REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time

Conference Paper · May 2014

DOI: 10.1109/ICRA.2014.6907233

3 authors, including:

Christian Forster
University of Zurich
20 PUBLICATIONS **4,446** CITATIONS
SEE PROFILE

Davide Scaramuzza
University of Zurich
290 PUBLICATIONS **19,873** CITATIONS
SEE PROFILE

REMODE: Probabilistic, Monocular Dense Reconstruction in Real Time

Matia Pizzoli, Christian Forster and Davide Scaramuzza[1]

***Abstract*— In this paper, we solve the problem of estimating dense and accurate depth maps from a single moving camera. A probabilistic depth measurement is carried out in real time on a per-pixel basis and the computed uncertainty is used to reject erroneous estimations and provide live feedback on the reconstruction progress. Our contribution is a novel approach to depth map computation that combines Bayesian estimation and recent development on convex optimization for image processing. We demonstrate that our method outperforms state-of-the-art techniques in terms of accuracy, while exhibiting high efficiency in memory usage and computing power. We call our approach REMODE (REgularized MOnocular Depth Estimation). Our CUDA-based implementation runs at 30Hz on a laptop computer and is released as open-source software.**

SUPPLEMENTARY MATERIAL

The accompanying video and source code are available at: `http://rpg.ifi.uzh.ch/software`.

I. INTRODUCTION

We present a method to compute an accurate, three-dimensional reconstruction of the scene observed by a moving camera and provide, in real time, information about the progress and the reliability of the ongoing estimation process. This problem is highly relevant in robot perception, where cameras are valuable and widespread sensors. From a single moving camera, it is possible to collect appearance and range information about the observed three-dimensional scene. In a multi-view stereo setting, the uncertainty on the depth measurement depends on the noise affecting image formation, on the camera poses, and the scene structure. Knowing how these factors affect the measurement uncertainty, it is possible to achieve arbitrarily high levels of confidence by collecting measurements from different vantage points. Such a capability is particularly valuable in robotics. For instance, if the camera is mounted on a robotic arm, the available high level of mobility can be exploited to disambiguate scene details and occlusions at a wide range of distances. The monocular setting is also an appealing sensing modality for Micro Aerial Vehicles (MAVs), where strict limitations apply on payload and power consumption. In this case, the high agility turns the platform into a formidable depth sensor, able to deal with a wide depth range and capable of achieving arbitrarily high confidence in the measurement. Inevitably, this high flexibility comes at a cost. The pose of the camera must be known and its accuracy influences the reconstruction quality. For a camera, information resides in the changing of the intensity gradient and this modality naturally fails in presence of low informative scenes that produce untextured images. It is therefore crucial to know how reliable each measurement is.

[1]The authors are with the Robotics and Perception Group, University of Zurich, Switzerland—`http://rpg.ifi.uzh.ch`. This research was supported by the CTI project number 14652.1, the Swiss National Science Foundation through project number 200021-143607 (“Swarm of Flying Cameras”) and the National Centre of Competence in Research Robotics.

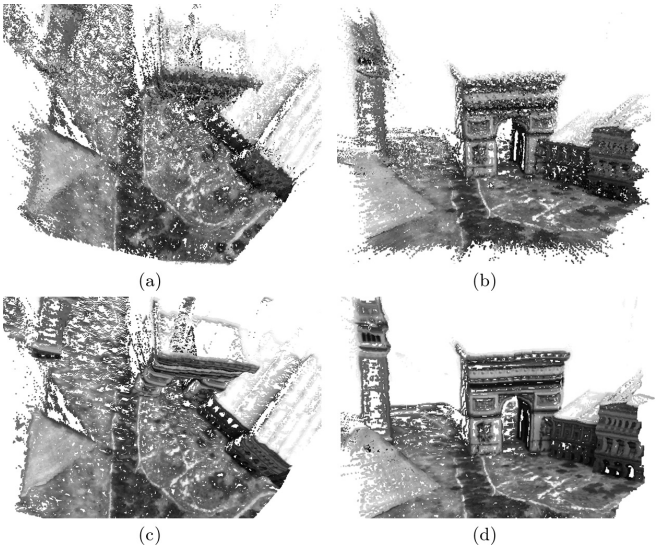


Fig. 1. In monocular dense reconstructions, the probabilistic approach to depth estimation produces compact and efficient representations. Highly parallelizable implementations are achieved by estimating the depth for every pixel independently. A smoothing step is nonetheless required to achieve robustness against noise and mitigate the effect of erroneous measurements. Figures (a) and (b) show the result of Bayesian depth estimation from multiple views; (c) and (d) show the same result after the de-noising step that we propose in this paper.

A. Related Work

The problem of reconstructing the scene from images collected by a moving camera has been studied for more than two decades and is known as Structure from Motion in computer vision [1] and Monocular SLAM in robotics [2]. The growing interest for dense reconstructions has renewed the attention in multi-view stereo techniques [3], [4], [5], [6], where the involved computational complexity used to prevent applications in robot perception. In robotics, the use of RGBD cameras is favouring the development of techniques for highly-detailed [7] and spatially-extended reconstructions [8], their applicability being limited to short range measurements and indoor environments. The literature in dense stereo is vast and we refer to [9] for a comparison. However, few relevant works have addressed real-time, dense reconstruction from a single moving camera and they shed light on some important aspects. Figure 1 illustrates the

problem we address in this paper. If, on one hand, estimating the depth independently for every pixel leads to efficient, parallel implementations, on the other hand the authors of [10], [11], [12] argued that, similar to other computer vision problems, such as image de-noising [13] and optical flow estimation [14], a smoothing step is required in order to deal with noise and spurious measurements. In [11], smoothness priors were enforced over the reconstructed scene by minimizing a regularized energy functional based on aggregating a photometric cost over different depth hypothesis and penalizing non-smooth surfaces. The authors showed that the integration of multiple images leads to significantly higher robustness to noise. A similar argument is put forth in [12], where the advantage of photometric cost aggregation [15] over a large number of images taken from nearby viewpoints is demonstrated. Regularized energy functionals also play an important role in recent methods for volumetric reconstruction [16], [17], [18], where the three-dimensional surface of a scene is generated by fusing several depth maps obtained from multi-view stereo. Depending on the scene appearance and the used stereo baselines, the computed depth maps are potentially noisy and a robust fusion method helps mitigate the effect of wrong depth estimations.

