Image Gradient-based Joint Direct Visual Odometry for Stereo Camera

Jianke Zhu[1,2]
[1]College of Computer Science, Zhejiang University, Hangzhou, China
[2]Alibaba-Zhejiang University Joint Research Institute of Frontier Technologies
jkzhu@zju.edu.cn

Abstract

Visual odometry is an important research problem for computer vision and robotics. In general, the feature-based visual odometry methods heavily rely on the accurate correspondences between local salient points, while the direct approaches could make full use of whole image and perform dense 3D reconstruction simultaneously. However, the direct visual odometry usually suffers from the drawback of getting stuck at local optimum especially with large displacement, which may lead to the inferior results. To tackle this critical problem, we propose a novel scheme for stereo odometry in this paper, which is able to improve the convergence with more accurate pose. The key of our approach is a dual Jacobian optimization that is fused into a multi-scale pyramid scheme. Moreover, we introduce a gradient-based feature representation, which enjoys the merit of being robust to illumination changes. Furthermore, a joint direct odometry approach is proposed to incorporate the information from the last frame and previous keyframes. We have conducted the experimental evaluation on the challenging KITTI odometry benchmark, whose promising results show that the proposed algorithm is very effective for stereo visual odometry.

1 Introduction

With the prevalence of development on driverless cars and unmanned aerial vehicles (UAV)s, visual odometry becomes an interesting problem in computer vision and robotics. In contrast to simultaneous localization and mapping (SLAM) [Klein and Murray, 2007], visual odometry aims at estimating the camera poses without resorting to loop-closure and global bundle adjustment.

Generally, stereo visual odometry is able to directly estimate the global scale using the baseline between two camera centers, while monocular visual odometry has to rely on some priors, e.g., the camera height from ground plane [Song *et al.*, 2016] or IMU measurement [Corke *et al.*, 2007]. In this paper, we focus our attention on the problem of stereo visual odometry, which is widely used in various applications.

Instead of using whole image [Comport *et al.*, 2010; Newcombe *et al.*, 2011], semi-dense visual odometry approaches [Engel *et al.*, 2014; Forster *et al.*, 2014] have recently shown the encouraging results using CPU implementation by sampling the points with large gradient magnitudes or corners, which are able to reduce the computational cost effectively. However, direct approaches usually suffer from the problem of tending to stuck at a local optimum especially with large displacement, which may lead to inferior results.

In KITTI odometry benchmark [Geiger *et al.*, 2012], nearly all the current top performers [Cvisic and Petrovic, 2015; Mur-Artal *et al.*, 2015] are based on the local feature matching besides those based on LiDAR sensor [Zhang *et al.*, 2014]. Although having achieved the promising performance, the feature-based visual odometry method cannot make full use of the information from whole image. Moreover, it may fail in the homogenous regions, i.e., highway scenario, which are lack of reliable salient feature points. Furthermore, the feature-based approach cannot directly yield dense 3D reconstruction while the direct visual odometry [Comport *et al.*, 2010; Engel *et al.*, 2014] is able to simultaneously recover the dense or semi-dense depth map.

In this paper, we address these current limitations by proposing a novel direct stereo visual odometry approach. Specifically, we suggest a dual Jacobian scheme for the multi-scale pyramid optimization. We find that the conventional Jacobian has wide convergence basin, however, the precision of camera pose estimation is unsatisfied. By replacing the ratio between the recovered 3D scene coordinates with image coordinates and focal length, an alternative Jacobian can obtain more accurate results near the optimal solution. Thus, we employ the conventional Jacobian at the coarse scale while the alternative Jacobian is used in the finest level. Moreover, we introduce the gradient-based feature representation to account for illumination changes. To incorporate more information from previous frames, we propose a joint direct stereo visual odometry method to boost the camera tracking performance.

In summary, the main contributions of this paper are: (1) a novel dual Jacobian optimization scheme to avoid the local optimum and improve the accuracy; (2) the gradient-based feature representation for direct visual odometry that enjoys the merit of being robust to illumination changes; (3) a joint direct odometry approach to incorporate the information from multiple frames; (4) experiments on the challenging KITTI odometry benchmark, comparing against the state-of-the-art approaches and obtaining the promising results.

2 Related Work

During the past fifteen years, considerable research effort has already been devoted to visual odometry and SLAM in computer vision and robotics. Most of previous visual odometry methods can be typically divided into three categories. The first group is based on the expensive LiDAR sensor [Zhang *et al.*, 2014], which registers scans between the different time stamps. The second one relies on local feature matching [Badino *et al.*, 2013; Mur-Artal *et al.*, 2015] across video frames. The last group directly minimizes the photometric error [Comport *et al.*, 2010; Engel *et al.*, 2014] between the current frame and reference keyframe.

