A variational approach for estimation of monocular depth and camera motion in autonomous driving

Huijuan Hu¹, Chuan Hu² and Xuetao Zhang³

Abstract

In this paper, a new direct computational approach to dense 3D reconstruction in autonomous driving is proposed to simultaneously estimate the depth and the camera motion for the motion stereo problem. A traditional Structure from Motion framework is utilized to establish geometric constrains for our variational model. The architecture is mainly composed of the texture constancy constraint, one-order motion smoothness constraint, a second-order depth regularize constraint and a soft constraint. The texture constancy constraint can improve the robustness against illumination changes. One-order motion smoothness constraint can reduce the noise in estimation of dense correspondence. The depth regularize constraint is used to handle inherent ambiguities and guarantee a smooth or piecewise smooth surface, and the soft constraint can provide a dense correspondence as initial estimation of the camera matrix to improve the robustness future. Compared to the traditional dense Structure from Motion approaches and popular stereo approaches, our monocular depth estimation results are more accurate and more robust. Even in contrast to the popular depth from single image networks, our variational approach still has good performance in estimation of monocular depth and camera motion.

Keywords

Monocular depth estimation, 3D reconstruction, optical flow, camera motion, structure from motion

Date received: 19 March 2021; accepted: 8 June 2021

Introduction

Structure from Motion (SfM) is a crucial challenge in computer vision for a long time. Most existing state-of-the-art systems are regarded to be involving several consecutive processing steps. An important goal of these pipes is the estimation of the structure and relative motion for consecutive image pairs. There are some intrinsic limitations in current implementations of this step. For example, it is common sense to estimate the camera motion before recovering the structure of the scene by means of dense correspondence matching. Therefore, erroneous estimation of camera motion can lead to incorrect depth prediction. In addition, the lower-level process that the camera motion is estimated from sparse correspondences matching is prone to outliers and in-effective in non-textured areas. Finally, all current existing SfM approaches are not applicable to the situation of small camera translation as it is difficult to integrate priors knowledge to obtain the rational solutions in those degenerate cases. SfM estimated in a small baseline setting is regarded unreliable due to bas-relief ambiguity and a large amount of rotation between two cameras is required for accurate reconstruction.

Presently, most work on estimation of depth and motion from image pairs follows Marr and Poggio’s computational vision framework¹,² which was biologically inspired by human vision, and therefore consists of two processes of feature detection (acting as the correspondence role) and 3D reconstruction from matched features. Longuet-Higgins¹ presented the first algorithm for reconstructing 3D scene from two views

¹CAS Key Laboratory for Quantitative Engineering Biology, Shenzhen Institute of Synthetic Biology, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, P.R. China
²Department of Mechanical Engineering, University of Alaska Fairbanks, Fairbanks, AK, USA
³The Department of Automation Science and Technology, Xi’an Jiaotong University, Xi’an, P.R. China

Corresponding author:
Chuan Hu, Department of Mechanical Engineering, University of Alaska Fairbanks, 1764 Tanana Loop, Fairbanks, AK 99775-5905, USA.
Email: chuan.hu.2013@gmail.com
where the first process of correspondence was assumed to have been solved and only the second process of 3D reconstruction was investigated. State-of-the-art systems\cite{4,5} are utilized for reconstructions of large scenes including big cities. They consist of a series of methods, starting with descriptor detecting and matching for searching sparse correspondences between image pairs, followed by solving the essential matrix to recover the camera motion. Before applying bundle adjustment to jointly optimize camera motions and structure of multiple images, the estimated accuracy mainly relies on the quality of the initial camera motion and structure between image pairs. Dense depth maps can be obtained by exploiting the epipolar geometry constraint after recovery of camera motion and sparse 3D point cloud. Such a divide-and-conquer strategy has been commonly adopted in computer vision community and a branch of multi-view geometry has been developed. Unlike above approaches, LSD-SLAM\cite{15} considers multiple consecutive images from a short temporal window by jointly optimizing semi-dense correspondences and depth maps. DTAM\cite{6} can estimate camera poses reliably by matching against dense depth maps. However, it still relies on classic structure and motion methods as an external depth initialization. Lately, CNN based optical flow methods\cite{7,8,9,10} have been proposed and show their good performances. These methods also can provide a dense correspondence to estimate the camera matrix for further 3D reconstruction.

Recently, SfM and optical flow (dense correspondences) share many connections. Estimation of Depth and camera pose from dense correspondence search has been presented by Valgaerts et al.\cite{11} with the epipolar constraint as an extra term into the objective function to build a direct link between optical flow and SfM problem. Lately, Aubry et al.\cite{12} presented a joint photometric and geometric variational model which focused on variational camera calibration to estimate camera extrinsics and dense correspondence for reconstruction. Other related work investigated in the framework of variational optimization include\cite{13,14}. In Becker et al.\cite{13}, the depth and camera motion were estimated by utilizing probabilistic inference in a video sequence. In Bagnato et al.\cite{14} the depth and camera motion were alternately optimized for omnidirectional image sequences. Roxas and Oishi\cite{15} proposed to separately optimize the camera extrinsics and 3D structure in existing SfM methods by adding the epipolar geometry as a soft constraint and refining the dense correspondence. However, all above referenced approaches are brute and two-stages for depth estimation. Therefore, extra error would be introduced into further recovering 3D structure by the incorrect camera motion estimation.

Meanwhile, Dense 3D reconstruction using motion stereo vision has been a hotspot research subject. Several methods have been proposed in successfully solving the 3D reconstruction problem.\cite{16,17,18} Stühmer et al.\cite{16} presented a TV-L1 minimization framework to handling the correspondence problem. Graber et al.\cite{17} solved the smoothness problem based on minimal-surface smoothness regularization. Galliani et al.\cite{18} proposed a multi-view PatchMatch Stereo approach with an additional 3D integration step. Recently, many depth estimation from single image networks\cite{20,21,22,23} have been presented to estimate the monocular depth for automatic drive and have showed their good performances. However, they all need extra information, that is, view synthesis or stereo images as the supervisory signal. Meanwhile, those methods are not suitable for multi-view 3D model reconstruction.