However, despite the ground-breaking results, these approaches present some limitations when addressing tasks in robot perception. Equally weighting measurements from small and large baselines, in close and far scenes, causes the aggregated cost to frequently present multiple or no minima. Depending on the depth range and sampling, these failures are not always recoverable by the subsequent optimization step. Furthermore, an inadequate number of images can lead to a poorly constrained initialization for the optimization and erroneous measurements that are hard to detect. It is not clear how many images should be collected, depending on the motion of the camera and the scene structure. Finally, the number of depth hypotheses controls the computational complexity, and the applicability is, thus, limited to scenes bounded in depth.

B. Contributions and Outline

The discussed limitations are overcome by probabilistic approaches handling measurement uncertainty. A compact representation and a Bayesian depth estimation from multi-view stereo were proposed in [19]. We build on their results for per-pixel depth estimation and introduce an optimization step to enforce spatial regularity over the recovered depth map. We propose a regularization term based on the weighted Huber norm but, differently from [12], we use the depth uncertainty to drive the smoothing and exploit a convex formulation for which a highly parallelizable solution scheme has been recently introduced [20]. The contributions of this paper are the following:

- a probabilistic depth map, in which the Bayesian scheme in [19] is integrated in a monocular SLAM algorithm to estimate per-pixel depths based on the live camera stream;
- a fast smoothing method that takes into account the measurement uncertainty to provide spatial regularity and mitigates the effect of noisy camera localization.

The outline of the paper follows. In Section II we detail our method for depth estimation from monocular views and in Section III we provide the implementation details. Section IV is dedicated to the discussion on the experimental evaluation. Finally, in Section V, we summarize our contribution and draw the conclusion.

II. Monocular Dense Reconstruction

A. Considerations

The solution we propose to compute a dense reconstruction from a single moving camera is motivated by the following considerations.

a) A measure of uncertainty is needed in robotic perception: many reconstruction pipelines previously proposed in computer vision and graphics literature aim at providing visually appealing maps. In contrast, we are interested in accurately mapping the environment in order to allow robotic tasks, such as autonomous navigation and exploration, active perception or situation awareness in the case of human-operated systems. As a passive sensing modality, measurement uncertainty in monocular multi-view stereo is related to the camera motion and the amount of visual information present in the scene (e.g. texture). A probabilistic depth map handles measure uncertainty, thus, allowing efficient updating, optimal sensor placement, and fusion with different sensors.

b) A dense reconstruction is needed to interact: sparse visual maps based on image features have been successfully used in robotics, e.g. to solve the SLAM problem. However, feature definitions change between sensing modalities and tasks; dense representations are, thus, required to actually solve the problem of registering data among largely different vantage points based on the three-dimensional structure [18]. When the task involves physical interaction with the environment—as in obstacle avoidance, path planning and manipulation—the highest achievable level of detail is desirable in order to estimate the surfaces involved in the interaction.

c) Perception must be fast: differently from many state-of-the-art systems, in order to be useful in robot perception our pipeline must run in real-time using the robot's on-board computing power. Depth estimation must be updated efficiently and the uncertainty in the estimation must improve according to the information conveyed by the image and the current camera pose.

In the designing of the monocular multi view stereo algorithm, these considerations naturally bring to the formulation of the following requirements: depth estimation must take into account the uncertainty arising from the scene and the camera pose and the estimation must be carried out on-line and updated sequentially. Bayesian estimation offers a natural way to deal with measure uncertainty, to handle sequential measurement updates and to reject unreliable estimations in an on-line fashion.

B. Depthmap from Multi View Stereo

We formulate the depth computation as a Bayesian estimation problem. Each observation provides a depth measurement by triangulating from the reference view and the last acquired view. The depth of a pixel is described by a parametric model that is updated on the basis of the current observation. Finally, smoothness on the resulting depth map is enforced by minimizing a regularized energy functional.

1) Bayesian Estimation: Let the rigid body transformation $\mathbf{T}_{k,w} \in SE(3)$ describe the pose of the camera acquiring the k-th view, i.e., $\mathbf{T}_{k,w}$ transforms scene points ${}_wp \in \mathbb{R}^3$ from the world frame to the frame of the k-th camera pose:

$$ {}_k\mathbf{p} = \mathbf{T}_{k,w}\ {}_w\mathbf{p}. \tag{1}$$

We denote the intensity image collected from the k-th camera pose as $I_k : \Omega \subset \mathbb{R}^2 \mapsto \mathbb{R}$, where Ω is the image domain. We denote by $\mathbf{u} \in \Omega$ a point in image coordinates.

An observation is a pair $\{I_k, \mathbf{T}_{k,w}\}$. A sequence of n observations is identified by the sequence of time steps $k = r, \dots, r+n$, in which the r-th observation is taken as reference. A depth hypothesis d_k is generated from the observation $\{I_k, \mathbf{T}_{k,w}\}$ by triangulating $\mathbf{u}$ from the views r and k.

