Appearance-based Active, Monocular, Dense Reconstruction for Micro Aerial Vehicles

Christian Forster, Matia Pizzoli, and Davide Scaramuzza
Robotics and Perception Group, University of Zurich, Switzerland
Email: {forster, pizzoli, sdavide}@ifi.uzh.ch

***Abstract*—In this paper, we investigate the following problem: given the image of a scene, what is the trajectory that a robot-mounted camera should follow to allow optimal dense depth estimation? The solution we propose is based on maximizing the information gain over a set of candidate trajectories. In order to estimate the information that we expect from a camera pose, we introduce a novel formulation of the measurement uncertainty that accounts for the scene appearance (i.e., texture in the reference view), the scene depth and the vehicle pose. We successfully demonstrate our approach in the case of real-time, monocular reconstruction from a micro aerial vehicle and validate the effectiveness of our solution in both synthetic and real experiments. To the best of our knowledge, this is the first work on *active, monocular dense* reconstruction, which chooses motion trajectories that minimize perceptual ambiguities inferred by the texture in the scene.**

I. Introduction

Recent advances in Structure-from-Motion and Visual SLAM made real-time, dense reconstruction from multiple views a viable alternative to laser range finders in robot perception tasks. Impressive results have been demonstrated in the context of Multi-View Stereo (MVS) [17, 26, 29], where the knowledge of the camera motion is used to estimate depth from different vantage points. Nonetheless, depending on the scene, camera motion plays a fundamental role in the quality of the obtained reconstruction.

When observing demonstrations of monocular dense reconstruction from hand-held cameras, such as [17, 19], one can notice the commonly used pattern of moving the camera in a *circular trajectory* around a reference view.[1] Intuitively, a circular trajectory constitutes a reasonable approach, as the generated epipolar lines span uniformly the images and increase the chances of reliable stereo matches. Now, suppose that monocular vision is used by a robot to estimate the depth. What radius should we use for the circular camera trajectory? Or more generally, *what is the camera trajectory that provides the best depth measurements?*

In practice, the *best* trajectory depends on different factors: (i) the depth estimate of the scene; (ii) the uncertainty of the current estimate; (iii) the appearance (texture) of the scene; (iv) the current robot pose. Based on the aforementioned considerations, in this paper we introduce a Bayesian formulation to estimate dense depth maps from a Micro Aerial Vehicle (MAV). The next best poses are computed as a function of the robot's current pose and motion as well as the expected depth uncertainty reduction due to predicted future measurements.

A video demonstrating the system is available on the author's website: http://rpg.ifi.uzh.ch.

A. Related Work

The problem of computing the optimal views to reconstruct an object or a scene has been studied for more than two decades and is known in the computer vision literature as active vision, View Path Planning (VPP), or Next-Best-View (NBV) [1, 2, 4, 7, 21]. Often, the sensor motion is restricted to a sphere and it is assumed that the object of interest is at all times located completely in the sensor frustum. Proposed algorithms reason about voxel occupancy, occlusion edges, and surface coverage [13, 15]. Schmid et al. [20] addressed view planning with an MAV. Similarly to our work, the authors compute a set of aerial views to be used in a multi-view stereo pipeline. However, differently from our approach, their system assumes an *a-priori* model of the scene of interest. Viewpoints are, thus, computed off-line on the pre-computed object hull and the most informative ones are selected on the basis of heuristics that aim at providing full scene coverage. In contrast, we provide an active depth estimation method operating in real-time and on-line.

In the robotics community, a related field to view planning is known as exploration. The first to close the loop between view planning and 3D reconstruction were Whaite and Ferrie [31]. The exploration of a depth-sensor attached to a robot arm was driven by uncertainty reduction of a probabilistic surface model. Feder et al. [10] proposed the first work on active SLAM that seeks to minimizes both vehicle and landmark uncertainties. Bourgault et al. [5] proposed to complement the sparse feature-based SLAM approach with an occupancy grid to provide means of integrating dense range measurements. The proposed exploration policy uses the entropy in the occupancy grid map to stimulate exploration while the uncertainty in the SLAM assures localization accuracy. This approach was extended to particle-filter SLAM [25] and recently to pose-graph SLAM [27].

While the previous works relied on depth sensors, Davison and Murray [8] were the first to take into account the effects of actions during *visual* SLAM. The goal was to select a fixation-point of a moving stereo head attached to a mobile

This research was supported by the Swiss National Foundation (project number 200021-143607, "Swarm of Flying Cameras"), the National Center of Competence in Research Robotics, the CTI project number 14652.1, and the Hasler Foundation (project number 13027).

[1] http://youtu.be/Df9WhgibCQA, http://youtu.be/QTKd5UWCG0Q

robot in order to minimize the motion drift along a predefined trajectory. Vidal Calleja et al. [28] demonstrated an active feature-based visual SLAM framework that provides real-time user-feedback to minimize both map and camera pose uncertainty. Bryson and Sukkarieh [6] demonstrated a similar visual and inertial EKF-SLAM formulation for active control of flying vehicles. The goal was to cover a predefined area with a camera sensor while maintaining an accurate estimation of both the map and the vehicle state. Extensive simulation results were provided of a MAV that is restricted to fly on a plane. Similar to [28, 6] the exploration in our algorithm is driven by a set of states (i.e., dense depth estimates in the reference view) that are initialized with high uncertainty at the start of the exploration. Our resulting map is spatially smaller but denser and exhibits more detail, which is crucial e.g., for path planning in cluttered environments. Furthermore, in [28, 6] the image is only used to extract features and subsequently neglected. On the other hand, our proposed approach is *direct* [14]—the intensity values in the image are directly used to reason about the next best view.

In [23], Soatto introduces the notion of *Actionable Information* that is the portion of data that is useful towards the accomplishment of a task and after discounting nuisance factors. In [24, Chapter 8], he describes a hypothetic greedy explorer that tries at every time instant to maximize the *Actionable Information Increment* (AIN). He argues that such an explorer can get stuck in a local minima where no control action yields any information and, therefore, suggests two improvements: firstly, to plan a trajectory that maximizes the AIN over a *finite horizon*. Secondly, to use the *memory* of past observations to build a representation of the environment and to plan the trajectory so as to minimize the uncertainty in this representation. Soatto recognizes that it is trivial to design an explorer that achieves complete exploration of a static environment as, for instance, a random explorer (Brownian motion) would asymptotically do so. However, the goal is to do so *efficiently*. In this work we present an implementation of such an explorer for monocular, dense depth estimation.