In this paper, we propose a new direct computational approach to solve dense 3D reconstruction problem that includes computation of depth, estimation of dense correspondences, and the camera motions between image pairs. Here, the “direct” means that the intermediate process of feature detection or correspondence is not needed in the proposed approach. Instead, as a “by-product”, the correspondence task is accomplished in the process of depth estimation. Specially, the admission solution for the depth is restricted in smooth surfaces, in order to overcome the ill-posedness of the binocular depth estimation where the depth in 3D space is to be inversely recovered from 2D images. The camera motion and dense 3D structure are estimated by minimizing an variational objective function. Different from Hu and Chen\cite{24} the texture constancy constraint, one-order motion smoothness constraint, a second-order depth regularize constraint, and a soft constraint are integrated in the objective function in the proposed method. The texture constancy constraint can improve the robustness of our method to illumination change, one-order motion smoothness constraint can smooth the calculated optical flow field and reduce the noise in estimation of dense correspondence. The depth regularize constraint is used to guarantee a smooth or piecewise smooth surface, and the soft constraint can provide a dense correspondence as initial estimate of the camera matrix. This work can effectively addresses the dense reconstruction problem, and as a result, improves the quality of depth prediction and camera motion estimation. In summary, we show a variational model for jointly estimating of depth, camera motion which effectively breaks through Marr and Poggio’s biologically inspired framework on computational vision. The experimental results show that the proposed approach can work well even on complex scene.

The rest of the paper is organized as follows. In section II and III, basic concepts and the direct variational approach are reviewed. The Implementation about occlusion handling and ambiguity and numerical solution are presented in section IV. Experimental results and conclusion are given in section V and VI.

Basic concepts

Let us start by defining image pairs of a static scene $\mathcal{I} = \{I_1, I_2\} : \Omega \rightarrow \mathbb{R}^+$. Assuming $x$ and $w$ indicate the
image lattice and flow field, respectively. The forward optical flow from $I_1$ to $I_2$ of a pixel $x = (x, y)$ in the image domain $\Omega$ is defined as $w = (u, v, 1)$ in homogeneous representation. Otherwise, we assume a calibrated settings. The 3D point corresponding to pixel $x$ is marked as $X \in \mathbb{R}^3$. Assume that $K_1$ and $K_2$ denote the corresponding calibration internal camera parameters, the camera matrices are expressed as $P = [I_1 | 0] \in \mathbb{R}^{3 \times 4}$ and $P' = [R | t] \in \mathbb{R}^{3 \times 4}$, respectively. $R$ is the rotation matrix characterized by three rotation angles $\theta = (\theta_x, \theta_y, \theta_z)$ and $t = (t_x, t_y, t_z)$ is the translation vector between two cameras. The reconstructed 3D coordinate corresponding to pixel $x$ of the first view is determined by its depth $q_{x,y}$: $(q_{x,y}x, q_{x,y}y, q_{x,y})$ and the perspective projection to the second view can be expressed as $(x'_{x,y}, y'_{x,y})$:

\[
\begin{align*}
  x'_{x,y} &= P'_i q_{x,y}x, \\
  y'_{x,y} &= P'_i q_{x,y}y
\end{align*}
\]

where $P'_i (i = 1, 2, 3)$ is the $i^{th}$ row vector of the second camera matrix $P'$. $q_{x,y}$ can be simplified as $q$. For a concise representation, as the cameras are calibrated, suppose camera matrix $P'$ and depths $q$ are embedded in a parameter vector $\hat{\theta}$ which can be expressed as $\hat{\theta}(\theta, t, q)$. The depth estimation can be considered as an energy function of $\hat{\theta}(\theta, t, q)$. In this way, we can see that six variables of the camera matrix which contains three rotation angles and the translation parameters, and $n$ variables of the depth $q$ construct the parameter vector $\hat{\theta}$ with several dimensions of $n + 6$, wherein $n$ represents the number of images pixels.

**Framework overview**

**Dense correspondence constraint**

The brightness constancy assumption states that the intensity of a pixel keeps constant in two views when the object or camera moves. The brightness constancy constraint has been used in most of optical flow approaches. It is usually expressed as $I_1(x) = I_2(x + w)$ With Taylor expansion, the brightness constancy constraint can be linearized as follows:

\[
I_1(x,y)u + I_2(x,y)v = -I_1(x,y)
\]

where $I_1(x,y) = \frac{\partial I_1(x,y)}{\partial x}$, $I_2(x,y) = \frac{\partial I_2(x,y)}{\partial y}$, and $I_1(x,y) = I_2(x,y) - I_1(x,y)$. In spite of that, Equation (2) is under-determined equation. It consists of only a single equation with two unknown variables. It can be solved only under the condition that other constraint is added. that is, smoothness constraint. In addition, the brightness constancy constraint does not work well when the illumination changes. In order to deal with this problem, the gradient constancy constraint has been leveraged in many optical flow approaches and this assumption states that image gradients stays constant. It does not change with the movement of the object or the camera.

Recently, image textures based on different types of descriptors, such as the histogram of oriented gradient,\cite{29} the modified local directional pattern,\cite{31} and the census signature\cite{32} have been successfully embedded in the optical flow estimation. The image textures are robust against illumination changes. Afterwards, the variational model has been modified to utilize the texture constancy assumption based on the extracted features. In our paper, the framework of dense correspondences which contains the texture constraint based on the histogram of oriented gradient\cite{32} and a one-order motion smoothness constraint is formulated as:

\[
f(w) = f_{data}(w) + f_{smooth}(w) = \sum_{i=1}^{m} \sum_{j=1}^{\Omega} \Psi \left( |S'(x + w) - S'(x)|^2 \right) + \alpha \Psi \left( |\nabla w|^2 \right),
\]

where $m$ states the number of color channels, $S'$ represents the extracted descriptors extracted, the modified $L_1$ norm $\Psi(x^2) = \sqrt{x^2 + \varepsilon}$ with a small positive constant $\varepsilon$ to deal with discontinuities in the flow field, and $\alpha$ is a constant. In practice, the summation will make sure the distance between the two descriptors as small as possible and the motion as smooth as possible.