The sequence of d_k for $k = r, \dots, r+n$ denotes a set of noisy depth measurements. We model the depth sensor as a distribution that mixes a good measurement (normally distributed around the true depth $\hat{d}$) and an outlier measurement (uniformly distributed in an interval $[d_{min}, d_{max}]$ which is known to contain the depth for the structure of interest):

$$ p(d_k|\hat{d},\rho) = \rho\mathcal{N}(d_k|\hat{d},\tau_k^2) + (1-\rho)\mathcal{U}(d_k|d_{min},d_{max}), \tag{2}$$

where ρ and τ_k^2 are the probability and the variance of a good measurement, respectively. Assuming independent observations, the Bayesian estimation for $\hat{d}$ on the basis of the measurements $d_{r+1}, \dots, d_k$ is given by the posterior

$$ p(\hat{d},\rho|d_{r+1},\dots,d_k) \propto p(\hat{d},\rho)\prod_k p(d_k|\hat{d},\rho), \tag{3}$$

with $p(\hat{d},\rho)$ being a prior on the true depth and the ratio of good measurements supporting it. A sequential update is implemented by using the estimation at time step $k-1$ as a prior to combine with the observation at time step k. To this purpose, the authors of [19] show that the posterior in (3) can be approximated by the product of a Gaussian distribution for the depth and a Beta distribution for the inlier ratio:

$$ q(\hat{d},\rho|a_k,b_k,\mu_k,\sigma_k^2) = Beta(\rho|a_k,b_k)\mathcal{N}(\hat{d}|\mu_k,\sigma_k^2), \tag{4}$$

where a_k and b_k are the parameters controlling the Beta distribution. The choice is motivated by the fact that the $Beta \times Gaussian$ is the approximating distribution minimizing the Kullback-Leibler divergence from the true posterior (3). Upon the k-th observation, the update takes the form

$$ \begin{aligned} p(\hat{d},\rho|d_{r+1},\dots,d_k) \approx\ & q(\hat{d},\rho|a_{k-1},b_{k-1},\mu_{k-1},\sigma_{k-1}^2) \\ & p(d_k|\hat{d},\rho)\ const \end{aligned} \tag{5}$$

and the authors of [19] approximated the true posterior (5) with a $Beta \times Gaussian$ distribution by matching the first and second order moments for $\hat{d}$ and ρ. The updates formulas for a_k, b_k, μ_k and σ_k^2 are thus derived and we refer to the original work in [19] for the details on the derivation.

2) Regularized Posterior: We now detail our solution to the problem of smoothing the depth map $D(\mathbf{u})$. For every pixel $\mathbf{u} \in \Omega$, the depth estimation and its confidence upon the k-th observation are given, respectively, by μ_k and σ_k^2 in (4). We formulate the problem of computing a de-noised depth map $F(\mathbf{u})$ as the following minimization:

$$ \min_F \int_\Omega \left\{ G(\mathbf{u}) \left\| \nabla F(\mathbf{u}) \right\|_\epsilon + \lambda \left\| F(\mathbf{u}) - D(\mathbf{u}) \right\|_1 \right\} d\mathbf{u}, \tag{6}$$

where λ is a free parameter controlling the trade-off between the data term and the regularizer, and $G(\mathbf{u})$ is a weighting function related to the "G-Weighted Total Variation", introduced in [21] in the context of image segmentation. We penalize non-smooth surfaces by making use of a regularization term based on the Huber norm of the gradient, defined as:

$$ \left\| \nabla F(\mathbf{u}) \right\|_\epsilon = \begin{cases} \frac{||\nabla F(\mathbf{u})||_2^2}{2\epsilon} & \text{if } ||\nabla F(\mathbf{u})||_2 \leq \epsilon, \\ ||\nabla F(\mathbf{u})||_1 - \frac{\epsilon}{2} & \text{otherwise .} \end{cases} \tag{7}$$

We chose the Huber norm because it allows smooth reconstruction while preserving discontinuities at strong depth gradient locations ([12]). The weighting function $G(\mathbf{u})$ influences the strength of the regularization and we propose to compute it on the basis of the measure confidence for $\mathbf{u}$:

$$ G(\mathbf{u}) = \mathbb{E}_\rho[q](\mathbf{u})\frac{\sigma^2(\mathbf{u})}{\sigma_{max}^2} + \left\{1 - \mathbb{E}_\rho[q](\mathbf{u})\right\}, \tag{8}$$

where we have extended the notation for the expected value of the inlier ratio $\mathbb{E}_\rho[q]$ and the variance σ^2 in (4) to account for the specific pixel $\mathbf{u}$. The weighting function (8) affects the strength of the regularization term: for measurements with a high expected value for the inlier ratio ρ the weight is controlled by the measurement variance σ^2; measurements characterized by a small variance (i.e. reliable measurements) will be less affected by the regularization; differently, the contribution of the regularization term will be heavier for measurements characterized by a small expected value for the inlier ratio or higher measurement variance.

The solution to the minimization problem (6) is computed iteratively based on the work in [20]. The algorithm exploits the primal dual formulation of (6),

$$ \begin{aligned} \min_F \max_{F^*} & \langle \text{diag}(G)\nabla F, F^* \rangle \\ & + \lambda||C - D||_1 - \delta_{F^*}(F^*) - \frac{\epsilon}{2}||F^*||_2^2, \end{aligned} \tag{9}$$

and proceeds by alternating gradient descent and ascent steps in the primal and dual variables, namely F and F^*. The indicator function $\delta_{F^*}(F^*)$ is such that, for each F^*, $\delta_{F^*}(F^*) = 0$ if $||F^*||_1 \leq 1$, and otherwise ∞. Let t and t^* be the time steps for the gradient descent-ascent with respect to the primal and dual variable. The update steps in the case of the Weighted-Huber de-noising model (6) take the form