B. Contributions and Outline

State-of-the-art approaches to active mapping [15, 5, 8, 25, 27, 22] retain only geometric information while discarding the scene appearance. As a result, a robot trying to perceive the depth of a white wall, would generate different camera trajectories in vain, eventually failing to reduce the uncertainty in the depth measurement [23]. By contrast, we propose a method to compute the measurement uncertainty and, thus, the expected information gain, on the basis of scene structure *and* appearance (i.e., texture). By doing so, surfaces characterized by uniform intensity yield high uncertainty in stereo computation, thus encoding the fact that there is no information to obtain from staring at white walls.

The contributions introduced by this paper can be summarized as follows.

- We propose a formulation of the uncertainty characterizing a depth measurement from multi-view stereo that takes into account the appearance of the scene, the motion of the camera, and the structure of the scene currently available. This formulation is used to evaluate candidate camera poses on the basis of the expected information gain.
- For applications to dense reconstruction from MAVs, we provide a strategy to compute a candidate sequence of viewpoints that lie on a feasible trajectory and that maximize the expected information gain.
- We detail both synthetic and experimental validation of the proposed system in closed loop and compare against four different control strategies: a random strategy, a circular motion, a greedy strategy and a Next-Best-View (NBV) strategy that iteratively selects the globally optimal view points.

The outline of the paper follows. In Section II we detail our method to compute probabilistic depth maps from a moving camera, introduce our evaluation method and optimality criterion. Section III presents different strategies for the generation of candidate trajectories and Section IV is dedicated to the discussion on the experimental evaluation. Finally, in Section V, we summarize our contribution and draw the conclusions.

II. Probabilistic Monocular Depth Estimation

In this section, we formalize the recursive Bayesian estimation of depth from multi-view stereo, focusing on the measurement uncertainty, which is the crucial factor for planning informative trajectories.

We denote the intensity image collected at time step k as $\mathbf{I}_k : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$, where Ω is the image domain. Let the rigid-body transformation $\mathbf{T}_{w,k} \in SE(3)$ describe the pose of the camera acquiring $\mathbf{I}_k$ in the *world* reference frame. The inverse depth $\hat{d}_{\mathbf{u}}$ of a pixel $\mathbf{u}$ in the *reference* camera pose $\mathbf{T}_{w,r}$ is a latent variable we infer from observations. An observation is a pair $\{\mathbf{I}_k, \mathbf{T}_{w,k}\}$, where we assume that $\mathbf{T}_{w,k}$ is computed by an accurate visual odometry algorithm [11]. A measurement $d_{\mathbf{u},k}$ of pixel $\mathbf{u}$ is obtained by the k-th observation by triangulating from $\mathbf{T}_{r,k} = \mathbf{T}_{w,r}^{-1} \cdot \mathbf{T}_{w,k}$ and we assume it normally distributed with mean $\mu_{\mathbf{u},k}$ and variance $\tau_{\mathbf{u},k}^2$:

$$p(d_{\mathbf{u},k}|\hat{d}_{\mathbf{u}}) = \mathcal{N}(d_{\mathbf{u},k}|\mu_{\mathbf{u},k}, \tau_{\mathbf{u},k}^2). \tag{1}$$

Given a prior $p(\hat{d}_{\mathbf{u}})$ and assuming independent and identically distributed measurements, the estimation proceeds recursively from the observations $k \in \{r+1, \dots, n\}$:

$$p(\hat{d}_{\mathbf{u}}|d_{\mathbf{u},r+1}, \dots, d_{\mathbf{u},n}) \propto p(\hat{d}_{\mathbf{u}}) \prod_{k=r+1}^{n} p(d_{\mathbf{u},k}|\hat{d}_{\mathbf{u}}). \tag{2}$$

Upon the k-th observation, the posterior (2) is normally distributed with parameters computed from the estimation at

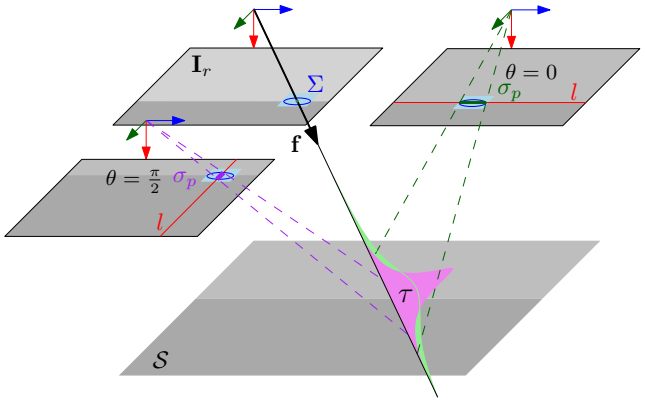


Figure 1. Disparity uncertainty. Depending on the image gradient, the camera motion influences the reliability of stereo matching and, thus, the uncertainty in the disparity computation σ_p^2.

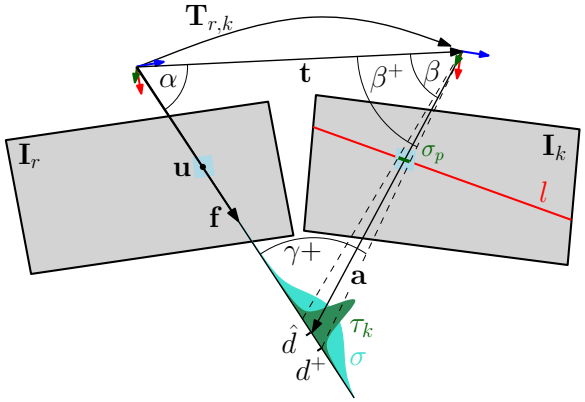


Figure 2. The uncertainty in depth measurement, τ_k^2, is computed by projecting the disparity uncertainty σ_p in image $\mathbf{I}_k$ on the pixel bearing-vector $\mathbf{f}$.

