Robust Depth Regularization Explicitly Constrained by Camera Motion

Nadège Zarrouati[1,2] , Emanuel Aldea[3] and Pierre Rouchon[1]
[1] *Mines-ParisTech, CAS, 60 boulevard Saint Michel, 75272 Paris Cedex, France*
[2]*DGA, 7-9 rue des Mathurins, 92220 Bagneux, France*
[3]*SYSNAV, 57 rue de Montigny, 27200 Vernon, France*
nadege.zarrouati@mines-paristech.fr

Abstract

The objective of our work is to reconstruct the dense structure of a static scene observed by a monocular camera system following a known trajectory. Our main contribution is representated by the proposition of a TV-L^1 energy functional that estimates directly the unknown depth field given the camera motion, thus avoiding to estimate as an intermediate step an optical flow field with additional geometric constraints. Our method has two main interests: we highlight a practical minimal parametrization for the given assumptions (static scene, known camera motion) and we solve the resulting variational problem using an efficient, discontinuity preserving formulation.

1. Introduction

We investigate in the present article the problem of estimating a dense depth field of a static scene observed by a monocular camera system following a known trajectory. The localization information may be provided by proprioceptive sensors, by linear guide, turntable or unconstrained motion capture systems, or if applicable by employing a sparse feature based pose estimation algorithm.

One of the fundamental challenges related to recovering scene structure from video is that depth observability is low for small camera displacements occurring between consecutive frames. SfM and SLAM systems often select a subset of the available data, denoted as keyframes, in order to improve the depth estimation for a sparse set of salient features; as a result, the estimated scene structure is also sparse, and the keyframe selection process is heuristic and application dependent [6]. The use of each available frame allows for the estimation of a dense displacement map, but the temporal integration of displacements in a depth estimation process must cope with a high level of noise, while at the same time being robust to outliers and other undesirable phenomena caused by occlusion and non-Lambertian surfaces. The spatial integration of depth estimates may be performed by an explicit probabilistic model [8], or by alternative methods which update the weights related to the surface position (usually involving truncated distance functions). In [11] we proposed an alternative method based on asymptotic observers, which ensures depth convergence from multiple observations within a robust mathematical framework.

The geometric constraint provided by the known camera displacement between consecutive frames must be taken into account for the temporal integration step, but also when computing the optical flow field between the frames. A variational approach is the method of choice for regularizing the flow and favoring piecewise smooth regions. However, in order to enforce geometric (epipolar) constraints in the case of static scenes, most proposed algorithms penalize the functional with additional data terms, or rely on over-parametrization (see for example [1, 9]). In the present article, we propose an algorithm which employs explicitly the general motion information of the camera into a robust TV-L^1 variational problem involving directly the instantaneous depth field.

2. Proposed method

We present in Figure 1 an overview of our depth estimation method. At each time step, we start by applying the global regularization scheme proposed in this article, based on the motion information, the current and the previous image. Then, the current depth map is merged with the previously estimated depths by the asymptotic observer which is described in detail in [11].

We continue by introducing the model we employed, the major assumptions and the formal problem statement.

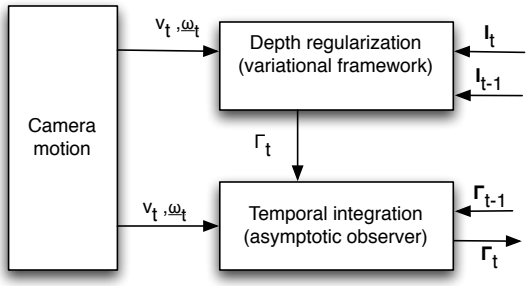

Figure 1. General view of our proposed method.

2.1 Modeling and framework assumptions

The model is based on geometric assumptions introduced in [2, 11], which are recalled in this sub-section. In a first step, we consider a spherical camera; its motion is given through the linear and angular velocities $v(t)$ and $\omega(t)$ expressed in the camera frame. More precisely, the position of the optical center in the reference frame $\mathcal{R}$ is denoted by $C(t)$.A pixel is labeled by the unit vector η in the camera frame: η belongs to the sphere $\mathbb{S}^2$ and receives the brightness $y(t, \eta)$.