**Second-order depth regularization**

It is a common sense that the second-order smoothness prior as depth constraint is more suitable for scenes with complex geometric characteristics to deal with ambiguity at object boundaries. Considering slanted surfaces in the reconstruction, we take advantage of the second-order depth regularization which allows to model linear depth changes instead of using a first-order depth constraint that inherently favors fronto-parallel surfaces. Then, a stair effect for slanted or highly curved surfaces is generated: $\Delta q = |q(x_0) - 2q(x) + q(x_1)|$.

The triple-clique $C = \{x_0, x, x_1\}$ can be obtained from all $3 \times 1$ and $1 \times 3$ patches in the nearest neighbor system of pixel $x$. To avoid penalizing depth discontinuities aligned with intensive edges, a weight for each patch is added:

\[
\omega(x) = \exp(-|I(x_0) - 2I(x) + I(x_1)|/\iota),
\]

where $\iota$ is a parameter controlling the significance of an edge. We choose the same $\iota = 20$ for all directions. $\omega(x)$ is simplified as $\omega$. 

This second-order depth smoothness prior can help to capture richer features of local structure and permit planar surfaces.

\[ f_{\text{depth}}(q) = \sum_{\Omega} \Psi(\omega^2 \| \Delta q \|^2). \]  

(5)

In order to overcome the tendency for large depths, the following regularization term is used:

\[ f_{\text{depth}}(q) = \sum_{\Omega} \Psi(\omega^2 \| \Delta q \|^2 / \| \Delta q \|). \]  

(6)

**Soft regularization**

As descriptor matching is a discrete technique that can not provide subpixel accuracy in optical flow estimation, to further improve the accuracy of depth estimation and the robustness in complex scene, we would like to combine some excellent optical flow matching with the variational model in its coarse-to-fine optimization.

\[ f_{\text{soft}}(\theta) = \sum_{\Omega} \delta(x) \Psi(\| w(\theta) - w_{\text{om}} \|^2). \]  

(7)

\[ \delta(x) = \begin{cases} 1, & \text{descriptor available at } x, \\ 0, & \text{otherwise.} \end{cases} \]  

(8)

where \( w_{\text{om}} \) is the sparse optical flow value. The auxiliary variable \( w_{\text{om}} \) is used to integrate sparse matching into our variational approach in the form of soft constraints to obtain preciser sparse correspondences. In this paper, FlowNet2-CSS34 and SIFT-flow are used for sparse matching in KITTI and Strecha dataset, respectively. Meanwhile, these two methods can perform sparse matching to solve an accurate pose estimation as the initial camera optimization variables.

**Variational model**

Above all, the flow field \( w \) is expressed as functions with respect to \( \theta \) as the parameter variables: \( w(\theta) \). Thus, the objective function (3) can be turned into \( f(\theta) = f(w(\theta)) \) in terms of \( \theta \).

As described above, we propose to impose a soft constraint on the optical flow and a second-order smoothness constraint directly on the depth. Therefore, the following objective function is used:

\[ f(\theta) = f_{\text{data}}(\theta) + \alpha f_{\text{smooth}}(\theta) + \alpha f_{\text{depth}}(q) + \alpha_{\text{soft}} f_{\text{soft}}(\theta), \]  

(9)

where the depth parameter \( q \) and the camera matrix \( P \) is encapsulated in \( \theta \). \( f_{\text{data}} \) states that two images are to be matched as best as possible. Three regularizers \( f_{\text{smooth}}, f_{\text{depth}}, \) and \( f_{\text{soft}} \) are used to resolve ambiguities between the unknowns to guarantee a smooth or piecewise smooth surface.

**Implementation**

**Occluded handling**

Occlusion estimation is a universally known and long-standing difficult issue that 3D reconstruction has been entangled with. Accurate priori knowledge of occluded regions is crucial for reliable 3D reconstruction. Yet, estimating accurate 3D reconstruction, conversely, is required for localizing occlusions reliably as it causes severe distortion around occlusion. Checking the left-right motion consistency in a subsequent post-processing step and extrapolating flow into inconsistent regions also helps resolving the motion mismatch in the occluded area. Occlusion is a challenging issue for our proposed method as 3D structure can’t be rebuilt for occluded parts. Here, our occlusion detection is based on the forward-backward consistency assumption. \(^{35}\) It means that we hope the forward flow should be equal to the inverse of the backward flow at the corresponding pixel in the second view for non-occluded pixels. If the mismatch between the two flows is too large, the pixels can be marked as occluded. Thus, occlusion in the forward direction can be set to be one wherever the constraint is violated, and zero otherwise.

The occlusion in the forward direction can be defined as:

\[ | w_L - w_R(2x_L + w_L) |^2 < \alpha_1( | w_R |^2 + | w_L(2x_R + w_R) |^2 ) + \alpha_2. \]  

(10)

The occlusion in the backward direction can be defined as the same way. In all of our experiments, \( \alpha_1 = 0.01, \alpha_2 = 0.5 \).

With the forward-backward consistency constraint, the pixels satisfy the bi-directional flow consistency are forced to be visible. Otherwise, they are identified as occlusions.

**Update of projection matrix in the coarse-to-fine framework**

In the coarse-to-fine optimizing stage for our approach, with a ratio \( \sigma < 1 \) scaling, an initial value of depth is obtained by interpolating the estimate of the next coarser level. For SfM in equation (9), except the interpolation of depth, care has to be taken to deal with the projection matrix \( P \). From level \( m \) to \( m - 1 \), the image scales \( \frac{1}{\sigma} \) times. It can be easily verified that the \((m - 1)^{th}\) level projection matrix of the second camera \( P_m^{m-1} \) is as follows, provided \( P_m \) for the \( m^{th} \) level

\[ P_m^{m-1} = \frac{1}{\sigma} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} P_m \begin{pmatrix} \sigma & 0 & 0 & 0 \\ 0 & \sigma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \]  

(11)
Ambiguities

As for depth estimation, in Euclidean/similarity reconstruction, there exist two types of inherent ambiguities: scale ambiguity and bas-relief ambiguity. Regardless of calibrated or un-calibrated cameras, the scale of the reconstructed scene can’t be determined due to the scale ambiguity. That is, when the translation vector between two cameras scales k times and the scene scales k times, no change on two views is resulted in.