The feature-based visual odometry approaches [Badino *et al.*, 2013; Mur-Artal *et al.*, 2015] are currently very popular. The key is to first estimate the camera poses along with 3D coordinates from the salient point correspondences through a robust estimator like RANSAC [Fischler and Bolles, 1981], where the local feature extraction algorithm plays a very important role. Then, bundle adjustment [Triggs *et al.*, 1999] is employed to simultaneously refine 3D scene structure and camera poses. Badino et al. [Badino *et al.*, 2013] extensively evaluate the different combinations of local feature detectors and descriptors. The features from multiple frames are integrated together to achieve the good results. Mur-Artal et al. [Mur-Artal *et al.*, 2015] propose a full SLAM approach based on binary feature, including bundle adjustment, loop-closure detection and relocalization.

Most of direct visual odometry methods are generally based on the Lucas-Kanade framework [Lucas and Kanade, 1981; Baker and Matthews, 2004], which is one of the most widely used techniques in computer vision. These approaches directly find the optimal geometric transformation by minimizing the photometric error between the input image and the warped reference frame. Comport et al. [Comport *et al.*, 2010] propose a quadrifocal approach to stereo visual odometry. Newcombe et al. [Newcombe *et al.*, 2011] present a realtime dense tracking and mapping algorithm for a handheld monocular camera. Engel et al. [Engel *et al.*, 2014; 2015] propose a semi-dense visual odometry method with the probabilistic depth map estimation and update scheme. Forster et al. [Forster *et al.*, 2014] present a semi-dense visual odometry approach based on corners, where the camera motions and scene structures are refined like the conventional bundle adjustment. The major limitation of these methods is that they tend to become stuck at a local optimum and hence require a good initialization.

Unlike existing direct visual odometry method, our proposed approach can effectively handle various challenging conditions, including large displacements as well as lighting variations in outdoor environment. Moreover, we demonstrate that the direct stereo visual odometry approach is able to achieve the state-of-the-art results comparing to the feature-based methods.

3 Joint Direct Stereo Visual Odometry

In this section, we present the proposed approach to direct stereo visual odometry. Firstly, we formulate the problem into a nonlinear least square minimization (Section 3.1). Secondly, we present a dual Jacobian optimization scheme to deal with the challenge of large displacement (Section 3.2). Thirdly, we suggest a gradient feature representation to deal with lighting changes (Section 3.3). Finally, we propose a joint optimization approach to incorporating the information from multiple frames (Section 3.4).

3.1 Direct Stereo Visual Odometry

As in [Ma *et al.*, 2004], the projection function $\pi : R^3 \rightarrow \Omega$ for a conventional pinhole-camera of the rectified stereo pair projects 3D point $X = (x_i, y_i, z_i)$ in the scene onto 2D point $\mathbf{x}_i = (u_i, v_i)$ in image space $\Omega \subset R^2$:

$$\mathbf{x}_i = \pi(X_i) = \begin{pmatrix} f_u \frac{x_i}{z_i} + c_u \\ f_v \frac{y_i}{z_i} + c_v \end{pmatrix}, \tag{1}$$

where (f_u, f_v) is the focal length, and (c_u, c_v) is the principal point. On the other hand, the inverse projection function π^{-1} recovers the 3D scene point from image coordinate $\mathbf{x}_i$ and its depth measurement d_i. In this paper, d_i is estimated from stereo matching, i.e., block matching using Sum of Absolute Difference (SAD) or Semi-Global Matching (SGM) [Hirschmüller, 2008]. Then, $\pi^{-1}(\mathbf{x}_i, d_i)$ can be written as follows:

$$\pi^{-1}(\mathbf{x}_i, d_i) = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}. \tag{2}$$

Problem Formulation

Let $\mathbf{T}$ denote a rigid transformation represented in Lie group $SE(3)$. Generally, visual odometry aims at estimating the camera pose $\mathbf{T} = (\mathbf{R}, \mathbf{t}) \in SE(3)$ from the image observations, where $\mathbf{R} \in SO(3)$ denotes a camera orientation matrix, and $\mathbf{t} \in R^3$ is the position vector. The unknown twisted coordinates $\theta = (\omega, \nu)^T \in se(3)$ are defined by the lie algebra $se(3)$, which corresponds to the tangent space of $SE(3)$ at the identity. ω is the angular velocity, and ν is the linear velocity. Moreover, the camera pose $\mathbf{T}(\theta)$ is computed by the exponential map of θ:

$$\mathbf{T}(\theta) = \exp\left(\begin{bmatrix} [\omega]_\times & \nu \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right), \tag{3}$$

where $[\omega]_\times$ denotes the skew symmetric matrix of the angular vector ω.

Let $\mathbf{T}_r$ denote the pose of reference frame. To estimate the current camera pose $\mathbf{T}_c$, direct visual odometry [Comport *et al.*, 2010] minimizes the photometric errors between the reference keyframe I_r and the current frame I_c:

$$\min_{\mathbf{T}_c} \sum_{\mathbf{x}_i \in \Omega} \Big(I_r(\mathbf{x}_i) - I_c(\pi(\mathbf{T}_c(\theta)\mathbf{T}_r^{-1}\pi^{-1}(\mathbf{x}_i, d_i))) \Big)^2. \tag{4}$$

Obviously, the above problem is a nonlinear least square minimization. It can be solved via an iterative Gauss-Newton algorithm $\theta = -(J^T J)^{-1} J^T \mathbf{r}$, where J is Jacobian matrix that will be discussed later. $\mathbf{r}$ denotes the residual vector.