The scene is modeled as a closed, C^1 and convex surface Σ of $\mathbb{R}^3$, diffeomorphic to $\mathbb{S}^2$. The camera is inside the domain $\Omega \subset \mathbb{R}^3$ delimited by $\Sigma = \partial\Omega$. The density of light emitted by a point $M(s) \in \Sigma$ does not depend on the direction of emission (Σ is a Lambertian surface) and is independent of t (the scene is static). This means that $y(t, \eta)$ depends only on s. The distance $\|\overrightarrow{C(t)M(s)}\|$ between the optical center and the object seen in the direction η is denoted by D, and its inverse by $\Gamma = 1/D$. Figure 2 illustrates the model and the notations. Under the assumptions that v and ω are C^0 functions of t, and y is a C^1 function of s, y and Γ obey to ([2, 11])

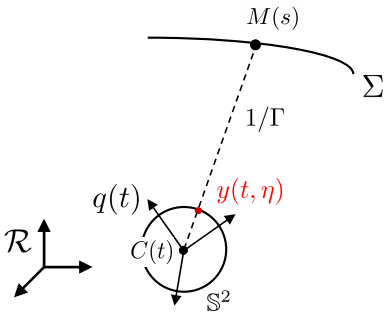

Figure 2. Model and notations of a spherical camera in a static environment [2, 11].

$$\dot{y} = -\nabla y \cdot (\eta \times (\omega + \Gamma\eta \times v)) \tag{1}$$

where ∇y and $\dot{y}$ denote the gradient on the Riemannian sphere $\mathbb{S}^2$ and the partial derivative with respect to t of y, respectively. The Euclidean scalar product of two vectors a and b in $\mathbb{R}^3$ is denoted by $a \cdot b$ and their wedge product by $a \times b$.

Equation (1) is another formulation of the optical flow constraint: instead of the usual two-dimensional disparity field, the only unknown in this equation is the depth field Γ.

2.2 Problem statement - the functional

In this paper, the input parameters are two images $y(t, \eta)$ and $y(t + dt, \eta)$ and the camera motion $v(t)$ and $\omega(t)$ performed between t and $t + dt$. The objective is to find the dense depth field $\Gamma(t, \eta)$ of the environment observed by the camera at time t. Inspired by the work of Chambolle and Pock ([3, 4, 10]), we define the functional J:

$$J(\Gamma) = \int_{\mathbb{S}^2} \left(|\nabla\Gamma| + \lambda|\dot{y} + \nabla y \cdot (\eta \times (\omega + \Gamma\eta \times v))|\right) d\sigma_\eta \tag{2}$$

where $d\sigma_\eta$ is the Riemannian infinitesimal surface element on $\mathbb{S}^2$ and λ is a parameter that weights between the data fidelity and the regularization. Both these terms deal with the L^1 norm, as opposed to previous functionals involved in optical flow estimation [5] or image denoising ([7]). The advantage of the choice of L^1-based total variation and data fidelity terms has been highlighted in [4], as it preserves discontinuities and avoids applying a diffusion scheme on the result.

3. The algorithm

In the following section, we propose a resolution strategy for the depth regularization problem stated above, and we provide the main steps of a practical implementation.

Proposition 1. *The minimizer of the functional (2) may be computed by dividing the main minimization task into two alternatively solved sub-tasks defined by the functionals J_1 and J_2:*

1. *For Λ being fixed, find the argument Γ of the minimum of*

$$J_1(\Gamma) = \int_{\mathbb{S}^2} \left(|\nabla\Gamma| + \frac{(\Gamma - \Lambda)^2}{2\theta}\right) d\sigma_\eta \tag{3}$$

2. *For Γ being fixed, find the argument Λ of the minimum of*

$$J_2(\Lambda) = \int_{\mathbb{S}^2} \left(\lambda \left| \rho(\Lambda) \right| + \frac{(\Gamma - \Lambda)^2}{2\theta} \right) d\sigma_\eta \quad (4)$$

where $\rho(\Lambda) = \dot{y} + \nabla y \cdot (\eta \times (\omega + \Lambda \eta \times v))$ is the depth residual.

Proof. This method has been proved to converge in [4] by the iterative alternation between dual and primal variables (Algorithm 1, Section 3). The proof stands in our case, as the function $G : \Gamma \mapsto \int_{\mathbb{S}^2} |\rho(\Gamma)| d\sigma_\eta$ is proper, convex and lower semicontinuous as the composition of the L^1 norm on $\mathbb{S}^2$ and an affine function:

$$\rho(\Gamma) = F_{\omega,t} + \Gamma F_v \quad (5)$$

with

$$\begin{cases} F_{\omega,t} = & \dot{y} + \nabla y \cdot (\eta \times \omega) \\ F_v = & \nabla y \cdot (\eta \times (\eta \times v)) \end{cases} \quad (6)$$

Under the assumptions of the model presented in 2.1, G admits a minimum.

