A note on
Real-Time Visual Odometry from Dense RGB-D Images

Jinbin Huang

Takeaways:

1. Rigorous mathematical formulation of the visual odemetry problem.
2. Taylor approximation of non-convex energy function

1 Mathematical Formulation

The geometries of 3d objects are represented by their surfaces. In this paper we consider all functions as differentiable for the sake of simplicity. Such assumption is sound at least in this research.
Let

$$I_{RGB} : \Omega \times \mathbb{R} \to [0,1]^3, \qquad (\vec{x}, t) \mapsto I_{RGB}(\vec{x}, t) \tag{1}$$

$$h : \Omega \times \mathbb{R} \to \mathbb{R}, \qquad (\vec{x}, t) \mapsto h(\vec{x}, t) \tag{2}$$

denote the color image data and height field on the image place $\Omega \subset \mathbb{R}^2$. From the height field, we can compute a surface S,

$$S : \Omega \to \mathbb{R}^3, \qquad x \mapsto S(x) \tag{3}$$

$$S(\vec{x}) = \left(\frac{(x + o_x) \cdot h(x)}{f_x}, \frac{(y + o_y) \cdot h(x)}{f_y}, h(x) \right)^T \tag{4}$$

where $\vec{x} = (x, y)^T$ is the 2d coordinate of a point in the image plane, $(o_x, o_y)^T$ denotes the principal point of the camera and f_x and f_y the focal lengths. Again, for simplicity, we consider only gray-scale image, i.e., we define $I = (I_R + I_G + I_G)/3$.

Given two images: $I(t_1), I(t_0)$ with surfaces $S(t_1), S(t_0)$, we now seek for the rigid body motion $g \in SE(3)$ of the camera between t_1 and t_0

The key idea is that **the rigid body motion $g_{l \to r}$ in combination with the surface S_l induces a unique mapping from pixels in I_l to pixels in I_r.**

1.1 Mathematical Induction of Energy Function

We represent the rigid body motion as a six-degree-of-freedom vector $\xi = (\omega_1, \omega_2, \omega_3, v_1, v_2, v_3)^T \in \mathbb{R}^6$. It defines a twist

$$\hat{\xi} = \begin{pmatrix} 0 & -\omega_3 & -\omega_2 & v1 \\ \omega_3 & 0 & -\omega_1 & v2 \\ -\omega_2 & \omega_1 & 0 & v3 \\ 0 & 0 & 0 & 0 \end{pmatrix} \tag{5}$$

Assuming that $\xi(t)$ is constant in the temporal interval $[t_0, t_1]$, the rigid body motion, $g(t_1)$ is given by the matrix exponential

$$g(t_1) = exp((t1 - t0)\hat{\xi})g(t_0) \tag{6}$$

where g is a 4×4 homogeneous matrix of the form

$$g = \begin{pmatrix} R & T \\ 0 & 1 \end{pmatrix}, \quad \text{with } R \in SO(3), T \in \mathbb{R}^3 \tag{7}$$

With simple calculus, we deduce

$$\frac{\mathrm{d}g}{\mathrm{d}t}(t) = \hat{\xi}(t)g(t) \tag{8}$$

Rigid body motions g of the camera give rise to respective transformation G of 3D points $P \in \mathbb{R}^3$

$$G : SE(3) \times \mathbb{R}^3 \to \mathbb{R}^3, \qquad G(g, P) = RP + T \tag{9}$$

The respectively transformed surface $G(g, P) = (G1, G2, G3)$ is then projected to the image plane by the projection map $\pi : \mathbb{R}^3 \to \Omega$, given by:

$$\pi(G) = \left(\frac{G_1 f_x}{G3} - o_x, \frac{G_2 f_y}{G_3} - o_y \right)^T \tag{10}$$

The warping operation $\omega_\xi(\vec{x}, t)$ is a combination of respective transformation and projection

$$\omega_\xi : \Omega \times \mathbb{R}_+ \to \Omega, \qquad (x, t) \mapsto \omega_\xi(x, t) \tag{11}$$

$$\omega_\xi(\vec{x}, t) = \pi \left(G(exp((t - t_0)\hat{\xi})g(t_0), S(\vec{x})) \right) \tag{12}$$

2 Error in terms of photoconsistency

We want to compute a twist ξ between t_1 and t_0 which minimizes the least-square error

$$E(\xi) = \int_\Omega \left[I(\omega_\xi(x, t_1), t_1) - I(\omega_\xi(x, t_0), t_0) \right]^2 dx \tag{13}$$