According to Lie algebra, the current pose estimation $\mathbf{T}_c$ is updated by a homogeneous update until convergence [Comport *et al.*, 2010]: $\mathbf{T}_c \leftarrow \mathbf{T}_c \mathbf{T}(\theta)$.

Robust Estimator

In order to effectively handle outliers, a robust estimator $\rho(t)$ is applied to the photometric error. Thus, the energy function can be derived as follows:

$$E = \min_{\mathbf{T}_c} \sum_{\mathbf{x}_i \in \Omega} \rho\Big(I_r(\mathbf{x}_i) - I_c(\pi(\mathbf{T}_c(\theta)\mathbf{T}_r^{-1}\pi^{-1}(\mathbf{x}_i, d_i)))\Big).$$

To deal with the large outliers, we choose the Tukey's biweight [Black and Rangarajan, 1996] loss function $\rho(t)$:

$$\rho(t) = \begin{cases} \frac{\kappa^2}{6}\Big[1 - (1 - (\frac{t}{\kappa})^2)^3\Big] & \text{for} \quad |t| \leq \kappa \\ \frac{\kappa^2}{6} & \text{for} \quad |t| > \kappa \end{cases}, \tag{5}$$

where κ is a tuning constant that is set to 4.6851 corresponding to the 95% asymptotic efficiency on the standard normal distribution. To take into account of the scale variations on residuals, all photometric errors are formed into a vector, which is further scaled by the inverse of its median value.

This problem is a weighted nonlinear least square minimization. Let W denote a diagonal matrix, in which the diagonal element is $\rho'(t)$ for each residual term. $\rho'(t)$ represents the first order derivative of Tukey's biweight loss function. Thus, we can obtain the following update equation for robust direct visual odometry:

$$\theta = -(J^T W J)^{-1} J^T W \mathbf{r}. \tag{6}$$

3.2 Optimization with Dual Jacobian

From the above formulation, the i-th row of Jacobian $J_i(\theta)$ is equal to the derivative of residual $\mathbf{r}_i$ with respect to the camera pose parameters θ. According to the chain-rule, we can derive the following equation:

$$J_i(\theta) = \frac{\partial \mathbf{r}_i}{\partial \theta} = \frac{\partial \mathbf{r}_i}{\partial \mathbf{x}_i} \frac{\partial \mathbf{x}_i}{\partial X_i} \frac{\partial X_i}{\partial \theta} \tag{7}$$

Conventional Jacobian J

In general, we can directly compute the derivative of residual $\mathbf{r}_i$ with respect to image coordinate $\mathbf{x}_i$ through image gradient:

$$\frac{\partial \mathbf{r}_i}{\partial \mathbf{x}_i} = \nabla I(\mathbf{x}_i), \tag{8}$$

where $\nabla I = [\ Ix \quad Iy\ ]^T$ denotes the gradient of image I.

Under Lucas-Kanade framework, there are several compositional update strategies. Inverse composition [Baker and Matthews, 2004] is able to take advantage of constant Hessian by reversing the role of template and input image. Moreover, efficient second order minimization (ESM) [Benhimane and Malis, 2004] shows promising convergence properties. Therefore, we employ ESM scheme in the following:

$$\frac{\partial \mathbf{r}_i}{\partial \mathbf{x}_i} = \nabla I_{esm}(\mathbf{x}_i) = \frac{1}{2}(\nabla I_c(\mathbf{x}_i) + \nabla I_r(\mathbf{x}_i)). \tag{9}$$

According to the perspective projection function defined in Eqn. 1, we can compute the derivative of image coordinates $\mathbf{x}_i$ with respect to the world point X_i as below:

$$\frac{\partial \mathbf{x}_i}{\partial X_i} = \begin{bmatrix} \frac{f_u}{z_i} & 0 & -f_u \frac{x_i}{z^2} \\ 0 & \frac{f_v}{z_i} & -f_v \frac{y_i}{z^2} \end{bmatrix}. \tag{10}$$

Since the camera pose $\mathbf{T}(\theta)$ is obtained by the exponential map in Eqn. 3, the derivatives can be calculated as follows:

$$\frac{\partial X_i}{\partial \theta} = [\ \mathbf{I} \quad [X_i]_\times\ ], \tag{11}$$

where $\mathbf{I}$ denotes a 3×3 identity matrix.