Mathematically, the objective function does not change when inherent ambiguities happen. Supposing \( \{\hat{\theta}, \hat{i}, \hat{q}\} \) is the estimated Euclidean solution, \( \{\hat{\theta}, k\hat{i}, k\hat{q}\} \) is another equivalent solution that produces the same value of the objective function, for any \( k > 0 \). Such a requirement explains the reason why the regularization term in equation (6) is not \( |\Delta q|^2 \), but \( |\Delta q|^2/\|q\|^2 \). That is, when \( q \) scales, the latter does not change and consequently, the regularization term in (6) does not affect the inherent ambiguities. It can also be seen from equation (9), the reconstructed 3D coordinates scale \( k \) times as the depth scales. Thus, only the scaled 3D structure can be estimated. The scale ambiguity can be overcome by knowing the translation vector \( t \) between two cameras, or \( \|t\| \), the distance between two cameras.

In addition to the scale ambiguity, the bas-relief ambiguity happens, when the rotation angle between two cameras is small enough and consequently the focal length can’t be reliably estimated in such settings. The effect of bas-relief ambiguity becomes clearer in a simplified model, where two parallel cameras are:

\[
\begin{bmatrix}
  f & 0 & 0 \\
 0 & f & 0 \\
 0 & 0 & 1
\end{bmatrix} \quad \text{and} \quad \begin{bmatrix}
  f & 0 & 0 \\
 0 & f & 0 \\
 0 & 0 & 1
\end{bmatrix}
\]

Similarly, when two parallel cameras have a \( k \) times depth direction, double views would remain same, as can be verified from the pinhole model.

Changing to the default camera representation, where the first camera projection matrix is \( \begin{bmatrix} f_3 & 0 \\ 0 & f_3 \\ 0 & 0 & 1 \end{bmatrix} \), the second camera projection matrix is

\[
\begin{bmatrix}
  1 & 0 & 0 \\
  0 & 1 & 0 \\
  0 & 0 & 1
\end{bmatrix}
\]

It can be verified by substitution that, if \( \{\hat{f}, \hat{i}, \hat{q}\} \) is a solution, \( \{k\hat{f}, \hat{i}, k\hat{q}\} \) is another equivalent solution for any \( k > 0 \). This ambiguity results in a stretching distortion in the depth direction: a 3D point \( (\hat{X}, \hat{Y}, \hat{Z}) \) in the \( \{\hat{f}, \hat{i}, \hat{d}\} \) reconstruction has coordinates of \( (\hat{X}, \hat{Y}, k\hat{Z}) \) in the \( \{k\hat{f}, \hat{i}, k\hat{q}\} \) reconstruction.

Scale ambiguity and bas-relief ambiguity are actually two particular sub-classes of projective ambiguities, with a projective transformation of \( H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & k \end{bmatrix} \) and

\[
\begin{bmatrix}
  1 & 0 & 0 & 0 \\
  0 & 1 & 0 & 0 \\
  0 & 0 & 1 & 1
\end{bmatrix}
\]

Sensitiveness to noise due to bas-relief ambiguity

In addition, the SFM estimation in a small-baseline setting is regarded unreliable due to the bas-relief ambiguity. Here, it should be stressed that the unreliability in SFM does not mean the bas-relief ambiguity itself, but the sensitiveness to the inaccuracy in correspondence or estimated disparity when the bas-relief ambiguity happens. The ambiguities happen because some factor, for example the scale of the focal length, can’t be visually determined, and consequently, the 3D reconstruction can be done up to a subclass of projective transformations.

The difficulty with the bas-relief ambiguity lies in the sensitiveness to the correspondences in accuracy in SFM when the bas-relief ambiguity (approximately) happens. Specifically, the inaccuracy in depth estimation due to the error in estimated disparity is approximately proportional to the square of the depth when the bas-relief ambiguity happens, as can be verified by the fact that the angular disparity approximately decreased with the square of the observation distance. Consequently, sparse SFM in a small baseline setting is still a challenge for existing algorithms, let alone propagation-based dense or quasi-dense SFM. Moreover, the resulted sensitiveness due to bas-relief ambiguity is aggravated by the current separate-manner SFM, to bas-relief ambiguity is aggravated by the current separate-manner SFM, where the depth estimation can be regarded as being separately accomplished feature by feature.

On the other hand, such sensitiveness to noise in a small-baseline SFM can be suppressed to a great degree in the proposed approach. As a result of the regularization technique, the depth estimation in the proposed approach is implemented, not in a separate one-by-one manner, but in a locally constrained way with its neighbors. Thus, noise in a single pixel or even in a local patch does not lead to unpredictable error. In addition, the advantage of the proposed approach can also be demonstrated in a complex scene, where a satisfactory 3D reconstruction is produced except around occlusions.

