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 time, monocular reconstruction from a micro aerial vehicle and 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 time, monocular reconstruction from a micro aerial vehicle and uncertainty that accounts for the scene appearance (i.e., texture in the reference view), the scene depth and the vehicle pose.

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. 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] [3] [4] [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
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] the image is only used to extract features and subsequently neglected. On the other hand, our proposed approach is 

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

$$p(d_{u,k}|\hat{d}_u) = \mathcal{N}(d_{u,k}|\mu_{u,k},\tau^2_{u,k}).$$  \hspace{1cm} (1)$$

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

$$p(d_{u}|d_{u,r+1}, \ldots, d_{u,n}) \propto p(\hat{d}_u) \prod_{k=r+1}^{n} p(d_{u,k}|\hat{d}_u).$$  \hspace{1cm} (2)$$

Upon the $k$-th observation, the posterior (2) is normally distributed with parameters computed from the estimation at
time \( k - 1 \):

\[
p(d_{a, k + 1}, \ldots, d_{a, k}) = N(d_{a, k}, \mu_{a, k}, \sigma_{a, k}^2),
\]

\[
\mu_{a, k} = \frac{\sigma_{a, k}^2 d_{a, k} - \sigma_{x}^2 d_{a, k - 1}}{\sigma_{x}^2 + \sigma_{a, k}},
\]

\[
\sigma_{a, k}^2 = \frac{\sigma_{x}^2}{\sigma_{a, k}^2 + \sigma_{x}^2},
\]

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 \( \tau_k \) related to a candidate camera motion \( T_{r, k} \), 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 \( \tau_k \).

The disparity error accounts for uncertainty in disparity measurement as shown in Figure 1, and the camera motion \( T_{r, k} \). 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 \( \sigma_p \) and subsequently in depth \( \tau \).

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 neighborhood of a pixel can be expressed by a zero mean bivariate normal distribution [16], with covariance matrix

\[
\Sigma = 2\sigma_i^2 (J J^T)^{-1},
\]

where we denote by \( \sigma_i^2 \) the variance of the image noise and by \( J = \sum_p (\partial I / \partial x, \partial I / \partial y) \) 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 \( T_{r, k} \). Let \( \theta \) 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

\[
\Sigma' = (R^T \Sigma^{-1} R)^{-1}, \quad R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}.
\]

The disparity error along the epipolar line follows the conditional distribution \( p(x | y = 0) \), which is Gaussian and characterized by the variance (cfr. [13, p.87])

\[
\sigma_p^2 = \Sigma'_{xx} - \Sigma'_{xy} \Sigma'^{-1} \Sigma'_{yx},
\]

where \( \Sigma'_{xx}, \Sigma'_{xy} \) and \( \Sigma'_{yy} \) are the entries of \( \Sigma' \).

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

\[
\sigma_p^2 = \frac{|\Sigma|}{\Sigma_{xx} \sin^2(\theta) + 2 \Sigma_{xy} \sin(\theta) \cos(\theta) + \Sigma_{yy} \cos^2(\theta)},
\]

where \( \Sigma_{xx}, \Sigma_{xy} \) and \( \Sigma_{yy} \) correspond to entries of \( \Sigma \) and \( |\Sigma| \) is the determinant of \( \Sigma \).

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 \( u \) in the image \( I_k \) is obtained by back-projecting the variance of the photometric disparity error \( \sigma_p^2 \). Referring to Figure 2, let \( d \) be the current depth estimation of pixel \( u \), the corresponding unit bearing vector is denoted as \( f \) and \( t \) denotes the translation component of the relative position \( T_{r,k} \). As proposed in [19], we can transform the measurement uncertainty \( \sigma_p^2 \) in the image to the depth uncertainty \( \tau_k^2 \) as follows:

\[
a = d \cdot f - t \tag{8}
\]

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

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

Let \( f \) be the camera focal length. The angle spanning \( \sigma_p \) pixels can be added to \( \beta \) in order to compute \( \gamma^+ \) and, thus, by applying the law of sines, recover \( d^+ \):

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

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

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

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 \( \sigma_p \) 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 \( I_r \). 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 III might be replaced in order to take into account, for instance, multiple depth hypotheses [30].

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

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

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

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

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

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

**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 \( \Delta t \in \mathbb{R}^3 \) 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 \( Z \) 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 \( T_{r,k} \) and the proposed method to measure the information gain of a measurement at the next pose \( \mathcal{I}_{k,k+1} = \mathcal{I}_{k,k+1}(T_{k,k+1}) \), which next pose \( T_{r,k+1} \in \mathcal{A}_k \) should be selected? The action space at time \( k \) is defined such that equidistant camera poses in the horizontal plane are selected:

\[
\mathcal{A}_k = \{ T \mid ||T_{r,k}^{-1} \cdot T||_2 = \Delta t \land T \in Z \}. \tag{18}
\]

**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.
fixed step size. Unless the underlying functional is convex
control always selects the viewpoint in the horizontal plane
but also for all feasible positions and orientations, the NBV
evaluated not only in the neighbourhood of the current pose
D. Next-Best-View Control
prone to get stuck in local minima. This approach is
promotes exploration of unknown areas [5], this approach is
maximize the expected information gain of the next measure-
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:
\[ \mathbf{T}_{r,k+1} = \arg \max_{\mathbf{T} \in \mathcal{A}_k} \mathcal{I}^*_{k,k+1}(\mathbf{T}). \] (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:
\[ \mathbf{T}_{r,k+1} = \arg \max_{\mathbf{T} \in \mathcal{Z}} \mathcal{I}^*_{k,k+1}(\mathbf{T}). \] (20)
Thus, the translation between poses is not limited to \( \Delta t \) 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 \( \{\mathbf{T}_{r,k+1}, \ldots, \mathbf{T}_{r,k+N}\} \) can be parametrized by the parameter vector \( \phi \), 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:
\[ \phi_k = \arg \max_\phi \sum_{i=k}^{k+N} \mathcal{I}^*_{i,i+1}(\phi). \] (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:
\[ \mathcal{I}^*_{k,k+N} = \mathcal{H}_k - \mathcal{H}_k^* + (\sigma^2_{k+1:N})^2, \] with
\[ \frac{1}{\sigma^2_{k+1:N}} = \frac{1}{\sigma^2_k} + \frac{1}{\tau^2_{k+1:N}(\phi_k)} + \cdots + \frac{1}{\tau^2_{k+1:N}(\phi_k)}, \] (22)
where \( \tau^2_{k+i}(\phi_k) \) is the predicted measurement uncertainty at pose \( \mathbf{T}_{r,k+i} \) that is a function of the trajectory parameters \( \phi_k \).
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 \( \phi \) 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.
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 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 $Z$ ($P_{3x}$ and $P_{3y}$), see Figure 4. In total the trajectory parametrization has three degrees of freedom $\phi = \{P_{2y}, P_{3x}, P_{3y}\}$. 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 $\Delta t$. The optimal trajectory in the three dimensional space can be found by a global optimization routine with the condition that the spline parameters $\phi$ must remain in the range $\pm 2N\Delta t$.

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 Figure 3 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 intuitively 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 restricted 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.
Figure 6. Synthetic scene 1.

Figure 7. Synthetic scene 2.

Figure 8. Synthetic scene 3.

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