Substitute Eqn. 9-11 into Eqn. 7, the conventional visual odometry approaches compute the derivates with respect to odometry parameter θ as follows:

$$J_i(\theta) = \nabla I_{esm}(\mathbf{x}_i) \cdot \tag{12}$$
$$\begin{bmatrix} \frac{f_u}{z_i} & 0 & -\frac{f_u x_i}{z_i^2} & -f_u \frac{x_i y_i}{z_i^2} & f_u \frac{x_i^2 + z_i^2}{z_i^2} & -\frac{f_u y_i}{z_i} \\ 0 & \frac{f_v}{z_i} & -\frac{f_v y_i}{z_i^2} & -f_v \frac{y_i^2 + z_i^2}{z_i^2} & f_v \frac{x_i y_i}{z_i^2} & \frac{f_v x_i}{z_i} \end{bmatrix}.$$

Alternative Jacobian $\tilde{J}$

For direct stereo visual odometry, the depth measurements from dense disparity map are usually quite noisy. Although the large outliers could be removed by the robust estimator, the performance is greatly degraded with the inaccurate observations. Thus, the feature-based methods dominate the odometry task with the large displacement and rapid motion, i.e., KITTI benchmark [Geiger *et al.*, 2012].

To tackle this critical issue, we present an alternative derivative as the complement for the visual odometry. Specifically, the following equality can be derived from the projection function in Eqn. 1:

$$\begin{cases} \frac{x_i}{z_i} = \frac{\tilde{u}_i}{f_u} \\ \frac{y_i}{z_i} = \frac{\tilde{v}_i}{f_v} \end{cases}, \tag{13}$$

where $\tilde{u}_i = u_i - c_u$, and $\tilde{v}_i = v_i - c_v$. Thus, we can obtain an alternative formulation of Jacobian $\tilde{J}_i(\theta)$ by substituting Eqn. 13 into Eqn. 12,

$$\tilde{J}_i(\theta) = \nabla I_{esm}(\mathbf{x}_i) \cdot \tag{14}$$
$$\begin{bmatrix} \frac{f_u}{z_i} & 0 & -\frac{\tilde{u}_i}{z_i} & -\frac{\tilde{u}_i \tilde{v}_i}{f_u} & \frac{f_u^2 + \tilde{u}_i^2}{f_u} & -\tilde{v}_i \\ 0 & \frac{f_v}{z_i} & -\frac{\tilde{v}_i}{z_i} & -\frac{f_v^2 + \tilde{v}_i^2}{f_v} & \frac{\tilde{u}_i \tilde{v}_i}{f_v} & \tilde{u}_i \end{bmatrix}.$$

Remark In contrast to the conventional Jacobian J in Eqn. 12, $\tilde{J}$ is computed from the ratio between the image coordinates and focal length rather than 3D scene coordinates. By taking advantage of the accurate offline camera calibration, $\tilde{J}$ is supposed to be more accurate than J if the initial pose is near to the optimal solution $\mathbf{T}_c^*(\theta)$.

Optimization Scheme with Dual Jacobian

To deal with the large displacements, we employ three important measures.

Firstly, an effective pose predictor is employed to initialize the nonlinear iterative minimization. In contrast to dead-reckoning or a constant acceleration model used in the previous approach [Persson *et al.*, 2015], we take advantage of

Kalman filter [Kalman, 1960] with six motion state variables to predict the current pose from the last observations, where measurement matrix is set to identity. Process noise is set to 10^{-8} for velocity and one for acceleration.

Secondly, an image pyramid is usually built to improve the convergence basin for direct visual odometry [Comport *et al.*, 2010]. Moreover, the estimated pose in the coarse level is employed to initialize the next level in order to avoid some local minima. The convergence criteria is either $|\theta| < 0.001$ or the maximum number of iterations $(25 \times (l+1))$ is reached.

Finally, we employ the conventional Jacobian J at the coarse scale while the alternative Jacobian $\tilde{J}$ is used at the finest level. We name this scheme as dual Jacobian method.

3.3 Illumination Variations

To deal with the complex illumination changes, the conventional approaches [Comport *et al.*, 2010; Engel *et al.*, 2014] usually employ a global gain a and bias b transform function $g(I(\mathbf{x})) = a \cdot I(\mathbf{x}) + b$ to account for the uniform lighting variations.

Instead of directly using raw pixel intensities, we introduce the gradient-based feature representation into visual odometry, which enjoys the merit of being robust to illumination variations. This is essential to the outdoor scenario, i.e., autonomous vehicles. Unlike descriptor fields [Crivellaro and Lepetit, 2014], we directly minimize the gradient difference between the current frame and reference frame as below:

$$\min_{\mathbf{T}} \sum_{\mathbf{x}_i \in \Omega} \rho\Big(\nabla I_r(\mathbf{x}_i) - \nabla I_c(\pi(\mathbf{T}(\theta)\pi^{-1}(\mathbf{x}_i, d_i)))\Big). \quad (15)$$

Besides we calculate the second order image gradient for both directions to compute Jacobian, the optimization for the above nonlinear least square is same as the one with raw intensities. Additionally, the gradient-based method does not require two extra affine lighting variables to deal with the illumination changes.

3.4 Joint Direct Stereo Visual Odometry

In most of previous direct visual odometry approaches [Comport *et al.*, 2010; Engel *et al.*, 2014], a single reference frame is typically employed during optimization. To boost the chances of homing in on the solution, we incorporate multiple frames into the energy function, which jointly estimate the current camera pose. Thus, we can obtain the following energy function:

$$\min_{\mathbf{T}_c} \sum_{k \in \mathcal{K}} \sum_{\mathbf{x}_i \in \Omega_k} \rho\Big(\nabla I_k(\mathbf{x}_i) - \nabla I_c(\pi(\mathbf{T}_c(\theta)\mathbf{T}_k^{-1}\pi^{-1}(\mathbf{x}_i, d_i)))\Big),$$

where $\mathcal{K}$ denotes a set of selected keyframes.