A new interpretation in nonlinear LS

Note that each term of the objective function (9) is convex. Above all, the objective function (9) can be rewrite as:
 \( f(P', q) = \sum_{i=1}^{m} \left( \psi \left( |S'(x + w) - S'(x)|^2 \right) + \alpha \nabla w \right) \)
\[ + \alpha \nabla (\omega^2 \Delta q^2 / ||q||^2) + \alpha \delta(x) \psi \left( |w(\theta) - w_{sm}|^2 \right). \]
\[ \tag{12} \]

wherein
\[ \nabla w = \frac{\partial w}{\partial q} \nabla q. \]
\[ \tag{13} \]

The objective function (12) is optimized by employing an iterative algorithm in nonlinear LS. We define an integral \( L \) which meet the standard form of the Euler-Lagrange constraint in calculus of variations as follow. (Omitting the camera matrix parameter \( P' \) temporarily, taking account of a depth parameter \( q \) only, \( q_x \) and \( q_y \) are the first order differentials of \( q \).)

\[ L(x, y, q, q_x, q_y) = \sum_{i=1}^{m} \psi \left( |S'(x + w) - S'(x)|^2 \right) + \alpha \nabla w \]
\[ + \alpha \nabla (\omega^2 \Delta q^2 / ||q||^2) + \alpha \delta(x) \psi \left( |w(\theta) - w_{sm}|^2 \right). \]
\[ \tag{14} \]

According to the Euler-Lagrange equation, for reaching an extreme value, the objective function (14) must meet:

\[ L_q = \frac{d}{dx} L_{q_x} - \frac{d}{dy} L_{q_y} = 0. \]
\[ \tag{15} \]

Particularly, the Euler-Lagrange equation (15) is:

\[ \sum_{i=1}^{m} \left( \psi' S'_i \left( \frac{\partial u}{\partial q} + \frac{\partial v}{\partial q} \right) \right) \]
\[ + \alpha \nabla w \left( \frac{\partial u}{\partial q} + \frac{\partial v}{\partial q} \right) \]
\[ \left( q_x^2 + q_y^2 \right) = \frac{\partial^2 u}{\partial q^2} q_x + \frac{\partial^2 v}{\partial q^2} q_y \]
\[ - \left( \frac{\partial u}{\partial q} \right)^2 + \left( \frac{\partial v}{\partial q} \right)^2 \]
\[ \text{div} \psi' \left( \frac{\partial u}{\partial q} \frac{\Delta q^2}{||q||^2} \frac{\partial v}{\partial q} \right) \]
\[ + \alpha \nabla w \left( \frac{\partial u}{\partial q} \frac{\Delta q^2}{||q||^2} \frac{\partial v}{\partial q} \right) \]
\[ + \alpha \nabla \psi' \left( u - u_{sm} \right) \frac{\partial u}{\partial q} \]
\[ + \left( v - v_{sm} \right) \frac{\partial v}{\partial q} = 0 \]
\[ \tag{16} \]

where the symbolic expressions are used as:

\[ \psi' \]
\[ \text{div} \left( \nabla \psi \right) + \alpha \nabla \psi' \left( \frac{\partial u}{\partial q} \frac{\Delta q^2}{||q||^2} \frac{\partial v}{\partial q} \right) \]
\[ + \alpha \nabla \psi' \left( u - u_{sm} \right) \frac{\partial u}{\partial q} \]
\[ + \left( v - v_{sm} \right) \frac{\partial v}{\partial q} = 0 \]

Meanwhile, according to the theorem that the derivative of the extremum is zero, the objective function (12) meets:

\[ \frac{\partial f}{\partial P'} = 0. \]
\[ \tag{18} \]

that is,

\[ \sum_{\Omega} \left\{ \sum_{i=1}^{m} \left( \psi' S'_i \left( \frac{\partial u}{\partial q} + \frac{\partial v}{\partial q} \right) \right) + \alpha \nabla w \left( \frac{\partial u}{\partial q} + \frac{\partial v}{\partial q} \right) \left( q_x^2 + q_y^2 \right) \frac{\partial^2 u}{\partial q^2} q_x + \frac{\partial^2 v}{\partial q^2} q_y \right\} \]
\[ + \alpha \nabla \psi' \left( u - u_{sm} \right) \frac{\partial u}{\partial q} + \left( v - v_{sm} \right) \frac{\partial v}{\partial q} \}
\[ = 0 \]
\[ \tag{19} \]

To this end, \( S'(x + w) \) of the formulas (12) and (14) is replaced with

\[ S'(x) + S'_x \left( \frac{\partial u}{\partial q} \delta P' + \frac{\partial u}{\partial q} \right) + S'_y \left( \frac{\partial v}{\partial q} \delta P' + \frac{\partial v}{\partial q} \right) \]
\[ \tag{20} \]

where \( \psi' \) in the data term (similarly \( \psi' \) for the other terms,) takes the simplified form of

\[ \psi' = \psi' \left( \left( S'_x \right)^2 + \gamma \left( S'_{xx} \right)^2 \right) \] \( q_x \) and \( q_y \) are approximated by using a first difference of \( q + \delta q; \ text{div}(\nabla q) \) is replaced with \( \text{div}(\nabla (q + \delta q)) \). In the iterative process, the current estimation of \( P' \) and \( q \) are updated as below:

\[ \theta \leftarrow \theta + \delta \theta. \]
\[ \tag{21} \]

Then, \( n + 6 \) linear equations with increments \( \delta \theta \) as variables can be established on the image lattices. Then, the incremental form of the formula (21) can be expressed as a linear equation as below:

\[ H \delta \theta + b = 0 \]
\[ \tag{22} \]

In this way, calculation on each iteration turns into solving
\[ \delta \theta = -H^{-1} \delta. \quad (23) \]

### Results and comparison

In this paper, we mainly focus on solving the reconstruction problem for two images. Although there are many existing multiview methods which can provide more accurate reconstruction results, we leave the extension to the multiview problem for future work.

To evaluate the performance of our algorithm, we make use of three public datasets: the KITTI dataset \(^{39}\), Strecha dataset \(^{40}\) and the official KITTI visual odometry split. Meanwhile, the forward-backward consistency assumption \(^{35}\) is taken to using for occlusion handling. For all experiments on KITTI dataset, the same parameters are set as: \([a_s, a_f, a_{so}, e] = [0.04, 0.11, 0.02, 0.001]\). For Strecha dataset, the same parameters are set as: \([a_s, a_f, a_{so}, e] = [0.08, 0.15, 0.02, 0.001]\). Meanwhile, the from-coarse-to-fine optimization strategy with a scale factor of \(\sqrt{3}\) is utilized in our experiments. The iterative optimization number is set as 15 at each resolution layer. As for the initialization, we use the initial dense correspondence of the FlowNet2-CSS and SIFT-flow to estimate the camera matrix \(P^\ast\). In this process, Least Median Squares (LMedS) method \(^{41}\) is used to optimize the fundamental matrix estimation. With the given camera intrinsic parameters \(K\), we estimate the essential matrix and decompose it to get the relative pose. Therefore, dense depth initialization can be estimated by utilizing interpolation method.

