Hybrid Control for Mobile Target Localization with Stereo Vision

Charles Freundlich, Philippos Mordohai and Michael M. Zavlanos

Abstract— **In this paper, we control image collection for a mobile stereo camera that is actively localizing a group of mobile targets. In particular, assuming that at least one pair of stereo images of the targets is available, we propose a novel approach to control the rotation and translation of the stereo camera so that the next observation of the targets will minimize their localization uncertainty. We call this problem the Next-Best-View problem for mobile targets (mNBV). The advantage of using a stereo camera is that, using triangulation, the two simultaneous images taken by the robot during a single observation can yield range and bearing measurements of the targets, as well as their uncertainty. A Kalman filter fuses the full state history and covariance estimates, as more measurements are acquired. Our solution to the mNBV problem determines the relative transformations between camera and targets that will minimize the fused uncertainty of the targets' locations. We determine a motion plan that realizes the mNBV while respecting field of view constraints. In particular, with every new observation, we compute a new mNBV in the frame relative to the camera and subsequently realize this view in global coordinates via a gradient descent algorithm that also respects field of view constraints. Integration of mNBV 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 our previous work [2], we have shown that allowing a mobile stereo camera to actively find the Next-Best-View of a group of static targets in 2D is advantageous in terms of its effect on localization accuracy. In this work, we control a 6-DOF mobile stereo camera that observes a group of mobile targets, which may themselves be other mobile robots.

The advantage of binocular stereo, 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 [3]–[5]. We leverage this "inherent" uncertainty of stereo vision to define the Next-Best-View (mNBV) as the position and orientation of a stereo camera that, given a sequence of observations of a group of possibly mobile targets, minimizes their localization uncertainty.

In the computer vision literature, the NBV problem has often been formulated as the selection of the next image from a finite data set using sampling or grid-based methods [6]–[12]. While these methods do obtain uncertainty estimates that depend on factors such as viewing distance and camera resolution, which improves accuracy in 3D reconstruction, they do not continuously guide the image collection process, consider dynamic environments or mobile targets.

Approaches capable of computing the mNBV position have been proposed in the cooperative localization literature [13]–[19]. They approximate the target location error covariance matrix, which captures the uncertainty, using abstract sensor models, often treating range and bearing measurement uncertainty as independent of range or independent of each other. Instead, in this work we show how to leverage the true sensing uncertainty for mNBV determination. 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. This is true for both the instantaneous covariance of one measurement and for the filtered covariance of the full sequence of measurements. As a result, our proposed controller guides the robot to more effective viewing positions compared to other approaches. Specifically, we first determine the mNBV, expressed as an optimal transformation between sensor and target, and then realize it by moving the camera via a gradient descent on an artificial potential that additionally respects field-of-view constraints. The robot moves until a next observation is made, which is used to determine a new next best viewing position to be realized. Integration of mNBV with continuous motion planning gives rise to a hybrid system that drives a robot in the direction that minimizes localization uncertainty of the mobile targets.

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

II. System Model & Problem Formulation

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

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

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

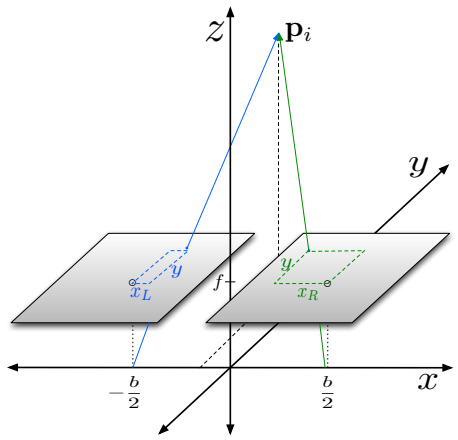

Fig. 1. Stereo geometry in 3D. Two rays from the camera centers to a target located at $\mathbf{p}_i$ creates a pair of image coordinates, (x_L, y) and (x_R, y).

orientation $R(t) \in SO(3)$, which is the special orthogonal group of dimension three, with respect to the global frame at time $t \geq 0$. The relative coordinate frame is anchored to the stereo camera. This frame, 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 (L) and (R). The (L) and (R) camera centers are located at $(-b/2, 0, 0)$ and $(b/2, 0, 0)$ in the relative coordinates, where b denotes the baseline.