time $k-1$:

$$
\begin{aligned}
p(\hat{d}_{\mathbf{u}}|d_{\mathbf{u},r+1},\dots,d_{\mathbf{u},k}) &= \mathcal{N}(\hat{d}_{\mathbf{u}}|\mu_{\mathbf{u},k},\sigma^2_{\mathbf{u},k}),\\
\mu_{\mathbf{u},k} &= \frac{\sigma^2_{\mathbf{u},k-1}\mu_{\mathbf{u},k}+\tau^2_{\mathbf{u},k}d_{\mathbf{u},k-1}}{\sigma^2_{\mathbf{u},k-1}+\tau^2_{\mathbf{u},k}},\\
\sigma^2_{\mathbf{u},k} &= \frac{\sigma^2_{\mathbf{u},k-1}\tau^2_{\mathbf{u},k}}{\sigma^2_{\mathbf{u},k-1}+\tau^2_{\mathbf{u},k}}.
\end{aligned}
\tag{3}
$$

A similar model to estimate the depth of a pixel is used in [19, 29]. To increase the robustness of this approach, it is proposed in [29] to explicitly model outliers. Furthermore, in [19], we showed how regularity in the depth map can be enforced by making use of a smoothness prior in regions characterized by high uncertainty.

A. Measurement uncertainty

A camera is a passive sensor and the measurement uncertainty is a function of the depth, the camera motion, and the scene texture. In this section, we detail how to compute the measurement uncertainty τ_k related to a candidate camera motion $\mathbf{T}_{r,k}$, starting from estimating the photometric disparity uncertainty $\sigma_{p,k}$, which accounts for possible ambiguities in epipolar matching (e.g., due to uniform texture), and propagating it through triangulation to the depth uncertainty τ_k.

The disparity error accounts for uncertainty in disparity measurement given the reference image appearance $\mathbf{I}_r$ and the camera motion $\mathbf{T}_{r,k}$. It encapsulates the fact that some motions are better than others to compute the disparity related to a pixel. Indeed, the camera motion determines the direction of the epipolar line l and the disparity measurement relies on comparison of intensity patches. Intuitively, matching is reliable for image patches characterized by strong intensity gradients; in the context of active vision, this means that the direction of the gradient in a region must be considered in order to select a motion that is suitable for the disparity estimation. For instance, when reconstructing regions characterized by a dominant gradient direction (see Figure 1), a camera motion resulting into epipolar lines that are parallel to the dominant gradient direction in the intensity image (e.g., motion to the right in Figure 1) will result in less reliable epipolar matches and, thus, higher uncertainty in the disparity σ_p and subsequently in depth τ.

More precisely, when the *sum of squared differences* (SSD) between image patches is used for stereo matching, the probability of a correct match in the neighbourhood of a pixel can be expressed by a zero mean bivariate normal distribution [16], with covariance matrix

$$
\mathbf{\Sigma} = 2\sigma_i^2(\mathbf{J}\mathbf{J}^\top)^{-1}, \tag{4}
$$

where we denote by σ_i^2 the variance of the image noise and by $\mathbf{J} = \sum_P (\partial\mathbf{I}/\partial x, \partial\mathbf{I}/\partial y)$ the sum of the image gradients over a patch P, centered at the pixel of interest.

We now take into account the camera motion and derive the uncertainty of disparity computation when matching is performed along the epipolar line generating from $\mathbf{T}_{r,k}$. Let θ be the angle formed by the epipolar line and the image x axis. We can transform the probability of a correct match to a reference system that has the x axis aligned with the epipolar line, which results in a covariance matrix

$$
\mathbf{\Sigma}' = \left(\mathbf{R}^\top\mathbf{\Sigma}^{-1}\mathbf{R}\right)^{-1}, \quad \mathbf{R} = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}. \tag{5}
$$

The disparity error along the epipolar line follows the conditional distribution $p(x|y=0)$, which is Gaussian and characterized by the variance (cfr. [3, p.87])

$$
\sigma_p^2 = \Sigma'_{xx} - \Sigma'_{xy}\Sigma'^{-1}_{yy}\Sigma'_{yx}, \tag{6}
$$

where Σ'_{xx}, Σ'_{xy} and Σ'_{yy} are the entries of $\mathbf{\Sigma}'$.

Thus, the disparity error is normally distributed along the epipolar line with variance

$$
\sigma_p^2 = \frac{|\mathbf{\Sigma}|}{\Sigma_{xx}\sin^2(\theta) + 2\Sigma_{xy}\sin(\theta)\cos(\theta) + \Sigma_{yy}\cos^2(\theta)}, \tag{7}
$$

where Σ_{xx}, Σ_{xy} and Σ_{yy} correspond to entries of $\mathbf{\Sigma}$ and $|\mathbf{\Sigma}|$ is the determinant of $\mathbf{\Sigma}$.

In the active vision context, we cannot compute the disparity error on the new image, as the image is not available at the

time we predict the measurement uncertainty. Therefore, we consider the epipolar line in the reference image and compute the disparity error therein. The assumption that the patch appearance can be predicted by the reference patch is valid for small viewpoint changes (i.e., neglecting distortions and occlusions).

The measurement variance of the depth at pixel $\mathbf{u}$ in the image $\mathbf{I}_k$ is obtained by back-projecting the variance of the photometric disparity error σ_p^2. Referring to Figure 2, let $\hat{d}$ be the current depth estimation of pixel $\mathbf{u}$, the corresponding unit bearing vector is denoted as $\mathbf{f}$ and $\mathbf{t}$ denotes the translation component of the relative position $\mathbf{T}_{r,k}$. As proposed in [19], we can transform the measurement uncertainty σ_p^2 in the image to the depth uncertainty τ_k^2 as follows:

$$\mathbf{a} = \hat{d} \cdot \mathbf{f} - \mathbf{t} \tag{8}$$

$$\alpha = \arccos\left(\frac{\mathbf{f} \cdot \mathbf{t}}{||\mathbf{t}||}\right) \tag{9}$$

$$\beta = \arccos\left(-\frac{\mathbf{a} \cdot \mathbf{t}}{||\mathbf{a}|| \cdot ||\mathbf{t}||}\right). \tag{10}$$