### KITTI dataset

Here, the KITTI dataset \(^{39}\) is used to evaluate the reconstruction performance by comparing the depths with Roxas and Oishi \(^{15}\) and Graber et al. \(^{17}\) using estimated pose. The qualitative depth results are shown in Figure 1 for sample images in terms of the estimated camera pose. The color-coding marks and visualizes outliers (> 3 px EPE) in red and inliers (< 3 px EPE) in blue on a logarithmic color scale. Table 1 shows the detailed comparison with the error metric measuring the percentage of erroneous pixels \(\tau > 3\) units in non-occluded areas (Out-Noc). The depth results state that our approach is significantly superior to Roxas and Oishi \(^{15}\), Graber et al. \(^{17}\) and Hu and Chen. \(^{29}\) Moreover, contrast to some current prominent CNN methods, \(^{20–28}\) our method uses stereo information and given ground truth pose, the results in Table 2 show that our approach can reach the performance of state-of-the-art methods. The estimated visualize depth results are presented in Figure 2.

To evaluate the dense correspondences, our simulation results of optical flow estimation on KITTI2012 are compared with the publicly available implementation DeepFlow, \(^{9}\) FlowNet2, \(^{34}\) and Roxas and Oishi. \(^{15}\)

---

**Table 1.** Depth estimation results (measuring Out-Noc metric \(\tau\)) from Roxals and Oishi \(^{15}\), Graber et al. \(^{17}\) and our approach on selected KITTI2012 dataset.

| Method                  | Out-Noc | Out-Noc | Out-Noc | Out-Noc |
|------------------------|---------|---------|---------|---------|
| Gruber et al. \(^{17}\) | 92.72   | 11.42   | 2.87    | 2.73    |
| Roxas and Oishi \(^{15}\)| 54.59   | 16.13   | 14.86   | 14.52   |
| Hu and Chen \(^{29}\)  | 33.81   | 17.70   | 11.56   | 9.52    |
| Our method              | 19.18   | 13.51   | 8.51    | 6.56    |
|                        | 29.90   | 12.43   | 8.53    | 7.85    |

**Table 2.** Quantitative evaluation of the depth result on KITTI by the split of Eigen et al. \(^{22}\) capped 80 m.

| Method                  | Abs Rel | Sq Rel | RMSE  |
|------------------------|---------|--------|-------|
| Eigen et al. \(^{22}\) | 0.203   | 1.548  | 6.307 |
| Liu et al. \(^{20}\)   | 0.202   | 1.614  | 6.523 |
| Godard et al. \(^{23}\) | 0.148   | 1.344  | 5.927 |
| Zhou et al. \(^{21}\)  | 0.208   | 1.768  | 6.856 |
| Mahjourian et al. \(^{25}\) | 0.163 | 1.240  | 6.220 |
| Yin and Shi \(^{24}\)  | 0.155   | 1.296  | 5.857 |
| Pilzer et al. \(^{26}\) | 0.142   | –      | 5.785 |
| Wong and Soatto \(^{27}\) | 0.133   | –      | 5.582 |
| Ye et al. \(^{28}\)    | 0.112   | –      | 4.978 |
| Ours                   | 0.132   | 1.302  | 5.315 |

---

*Figure 1.* Depth (normalized color), optical flow and their Out-Noc results on the KITTI2012 for our approach using estimated pose. From top to bottom: 000068, 000081, 000090, 000109, and 000134.
We compare the computed optical flow with previous traditional methods by EPE in Table 3. The estimated visualize results are shown in Figure 1 as same as the depth color coding results by measuring the percentage of erroneous pixels. All the above experimental results verify that our approach is comparable in both accuracy and efficiency in optical flow estimation. But with the added errors in pose estimation which caused by inaccuracy corresponding estimation, our method has slightly lower accuracy compared to FlowNet2-CSS, but still performs better than DeepFlow, Roxas and Oishi, and Hu and Chen.

Strecha dataset
Our approach also are applicable to model reconstruction, we use two consecutive frames of Fountain-P11 and the Herz-jesu-P8 subdataset in the strecha dataset with the approximate ground truth 3D model captured with laser radar system (LIDAR). Since the recovery of fine details heavily depends on the sharpness of the involved images, we downsample the initial slightly blurred frames to halve its resolution before rebuilding the scenes from only two views for our experiment.

We compare our reconstruction results with two recent proposed stereo approaches which can be applied to arbitrary camera settings. The variational method of Graber et al. with a surface smoothness constraint has a significant improvement on the performance compared to stand TV-regularization. The basic approach of Galliani et al. which is considered as the multi-view variant of PatchMatch Stereo would omit the additional 3D integration step in our experiment because the depth map only need to be evaluated from the reference camera. Qualitative results for the Fountain-P11 and the Herz-Jesu-P8 images are depicted in Figure 3. For the Fountain-P11 dataset, we

Table 3. AEE (average endpoint error) on the KITTI2012 dataset comparison of the optical flow results.

| Method          | DeepFlow 9 | FlowNet 2.0-CSS 34 | Roxas and Oishi 15 | Hu and Chen 29 | Ours |
|-----------------|------------|--------------------|--------------------|----------------|------|
| EPE             | 4.48       | 3.55               | 4.21               | 4.02           | 3.57 |

Figure 2. Qualitative results on the KITTI Eigen Split. The ground truth velodyne depth being very sparse, we interpolate it for visualization purpose.