The position of target i with respect to the relative camera frame is

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

where f denotes the focal length of the camera lens, measured in pixels, and x_{Li}, x_{Ri}, and y_i denote the coordinates of target i, measured in pixels, on the left and right camera images, as in Fig. 1, noting that y_i is equal in each image by the epipolar constraint. 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 $\hat{x}_{Li}$, $\hat{x}_{Ri}$, and $\hat{y}_i$, 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 [3], [5]. For convenience, we follow [4], [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 $\hat{\mathbf{p}}_i = \mathbf{p}(\hat{x}_{Li}, \hat{x}_{Ri}, \hat{y}_i)$ and covariance $\Sigma_i \in \mathbb{S}^3_+$ in global coordinates, where $\mathbb{S}^3_+$ denotes the set of 3×3 symmetric positive definite matrices. An analytical representation of $\Sigma_{i,k}$ is given in Section III.

A. Kalman Filtering for Mobile Targets

Now, assume that the stereo camera has made a sequence of observations of the mobile targets. Introduce an index $k \geq 0$ associated with every observation such that $\mathbf{y}_{i,k} \in \mathbb{R}^3$ denotes the observation and $\Sigma_{i,k} \in \mathbb{S}^3_+$ denotes its associated covariance, which is always in the global frame.

The goal of localization is to create accurate state information for a group of targets based on these observations. We consider the history of measurements with a Kalman filter (KF), which is an efficient information filter that incorporates noisy observations within a system model to create accurate state estimates [18].

Let $\mathbf{z}_i = [\mathbf{x}_i^T \ \dot{\mathbf{x}}_i^T \ \ddot{\mathbf{x}}_i^T]^T$ be the true state of target i. Since observations are discrete events, we model the continuous time evolution of $\mathbf{z}_i$ with the discrete time linear system

$$\mathbf{z}_{i,k} = \Phi \mathbf{z}_{i,k-1} + \mathbf{u}_{i,k-1}, \tag{2a}$$
$$\mathbf{y}_{i,k} = H\mathbf{z}_{i,k} + \mathbf{v}_{i,k}, \tag{2b}$$

where $H = [I_{3\times 3} \ \mathbf{0}_{3\times 6}]$, Φ is the state transition matrix, $\mathbf{u}_{i,k-1}$ and $\mathbf{v}_{i,k}$ are noise terms, and $\mathrm{cov}(\mathbf{v}_{i,k}) = \Sigma_{i,k}$. The specific nature of Φ and $\mathbf{u}$, including time-based adaptations and initialization procedures, is well studied from the perspective of mobile target tracking [21], [22]. In Section V, we use a constant acceleration model for Φ. The term $\mathbf{u}_{i,k-1}$ can be used to account for variations in the acceleration; see [21]. If *a priori* knowledge of the target trajectories is available, more specific models of Φ can be used. For any target motion model, denote by $\hat{\mathbf{z}}_i$ the estimate of $\mathbf{z}_i$. Also, denote the covariance of $\hat{\mathbf{z}}_i - \mathbf{z}_i$ by U_i. Given prior estimates $\hat{\mathbf{z}}_{i,k-1|k-1}$ and $U_{i,k-1|k-1}$, the Kalman Filter for target i is

$$\hat{\mathbf{z}}_{i,k|k-1} = \Phi \hat{\mathbf{z}}_{i,k-1|k-1}, \tag{3a}$$
$$U_{i,k|k-1} = \Phi U_{i,k-1|k-1} \Phi^T + W_k, \tag{3b}$$
$$K_k = U_{i,k|k-1} H^T \left[H U_{i,k|k-1} H^T + \Sigma_{i,k} \right]^{-1}, \tag{3c}$$
$$\hat{\mathbf{z}}_{i,k|k} = \hat{\mathbf{z}}_{i,k|k-1} + K_k \left[\mathbf{y}_{i,k} - H\hat{\mathbf{z}}_{i,k|k-1} \right], \tag{3d}$$
$$U_{i,k|k} = U_{i,k|k-1} - K_k H U_{i,k|k-1}, \tag{3e}$$

where W_k is process noise related to $\mathbf{u}_{i,k}$, which is explicitly given in [21], along with a simple initialization protocol. From equation (3e) and the results of [23], a closed form expression for the fused covariance estimate follows in the form of a Lemma. The proof is omitted.