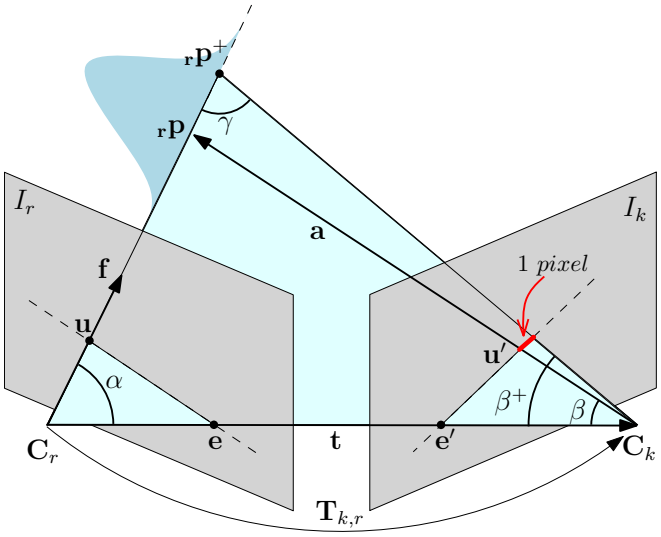


Fig. 2. Computation of the measurement uncertainty. The camera poses acquiring the views I_r and I_k are related by the transformation $\mathbf{T}_{k,r}$. The camera centres C_r, C_k and the current estimation of the scene point $_r\mathbf{p}$ lie on the epipolar plane.The variance corresponding to one pixel along the epipolar line passing through $\mathbf{e}'$ and $\mathbf{u}'$ is computed as $\tau_k^2 = \left(\|_r\mathbf{p}^+\| - \|_r\mathbf{p}\|\right)^2$.

$$
\begin{aligned}
F_{n+1}^* &= \text{prox}\left(\frac{F_n^* + t^*(\text{diag}(G)\nabla)\bar{F}}{1 + t^*\epsilon}\right), \\
F_{n+1} &= \text{shrink}\left(F_n - t\left(\nabla^T \text{diag}(G)\right) F_{n+1}^*\right), \\
\bar{F}_{n+1} &= 2F_{n+1} - F_n,
\end{aligned} \tag{10}
$$

where the resolvent operators are

$$
\text{prox}(\tilde{f}^*) = \frac{\tilde{f}^*}{\max(1, |\tilde{f}^*|)}, \tag{11}
$$

$$
\text{shrink}(\tilde{f}) = \begin{cases} \tilde{f} - t\lambda & \text{if } \tilde{f} - d > t\lambda \\ \tilde{f} + t\lambda & \text{if } \tilde{f} - d < -t\lambda \\ d & \text{if } |\tilde{f} - d| \leq t\lambda \end{cases} \tag{12}
$$

and d is the noisy depth value corresponding to a specific pixel.

III. Implementation Details

The monocular reconstruction pipeline is designed to run in real time on a commodity laptop, using a CPU and a GPU. The proposed probabilistic depth map and convex optimization lead to highly parallel algorithms and we based our implementation on CUDA[1].

A. Camera pose estimation

At every time step k, the pose of the camera $\mathbf{T}_{k,r}$ in the depth map reference frame r is computed by a visual odometry routine that is based on recent advancement on semi-direct methods for camera localization [22]. The algorithm operates directly on the image intensity, eliminating the need for costly feature extraction and resulting in sub-pixel accuracy at high frame-rates. Three-dimensional map points are estimated making use of the probabilistic method described in Section II-B, which proved at the same time highly robust, accurate and computationally efficient. Our implementation is characterized by an average drift in pose of 0.0038 metres per second for an average depth of 1 metre and a computing time of 3.3 milliseconds per acquired image on the experimental platform detailed in Section IV. The visual odometry algorithm is run by the CPU, and its accuracy and efficiency support the simultaneous execution of the monocular reconstruction pipeline.

[1] `http://www.nvidia.com`

B. Measurement update

The parametric model in (4) is a compact representation, as it stores our confidence in the depth measurement corresponding to a pixel in only four parameters: a, b, μ and σ. When a reference frame is taken, the estimation for every pixel is initialized and updated with every subsequent view. We set the initial parameters $a_0 = 10$, $b_0 = 10$, $\mu_0 = 0.5(d_{min} + d_{max})$ and $\sigma_0 = \sigma_{max}$, where σ_{max} is such that 99% of the probability mass lies in the interval $[d_{min}, d_{max}]$. Upon the acquisition of the k-th view, the update introduced in [19] is performed for every pixel of the reference view. We perform the update until the depth estimation converges or diverges. At this point, we can either consider the measurement reliable or discard it. We check the convergence and divergence conditions by looking at the variance of the depth posterior σ_k^2 and the estimated inlier ratio ρ_k. Let η_{inlier} and $\eta_{outlier}$ be thresholds on the estimated inlier ratio and σ_{thr} be a threshold on the variance of the depth posterior. We have three cases:

- if $\mathbb{E}_\rho[q] > \eta_{inlier}$ and $\sigma_k^2 < \sigma_{thr}^2$, then the estimation has converged;
- else if $\mathbb{E}_\rho[q] < \eta_{outlier}$, then the estimation has diverged;
- otherwise, the estimation continues.

The parameters η_{inlier}, $\eta_{outlier}$ and σ_{thr} control the estimation convergence and can be set according to the accuracy and robustness requirements for the application at hand.

In order to deal with higher depth ranges, we base our implementation on the inverse depth [23] and use the currently estimated variance to limit the search for correspondence on the epipolar line.

C. Measurement uncertainty

When triangulating matched points to estimate the depth from multiple views, frames taken from nearby vantage points are less affected by occlusions and allow high quality matches. On the other hand, a large baseline enables a more reliable depth estimation but with a higher chance to incur in occluded regions.

Referring to Figure 2, let $_r\mathbf{p}$ be the current estimation of the scene point corresponding to the pixel $\mathbf{u}$ in the image I_r. The variance on the position of $_r\mathbf{p}$ is obtained by back-projecting a constant variance of one pixel in the image I_k.