Let f be the camera focal length. The angle spanning σ_p pixels can be added to β in order to compute γ^+ and, thus, by applying the law of sines, recover d^+:

$$\beta^+ = \beta + 2\tan^{-1}\left(\frac{\sigma_p}{2f}\right) \tag{11}$$

$$\gamma^+ = \pi - \alpha - \beta^+ \tag{12}$$

$$d^+ = ||\mathbf{t}|| \frac{\sin\beta^+}{\sin\gamma^+}. \tag{13}$$

Therefore, the measurement uncertainty is computed as:

$$\tau_k^2 = \left(d^+ - \hat{d}\right)^2. \tag{14}$$

The derivation of the depth uncertainty reported in Equations (8) - (14) is similar to the one presented in [19], however with one critical difference that occurs in Equation (11). In the present paper, the disparity uncertainty σ_p is a function of the *appearance* (i.e., texture) in the scene. In contrast, in [19] this was simply set to 1, meaning that the uncertainty was assumed independent of the scene appearance.

B. The Information Gain of a Measurement

We now demonstrate how the proposed probabilistic depth map representation and update method can be applied to the problem of selecting the next best placements for the camera.

Suppose that we are computing the depth map for a given reference image $\mathbf{I}_r$. We describe the uncertainty in the depth map estimate at time k with the entropy $\mathcal{H}_k$. In such a way, the treatment is independent on the actual model and the parametric formulation described in Section II might be replaced in order to take into account, for instance, multiple depth hypotheses [30].

Since, for every pixel $\mathbf{u} \in \Omega$, the depth estimation proceeds independently, $\mathcal{H}_k$ can be computed as (see, for instance, [3])

$$\mathcal{H}_k = \frac{1}{2}\sum_{\mathbf{u}\in\Omega} \ln(2\pi e\ \sigma_{\mathbf{u},k}^2), \tag{15}$$

where $\sigma_{\mathbf{u},k}^2$) denotes the depth uncertainty of pixel $\mathbf{u}$ at time k (see Eq. (3)).

Upon the acquisition of a measurement from the $(k+1)$-th camera pose $\mathbf{T}_{r,k+1}$, the variance of the estimated depth for the pixel $\mathbf{u}$ is updated to take into account the measurement uncertainty $\tau_{\mathbf{u},k+1}^2$.

We define the *information gain* as the difference

$$\mathcal{I}_{k,k+1} = \mathcal{H}_k - \mathcal{H}_{k+1}, \tag{16}$$

which, plugging (3) into (15), yields

$$\mathcal{I}_{k,k+1} = \frac{1}{2}\sum_{\mathbf{u}\in\Omega} \ln\left\{\frac{\tau_{\mathbf{u},k+1}^2 + \sigma_{\mathbf{u},k}^2}{\tau_{\mathbf{u},k+1}^2}\right\}. \tag{17}$$

III. Solution Strategies

In this section, we describe five different control strategies for the active depth-map estimation problem. The control strategies range from random, heuristic, and greedy methods to a model-predictive control approach that optimizes the next N views to maximize the information gain. In Section IV, we will evaluate the proposed methods in synthetic and real-world experiments.

We simplify the problem by assuming that the camera moves at constant speed and takes measurements at fixed frame rate. This results in equidistant measurements with a relative distance $\Delta\mathbf{t} \in \mathbb{R}^3$ that is fixed a priori. The proposed system can be extended to incorporate the inertia, controllability, and the dynamics of the camera-equipped robot.

One can obtain more precise, thus more informative, measurements closer to the surface. Therefore, an optimal control strategy eventually would make the robot approach the surface (see Figure 3 (b)). To avoid collisions in our envisioned MAV application, we additionally restrict the motion to the horizontal plane $\mathcal{Z}$ at the height of the reference view. Nevertheless, all proposed solution strategies can be extended to the 3D space with increased computational cost that comes with the enlarged action space.

With these assumptions we can formalize the problem as follows: given the current pose relative to the reference view $\mathbf{T}_{r,k}$ and the proposed method to measure the information gain of a measurement at the next pose $\mathcal{I}_{k,k+1} = \mathcal{I}_{k,k+1}(\mathbf{T}_{k,k+1})$, which next pose $\mathbf{T}_{r,k+1} \in \mathcal{A}_k$ should be selected? The action space at time k is defined such that equidistant camera poses in the horizontal plane are selected:

$$\mathcal{A}_k = \left\{\mathbf{T} \;\middle|\; ||\mathbf{T}_{r,k}^{-1} \cdot \mathbf{T}||_2 = \Delta\mathbf{t} \ \wedge\ \mathbf{T} \in \mathcal{Z}\right\}. \tag{18}$$

A. Random Walk Control

Similar to [23], we use as a baseline a random walk strategy that at every measurement k selects randomly the next pose from the action space $\mathcal{A}_k$. This approach is completely blind, hence should perform worse than all of the following strategies.

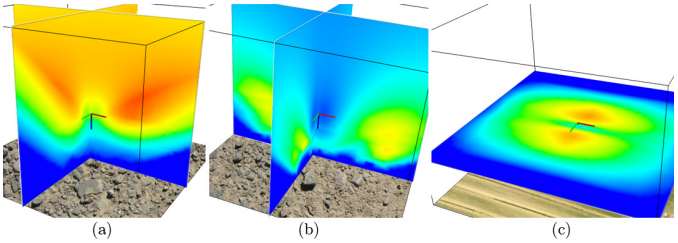


Figure 3. Information gain for the NBV strategy. The distributions are visualized as heat-maps (red means high information gain, blue low). Figure (a) shows the information gain before the first measurement in an environment of isotropic gravel texture. Figure (b) shows the information gain after the 10th measurement in the same scene. Figure (c) shows the information gain in an environment with a dominant gradient direction in the texture.

B. Circular Heuristic Control

A circular trajectory guarantees that the epipolar line sweeps over all directions. Thereby, depth uncertainties that arise from the aperture problem during triangulation can be disambiguated. For this reason, a circular trajectory is intuitively a good heuristic and typically used in demonstrations of monocular multi-view stereo systems [17]. However, the radius of the circle must be tuned to the depth of the scene. The radius should trade off accuracy through increased base-line versus visibility of the reconstructed surface area $\mathcal{S}$. In the synthetic experiments we selected the radius to give the best results in the first scene and kept the radius fixed for the other experiments.