Lemma 2.1: Let $U_{i,k|k-1}$ denote the fused 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

$$\Xi_{i,k} \triangleq H U_{i,k|k} H^T = \left[\left(H U_{i,k|k-1} H^T \right)^{-1} + \Sigma_{i,k}^{-1} \right]^{-1}. \tag{4}$$

B. The Mobile Next Best View Problem

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

$$H U_{s,k|k-1} H^T \text{ with } s = \underset{j \in \mathcal{N}}{\mathrm{argmax}} \left\{ \mathbf{tr} \left[H U_{j,k|k-1} H^T \right] \right\} \tag{5}$$

denote the predicted covariance of the worst localized target and

$$H U_{c,k|k-1} H^T = \frac{1}{n} H \sum_{i \in \mathcal{N}} U_{i,k|k-1} H^T. \tag{6}$$

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 $HU_{s,k|k-1}H^T$ (respectively, the average of the targets' predicted covariances $HU_{c,k|k-1}H^T$) and the predicted next location $\mathbf{z}_{s,k|k-1}$ of target s (respectively, the average of the targets' predicted locations $\mathbf{z}_{c,k|k-1}$), determine $\mathbf{p}_{s,k}$ (respectively, $\mathbf{p}_{c,k}$) so that $\mathbf{tr}[\Xi_{s,k}]$ (respectively, $\mathbf{tr}[\Xi_{c,k}]$) is minimized.

In problem 1, we have chosen the trace as a measure of uncertainty among other choices, such as the determinant or the maximum eigenvalue. It is shown in [24] 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*. Clearly, $\Xi_{s,k}$ 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 $\mathbf{p}_{i,k} \forall i \in \mathcal{N}$. Attempting to find a $\mathbf{p}_{c,k}$ 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 place a virtual target at the centroid, $\mathbf{p}_c = \frac{1}{n}\sum_{i\in\mathcal{N}} \mathbf{p}_i$. The centroid serves as a proxy for all targets.

In what follows, we decompose optimization of the above objectives in the relative camera frame and the global frame. 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 $k-1$ 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 location proxies $\mathbf{p}_{s,k}$ or $\mathbf{p}_{c,k}$ on the relative camera frame so that if a new observation is made at time t_k 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 Σ_i of target i as a function of its relative position $\mathbf{p}_i$ for any time $t \in [t_{k-1}, t_k]$. Let

$$Q = \mathrm{cov}([\hat{x}_{Li} \quad \hat{x}_{Ri} \quad \hat{y}_i]) \approx \mathrm{diag}\begin{bmatrix}\sigma_L^2 & \sigma_R^2 & \sigma_y^2\end{bmatrix} \tag{7}$$

denote the approximate covariance of the error in target coordinates image frames, respectively, where σ_L^2, σ_R^2, and σ_y^2 denote the associated variances.[1] Also let J_i be the Jacobian of $\mathbf{p}_i \triangleq \mathbf{p}(x_{Li}, x_{Ri}, y_i)$ evaluated at the point $(\hat{x}_{Li}, \hat{x}_{Ri}, \hat{y}_i)$. Then, the first order (linear) approximation of $\mathbf{p}_i = \mathbf{p}(x_{Li}, x_{Ri}, y_i)$ about the point $(\hat{x}_{Li}, \hat{x}_{Ri}, \hat{y}_i)$ is

$$\mathbf{p}(x_{Li}, x_{Ri}, y_i) \approx \mathbf{p}(\hat{x}_{Li}, \hat{x}_{Ri}, \hat{y}_i) + J_i[x_{Li} \quad x_{Ri} \quad y_i]^T. \tag{8}$$

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

Since $\mathbf{p}_i(\hat{x}_{Li}, \hat{x}_{Ri}, \hat{y}_i)$ corresponds to the current mean estimate of target coordinates, it is constant in (8). Therefore, the covariance of $\mathbf{p}_i$ in the relative camera frame is $J_i Q J_i^T$. Fusing covariance matrices as in Lemma 2.1 requires that they are represented in the same coordinate system. 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 $[t_{k-1}, t_k]$, we may approximate the camera's rotation $R(t)$ at time $t \in [t_{k-1}, t_k]$ by its initial rotation $R(t_{k-1})$. Dropping the time index, i.e. $R(t_{k-1}) \approx R$, the instantaneous covariance of $\mathbf{p}_i$ at any time instant $t \in [t_{k-1}, t_k]$ can be approximated by

$$\Sigma_i = \mathrm{cov}[\mathbf{p}(\hat{x}_{Li}, \hat{x}_{Ri}, \hat{y}_i)] \approx R J_i Q J_i^T R^T. \tag{9}$$

In view of (1), the covariance in (9) is clearly a function of the target coordinates on the relative image frame.