Obviously, the computational cost increases along with the total number of keyframes $|\mathcal{K}|$. In our empirical study, we find that it is very effective to only use the previous frame and a single selected keyframe. Thus, we can derive the energy function for joint direct visual odometry as below:

$$\min_{\mathbf{T}_c} \sum_{\mathbf{x}_i \in \Omega_p} \rho\Big(\nabla I_p(\mathbf{x}_i) - \nabla I_c(\pi(\mathbf{T}_c(\theta)\mathbf{T}_p^{-1}\pi^{-1}(\mathbf{x}_i, d_i)))\Big) + \sum_{\mathbf{x}_i \in \Omega_k} \rho\Big(\nabla I_k(\mathbf{x}_i) - \nabla I_c(\pi(\mathbf{T}_c(\theta)\mathbf{T}_k^{-1}\pi^{-1}(\mathbf{x}_i, d_i)))\Big),$$

where I_p and $\mathbf{T}_p$ denote the previous frame and its pose respectively.

For the keyframe I_k, we build a circular queue of previous k frames. Therefore, the frame at the rear of queue is chosen as the keyframe. In our empirical study, this simple strategy is effective in the case of ego-motion like KITTI benchmark.

4 Experiment

In this section, we give details of our experimental implementation and discuss the results of visual odometry.

4.1 Experimental Testbed

To examine the empirical efficacy of the proposed stereo visual odometry approach, we conduct the experiments for comprehensive performance evaluations on KITTI odometry benchmark [Geiger *et al.*, 2012]. KITTI dataset is composed of the captured videos and laser scans along with the very accurate GPS/INS for ground truth. The rectified stereo images are with the size around 1230×370, which are recorded at the frequency of 10Hz. Thus, the displacement of cameras between the consequent frames are quite large. There are 11 sequences (00-10) with ground truth poses for training, and 11 sequences (11-21) for testing.

The average relative translation error t_{rel} (%) and rotation error r_{rel} (deg/100m) are employed as the performance metrics, which are usually used to evaluate the odometry method [Sturm *et al.*, 2012]. To further illustrate the efficacy of our proposed approach, we also include the absolute translation error t_{abs} (m) in [Sturm *et al.*, 2012] that is used to evaluate the SLAM algorithms with loop-closures.

To facilitate the multi-scale optimization, we build a 4-level image pyramid. Since stereo matching is very time consuming, we compute the disparity map at the $\frac{1}{4}$ size of the original resolution. Moreover, the disparity map is calculated by 5×5 block matching with SAD. For the finest resolution, the disparity values are obtained by nearest neighbor interpolation. Furthermore, only the pixels with larger gradient magnitude ($\|\nabla I\|^2 > 18$) are selected for computation.

For the proposed joint direct stereo visual odometry approach, we retain a circular queue of the previous 12 frames. In contrast to full SLAM, we do not cache lots of keyframes for loop-closure. Therefore, our approach requires less memory, which can be adapted to the resource limited devices. All of our experiments were carried out on a PC with Intel Core i7-3770 3.8GHz processor and 8GB RAM using single thread.

4.2 Evaluation on Different Settings

We now evaluate the performance of stereo odometry using different settings, including Jacobian for Gauss-Newton, feature representation, and joint direct odometry scheme. We conduct the quantitative experiments on training set of KITTI odometry benchmark with grayscale image. For simplicity, 'DVO' denotes the method using pixel intensity with conventional Jacobian, which is equivalent to D6DVO [Comport *et al.*, 2010] in KITTI benchmark. 'Pixel' represents the dual Jacobian approach with intensity and 'Gradient' is the dual Jacobian approach with two gradient channels. Our proposed joint direct visual odometry method is denoted as 'Joint'.