C. Greedy Control

A greedy controller tries to take control actions so as to maximize the expected information gain of the next measurement [10]. The greedy control can then be written as follows:

$$\mathbf{T}_{r,k+1} = \arg\max_{\mathbf{T}\in\mathcal{A}_k} \mathcal{I}^*_{k,k+1}(\mathbf{T}). \tag{19}$$

This control law is equal to a gradient descent algorithm with fixed step size. Unless the underlying functional is convex or the cost is extended with an additional *curiosity*-term that promotes exploration of unknown areas [5], this approach is prone to get stuck in local minima.

D. Next-Best-View Control

Since the information gain proposed in Section II-B can be evaluated not only in the neighbourhood of the current pose but also for all feasible positions and orientations, the NBV control always selects the viewpoint in the horizontal plane $\mathcal{Z}$ that provides the highest information gain, independently of the current pose:

$$\mathbf{T}_{r,k+1} = \arg\max_{\mathbf{T}\in\mathcal{Z}} \mathcal{I}^*_{k,k+1}(\mathbf{T}). \tag{20}$$

Thus, the translation between poses is not limited to $\Delta\mathbf{t}$ anymore. Since there is no guarantee that subsequent measurements are spatially close, the travel distance of this approach between two measurements will be high.

E. Receding-Horizon Control

Let us assume that the position of the next N poses $\{\mathbf{T}_{r,k+1},\dots,\mathbf{T}_{r,k+N}\}$ can be parametrized by the parameter vector $\boldsymbol{\phi}_k$ such that each pose lies in the action space of the previous pose: $\mathbf{T}_{r,k+i}\in\mathcal{A}_{k+i-1}$. We can improve the greedy control strategy by considering the information gain over the finite horizon N as proposed in [12, 24]. Given the current frame k, the receding-horizon control maximizes the expected information gain over the course of the next N views:

$$\boldsymbol{\phi}_k = \arg\max_{\boldsymbol{\phi}} \sum_{i=k}^{k+N} \mathcal{I}^*_{i,i+1}(\boldsymbol{\phi}). \tag{21}$$

One can predict the probability of a measurement at time $k+1$ based on the uncertainty in the current depth-map. To compute the expected measurement at time $k+2$ would require to integrate over all possible depth-maps that can result from the update at $k+1$. This problem can be formulated with a partially observable Markov decision process (POMDP [18]) which becomes intractable with high state- and action-spaces.

However, as proposed in [12], we can make the assumption that the next measurements do not provide any new *evidence*, meaning that the prediction coincides with the measurement and thus, the mean of the estimate does not change. With this assumption it is straightforward to compute the information gain over the next N measurements:

$$\begin{aligned} &\mathcal{I}^*_{k,k+N} = \mathcal{H}_k - \mathcal{H}^*_{k+N}({\sigma^*}^2_{k+N}), \qquad \text{with} \\ &\frac{1}{{\sigma^*}^2_{k+N}} = \frac{1}{\sigma^2_k} + \frac{1}{{\tau^*_{k+1}}^2(\boldsymbol{\phi}_k)} + \dots + \frac{1}{{\tau^*_{k+N}}^2(\boldsymbol{\phi}_k)}, \end{aligned} \tag{22}$$

where ${\tau^*_{k+i}}^2(\boldsymbol{\phi}_k)$ is the predicted measurement uncertainty at pose $\mathbf{T}_{r,k+i}$ that is a function of the trajectory parameters $\boldsymbol{\phi}_k$.

Increasing the prediction horizon N in this formulation makes sense only when the depth uncertainty is not too high, since this approach is based on the assumption that the mean of the current depth estimate does not change over the next N measurements. Further, note that similarly to the greedy approach, there is no guarantee that this approach does not fall into a local minima.

A heuristic that we apply in order to increase the prediction accuracy in uncertain depth maps and to avoid local minima is to start with a short prediction horizon $N=3$ when the map is uncertain and to increase the prediction horizon when the predicted information gain $\mathcal{I}^*_{k,k+N}$ falls below some threshold in order to escape local minima. Furthermore, since the depth estimate changes as soon as the $(k+1)$-th measurement is acquired, the trajectory until measurement $N+1$ is replanned immediately.

The computational demand of the prediction grows exponentially with the degrees of freedom of the trajectory parameters $\boldsymbol{\phi}$ and linearly with the prediction horizon N.

F. Implementation Details

In this section, we provide more details on our implementation of the receding-horizon control strategy and the information gain computation.

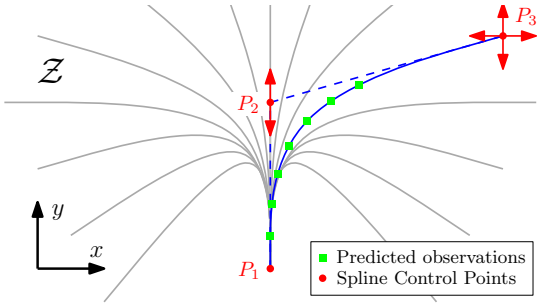


Figure 4. B-Spline trajectory parametrization. P_1, P_2 and P_3 are the control points. The candidate camera poses are visualized in green.

To favor the dynamics of the MAV, we reduce the dimensionality of the action space by enforcing the continuity of the trajectory and by setting the tangent at the current position to the current direction of motion. Additionally, we prohibit yaw camera rotations in order to minimize motion blur. We chose to parametrize the trajectory with a B-spline [9] of third-order with three control points (see Figure 4). B-splines are piecewise polynomial functions with local support and simple derivatives. However, any other temporal basis function could be used. The first control point P_1 of the B-spline is set fixed to the current position of the camera, the second control point P_2 has one degree of freedom (P_{2y}) along the current direction of motion of the MAV and the third control point P_3 has two degrees of freedom in the horizontal plane $\mathcal{Z}$ (P_{3x} and P_{3y}, see Figure 4). In total the trajectory parametrization has three degrees of freedom $\boldsymbol{\phi} = \{P_{2y}, P_{3x}, P_{3y}\}$. By setting constraints on the position of $\{P_1, P_2\}$, it is possible to enforce the dynamic constraints of the MAV on the trajectory. The predicted observations are located along the trajectory with equal distance $\Delta\mathbf{t}$. The optimal trajectory in the three dimensional space can be found by a global optimization routine with the condition that the spline parameters $\boldsymbol{\phi}$ must remain in the range $\pm 2N\Delta\mathbf{t}$.