To determine the vector $\mathbf{p}_{o,k}$ that minimizes localization uncertainty, we define the uncertainty potential,

$$h(\mathbf{p}_{o,k}) = \mathbf{tr}\big[\Xi_{o,k}\big], \tag{10}$$

where o stands for 'objective' and can be either s or c, depending on the objective that is used to obtain the next best view [cf. (5) and (6)]. Then, a gradient descent step for the minimization of h is

$$\mathbf{p}_{o,k} = \mathbf{p}_{o,k-1} - \int_0^T \nabla h\left(\mathbf{p}_o(\tau)\right) d\tau. \tag{11}$$

The length $T > 0$ of the integration interval is chosen so that the distance between $\mathbf{p}_{o,k}$ and $\mathbf{p}_{o,k+1}$ 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 (11). Here, the relative velocities of the targets play a key role in determining $\frac{\partial \Sigma_o}{\partial \mathbf{p}_o}$.

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

$$\left[\nabla h\left(\mathbf{p}_o\right)\right]_j = \mathbf{tr}\left\{\Sigma_o^{-1}\Xi_{o,k}^2\Sigma_o^{-1}\left[\frac{\partial \Sigma_o}{\partial \mathbf{p}_o}\right]_j\right\}, \tag{12}$$

where $\left[\frac{\partial \Sigma_o}{\partial \mathbf{p}_o}\right]_j$ is the j-th coordinate of the gradient of Σ_o with respect to the individual coordinates of $\mathbf{p}_o$, and $j = 1, 2, 3$, corresponding to the three coordinates of $\mathbf{p}_o$.
The proof of Proposition 3.1 is omitted.

IV. Controlling the Global Frame

The update in (11) provides the relative target coordinates on the camera frame where, if the next observation of target o at time t_k is at $\mathbf{p}_{o,k}$, the localization uncertainty associated with objective "o" is minimized. Our goal in this section is to determine a new camera position $\mathbf{r}(t_k)$ and orientation $R(t_k)$ in space and that realizes the mNBV defined by $\mathbf{p}_{o,k}$ from (11). For this, let

$$\psi\left(\mathbf{r}, R\right) = \|R\mathbf{p}_o - \hat{\mathbf{x}}_o + \mathbf{r}\|_F^2, \tag{13}$$

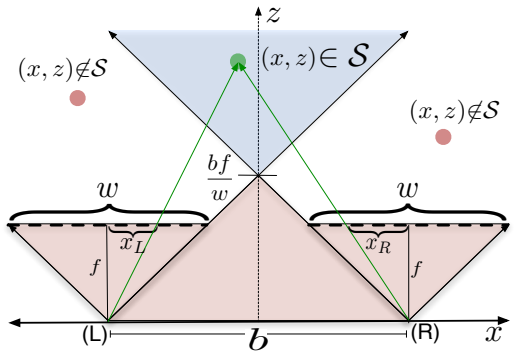

Fig. 2. The field of view for a stereo camera in the xz plane. The field of view in the yz plane is similar.

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 Frobenious norm and to simplify notation we have dropped dependence of $\hat{\mathbf{x}}_{o,k}$ and $\mathbf{p}_{o,k}$ on the observation index k.

The stereo setup is not omnidirectional. For a 3D point to appear in a given image, the point must lie within the field of view of the stereo pair as the robot rotates and translates in an effort to minimize (13). The two cameras are identical to each other; they have square images with the same field of view, which is determined by the image width w and the focal length, f. Denote by $\mathcal{S}$ the set of points in relative coordinates that are visible to both cameras in the pair. In Fig. 2, this set is given by the blue shaded polyhedron

$$\mathcal{S}=\left\{(x,y,z)\in\mathbb{R}^3\,:\,|x|\leq\frac{wz-bf}{2f},\,|y|\leq\frac{zw}{2f},\,z>\frac{bf}{w}\right\}, \tag{14}$$

which is the intersection of two pyramids facing the positive z direction with vertices located at the two camera centers. 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.