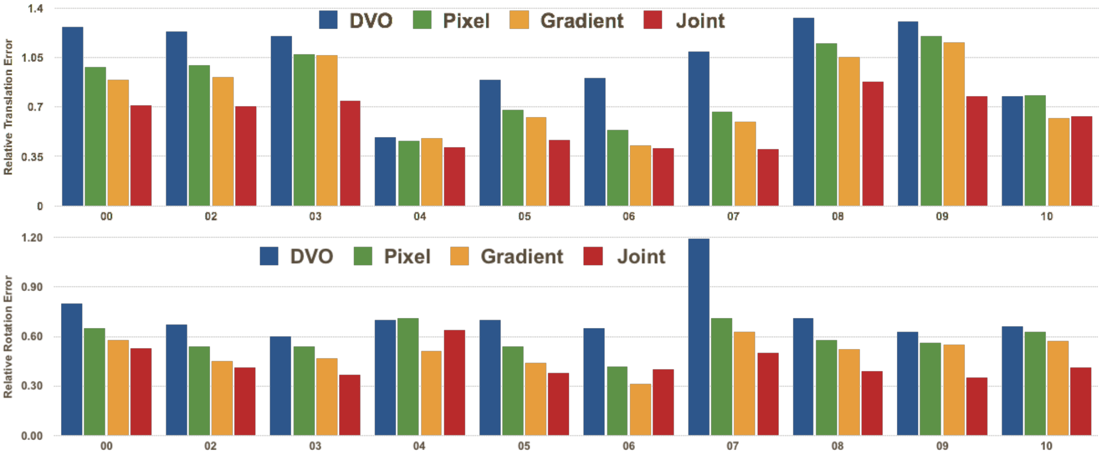


Figure 1: Performance evaluation on the training set of KITTI odometry dataset with various settings.

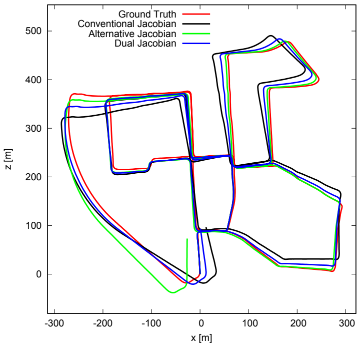


Figure 2: Camera path for the stereo visual odometry with different Jacobians on 'sequence 00' of KITTI dataset.

Jacobian We evaluate the performance of different Jacobians, including the conventional method in Eqn. 12, alternative solution in Eqn. 14, and the proposed dual Jacobian optimization scheme. In our experiments, the alternative Jacobian approach fails for 4 out of 11 sequences in KITTI training set. Both conventional method and the dual Jacobian scheme succeed all the sequences except large translation errors occur at 'sequence 01' with highway environment. Therefore, Fig. 1 shows the comparison results for 10 sequences except 'sequence 01'. It can be seen the dual Jacobian scheme outperforms the conventional DVO method at a very large margin on the accuracy of both translation and rotation. Moreover, we show the qualitative comparison in Fig. 2. It can be clearly seen that the alternative Jacobian method can obtain quite accurate trajectory, however, it is easy to stuck at local optimum. On the other hand, the conventional DVO method is very robust but less accurate. Thus, the proposed approach employs the conventional method at the coarse scale to initialize the alternative Jacobian at the finest level.

Feature We compare the results with different feature representations. In Fig. 1, it can be observed that the gradient feature greatly improves the visual odometry performance for almost all the sequences, which demonstrates its effectiveness on handling complex outdoor illuminations. Additionally, we evaluate other features like descriptor fields [Crivellaro and Lepetit, 2014] of 4-channel and fusing the gradient feature with raw intensity of 3-channel. However, there is no noticeable improvement found on KITTI odometry benchmark.

Table 1: Comparison of performance on KITTI Odometry training dataset.

Data	GDVO			LSD-VO		LSD-SLAM		
#	t_{rel}	r_{rel}	t_{abs}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{abs}
00	0.71	0.53	4.9	1.09	0.42	0.63	0.26	1.0
01	1.00	0.65	5.2	2.13	0.37	2.36	0.36	9.0
02	0.70	0.41	6.1	1.09	0.37	0.79	0.23	2.6
03	0.75	0.37	0.3	1.16	0.32	1.01	0.28	1.2
04	0.42	0.64	0.2	0.42	0.34	0.38	0.31	0.2
05	0.47	0.38	1.8	0.90	0.34	0.64	0.18	1.5
06	0.41	0.40	1.5	1.28	0.43	0.71	0.18	1.3
07	0.40	0.50	0.8	1.25	0.79	0.56	0.29	0.5
08	0.88	0.39	2.4	1.24	0.38	1.11	0.31	3.9
09	0.77	0.35	2.2	1.22	0.28	1.14	0.25	5.6
10	0.63	0.41	1.1	0.75	0.34	0.72	0.33	1.5
00-10	0.71	0.44	2.4	1.14	0.40	0.91	0.27	2.6
11-21	0.86	0.31	-	1.40	0.36	1.21	0.35	-

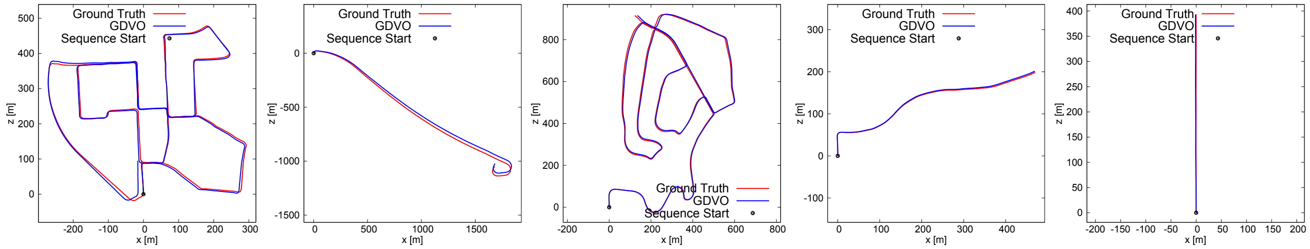
Joint optimization As shown in Fig. 1, the proposed joint visual odometry significantly reduces the translation error in 9 out of 10 sequences. Specifically, the translation error for 'sequence 10' is slightly larger than the method using previous frame only. For the rotation error, the joint approach outperforms the single frame method in 7 out of 10 sequences. Moreover, our proposed approach succeeds in 'sequence 01' with low pose prediction errors, which is difficult highway scene with fast motion and few reliable features.