Let $\mathbf{t}$ be the translation component of $\mathbf{T}_{k,r}$ and $\mathbf{f} = \frac{{}_r\mathbf{p}}{||{}_r\mathbf{p}||}$, then

$$\mathbf{a} = {}_r\mathbf{p} - \mathbf{t} \tag{13}$$

$$\alpha = \arccos\left(\frac{\mathbf{f}\cdot\mathbf{t}}{||\mathbf{t}||}\right) \tag{14}$$

$$\beta = \arccos\left(-\frac{\mathbf{a}\cdot\mathbf{t}}{||\mathbf{a}||\cdot||\mathbf{t}||}\right). \tag{15}$$

Let f be the camera focal length. The angle spanning one pixel can be added to β in order to compute γ and, thus, by applying the law of sines, recover the norm of ${}_r\mathbf{p}^+$:

$$\beta^+ = \beta + 2\tan^{-1}\left(\frac{1}{2f}\right) \tag{16}$$

$$\gamma = \pi - \alpha - \beta^+ \tag{17}$$

$$||{}_r\mathbf{p}^+|| = ||\mathbf{t}||\frac{\sin\beta^+}{\sin\gamma}. \tag{18}$$

Therefore, the measurement uncertainty is computed as

$$\tau_k^2 = \left(||{}_r\mathbf{p}^+|| - ||{}_r\mathbf{p}||\right)^2. \tag{19}$$

IV. Experimental Evaluation

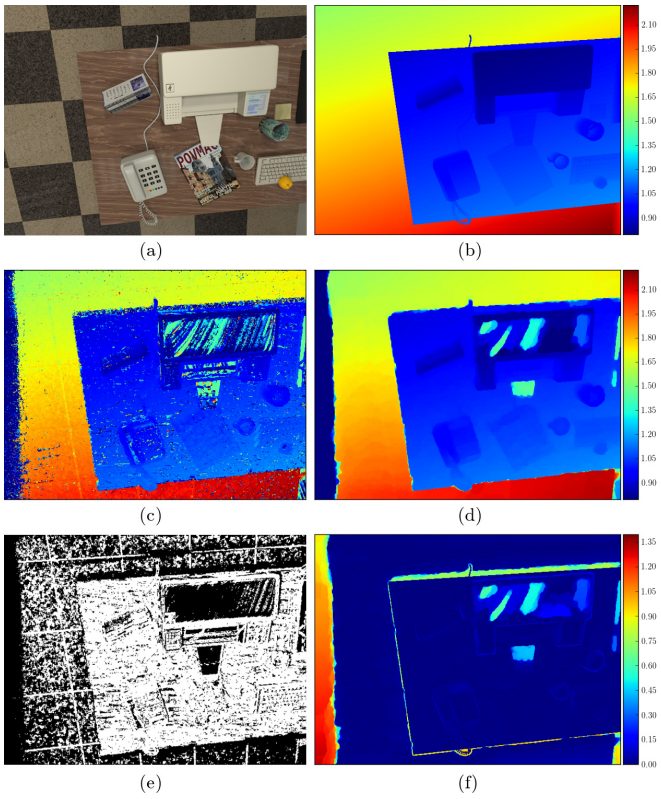
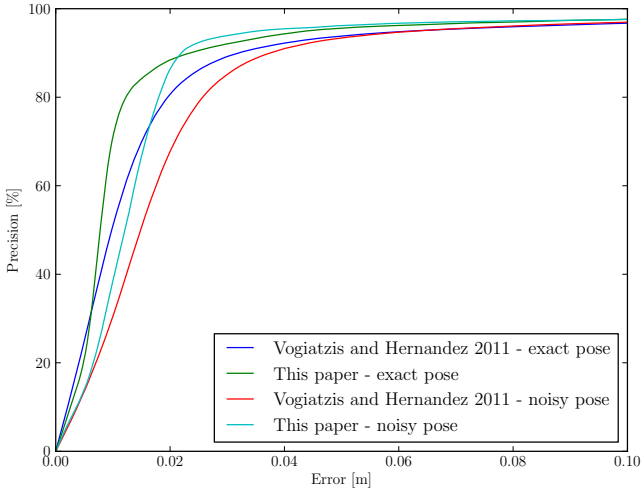
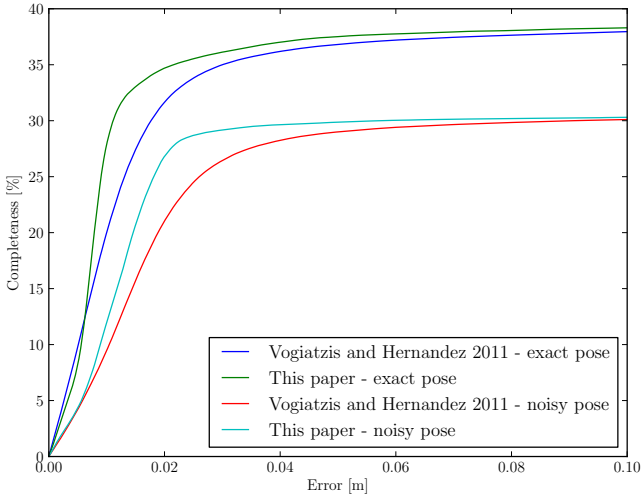


Fig. 3. The *over table* evaluation sequence. (a): the reference view. (b): ground truth depth map. (c): depth map based on [19]. (d): depth map computed by the proposed method. (e): map of reliable measurement according to Section III-B. (f): error for the proposed method.

The platform we used for the experimental evaluation of the proposed monocular reconstruction method is an Intel i7-3720QM based laptop, equipped with 15 GB of RAM, and an NVIDIA Quadro K2000M GPU with 384 CUDA cores.



(a)



(b)

Fig. 4. Quantitative evaluation on the *over table* sequence. In (a) the precision is plotted, namely the percentage of converged estimations that are within a certain error from the ground truth. In (b) the completeness is plotted, namely the percentage of ground truth measurements that are within a certain error from the converged estimations.