In (14), any feasible target location in relative coordinates can be written as

$$(x_i,y_i,z_i)^T\triangleq\mathbf{p}_i(\mathbf{r},R)=R^T(\hat{\mathbf{x}}_i-\mathbf{r}) \tag{15}$$

In view of (14) and (15), we constrain the control variables $(\mathbf{r},R)$ during the minimization of $\psi(\mathbf{r},R)$ to the set

$$\mathcal{D}=\{(\mathbf{r},R)\in\mathbb{R}^3\times SO(3)\,:\,(x_i,y_i,z_i)\in\mathcal{S}\,\forall\,i\in\mathcal{N}\}. \tag{16}$$

Define, further, the functions

$$\phi_{i1}(\mathbf{r},R)=\left(\frac{wz_i-bf}{2f}\right)^2-x_i^2, \tag{17}$$

$$\phi_{i2}(\mathbf{r},R)=\left(\frac{wz_i}{2f}\right)^2-y_i^2,\quad \phi_{i3}(\mathbf{r},R)=z_i^2-\left(\frac{bf}{w}\right)^2,$$

that are positive if $(\mathbf{r},R)\in\mathcal{D}$, where x_i,y_i and z_i are given in (15) as functions of $\mathbf{r}$ and R. Given a vector $\mathbf{p}_o$, an initial pose $\big(\mathbf{r}(t_{k-1}),R(t_{k-1})\big)$, and an estimate of target locations $\hat{\mathbf{x}}_i$ for $i=1,\ldots,n$, our goal is to solve the problem

$$\min_{(r,R)}\psi\left(\mathbf{r},R\right)\ \text{s.t.}\ \phi_{ij}(\mathbf{r},R)\geq 0\quad\forall i\in\mathcal{N},j=1,2,3, \tag{18}$$

where $(\mathbf{r},R)\in\mathbb{R}^3\times SO(3)$. We solve problem (18) by designing a gradient flow on the space of rotations and translations that minimizes ψ while respecting the field of view constraints. We enforce the field of view constraints using a barrier method, which incorporates them into the objective function. Specifically, since the initial pose satisfies $(\mathbf{r}(t_{k-1}),R(t_{k-1}))\in\mathcal{D}$, we choose $\frac{1}{\phi_{ij}}$ as suitable barriers, since $\frac{1}{\phi_{ij}}\to+\infty$ if any $\phi_{ij}\to 0$ from the right. This gives rise to the objective function,

$$\hat{\psi}\left(\mathbf{r},R\right)=\psi\left(\mathbf{r},R\right)+\frac{\rho}{n}\sum_{i\in\mathcal{N}}\sum_{j=1}^{3}\frac{1}{\phi_{ij}(\mathbf{r},R)}, \tag{19}$$

where $\rho>0$ is a penalty parameter and multiplication by $1/n$ ensures that the number of targets does not affect the strength of the penalty.

Letting $t_{k-1}>0$ denote the time instant associated with observation $k-1$ and for all time $t\in[t_{k-1},t_k]$, we define the gradient flow

$$\dot{\mathbf{r}}=-K\nabla_{\mathbf{r}}\hat{\psi}(\mathbf{r},R), \tag{20a}$$

$$\dot{R}=-R\nabla_R\hat{\psi}(\mathbf{r},R), \tag{20b}$$

on the joint space of camera positions $\mathbb{R}^3$ and orientations $SO(3)$. K is a positive gain for numerical purposes. Equations (20), ensure that if $R(t_{k-1})\in SO(3)$ and $R(t)$ evolves as in (20b) and $\nabla_R\hat{\psi}(\mathbf{r},R)$ is skew-symmetric, then $R(t)\in SO(3)$ for all time $t\in[t_{k-1},t_k]$, see, e.g., [25].

A. Gradients of the Global Potential Function

In the remainder of this section we provide analytic expressions for the gradients in (20). In particular, the gradient of ψ with respect to $\mathbf{r}$ and R are given by

$$\nabla_{\mathbf{r}}\psi=2(R\mathbf{p}_o-\hat{\mathbf{x}}_o+\mathbf{r}) \tag{21a}$$

$$\nabla_R\psi=R^T(\mathbf{r}-\hat{\mathbf{x}}_o)\mathbf{p}_o^T-\mathbf{p}_o(\mathbf{r}-\hat{\mathbf{x}}_o)^TR. \tag{21b}$$

Equation (21b) is obtained by taking the first order approximation of ψ in a neighborhood of R, defined as $R(I+\Omega)$ with Ω a skew-symmetric matrix and using the fact that the matrix inner product is defined by $\langle A,B\rangle=\mathbf{tr}(A^TB)$.