Figure 3. Strecha dataset Fountain-P11 and Herz-Jesu-P8 images. Two different views. From Top to bottom: Reference image, ground truth, our method, Graber et al., and Galliani et al.
can see that the corresponding reconstruction of Gallinai et al. can’t generate fine details and contains many significant outliers in occlusion areas. The method of Graber et al. recovers a more detailed reconstruction with considerable noise. Compared with these two stereo approaches, the reconstruction of our approach is quite accurate in depth estimation. While flat surfaces are almost noise free in non-occlusion areas, details of the fountain and the wall are more pronounced. Also, the results of Herz-Jesu-P8 dataset is also visually appealing. The quantitative numbers of the results on the strecha dataset are shown in Table 4. It reveals that the RMS errors of our approach are significantly lower than those of the above two stereo methods both for the Fountain-P11 and the Herz-Jesu-P8 images.

### Pose estimation

In this paper, sequences 09–10 of the official KITTI odometry split are applied to our algorithm to evaluate the performance of our approach in camera pose estimation. The common 5-point Absolute Trajectory Error (ATE)\(^2^0\)–\(^2^5\) metric measures local agreement between the estimated trajectories and the respective ground truth. Moreover, we compare our method with a traditional representative SLAM framework: ORB-SLAM.\(^4^2\) It involves global optimization steps such as loop closure detection and bundle adjustment. As shown in Table 5, our method significantly outperforms the unsupervised methods by Liu et al.,\(^2^0\) Zhou et al.,\(^2^1\) Eigen et al.,\(^2^2\) Godard et al.,\(^2^3\) and Hu and Chen,\(^2^9\) but falls short of Yin and Shi\(^2^4\) and Mahjourian et al.\(^2^5\)

### Table 4. Comparison to different stereo methods in terms of the root mean square (RMS) error of the surface on Strecha dataset.

| Methods                | Fountain-P11 | Herz-Jesu-P8 |
|------------------------|--------------|--------------|
|                        | Seq.09       | Seq.10       |
| ORB-SLAM (full)        | 0.014±0.008  | 0.012±0.011  |
| ORB-SLAM (short)       | 0.064±0.141  | 0.064±0.130  |
| Mean Odom              | 0.032±0.026  | 0.028±0.023  |
| Zhou et al.\(^2^1\)   | 0.021±0.017  | 0.020±0.015  |
| Mahjourian et al.\(^2^5\)| 0.013±0.010  | 0.012±0.011  |
| Yin and Shi\(^2^4\)   | 0.011±0.007  | 0.012±0.009  |
| Hu and Chen\(^2^9\)   | 0.014±0.005  | 0.013±0.008  |
| Our method             | 0.013±0.010  | 0.012±0.010  |

### Table 5. Absolute Trajectory Error (ATE) on the KITTI odometry dataset averaged over all multi-frame snippets.

| Methods                | Seq.09  | Seq.10  |
|------------------------|---------|---------|
| Gruber et al.\(^1^7\) | 0.0688  | 0.2217  |
| Graber et al.\(^1^7\) | 0.0264  | –       |
| Galliani et al.\(^1^8\)| 0.6124  | 3.2813  |
| Our method             | 0.0234  | 0.0826  |

### Conclusion

In this paper, a direct dense 3D reconstruction approach is proposed by jointly employing the framework of multi-view geometry and the regularization strategies involving the texture constancy constraint, one-order motion smoothness constraint, depth regularize constraint and soft constraint. The texture constancy constraint is conducive to improve the robustness against illumination changes. One-order motion smoothness constraint is used to guarantee smooth dense correspondences. The depth regularize constraint is taken to handle inherent ambiguities problem, and the soft constraint can provide a more accurate dense correspondence to improve the robustness future. The results showed our approach have a good performance in depth estimation. It matches or outperforms the excellence variational depth estimation methods and state-of-the-art CNN models. We also achieved comparable results with existing variational and learning-based camera pose estimation methods. Our approach is still limited by the occlusion problem and cannot account for large illumination change. In future work, we will focus on these problems to create a more robust and accurate system that can handle occlusion problem and be applied to more complex scenes.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### ORCID iDs

Huijuan Hu [https://orcid.org/0000-0001-8454-0335](https://orcid.org/0000-0001-8454-0335)
Chuan Hu [https://orcid.org/0000-0001-5379-1561](https://orcid.org/0000-0001-5379-1561)

### References

1. Longuet-Higgins HC. A computer algorithm for reconstructing a scene from two projections. *Nature* 1981; 293(5828): 133–135.
2. Sun K and Wenbing T. A constrained radial agglomerative clustering algorithm for efficient structure from motion. *IEEE Signal Process Lett* 2018; 25(7): 1089–1093.
3. Frahm JM, Fite-George P, Gallup D, et al. Building rome on a cloudless day. In: *European conference on computer vision*, Heraklion, Crete, Greece, 5–11 September 2010, pp.834–849. Berlin: Springer.
4. Wu C. Towards linear-time incremental structure from motion. In: 2013 international conference on 3D Vision (3DV), Seattle, WA, 29 June–1 July 2013, pp.127–134. New York, NY: IEEE.

5. Engel J, Schöps T and Cremers D. LSD-SLAM: Large-scale direct monocular SLAM. In European conference on computer vision, Zurich, Switzerland, September 6–12, 2014; 834–849. Springer, Cham.

6. Newcombe RA, Lovegrove SJ and Davison AJ. DTAM: dense tracking and mapping in real-time. In: 2013 IEEE international conference on computer vision, Sydney, Australia, 1–8 December 2013, pp.127–134.

7. Xu J, Ranftl R and Kolten V. Accurate optical flow via direct cost volume processing. In: 2017 IEEE conference on computer vision and pattern recognition, Honolulu, HI, 21–26 July 2017, pp.1289–1297. New York, NY: IEEE.

8. Wulff J, Sevilla-Lara L and Black MJ. Optical flow in mostly rigid scenes. In: IEEE conference on computer vision and pattern recognition, Honolulu, HI, 21–26 July 2017, pp.4671–4680. New York, NY: IEEE.

9. Revaud J, Weinzaepfel P, Harchaoui Z, et al. Deep matching: hierarchical deformable dense matching. Int J Comput Vis 2016; 120(3): 300–323.

10. Maurer D, Marniok N, Goldluecke B, et al. Structure-from-motion-aware patchmatch for adaptive optical flow estimation. In: Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018, pp.575–592.