Table 2: Comparison of state-of-the-art methods on KITTI odometry benchmark.

Rank	Method	Category	t_{rel}	r_{rel}
1	V-LOAM [Zhang and Singh, 2015]	Laser	0.68	0.16
4	GDVO (Proposed)	Direct	0.86	0.31
6	SOFT [Cvisic and Petrovic, 2015]	Feature	0.88	0.22
13	DEMO [Zhang *et al.*, 2014]	Laser	1.14	0.49
14	ORB-SLAM2 [Mur-Artal *et al.*, 2015]	Feature	1.15	0.27
18	S-LSD-SLAM [Engel *et al.*, 2015]	Direct	1.20	0.33
33	D6DVO [Comport *et al.*, 2010]	Direct	2.04	0.51

4.3 KITTI Odometry Benchmark

We compare the proposed method with the state-of-the-art methods on KITTI odometry Benchmark. Table 2 shows the performance evaluation for each sequence in the training set.



(a) Sequence 00 (b) Sequence 01 (c) Sequence 02 (d) Sequence 03 (e) Sequence 04

Figure 3: Camera path for our proposed method on KITTI odometry dataset.

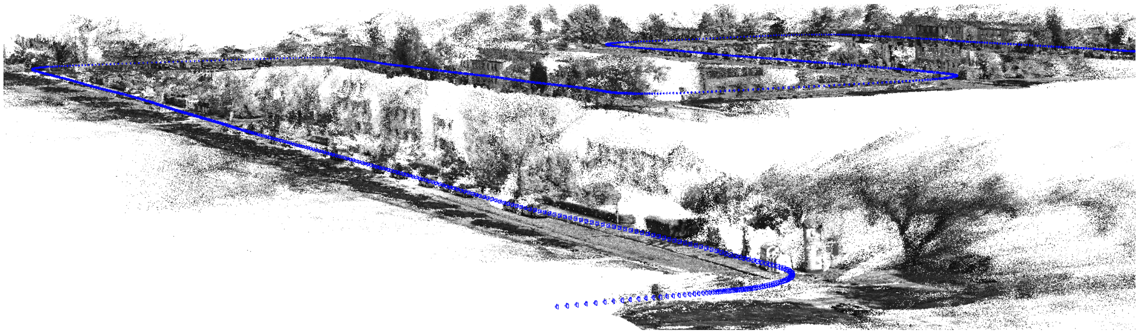


Figure 4: 3D reconstruction results on KITTI odometry dataset.

To make it clear, we denote the proposed joint direct stereo odometry approach as 'GDVO'. Stereo LSD-SLAM [Engel *et al.*, 2015] is the state-of-the-art direct SLAM method while LSD-VO represents LSD-SLAM without loop-closure. It can be clearly seen that GDVO achieves better translation estimation results comparing to both LSD-VO and LSD-SLAM. By taking advantage of both the static and temporal stereo, LSD-VO achieves lower rotation estimation on the training set, which can be further reduced by loop-closure. It is interesting to find that our proposed approach outperforms LSD-VO and LSD-SLAM at a large margin on the testing set, which is even better than the laser-based method [Zhang *et al.*, 2014]. In addition to the relative pose errors, we calculate the absolute translation error to compare against LSD-SLAM. Our proposed approach achieves the lower prediction error and outperforms LSD-SLAM on 6 out of 11 sequences.

Currently, our proposed GDVO method ranks 2^{nd} among vision only algorithms on KITTI odometry benchmark[1], which indicates that direct approach is also promising for stereo visual odometry. As shown in Table 2, our method significantly improves the conventional direct stereo odometry technique [Comport *et al.*, 2010].

Computational Efficiency For our implementation, the whole system include rendering and loading images runs around 12 fps using single thread on CPU. Both block matching stereo and Gauss-Newton optimization are speeded up by SSE instructions. Specifically, our proposed method requires about 75ms to process one stereo frame. It takes 5ms to build the image pyramid and compute gradients for each level. Moreover, it requires 15ms to compute the disparity map and calculate the 3D coordinates. Joint optimization takes around 55ms for each stereo pair.

3D Scene Reconstruction We can take advantage of our proposed stereo odometry approach for 3D scene reconstruction. We remove the large outliers by checking the 3D consistency using the re-projected disparity error between the successive frames. Fig. 4 illustrates the reconstruction result for Sequence '08' of KITTI dataset.