The gradients of $\frac{1}{\phi_{ij}}$ with respect to $\mathbf{r}$ and R are available through the chain rule:

$$\nabla_{\mathbf{r}}\frac{1}{\phi_{ij}}=\frac{-1}{\phi_{ij}^2}\frac{\partial\phi_{ij}}{\partial x_i}\nabla_{\mathbf{r}}x_i-\frac{1}{\phi_{ij}^2}\frac{\partial\phi_{ij}}{\partial y_i}\nabla_{\mathbf{r}}y_i-\frac{1}{\phi_{ij}^2}\frac{\partial\phi_{ij}}{\partial z_i}\nabla_{\mathbf{r}}z_i, \tag{22a}$$

$$\nabla_R\frac{1}{\phi_{ij}}=\frac{-1}{\phi_{ij}^2}\frac{\partial\phi_{ij}}{\partial x_i}\nabla_Rx_i-\frac{1}{\phi_{ij}^2}\frac{\partial\phi_{ij}}{\partial y_i}\nabla_Ry_i-\frac{1}{\phi_{ij}^2}\frac{\partial\phi_{ij}}{\partial z_i}\nabla_Rz_i. \tag{22b}$$

The coefficient derivatives in (22) are elementary. Next, the gradients of x_i,y_i, and z_i with respect to R are given by the skew symmetric matrices

$$\nabla_Rx_i=(1/2)\left[R^T(\hat{\mathbf{x}}_i-\mathbf{r})\mathbf{e}_1^T-\mathbf{e}_1(\hat{\mathbf{x}}_i-\mathbf{r})^TR\right], \tag{23a}$$

$$\nabla_Ry_i=(1/2)\left[R^T(\hat{\mathbf{x}}_i-\mathbf{r})\mathbf{e}_2^T-\mathbf{e}_2(\hat{\mathbf{x}}_i-\mathbf{r})^TR\right], \tag{23b}$$

$$\nabla_Rz_i=(1/2)\left[R^T(\hat{\mathbf{x}}_i-\mathbf{r})\mathbf{e}_3^T-\mathbf{e}_3(\hat{\mathbf{x}}_i-\mathbf{r})^TR\right], \tag{23c}$$

where $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$ are unit vectors of the standard basis. Finally, the gradients of x_i,y_i, and z_i with respect to $\mathbf{r}$ are

$$\nabla_{\mathbf{r}}x_i=-R\mathbf{e}_1,\quad\nabla_{\mathbf{r}}y_i=-R\mathbf{e}_2,\quad\text{and}\quad\nabla_{\mathbf{r}}z_i=-R\mathbf{e}_3. \tag{24}$$

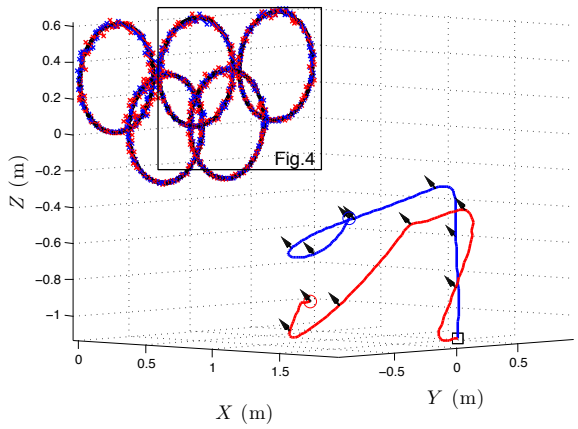

Fig. 3. Camera trajectories from one of the simulations used to create Figs. 5, 6, and 7. Colors correspond to these figures. The KF location outputs are plotted on top of the target trajectories with corresponding colors.

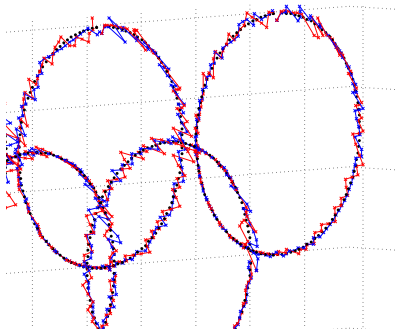

Fig. 4. A closeup of the trajectories Fig. 3. The final KF-output locations are plotted as $\times$ with lines to connect them to guide the eye. The ground truth target locations are plotted as $\bullet$. Each of the camera's outputs are plotted, and the colors correspond to the legend in Figs. 5, 6, and 7.

Combining (21), (22), (23), and (24), the gradients of $\hat{\psi}(\mathbf{r}, R)$ as required in (20) are

$$\nabla_{\mathbf{r}}\hat{\psi}\left(\mathbf{r}, R\right) = \nabla_{\mathbf{r}}\psi + \frac{\rho}{n}\sum_{i\in\mathcal{N}}\sum_{j=1}^{3}\nabla_{\mathbf{r}}\frac{1}{\phi_{ij}}, \tag{25a}$$

$$\nabla_{R}\hat{\psi}\left(\mathbf{r}, R\right) = \nabla_{R}\psi + \frac{\rho}{n}\sum_{i\in\mathcal{N}}\sum_{j=1}^{3}\nabla_{R}\frac{1}{\phi_{ij}}. \tag{25b}$$

Note that (25b) is a skew-symmettric matrix.