Without no loss of generality, we assume $g(t_0) = id$ and hence the second term of (11) reduces to

$$I(\omega_\xi(x, t_0), t_0) = I(x, t_0) \tag{14}$$

2.1 Linearization of Energy

The energy function given by (11) is non-convex in the parameter ξ and therefore finding the minimum is non-trivial. To overcome this limitation, we approximate both the image at time $t_1 : I(t1)$ and the corresponding warp w by first-order Taylor approximations.

$$I(\omega_\xi(\vec{x}, t_1), t_1) \approx I(\vec{x}, t_1) + (\omega_\xi(x, t_1) - x) \cdot \nabla I(\vec{x}, t_1) \tag{15}$$

and

$$\omega_\xi(\vec{x}, t_1) \approx \vec{x} + (t_1 - t_0) \cdot \underbrace{\frac{\mathrm{d}(\pi \circ G \circ g)}{\mathrm{d}t}}_{=\frac{\mathrm{d}w}{\mathrm{d}t}} \Big|_{(x,t_0)} \tag{16}$$

By using the approximations (15) and (16), we get

$$E_l(\xi) = \int_\Omega \left(I(x, t_1) - I(x, t_0) + \nabla I(x, t_1) \cdot (t_1 - t_0) \cdot \frac{\mathrm{d}w}{\mathrm{d}t}(x, t_0) \right) \tag{17}$$

WLOG, we set $(t_1 - t_0) = 1$, since it is only a scalar factor to the minimizing ξ of the linearized energy. Additionaly, we can assume that the temporal derivative of the image is constant between t_0 and t_1, so we can substitute $I(x, t_1) - I(x, t_0) = \frac{\partial I}{\partial t}$ and obtain

$$E_l(\xi) = \int_\Omega \left(\frac{\partial I}{\partial t} + \nabla I(x, t_1) \cdot \frac{\mathrm{d}w}{\mathrm{d}t}(x, t_0) \right) dx \tag{18}$$

By means of the chain rule, we can express the toatal derivative $\frac{\mathrm{d}w}{\mathrm{d}t}$ in (18) as the product of several total derivations:

$$\frac{\mathrm{d}w}{\mathrm{d}t} = \frac{\mathrm{d}\pi}{\mathrm{d}G}\Big|_{\pi(G(g(t_0),S(\vec{x}))} \cdot \frac{\mathrm{d}G}{\mathrm{d}g}\Big|_{G(g(t_0),S(\vec{x}))} \cdot \frac{\mathrm{d}g}{\mathrm{d}t}\Big|_{t_0} \tag{19}$$

With this, the energy becomes

$$E_l(\xi) = \int_\Omega \left(\frac{\partial I}{\partial t} + \nabla I \cdot \frac{\mathrm{d}\pi}{\mathrm{d}G} \cdot \frac{\mathrm{d}G}{\mathrm{d}g} \cdot \frac{\mathrm{d}g}{\mathrm{d}t} \right)^2 dx \tag{20}$$

Where the evaluation points are neglected to improve readability. Next, we plug (8) into (20) and get

$$E_l(\xi) = \int_\Omega \left(\frac{\partial I}{\partial t} + \nabla I \cdot \frac{\mathrm{d}\pi}{\mathrm{d}G} \cdot \frac{\mathrm{d}G}{\mathrm{d}g} \cdot \hat{\xi} \cdot g(t) \right)^2 dx \tag{21}$$

Note that $\hat{\xi} \cdot g(t)$ is a 4×4 matrix and hence the derivative $\frac{\mathrm{d}G}{\mathrm{d}g}$ is a $3 \times 4 \times 4$ tensor. Since the elements in the last row of $\hat{\xi}$ are all zeros, the matrix $\hat{\xi} \cdot g(t)$ can be represented as a vector $stack(\hat{\xi} \cdot g(t))$ in $\mathbb{R}^{12}$. One can easily verify that there exsits one 4×4 matrix M_g such that $stack(\hat{\xi} \cdot g(t)) = M_g \cdot g(t)$ (Note that all elements of $stack(\hat{\xi} \cdot g(t))$ are linear combination of $(\omega_1, \omega_2, \omega_3, v_1, v_2, v_3)^T$).

Representing $\hat{\xi} \cdot g(t)$ by $M_g \cdot \xi$, we get the final form of our energy function:

$$E_l(\xi) = \int_\Omega \left(\frac{\partial I}{\partial t} + \left(\underbrace{\nabla I \cdot \frac{\mathrm{d}\pi}{\mathrm{d}G} \cdot \frac{\mathrm{d}G}{\mathrm{d}g} \cdot M_g}_{=:C(x,t_0)} \Big|_{(x,t_0)} \cdot \xi \right) \right)^2 dx \tag{22}$$

References

Steinbrücker, Frank, Jürgen Sturm, and Daniel Cremers. "Real-time visual odometry from dense RGB-D images." 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, 2011.