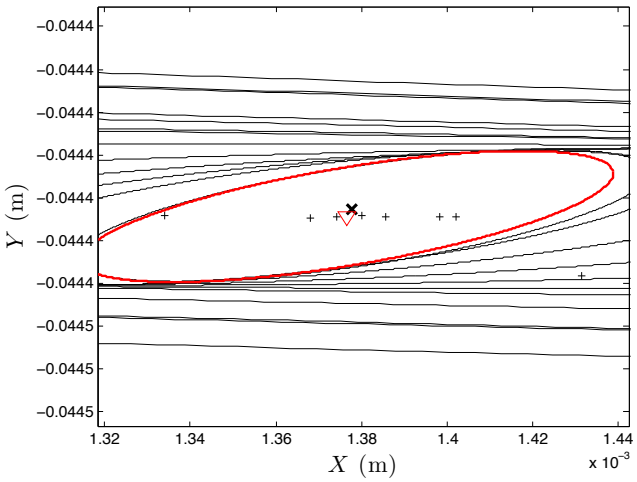

Fig. 5. The 3σ confidence intervals produced by the *supremum objective* for the area marked in Fig. 3, printed every 750 observations. The final KF-output location is a red triangle and a red ellipse, while the leading sequence are black crosses and ellipses. The **x** is ground truth.

of the target as well as successive estimates of its location and its uncertainty, represented by confidence intervals for the *supremum objective*.

VI. Conclusions

In this paper, we presented a novel solution to the Next-Best-View problem using mobile stereo vision. Our approach relied on a novel control decomposition in the relative camera frame and the global space. In the relative frame, we explicitly modeled uncertainty in target localization. This allowed us to obtain the next best view using gradient descent on appropriately defined potentials, without sampling the pose space or having to select from a set of previously recorded image pairs. This next best view was realized in the global space as a result of the camera's motion. Motion control was due to artificial potentials that jointly controlled the camera's rotation and translation in order to match a sequence of desired next best views. The integrated hybrid system was shown to exhibit superior localization properties compared to baseline methods operating under the same conditions. Compared to previous gradient-based approaches, our formulation is more precise since we take into account the correlation between errors in range and bearing, which are both due to quantization noise in the images, instead of treating them as independent. Furthermore, we do not assume omnidirectional sensors, but impose field of view constraints.

References

[1] N. Michael, D. Mellinger, Q. Lindsey, and V. Kumar, "The grasp multiple micro-uav testbed," *Robotics Automation Magazine, IEEE*, vol. 17, no. 3, pp. 56 –65, 2010.

[2] K. Zhou and S. Roumeliotis, "Multirobot active target tracking with combinations of relative observations," *IEEE Transactions on Robotics*, vol. 27, no. 4, pp. 678 –695, 2011.

[3] E. Xu, Z. Ding, and S. Dasgupta, "Target tracking and mobile sensor navigation in wireless sensor networks," *IEEE Transactions on Mobile Computing*, no. 99, p. 1, 2009.

[4] S. D. Blostein and T. S. Huang, "Error analysis in stereo determination of 3-d point positions," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 9, no. 6, pp. 752–766, 1987.

[5] L. H. Matthies and S. A. Shafer, "Error modelling in stereo navigation," *IEEE Journal of Robotics and Automation*, vol. 3, no. 3, pp. 239–250, 1987.

[6] C. C. Chang, S. Chatterjee, and P. R. Kube, "A quantization error analysis for convergent stereo," in *ICIP*, 1994, pp. II: 735–739.

[7] N. Massios and R. Fisher, "A best next view selection algorithm incorporating a quality criterion," in *British Machine Vision Conference*. Citeseer, 1998, pp. 780–789.

[8] R. Pito, "A solution to the next best view problem for automated surface acquisition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 10, pp. 1016 –1030, oct 1999.

[9] E. Dunn, J. van den Berg, and J.-M. Frahm, "Developing visual sensing strategies through next best view planning," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, oct. 2009, pp. 4001 –4008.

[10] C. Munkelt, A. Breitbarth, G. Notni, and J. Denzler, "Multi-view planning for simultaneous coverage and accuracy optimisation," in *British Machine Vision Conference*, 2010.

[11] M. Trummer, C. Munkelt, and J. Denzler, "Online next-best-view planning for accuracy optimization using an extended e-criterion," in *Int. Conf. on Pattern Recognition*, 2010, pp. 1642–1645.

[12] S. Wenhardt, B. Deutsch, J. Hornegger, H. Niemann, and J. Denzler, "An information theoretic approach for next best view planning in 3-d reconstruction," in *IEEE. Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2007, pp. 103 –106.

[13] A. Hornung, B. Zeng, and L. Kobbelt, "Image selection for improved multi-view stereo," in *IEEE. Conf. on Computer Vision and Pattern Recognition*, 2008.

[14] D. Fox, W. Burgard, H. Kruppa, and S. Thrun, "A probabilistic approach to collaborative multi-robot localization," *Autonomous Robots*, vol. 8, no. 3, pp. 325–344, 2000.

[15] S. Roumeliotis and G. Bekey, "Distributed multirobot localization," *IEEE Transactions on Robotics and Automation*, vol. 18, no. 5, pp. 781 – 795, 2002.

[16] A. W. Stroupe and T. Balch, "Value-based action selection for observation with robot teams using probabilistic techniques," *Robotics and Autonomous Systems*, vol. 50, no. 2-3, pp. 85 – 97, 2005.

[17] T. Chung, J. Burdick, and R. Murray, "A decentralized motion coordination strategy for dynamic target tracking," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2006, pp. 2416 – 2422.

[18] P. Yang, R. Freeman, and K. Lynch, "Distributed cooperative active sensing using consensus filters," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2007, pp. 405 –410.

[19] S. Ponda and E. Frazzoli, "Trajectory optimization for target localization using small unmanned aerial vehicles," in *AIAA Conf. on Guidance, Navigation, and Control*, Chicago, IL, 2009.

[20] W. Förstner, "Uncertainty and projective geometry," in *Handbook of Geometric Computing*, E. Bayro-Corrochano, Ed. Springer, 2005, pp. 493–534.

[21] M. W. M. G. Dissanayake, P. Newman, S. Clark, H. Durrant-Whyte, and M. Csorba, "A solution to the simultaneous localization and map building (slam) problem," *IEEE Transactions on Robotics and Automation*, vol. 17, no. 3, pp. 229 –241, 2001.

[22] S. Thrun, *Robotic mapping: A survey*. Morgan Kaufmann, 2002.

[23] H. F. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: part i," *Robotics Automation Magazine, IEEE*, vol. 13, no. 2, pp. 99 –110, 2006.

[24] G. Welch and G. Bishop, "An introduction to the kalman filter," In SIGGRAPH Courses, 2001.

[25] S. Wenhardt, B. Deutsch, E. Angelopoulou, and H. Niemann, "Active visual object reconstruction using d-, e-, and t-optimal next best views," in *IEEE. Conf. on Computer Vision and Pattern Recognition*, 2007.

[26] M. M. Zavlanos and G. J. Pappas, "A dynamical systems approach to weighted graph matching," *Automatica*, vol. 44, no. 11, pp. 2817–2824, Nov. 2008.