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*. All simulations were performed using image width equal to 1024 pixels and a baseline (b from Fig. 2) of 10 cm. The standard deviation of the Gaussian approximation to quantization noise was set equal to 1 pixel.

In each simulation, the stereo cameras localize a group of mobile targets that fly in the Olympic ring pattern, which represents a difficult maneuvering task for unmanned aerial vehicles, in which precise localization would be critical. The three upper target trajectories move counterclockwise while the two lower trajectories move clockwise at 0.5 m/s. Although the group's motion is not *a priori* known to the

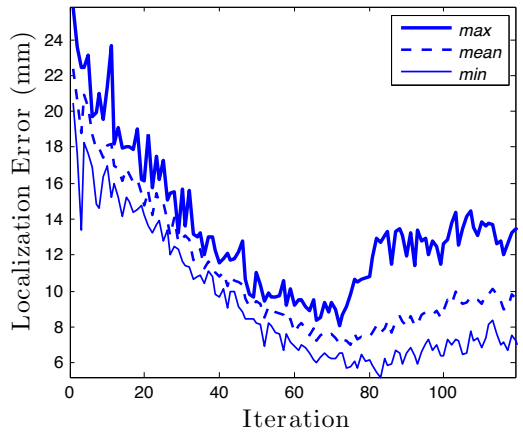

Fig. 5. The minimum, mean, maximum errors localization from the *centroid objective*, averaged over 100 simulations.

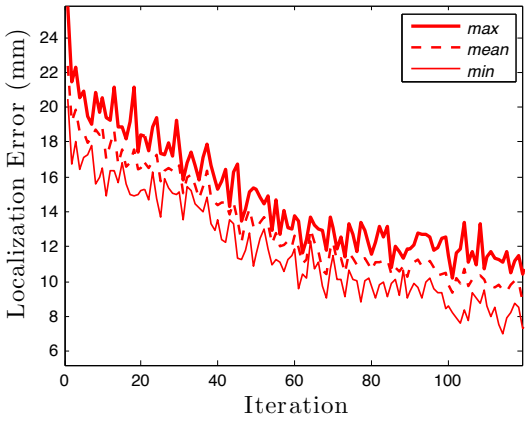

Fig. 6. The minimum, mean, maximum errors localization from the *supremum objective*, averaged over 100 simulations.

cameras, the cameras use the constant acceleration model of target motion. This model is integrated into the respective Kalman Filters. The penalty parameter, $\rho = 1\text{e}{-2}$, ensured that all targets remained within the 70° field of view.

Given a frame rate and sensor speed, which for our simulations were set equal to 30 fps and 1 m/s, we set the integration time interval T so that the distance between $\mathbf{p}_{o,k+1}$ and $\mathbf{p}_{o,k}$ is the distance the camera travels before taking a new measurement. Once the new mNBV $\mathbf{p}_{o,k+1}$ has been determined in the relative frame, the camera drives (in the global coordinates) to realize $\mathbf{p}_{o,k+1}$. The gain for the global update was $K = 0.1$. The camera moves until one of two events occurs. Either the next best view is successfully realized, or the robot moved the maximum distance.

We show the results of 100 simulations. The initial camera location in each simulation was a random point on a sphere centered at the origin of the XYZ coordinate system shown in Figure 3. The orientation was initialized toward the centroid of the targets. 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 plots the KF location outputs on top of the actual target trajectories from one of the 100 simulations that were used to create Figs. 5, 6, and 7. Figures 5 and 6 show the localization error per target for

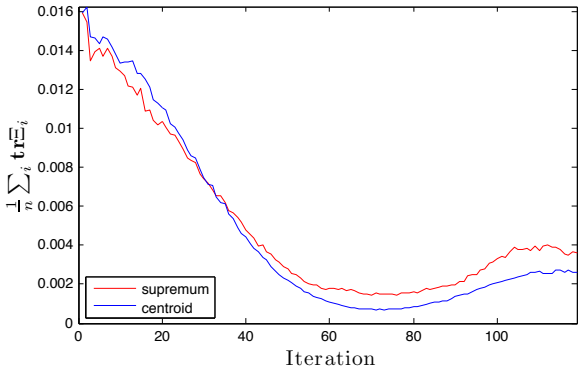

Fig. 7. The trace of the average of the KF-updated covariance for each target, averaged over all simulations.

the *centroid* and *supremum objectives*, averaged over 100 simulations. Figure 7 shows the values of the trace of the average of all target covariance matrices.