We chose the dataset presented in [24] in order to quantitatively evaluate our approach. The dataset consists of views generated through ray-tracing from a three-dimensional synthetic model. Along with each view, the related exact camera pose and depth maps are made available. Table I summarizes the details for the sequences used in the evaluation.

Over table identifies a sequence of views collected down-looking on a desktop scenario. The sequence is characterized by a frame rate of 30 frames per second and smooth camera motion. The sequence identified as *fast motion* is a collection of views generated at 60 frames per second with large and sudden changes of vantage point. The evaluation is based on comparison with the ground truth depth map corresponding to the view taken as reference in the reconstruction process. Two depth maps are compared by computing the sum of the per-pixel absolute difference. Since we are interested in

TABLE I

DATASETS FOR COMPARISON AGAINST GROUND TRUTH.

	Frames	Range [m]	Mean [m]	Motion [m]	Speed [m/s]
Over table	200	0.827-2.84	1.531	4.576	0.686
Fast motion	900	0.971-6.802	2.015	21.6	1.61

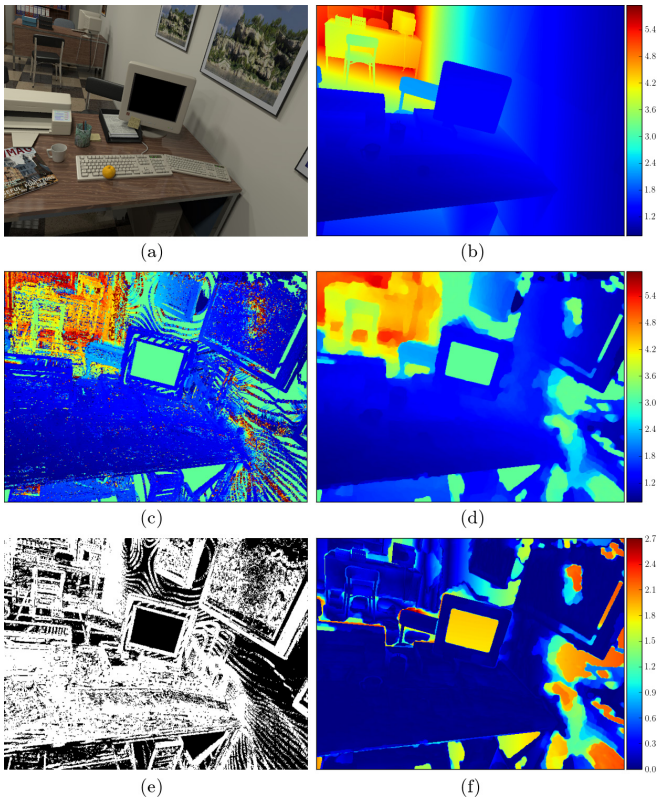


Fig. 5. The *fast motion* evaluation sequence. (a): the reference view. (b): ground truth depth map. (c): depth map based on [19]. (d): depth map computed by the proposed method. (e): map of reliable measurement according to Section III-B. (f): error for the proposed method.

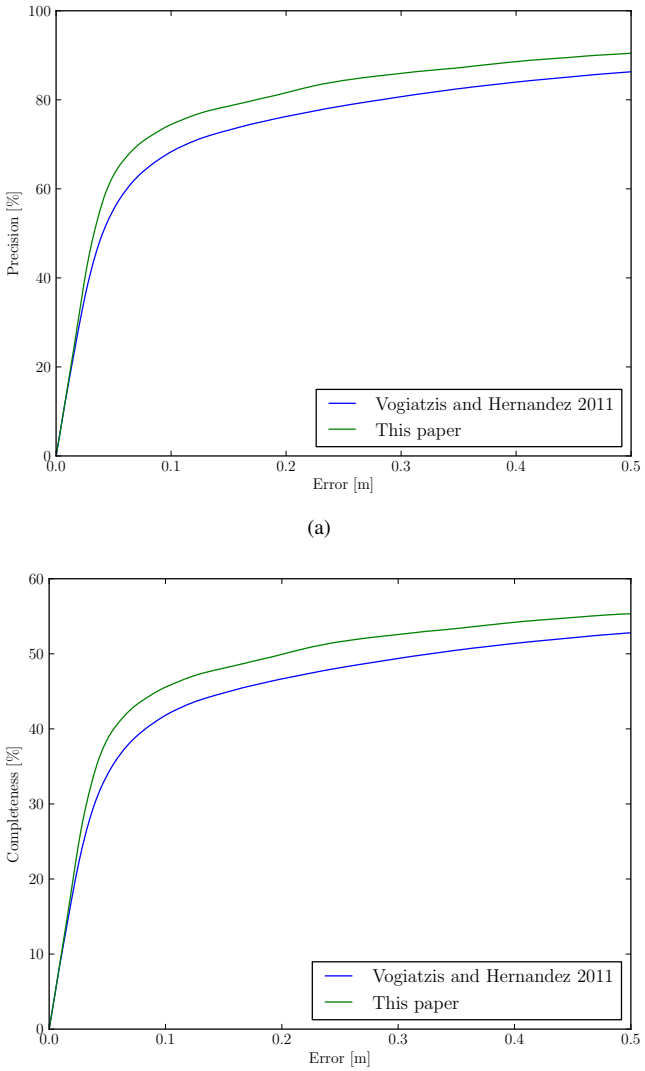


Fig. 6. Quantitative evaluation on the *fast motion* sequence. In (a) the precision is plotted, namely the percentage of converged estimations that are within a certain error from the ground truth. In (b) the completeness is plotted, namely the percentage of ground truth measurements that are within a certain error from the converged estimations.

evaluating the depth measurements that have been identified as reliable by our algorithm, we only take into account those measurements that have converged according to Section III-B. We therefore use the converged measurements to create the masks (e) in Figure 3 and Figure 5, which are used in the comparison. We define two evaluation metrics: *precision*, namely the percentage of converged measurements that fall below a certain error when compared to the relative ground truth, and *completeness*, namely the percentage of ground truth depths that have been estimated by the proposed method within a certain error. In order to show the effectiveness of our approach, we compare our results with depth maps computed according to the state-of-the-art method introduced in [19]. This work is at the basis of our probabilistic treatment and, so far, its applicability has been demonstrated only for reconstruction of small objects. For our comparison using the ground truth sequences, the parameters defining reliable measures have been set at $\eta_{inlier} = 0.6$, $\eta_{outlier} = 0.05$ and $\sigma_{thr} = \sigma_{max}/10^3$. The parameters governing the optimization were set at $\epsilon = 10^{-4}$ and $\lambda = 0.3$, and 200 iterations of the primal-dual update in (10) were run.