5 Conclusion and Future Work

This paper proposed a novel direct stereo visual odometry approach, which is capable of dealing with large displacement and challenging environments. The proposed method takes advantage of dual Jacobian scheme converging to more accurate results. Moreover, a gradient-based feature representation is employed to account for the illumination variations. Furthermore, we present a joint direct stereo visual odometry to incorporate the information from previous frames, which greatly improves the performance. We have conducted extensive evaluations on KITTI odometry benchmark. The encouraging experimental results showed that our proposed method is on par with the current state-of-the-art feature-based methods while offering dense reconstruction without extra cost.

Despite the promising results, some limitations and future work need to be addressed. At present, we only take consideration of the static stereo. Besides, we have yet to investigate bundle adjustment, loop-closure and IMU measurements. In future work, we will address these issues by incorporating the temporal stereo and pose graph optimization for a full SLAM implementation.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (No. 2016YFB1001501).

[1] `http://www.cvlibs.net/datasets/kitti/eval_odometry.php`, accessed at Feb. 14th, 2017.

References

[Badino *et al.*, 2013] Hernan Badino, Akihiro Yamamoto, and Takeo Kanade. Visual odometry by multi-frame feature integration. In *International Workshop on Computer Vision for Autonomous Driving (CVAD 13)*, 2013.

[Baker and Matthews, 2004] Simon Baker and Iain Matthews. Lucas-kanade 20 years on: A unifying framework. *Int'l J. Computer Vision*, 56(3):221–255, March 2004.

[Benhimane and Malis, 2004] Selim Benhimane and Ezio Malis. Real-time image-based tracking of planes using efficient second-order minimization. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 943–948, 2004.

[Black and Rangarajan, 1996] Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Int'l J. Computer Vision*, 19(1):57–91, 1996.

[Comport *et al.*, 2010] Andrew I. Comport, Ezio Malis, and Patrick Rives. Real-time quadrifocal visual odometry. *I. J. Robotics Res.*, 29(2-3):245–266, 2010.

[Corke *et al.*, 2007] Peter Corke, Jorge Lobo, and Jorge Dias. An introduction to inertial and visual sensing. *I. J. Robotics Res.*, 26(6):519–535, 2007.

[Crivellaro and Lepetit, 2014] Alberto Crivellaro and Vincent Lepetit. Robust 3d tracking with descriptor fields. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014*, pages 3414–3421, 2014.

[Cvisic and Petrovic, 2015] Igor Cvisic and Ivan Petrovic. Stereo odometry based on careful feature selection and tracking. In *2015 European Conference on Mobile Robots, ECMR 2015*, pages 1–6, 2015.

[Engel *et al.*, 2014] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: large-scale direct monocular SLAM. In *Computer Vision - ECCV 2014 - 13th European Conference*, pages 834–849, 2014.

[Engel *et al.*, 2015] Jakob Engel, Jörg Stückler, and Daniel Cremers. Large-scale direct SLAM with stereo cameras. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2015*, pages 1935–1942, 2015.

[Fischler and Bolles, 1981] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *CACM*, 24(6):381–395, 1981.

[Forster *et al.*, 2014] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. SVO: fast semi-direct monocular visual odometry. In *2014 IEEE International Conference on Robotics and Automation, ICRA 2014*, pages 15–22, 2014.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[Hirschmüller, 2008] Heiko Hirschmüller. Stereo processing by semiglobal matching and mutual information. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(2):328–341, 2008.

[Kalman, 1960] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[Klein and Murray, 2007] Georg Klein and David Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan, November 2007.

[Lucas and Kanade, 1981] Bruce D. Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI81*, pages 674–679, 1981.

[Ma *et al.*, 2004] Yi Ma, Stefano Soatto, Jana Kosecka, and S. Shankar Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer-Verlag, New York, 2004.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015.

[Newcombe *et al.*, 2011] Richard A. Newcombe, Steven Lovegrove, and Andrew J. Davison. DTAM: dense tracking and mapping in real-time. In *IEEE International Conference on Computer Vision*, pages 2320–2327, 2011.

[Persson *et al.*, 2015] Mikael Persson, Tommaso Piccini, Michael Felsberg, and Rudolf Mester. Robust stereo visual odometry from monocular techniques. In *2015 IEEE Intelligent Vehicles Symposium, IV 2015*, pages 686–691, 2015.

[Song *et al.*, 2016] Shiyu Song, Manmohan Chandraker, and Clark C. Guest. High accuracy monocular SFM and scale correction for autonomous driving. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(4):730–743, 2016.

[Sturm *et al.*, 2012] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2012*, pages 573–580, 2012.

[Triggs *et al.*, 1999] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment - A modern synthesis. In *Vision Algorithms: Theory and Practice, International Workshop on Vision Algorithms, held during ICCV '99*, pages 298–372, 1999.

[Zhang and Singh, 2015] Ji Zhang and Sanjiv Singh. Visual-lidar odometry and mapping: low-drift, robust, and fast. In *IEEE International Conference on Robotics and Automation, ICRA 2015*, pages 2174–2181, 2015.

[Zhang *et al.*, 2014] Ji Zhang, Michael Kaess, and Sanjiv Singh. Real-time depth enhanced monocular odometry. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4973–4980, 2014.