The computation of the depth-map entropy, which is evaluated multiple times in every control iteration according to (15), requires summation over all pixels in the image. To maintain real-time performance, we were required to select a subset of pixels for which the information gain is computed. In practice we compute the information gain, thus, the trajectory, based on 400 uniformly distributed pixels with high gradient magnitude.

IV. Experimental Evaluation

A. Simulation Experiments

We evaluated the proposed control strategies in three different synthetic environments (Figure 6 to 8). The scenes vary in both the texture and shape of the surface. Scenes 1 and 2 contain isotropic gravel texture while the texture of Scene 3 exhibits a dominant gradient direction. The surface in Scene 1 and 3 is planar and in Scene 3 there is a step.

To give an intuition of the information distribution, we sampled the information gain regularly in a cube around the reference view and display the results in Figure 3. The information density before the first measurement in Scene 1 is displayed in Figure 3 (a) and after the 10th measurement in Figure 3 (b). The coordinate frame displayed in the center of the figures illustrates the position of the downward-looking reference view. Hot (red) colors indicate relative positions with high expected information gain and cold (blue) colors positions with low potential. Neglecting the restriction of the motion to a horizontal plane for now, one can observe that for the first measurement a horizontal and vertical motion would be optimal. Moving horizontally increases the baseline and moving vertically ensures that the whole surface remains within the field of view. This illustrates intuitvely why planning multiple steps ahead is superior to next-best-view planning: rather than moving upwards and ensuring that the whole depth-map is within the field of view, two close measurements—each updating one side of the depth-map—would result in higher uncertainty reduction. Figure 3 (b) shows that after a 10 measurements, the information-gain is generally lower and that it is advantageous to move closer to the surface. Figure 3 (c) shows the initial cost in the horizontal plane of Scene 3. Scenes with isotropic texture exhibit a circular region around the reference view with high information gain. However, since Scene 3 is textured with a dominant gradient direction, the photometric disparity error is higher for motions along the gradient direction (aperture problem). This reflects in the information gain computation and thus motions rectangular to the gradient directions are favoured.

Figure 9 shows the information gain in the horizontal plane centered two meters above a horizontally striped surface. When neglecting the texture (i.e., $\sigma_p^2 = 1$), the robot would prefer a horizontal motion since less pixels move out of the field of view. However, when considering the appearance, a motion in x direction does not provide any information due to the aperture problem.

The plots in Figures 6 to 8 compare the proposed control strategies for each of the synthetic environments. The simulation of all control strategies was run until an accuracy of less than 1 mm in the depth-map was reached. The red plane in each rendering illustrates the altitude to which the camera was resticted to move. The reference view is acquired in the center of each red plane with a downward-looking camera. Plot (b) in each figure shows the resulting trajectories on the horizontal plane for all control strategies while Plot (c) shows the entropy reduction over travelled distance. When comparing the information gain over the travelled distance in Plot (c), the greedy approach performs similar to the spline-based method in terms of entropy reduction over travelled distance in the first environment. However, in the second and third environment, the greedy approach gets stuck in a local minimum. The spline-based receding-horizon control requires in all environments the least motion to achieve the predefined accuracy level. The results of the random strategy are averaged over 100 measurements of which we display only one in the trajectory plots.

In Figure 8 (b) it is clearly visible how the photometric disparity uncertainty drives the receding-horizon control to select views which do not suffer from the aperture problem. After moving in positive y direction, the MAV seems to get stuck in a local minimum, however, by increasing the prediction horizon it finds the path towards the other side of the map.

B. Real-World Experiments

In Figure 5(a), we show the setup of the real experiments. The MAV is equipped with a downward-looking camera and embedded processor. A vision-based SLAM algorithm [11] runs onboard to estimate the egomotion and stabilize the vehicle. To achieve real-time performance, we run the dense reconstruction and path planning off-board on an Intel i7 laptop. Therefore, the MAV streams video and estimated poses to a ground-station where the proposed algorithms compute and return in real time the trajectory commands. A video of the experiment can be viewed on the author's website: http://rpg.ifi.uzh.ch.

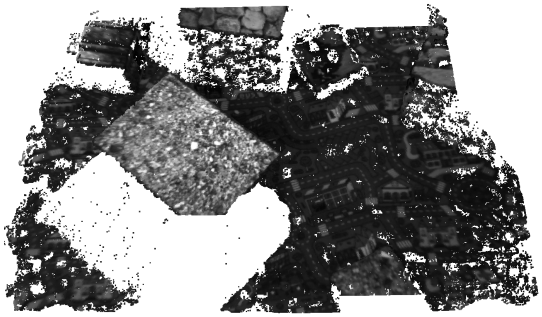
We compared the three best performing control strategies and report the results Figure 5(e). In Figure 5(d), the resulting trajectories are shown, where we additionally display the B-splines that are computed at every iteration. The final depth-map of the spline strategy is shown in Figures 5(b) and 5(c).

A comparison of the control strategies in real experiments is more challenging than in simulation since the reference view must be taken exactly at the same location, which is almost impossible. For this reason, the comparison of the convergence speed must be analyzed with caution. The greedy method fell in a local minimum and approached a wall when the experiment had to be stopped. For the circle strategy we tuned the radius to give best performance in this scenario. Indeed, it converges slightly faster than the receding-horizon (spline) strategy. The advantage of the spline strategy, however, is that it must not be adapted to the environment height, shape and appearance.

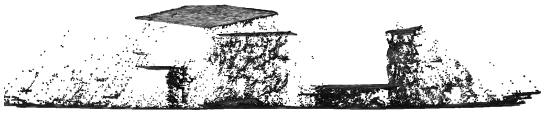
V. Conclusion and Future Work

In this paper, we proposed an approach to actively acquire informative views for monocular dense depth estimation. In evaluating a candidate camera trajectory, we proposed to take into account the texture of the scene, and we contributed a novel formulation of the depth measurement uncertainty based on propagating the uncertainty in photometric stereo disparity to triangulation. We evaluated different strategies in both simulation and real scenarios and we showed how the camera trajectories emerging from the information maximization problem are, at the same time, *informative*, in terms of depth estimation, and *parsimonious*, in terms of traveled distance. For applications to Micro Aerial Vehicle (MAV) perception, we reduced the dimensionality of the search space by enforcing continuity on the trajectory. To the best of our knowledge, this is the first work on active, monocular *dense* reconstruction demonstrated on a robot.