11. Valgaerts L, Bruhn A, Mainberger M, et al. Dense versus sparse approaches for estimating the fundamental matrix. Int J Comput Vis 2012; 96(2): 212–234.

12. Aubry M, Kolev K, Goldlücke B, et al. Decoding photometry and geometry in dense variational camera calibration. In: 2011 International conference on computer vision, Barcelona, Spain, 6–13 November 2011, pp.1411–1418. New York, NY: IEEE.

13. Becker F, Lenzen F, Kappes J, et al. Variational recursive joint estimation of dense scene structure and camera motion from monocular high speed traffic sequences. Int J Comput Vis 2013; 105(3): 269–297.

14. Bagnato L, Frossard P and Vandergheynst P. A variational framework for structure from motion omnidirectional image sequences. J Math Imag Vis 2011; 41(3): 182–193.

15. Rojas M and Oishi T. Real-time simultaneous 3D construction and optical flow estimation. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, 12–15 March 2018, pp.885–893. New York, NY: IEEE.

16. Stühmer J, Gunhold S and Cremers D. Real-time dense geometry from a handheld camera. In: Joint pattern recognition symposium. Berlin: Springer, 2010, pp.11–20.

17. Graber G, Balzer J, Soatto S, et al. Efficient minimal-surface regularization of perspective depth maps in variational stereo. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Boston, MA, 7–12 June 2015, pp.511–520. New York, NY: IEEE.

18. Galliani S, Lasansky K and Schindler K. Massively parallel multiview stereopsis by surface normal diffusion. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015, pp.873–881.

19. Shin S, Im S, Shim I, et al. Geometry guided three-dimensional propagation for depth from small motion. IEEE Signal Process Lett 2017; 24(12): 1857–1861.

20. Liu F, Shen C, Lin G, et al. Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans Pattern Anal Mach Intell 2015; 38(10): 2024–2039.

21. Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017, pp.6612–6621. New York, NY: IEEE.

22. Eigen D, Puhrsch C and Fergus R. Depth map prediction from a single image using a multi-scale deep network. arXiv preprint, arXiv:1406.2283. 2014 Jun 9.

23. Godard C, Aodha OM and Brostow GJ. Unsupervised monocular depth estimation with left-right consistency. In: IEEE conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 21–26 July 2017, pp.6602–6611.

24. Yin Z and Shi J. Geonet: unsupervised learning of dense depth, optical flow and camera pose. In: IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City, UT, 18–23 June 2018, pp.9768–9777. New York, NY: IEEE.

25. Mahjourian R, Wicke M and Angelova A. Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints. In: IEEE/CVF conference on computer vision and pattern recognition, Salt Lake City, UT, 18–23 June 2018, pp.5667–5675. New York, NY: IEEE.

26. Pilzer A, Lathuiliere S, Sebe N, et al. Refine and distill: exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation. In: IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 15–20 June 2019, pp.9768–9777. New York, NY: IEEE.

27. Wong A and Soatto S. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction. In: IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 15–20 June 2019, pp.9768–9777. New York, NY: IEEE.

28. Ye X, Chen S and Xu R. Dpnet: detail-preserving network for high quality monocular depth estimation. Pattern Recognit 2021; 109: 107578.

29. Hu H and Chen P. Direct optical-flow-aware computational framework for 3D reconstruction. IEEE Access 2019; 7(10): 18–27.

30. Mohamed MA, Rashwan HA, Mertsching B, et al. Illumination-robust optical flow using a local directional pattern. IEEE Trans Circuits Syst Video Technol 2014; 24(9): 1499–1508.

31. Müller T, Rabe C, Rannacher J, et al. Illumination-robust dense optical flow using census signatures. In: Mester R and Felsberg M (eds) Pattern recognition. DAGM 2011. Lecture Notes in computer science, vol. 6835. Berlin: Springer, pp.236–245.

32. Rashwan HA, Mohamed MA, García MA, et al. Illumination robust optical flow model based on histogram of oriented gradients. In: Pattern recognition: 35th German Conference, Saarbrücken, Germany, 3–6 September 2013, pp.354–363. Berlin: Springer.

33. Schroers C, Hafner D and Weickert J. Multiview depth parameterisation with second order regularisation. In: International conference on scale space and variational methods in computer vision, Lège-Cap Ferret, France, May 31–June 4, 2015, pp.551–562. Cham: Springer.
34. Ilg E, Mayer N, Saikia T, et al. Flownet 2.0: evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, Honolulu, HI, 21–26 July 2017, pp.2462–2470.

35. Sundaram N, Brox T and Keutzer K. Dense point trajectories by gpu-accelerated large displacement optical flow. In: European conference on computer vision, Heraklion, Crete, Greece, September 2010, pp.438–451.

36. Oliensis J. A new structure-from-motion ambiguity. IEEE Trans Pattern Anal Mach Intell 2000; 105(3): 269–297.

37. Furukawa Y and Ponce J. Accurate, dense, and robust multiview stereopsis. IEEE Trans Pattern Anal Mach Intell 2010; 32(8): 1362–1376.

38. Lhuillier M and Quan L. A quasi-dense approach to surface reconstruction from uncalibrated images. IEEE Trans Pattern Anal Mach Intell 2005; 27(3): 418–433.

39. Geiger A, Lenz P and Urtasun R. Are we ready for autonomous driving? The kitti vision benchmark suite. In: IEEE conference on computer vision and pattern recognition, Providence, RI, 16–21 June 2012, pp.3354–3361. New York, NY: IEEE.

40. Strecha C, Hansen WV, Gool LV, et al. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In: IEEE conference on computer vision and pattern recognition, Anchorage, AK, 23–28 June 2008, pp.1–8. New York, NY: IEEE.

41. Zhang Z. Determining the epipolar geometry and its uncertainty: a review. Int J Comput Vis 1998; 27(2): 161–195.

42. Mur-Artal R, Montiel JMM and Tardos JD. Orb-slam: a versatile and accurate monocular slam system. IEEE Trans Rob 2015; 31(5): 1147–1163.