□

The main algorithm steps are given by the following:

Algorithm: (v, ω) driven TV-L^1 depth regularization

- Initialization: $\Gamma^0 = 0$, $\Lambda^0 = 0$, $\bar{\Gamma}^0 = 0$ $\mathbf{p}^0 = 0$, τ and $\sigma \in [0, 1]$ s.t. $\tau\sigma L^2 \leq 1$ (L being the induced norm of ∇)
- Iterations:
  1) $\mathbf{p}^{n+1} = \frac{\mathbf{p}^n + \sigma \nabla \bar{\Gamma}^n}{1 + \sigma |\nabla \Gamma^n|}$
  2) $\mathbf{p}^{n+1} = \frac{\mathbf{p}^{n+1}}{\max(1, ||\mathbf{p}^{n+1}||_2)}$
  3) $\Gamma^{n+1} = \Gamma^n + \tau \text{div} \mathbf{p}^{n+1}$
  4) $\Lambda^{n+1} = \Lambda^n + \begin{cases} -\frac{\rho(\Gamma)}{F_v}, \text{if } |\rho(\Gamma)| \leq \tau\lambda F_v^2 \\ \lambda\tau F_v, \text{if } \rho(\Gamma) < -\tau\lambda F_v^2 \\ -\lambda\tau F_v, \text{if } \rho(\Gamma) > \tau\lambda F_v^2 \end{cases}$
  5) $\bar{\Gamma}^{n+1} = 2\Lambda^{n+1} - \Lambda^n$

where $\mathbf{p} = (p_1, p_2)$. The first three steps of the iteration process result from the minimization of J_1, and the fourth step is the thresholding method for minimizing J_2 applied to our specific residual. More precisely, it is based on the first order stationary condition in the cases where $\rho(\Lambda) = 0, \rho(\Lambda) < 0$ and $\rho(\Lambda) > 0$, respectively. For example, when $\rho(\Lambda) > 0$, the minimum may only be reached if Λ satisfies

$$\lambda F_v - \frac{(\Gamma - \Lambda)}{\theta} = 0 \quad (7)$$

which yields $\Lambda = \Gamma - \theta\lambda F_v$. Thus, $\rho(\Lambda) > 0$ reads $\rho(\Gamma) > \theta\lambda F_v^2$.

3.1 Adaptation to a pinhole camera model

Note that the abstract model of a spherical camera can be easily adapted to common camera models through a basic correspondence between the unit vector $\eta \in \mathbb{S}^2$ and the coordinates of a pixel in the considered model. For simulations and experimentations, we consider a pinhole camera model: the pixel of coordinates (z_1, z_2) corresponds to the unit vector $\eta \in \mathbb{S}^2$ of coordinates in $\mathbb{R}^3$: $\left(1 + z_1^2 + z_2^2\right)^{-1/2} (z_1, z_2, 1)^T$. The optical camera axis (pixel $(z_1, z_2) = (0, 0)$) corresponds here to the direction z_3. Directions 1 and 2 correspond respectively to the horizontal axis from left to right and to the vertical axis from top to bottom on the image frame. In this camera frame, linear and angular velocities components are denoted (v_1, v_2, v_3) and $(\omega_1, \omega_2, \omega_3)$, respectively. The gradient ∇y must be expressed with respect to z_1 and z_2. Firstly, ∇y is tangent to $\mathbb{S}^2$, thus $\nabla y \cdot \eta = 0$. Secondly, the differential dy corresponds to $\nabla y \cdot d\eta$ and to $\frac{\partial y}{\partial z_1} dz_1 + \frac{\partial y}{\partial z_2} dz_2$. By identification, we get the Cartesian coordinates of ∇y in $\mathbb{R}^3$. Injecting these coordinates in (6), we get:

$$\begin{aligned} F_{\omega,t}(z_1, z_2) &= \dot{y}(z_1, z_2) \\ &+ \frac{\partial y}{\partial z_1} \left[z_1 z_2 \omega_1 - (1 + {z_1}^2)\omega_2 + z_2 \omega_3 \right] \\ &+ \frac{\partial y}{\partial z_2} \left[(1 + {z_2}^2)\omega_1 - z_1 z_2 \omega_2 - z_1 \omega_3 \right] \\ F_v(z_1, z_2) &= \sum_{i=1,2} \frac{\partial y}{\partial z_i} \left[\sqrt{1 + z_1^2 + z_2^2} (-v_i + z_i v_3) \right] \\ d\sigma_\eta &= (1 + z_1^2 + z_2^2)^{-3/2} dz_1 dz_2 \end{aligned} \quad (8)$$

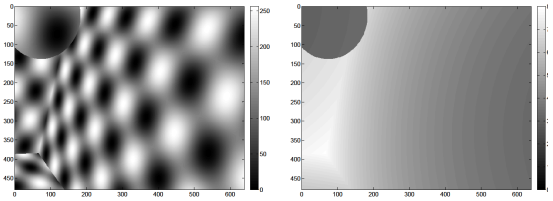

Figure 3. An image of the synthetic sequence and the associated depth field of the environment.