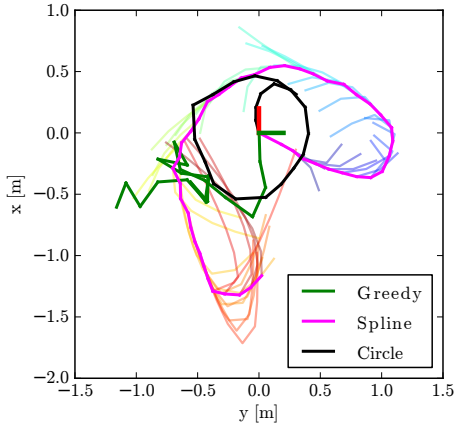
(a) Experiment Setup



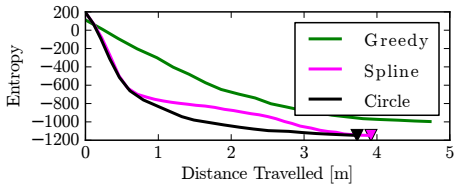
(b) Reconstruction result



(c) Reconstruction result



(d) Reconstruction trajectory



(e) Entropy reduction over travelled distance

Figure 5. Real world experiment and reconstruction results.

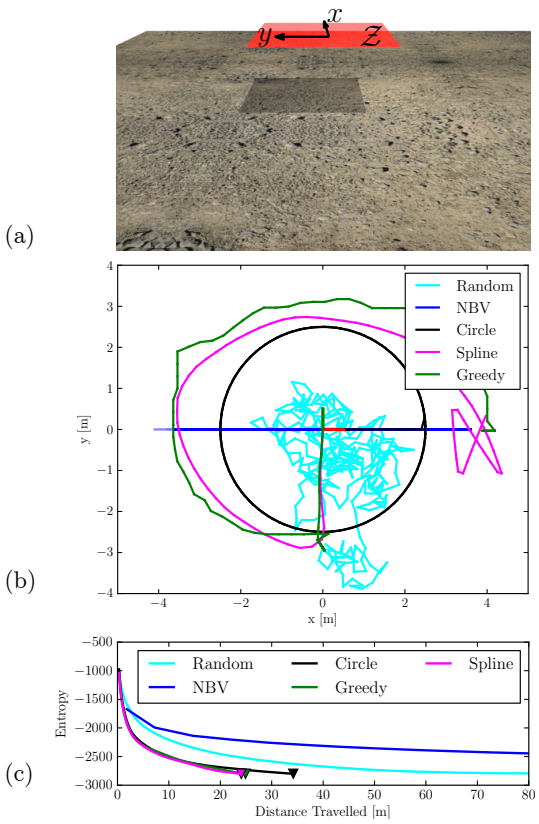


Figure 6. Synthetic scene 1.

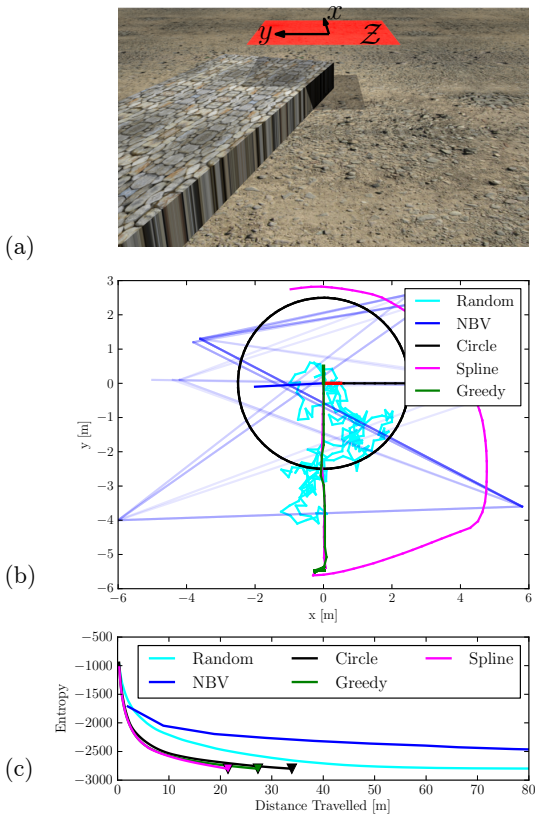


Figure 7. Synthetic scene 2.

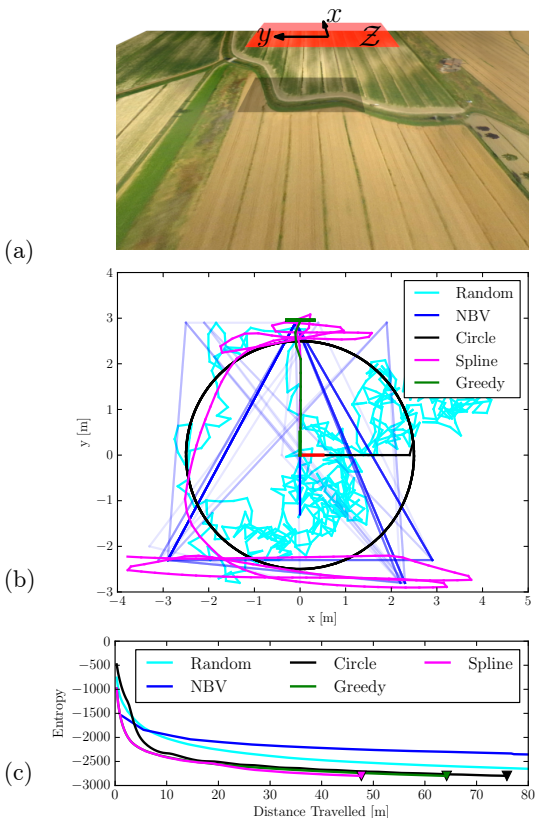
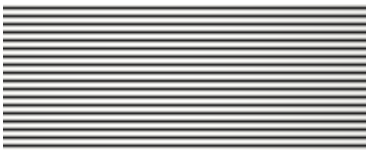
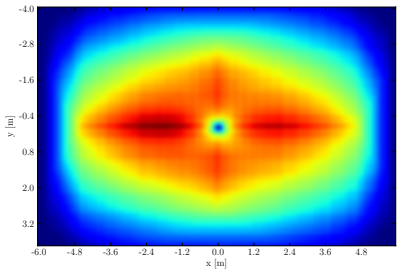