Figure 4 reports the result of the evaluation on the *over table* sequence. Our approach is capable to recover a number of erroneous depth estimations, thus yielding a sensible improvement in terms of accuracy and completeness. To verify the robustness against noisy camera pose estimation, we corrupted the camera position with Gaussian noise, with zero mean and one centimetre standard deviation on each coordinate. The results show that the completeness drops. This is inevitable due to the smaller number of converged estimations. However, the computation of the depth map takes advantage of the de-noising step. This trend is even more evident in the *fast motion* sequence, depicted in Figure 5. Here, according to the results in Figure 6, the

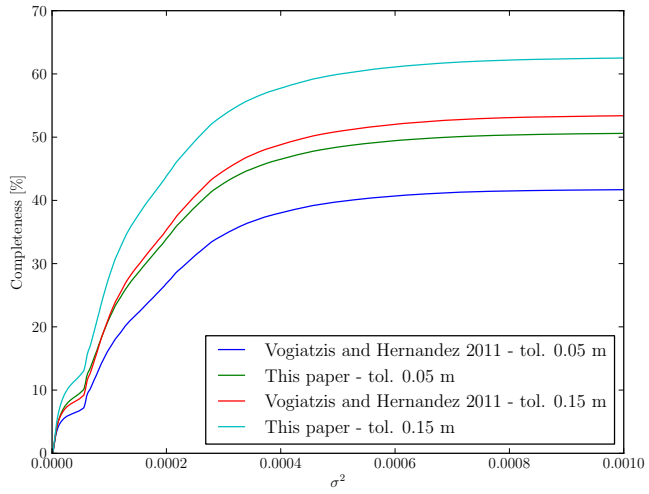


Fig. 7. The percentage of ground truth measurements that are within an error of 5 and 15 centimetres is plotted as a function of the measurement variance σ^2.

advantage of our approach is clearly demonstrated in terms of both precision and completeness. Handling measurement uncertainty, the probabilistic treatment of depth allows us to select the optimal trade-off between precision and accuracy by varying the σ_{thr} parameter. Figure 7 shows how, for a given error tolerance, the completeness varies as a function of the variance σ^2 that characterizes a reliable measurement. We can see, for instance, that using a threshold $\sigma_{thr} = 6 \times 10^{-4}$, which is approximately 2×10^3 times the initialization value σ_{max}, more than 60% of the depth measurements computed by our method are affected by an error up to 15 centimetres, that is approximately 2.6% of the full depth range.

TABLE II
COMPUTING TIME FOR THE EVALUATION DATA

	Update time [s]		Optimization time [s]	
	Mean	Variance	Mean	Variance
Over table	0.0382	0.0025	0.1107	0
Fast motion	0.0499	0.0035	0.1149	0
Live acquisition	0.0301	0.0011	0.1122	0.0044

In order to demonstrate the effectiveness of the proposed approach on real time reconstructions, we present our results on the *City of Sights* stage set [25]. We computed a point cloud from different depth maps acquired by a single hand-held camera. Our reconstruction pipeline was fed with images and camera poses computed by the underlying visual odometry (cfr. Section III-A) at 30 frames per second.

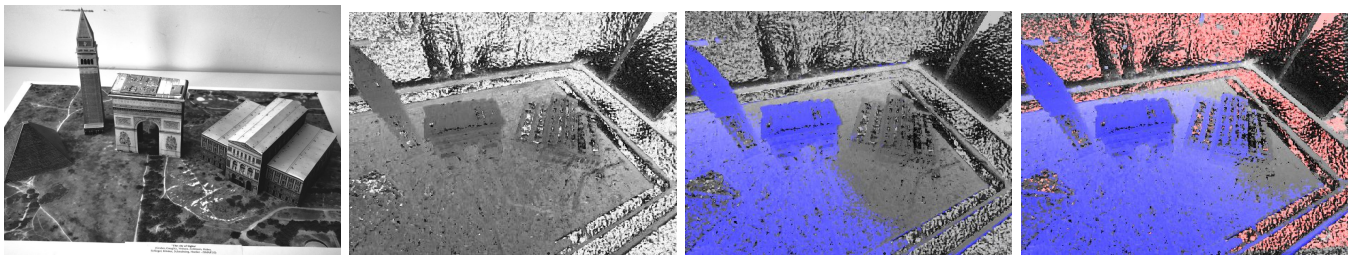
Figure 8 depicts the process of a live depth map acquisition. During the reconstruction, the convergence and divergence of estimations are displayed as a live feedback for the user (blue and red respectively in the figure), guiding the motion of the camera to acquire portions of the scene for which the estimation has not yet converged or diverged. A qualitative evaluation of the results can be drawn from Figures 8 and 9. The minimization in (6) imposes a smoothness constraint on the resulting surface and acts as a prior when the estimation is uncertain. Wrong depth computations, caused by shadows or matching errors (see Figure 8b), cause the respective estimations to diverge (red points in Figure 8d). The de-noising step propagates the depth value produced by converged measurements to those neighbours yielding low confidence, which are characterized by diverged measurements. The final result, in the form of a coloured point cloud rendered from two different viewpoints, is depicted in Figure 9.