4 Simulations

We test the proposed method on a synthetic sequence of VGA images; the frame rate of the sequence is 60 Hz and the field of view is 50 deg by 40 deg. The motion of the virtual camera consists of a realistic human motion combining translations and rotations. The virtual

scene features a circular panel placed in a room with walls, floor and ceiling; the observed surfaces are textured by a sine varying gray pattern. A normally distributed noise $\mathcal{N}(0, 1)$ may be added to each image. An image of the sequence, and the corresponding ground truth depth field are represented in Figure 3.

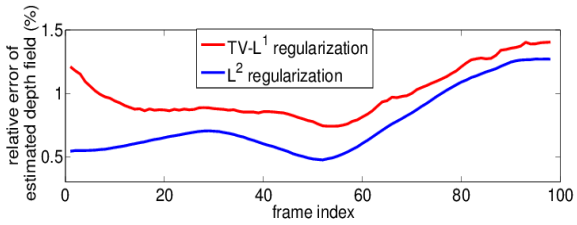

Figure 4. Comparison of regularizing the depth field using a L2 method (red) and our proposed approach (blue)

First, the proposed method is tested on a noisy sequence of images, and the results are compared to those obtained by a method relying on a more straightforward depth regularization in L^2 norm, introduced in [11]. The results are plotted in Figure 4, as the cumulated relative errors of estimation over the entire field of view. TV-L^1 regularization shows a better performance than L^2 regularization as depth discontinuities are better preserved.

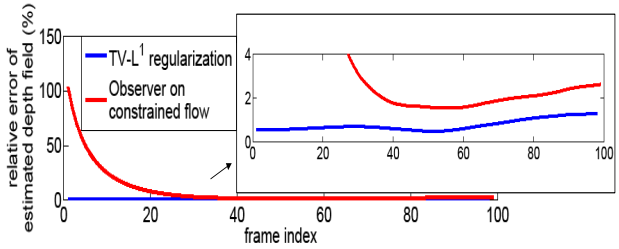

Figure 5. Comparison of filtering the depth field using an observer for the constrained flow provided by [9] (red) and the proposed approach for TV-L^1 depth (blue)

In order to highlight the advantage of direct depth estimation over optical flow filtering in an equivalent motion-constrained framework, we apply the observer described in [11] to the optical flow obtained by TV-L^1 regularization and constrained by a fundamental matrix prior [9]. The errors are plotted in Figure 5, and show that the proposed method is more accurate and provides instantaneous estimates.

5. Concluding remarks

We have considered a depth map estimation problem defined for a moving monocular system observing a static scene. We have proposed a variational method which takes explicitly into account camera dynamics information in order to constrain the depth regularization by means of a robust TV-L^1 functional. The results show that this approach performs better than an equivalent solution based on optical flow estimation, and that an asymptotic observer employed for temporal data fusion improves the overall stability of the system, while avoiding batch optimizations. We intend to extend this work by coupling it with a sparse keypoint based SLAM in order to recover a dense depth field on top of a robust localization system.

References

[1] D. Bitton, G. Rosman, T. Nir, A. M. Bruckstein, A. Feuer, and R. Kimmel. Over-parameterized optical flow using a stereoscopic constraint. In *SSVM*, 2011.

[2] S. Bonnabel and P. Rouchon. Fusion of inertial and visual : a geometrical observer-based approach. *CISA*, 1107:54–58, 2009.

[3] A. Chambolle. An algorithm for total variation minimization and applications. *JMIV*, 20(1):89–97, 2004.

[4] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *JMIV*, pages 1–26, 2010.

[5] B. Horn and B. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[6] C. Mei, G. Sibley, M. Cummins, P. Newman, and I. Reid. Rslam: A system for large-scale mapping in constant-time using stereo. *IJCV*, 94:198–214, 2010.

[7] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[8] G. Vogiatzis and C. Hernández. Video-based, real-time multi-view stereo. *Image Vision Comput.*, 2011.

[9] A. Wedel, T. Pock, J. Braun, U. Franke, and D. Cremers. Duality TV-L^1 flow with fundamental matrix prior. In *IVCNZ 2008*, 2008.

[10] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L^1 optical flow. *Pattern Recognition*, pages 214–223, 2007.

[11] N. Zarrouati, E. Aldea, and P. Rouchon. SO(3)-invariant asymptotic observers for dense depth field estimation based on visual data and known camera motion. In *American Control Conference, Montreal*, 2012. (arXiv:1103.2539v2).