Figures 5 and 6 suggest that both objectives achieve localization resolution of $<$1.5 cm of fast and nonrigidly moving point targets. Because of the symmetry in the data set, the *centroid objective* is advantaged, which is why it outperforms the *supremum objective* in Figure 7. We also performed simulations with asymmetric data sets and outliers, which favored the *supremum objective*. Note that, in Figures 5 and 6, the difference between the maximum and mean errors for the *centroid objective* is greater than it is for the *supremum objective*. This is because the *centroid objective* algorithm gives less weight to the poorest localized target in favor of the majority. Any nondecreasing properties in Figures 5, 6, and 7 appear due to quantized observations. The correlation coefficient between the mean target error in Figures 5 and 6 and the average of the traces of the updated covariances from Figure 7 is 0.98 for the *centroid objective* and 0.90 for the *supremum objective*. Overall localization accuracy could be further improved by *a priori* knowledge of motion model, on-line adaptive modeling [22], and multiple observers.

VI. Conclusions

In this paper, we posed the localization problem for moving targets in 3D as the mobile Next-Best-View (mNBV) problem. Our approach relied on a novel control decomposition that was designed to iteratively reduce sensing uncertainty. In the relative frame, we explicitly modeled uncertainty in target localization with stereo vision. We integrated this information via Kalman Filtering, which provided accurate covariances and target location predictions. This allowed us to obtain the mNBV using gradient descent on appropriately defined potentials, without sampling the pose space or having to select from a set of previously recorded image pairs. The camera's motion in the global space realized the mNBV via the potentials, jointly guiding the camera location and orientation to match a sequence of desired next best views. Compared to previous gradient-based approaches, our integrated hybrid system 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 *et al.*, "The grasp multiple micro-uav testbed," *IEEE Robotics Automation Magazine*, vol. 17, no. 3, pp. 56 –65, 2010.

[2] C. Freundlich *et al.*, "A hybrid control approach to the next-best-view problem using stereo vision," in *IEEE Int. Conf. on Robotics and Automation*, 2013, pp. 4478 –4483.

[3] 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.

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

[5] C. C. Chang *et al.*, "A quantization error analysis for convergent stereo," in *Int. Conf. on Image Processing*, 1994, pp. II: 735–739.

[6] 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.

[7] 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.

[8] E. Dunn *et al.*, "Developing visual sensing strategies through next best view planning," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, oct. 2009, pp. 4001 –4008.

[9] C. Munkelt *et al.*, "Multi-view planning for simultaneous coverage and accuracy optimisation," in *British Machine Vision Conference*, 2010.

[10] M. Trummer *et al.*, "Online next-best-view planning for accuracy optimization using an extended e-criterion," in *Int. Conf. on Pattern Recognition*, 2010, pp. 1642–1645.

[11] S. Wenhardt *et al.*, "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.

[12] A. Hornung *et al.*, "Image selection for improved multi-view stereo," in *IEEE. Conf. on Computer Vision and Pattern Recognition*, 2008.

[13] D. Fox *et al.*, "A probabilistic approach to collaborative multi-robot localization," *Autonomous Robots*, vol. 8, no. 3, pp. 325–344, 2000.

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

[15] 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.

[16] T. Chung *et al.*, "A decentralized motion coordination strategy for dynamic target tracking," in *IEEE Int. Conf. on Robotics and Automation*, 2006, pp. 2416 –2422.

[17] P. Yang *et al.*, "Distributed cooperative active sensing using consensus filters," in *IEEE Int. Conf. on Robotics and Automation*, 2007, pp. 405 –410.

[18] 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.

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

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

[21] R. Singer, "Estimating optimal tracking filter performance for manned maneuvering targets," *IEEE Trans. Aerospace and Electronic Systems*, vol. AES-6, no. 4, pp. 473 –483, july 1970.

[22] X. Rong Li and V. Jilkov, "Survey of maneuvering target tracking. part I. dynamic models," *IEEE Trans. Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333 – 1364, oct. 2003.

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

[24] S. Wenhardt *et al.*, "Active visual object reconstruction using d-, e-, and t-optimal next best views," in *IEEE. Conf. on Computer Vision and Pattern Recognition*, 2007.

[25] 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.