Finally, the proposed method is suitable for real time execution, as can be seen in Table II, where we have reported the computing time for the evaluation sequences. The computational cost of the proposed method is dominated by the search for correspondences on the epipolar line. When the motion of the camera is smooth, like in the cases of the *over table* dataset and live acquisition, the region selected for the search is small; when the camera motion forms large baselines, then the candidate search area is wider, affecting the computing time as in the case of the *fast motion* dataset. The depth range characterizing the volume of interest for the reconstruction also plays an important role, as the measurement uncertainty is higher for distant points (cfr. Section III-C). This causes the depth estimation to require a larger number of views to converge. Nonetheless, the estimation update runs in real time on the live 30 fps camera stream, for a camera resolution of 752×480 pixels. The computational cost of the optimization step depends only on the image size and number of iterations, and is thus constant among an evaluation sequence. Optimization was run several times during the *live acquisition*, triggered by the instantiation of new reference frames, while for the ground truth sequences the single optimization step that is performed motivates the 0 variance entries in Table II.

An open-source implementation, along with a video demonstrating the reconstruction of scenes acquired by a hand-held camera and a flying robot, are available at the website `http://rpg.ifi.uzh.ch/software`.

V. CONCLUSION

In this paper we presented REMODE, a probabilistic approach to monocular dense reconstruction for robot perception. Our method computes depth maps by combining Bayesian estimation and recent developments in convex optimization for image processing. We showed how a probabilistic update scheme can produce a compact and efficient representation of a depth map and its related uncertainty. In order to achieve real time execution on a live camera stream, we parallelized the computation of a depth map by considering each pixel independently. Afterwards, we introduced a fast smoothing step that takes into account the measurement uncertainty to enforce spatial regularity and mitigates the effect of noisy camera localization. We evaluated our method in terms of accuracy and completeness, showing a sensible improvement with respect to the current state-of-the-art. By handling measurement uncertainty, our method provides real time information about the progress



(a) (b) (c) (d)

Fig. 8. Depth map computation for the *City of Sights* stage set [25]. Dark points are close, bright points are far. Blue and red identify converged and diverged estimations, respectively.

and the reliability of the ongoing reconstruction process. This information is highly valuable to drive the reconstruction, that is, to determine what views are most informative for the task at hand.

Fig. 9. Reconstructed point clouds for the *City of Sights* stage set [25].

REFERENCES

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[2] A. Davison, "Real-time simultaneous localisation and mapping with a single camera," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2003.

[3] L. Matthies, R. Szeliski, and T. Kanade, "Incremental estimation of dense depth maps from image sequences," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1988.

[4] S. B. Kang, R. Szeliski, and J. Chai, "Handling occlusions in dense multi-view stereo," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2001.

[5] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[6] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 32, no. 8, pp. 1362–1376, 2010.

[7] M. Meilland and A. Comport, "Super-resolution 3D Tracking and Mapping," in *IEEE Intl. Conf. on Robotics and Automation*, 2013.

[8] T. Whelan, H. Johannsson, M. Kaess, J. Leonard, and M. J.B., "Robust real-time visual odometry for dense RGB-D mapping," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2013.

[9] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 31, no. 9, 2009.

[10] D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-time plane-sweeping stereo with multiple sweeping directions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2007.

[11] J. Stuehmer, S. Gumhold, and D. Cremers, "Real-time dense geometry from a handheld camera," in *DAGM Symposium on Pattern Recognition*, 2010, pp. 11–20.

[12] R. A. Newcombe, S. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. IEEE Intl. Conf. on Computer Vision*, 2011, pp. 2320–2327.

[13] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 14, 1992.

[14] M. Werlberger, T. Pock, and H. Bischof, "Motion estimation with non-local total variation regularization," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2010.

[15] R. Szeliski and D. Scharstein, "Sampling the disparity space image," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 3, 2004.

[16] C. Zach, "Fast and high quality fusion of depth maps," in *Proc. of 3DPVT*, 2008.

[17] G. Graber, T. Pock, and H. Bischof, "Online 3d reconstruction using convex optimization," in *Proc. Workshops of IEEE Intl. Conf. on Computer Vision*, 2011.

[18] C. Forster, M. Pizzoli, and D. Scaramuzza, "Air-ground localization and map augmentation using monocular dense reconstruction," in *Proc. Intl. Conf. on Intelligent Robots and Systems*, 2013.

[19] G. Vogiatzis and C. Hernández, "Video-based, real-time multi-view stereo," *Image and Vision Computing*, vol. 29, no. 7, 2011.

[20] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *Journal of Mathematical Imaging and Vision*, vol. 40, no. 1, 2011.

[21] X. Bresson, S. Esedoglu, P. Vandergheynst, J.-P. Thiran, and S. Osher, "Fast global minimization of the active contour/snake model," *Journal of Mathematical Imaging and Vision*, vol. 28, no. 2, 2007.

[22] C. Forster, M. Pizzoli, and D. Scaramuzza, "SVO: Fast Semi-Direct Monocular Visual Odometry," in *Proc. IEEE Intl. Conf. on Robotics and Automation*, 2014.

[23] J. Civera, A. Davison, and J. Montiel, "Inverse depth parametrization for monocular slam," *IEEE Trans. on Robotics*, vol. 24, no. 5, 2008.

[24] A. Handa, R. Newcombe, A. Angeli, and A. Davison, "Real-time camera tracking: When is high frame-rate best?" in *Proc. IEEE European Conf. on Computer Vision*, 2012.

[25] L. Gruber, S. Gauglitz, J. Ventura, S. Zollmann, M. Huber, M. Schlegel, G. Klinker, D. Schmalstieg, and T. Höllerer, "The city of sights: Design, construction, and measurement of an augmented reality stage set," in *Proc. IEEE Intl. Symposium on Mixed and Augmented Reality (ISMAR)*, 2010.