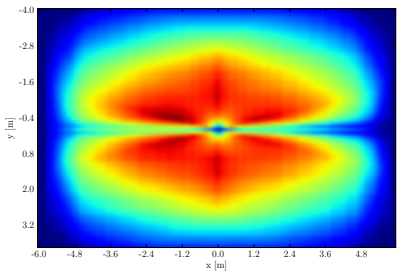
Figure 8. Synthetic scene 3.



(a) Horizontally striped texture.



(b) Information gain neglecting the texture.



(c) Information gain using the texture.

Figure 9. Influence of striped texture on the information gain.

REFERENCES

[1] John Aloimonos, Isaac Weiss, and Amit Bandyopadhyay. Active Vision. *International Journal of Computer Vision*, 1(4), 1988.

[2] Ruzena Bajcsy. Active Perception. In *Proceedings of the IEEE*, volume 76, 1988.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., 2006.

[4] Andrew Blake and Alan Yuille. *Active Vision*. the MIT Press, 1988.

[5] Frédéric Bourgault, Alexei A. Makarenko, Stefan B. Williams, Ben Grocholsky, and Hugh f. Durrant-Whyte. Information Based Adaptive Robotic Exploration. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, 2002.

[6] Mitch Bryson and Salah Sukkarieh. Observability Analysis and Active Control for Airborne SLAM. *IEEE Transactions on Aerospace and Electronic Systems*, 44 (1), 2008.

[7] Shengyong Chen, Youfu Li, and Ngai Ming Kwok. Active vision in robotic systems: A survey of recent developments. *International Journal of Robotics Research*, 30(11), 2011.

[8] Andrew J. Davison and Richard M. Murray. Simultaneous Localization and Map-Building Using Active Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7), 2002.

[9] C. de Boor. *A practical guide to splines*. Springer Verlag New York, 2001.

[10] Hans Jacob S. Feder, John J. Leonard, and Christopher M. Smith. Adaptive Mobile Robot Navigation and Mapping. *The International Journal of Robotics Research*, 18(7):650–558, 1999.

[11] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: Fast Semi-Direct Monocular Visual Odometry. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2014.

[12] Shoudong Huang, N. M. Kwok, Gamini Dissanayake, Q. P. Ha, and Gu Fang. Multi-Step Look-Ahead Trajectory Planning in SLAM: Possibility and Necessity. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2005.

[13] Xiaoxia Huang, Ian Walker, and Stan Birchfield. Occlusion-Aware Reconstruction and Manipulation of 3D Articulated Objects. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2012.

[14] M. Irani and P. Anandan. All About Direct Methods. In *Proc. Workshop Vis. Algorithms: Theory Pract.*, 1999.

[15] Simon Kriegel, Tim Bodenmüller, Michael Suppa, and Gerd Hirzinger. A Surface-Based Next-Best-View Approach for Automated 3D Model Completion of Unknown Objects. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2011.

[16] Larry Matthies, Richard Szeliski, and Takeo Kanade. Incremental estimation of dense depth maps from image sequences. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 1988.

[17] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. DTAM: Dense Tracking and Mapping in Real-Time. *Proc. IEEE Int. Conf. on Computer Vision*, 2011.

[18] Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101, 1998.

[19] Matia Pizzoli, Christian Forster, and Davide Scaramuzza. REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2014.

[20] Korbinian Schmid, Heiko Hirschmüller, Andreas Dömel, Iris Grixa, Michael Suppa, and Gerd Hirzinger. View planning for multi-view stereo 3d reconstruction using an autonomous multicopter. *Journal of Intelligent and Robotic Systems*, 65(1-4), 2012.

[21] William R. Scott, Gerhard Roth, and Jean-François Rivest. View Planning for Automated Three-dimensional Object Reconstruction and Inspection. *ACM Computing Surveys*, 35(1), 2003.

[22] Robert Sim and Nicholas Roy. Global A-Optimal Robot Exploration in SLAM. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2005.

[23] Stefano Soatto. Actionable Information in Vision. In *Proc. IEEE Int. Conf. on Computer Vision*, 2009.

[24] Stefano Soatto. Steps Towards a Theory of Visual Information: Active Perception, Signal-to-Symbol Conversion and the Interplay Between Sensing and Control. *ArXiv e-prints*, 2011.

[25] Cyrill Stachniss, Giorgio Grisetti, and Wolfram Burgard. Information Gain-based Exploration Using Rao-Blackwellized Particle Filters. In *Robotics: Science and Systems*, 2005.

[26] Jan Stühmer, Stefan Gumhold, and Daniel Cremers. Real-Time Dense Geometry from a Handheld Camera. In *Pattern Recognition (Proc. DAGM)*, 2010.

[27] Rafael Valencia, Jaime Valls Miró, Gamini Dissanayake, and Juan Andrade-Cetto. Active Pose SLAM. In *Proc. IEEE Int. Conf. on Robotics and Automation*, 2012.

[28] Teresa A. Vidal-Calleja, Alberto Sanfeliu, and Juan Andrade-Cetto. Action Selection for Single-Camera SLAM. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(6), 2010.

[29] George Vogiatzis and Carlos Hernández. Video-based, Real-Time Multi View Stereo. *Image and Vision Computing*, 29(7), 2011.

[30] Andreas Wendel, Michael Maurer, Gottfried Graber, Thomas Pock, and Horst Bischof. Dense reconstruction on-the-fly. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[31] Peter Whaite and Frank P. Ferrie. Autonomous Exploration: Driven by Uncertainty. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(3):193–205, 1997.