

Direct Sparse Visual-Inertial Odometry with Stereo Cameras

Ziqiang Wang¹, Chengcheng Guo¹, Lin Zhao¹, Mei Li¹, Xinyu Qi¹

Abstract—We present Stereo VI-DSO, a novel tightly-coupled approach for visual-inertial odometry, which jointly optimizes all the model parameters within the active window, including the IMU pose, velocity, biases, affine brightness parameters of all keyframes and the depth values of all selected pixels. The visual part of the system is integrated constraints from static stereo into the bundle adjustment pipeline of dynamic multi-view stereo, but unlike keypoint based systems it directly minimizes the photometric error. Stereo-VI method can initialize faster than mono-VI. IMU information is accumulated between keyframes using measurement pre-integration, and it is inserted into the optimization as an additional constraint between keyframes. Quantitative evaluation demonstrates that the proposed Stereo VI-DSO is superior to Stereo DSO both in terms of tracking accuracy and robustness. In addition, we introduce a simulation platform developed on Unreal Engine 4, it can output raw data of most sensors used in the field of autonomous driving. We evaluate our method with absolute ground-truth value base on simulation data.

I. INTRODUCTION

Motion estimation is a key task for robots, as it can enable emerging technologies such as (semi)-autonomous driving, augmented and virtual reality, robot or drone navigation. Lidar, Radar, RGB-D cameras, differential GPS(DGPS) and other sensors can be used. LOAM is a low-drift odometry in term of 3D laser rangefinders point clouds[18]. However, it is almost inevitable that single sensor odometry fail in certain scenarios. For example, scan matching in pure Lidar odometry may get wrong match results in degenerate scenes such as those dominated by planar areas. Hence, sensor fusion aroused great interest among the researchers which explores advantages of each sensor and compensates for drawbacks from other sensors. In [19], an inertial measurement unit (IMU) provides a motion prior and mitigate for gross, high-frequency motion.

Since cameras are cheap, lightweight, small and easy to mass produce sensors, they have drawn a large attention of the robotics community[1][3]. However, current pure visual odometry methods require moderate lighting condition and fail when confronted with low textured areas or fast maneuvers. Visual odometry also meets a big challenge by objects moving in front of cameras. IMU that provides sensor body-self rotation and position information, unlike vision, is not impacted by dynamic objects environment. Usually, the IMU frequency is hundreds of Hertz, which captures accurate short-term ego-motion constraints. However, IMU has drastic drift affected by slowly time-varying sensor bias and the presence of measurement noise. Pose results will diverge quickly if only integrating IMU measurements.

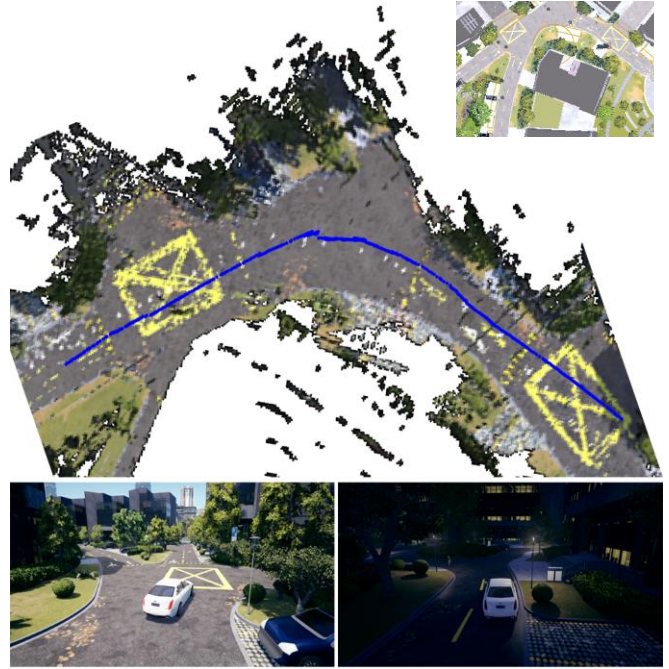


Figure 1. Bottom: Example images from the simulation platform dataset: A Cadillac CT6 is mounted with Lidar, camera, IMU, GNSS and other sensors. Strong motion, low illumination are significant challenges for odometry estimation. Our method is able to work with a rmse of less then 0.33m. Top: Overlay estimated trajectory (bule) in Unreal Editor. Top right is a topview in same road in Google map.

In this paper, we propose a tightly coupled direct sparse stereo visual-inertial odometry. Associating a stereo camera with an IMU, the method estimates accurate motion, which is based on Stereo Direct Sparse Odometry (Stereo DSO) [6]. Combing IMU measurements into the error function, a bundle-adjustment simultaneously optimizes poses, velocity, IMU biases, camera affine brightness parameters and points depth in a combined energy function. A drawback of monocular visual-inertial odometry is that metric scale needs to be properly initialized, otherwise optimization might diverge. Adding a stereo camera enables us to speed up system initialization.

Quantitative evaluation on the indoor and outdoor dataset demonstrates that we can reliably determine sensor motion from a stereo visual-inertial system on a rapidly moving unmanned ground vehicle (UGV) and EuRoC. Besides, we evaluate our method on an autonomous driving simulation platform based on the stereo image stream, IMU raw data and ground-truth value.

¹The authors are with the UISEE(Shanghai) Automotive Technologies LTD, 201800 Shanghai, China. Email: {ziqiang.wang, chengcheng.guo, lin.zhao, mei.li, xinyu.qi}@uisee.com

II. RELATED WORK

In the last few decades, motion estimation using cameras and IMUs has been a popular research topic. In this section we will give an overview of visual and visual-inertial odometry methods.

Visual odometry was introduced in the work of Nister et al. [20]. The main idea underlying these structure and motion techniques is to select a set of keypoints (typically corner-like structures), especially like MonoSLAM [1], a real-time capable EKF-based method. Another famous example is PTAM [21], which combines a bundle-adjustment backend for mapping with real-time capable tracking of the camera relative to the constructed map. Recently, based on ORB features and innovations such as map reuse, ORB-SLAM [3][4] introduced an efficient visual SLAM solution. It has gained a lot of popularity due to its robustness and high tracking accuracy among those state-of-the-art methods for visual SLAM.

Different from the feature-based methods, direct methods use unprocessed intensities in the image to estimate the motion of the camera. The first real-time capable direct approach for stereo cameras was presented [22]. Certain methods for motion estimation for RGB-D cameras were developed by Kerl et al. [12]. More recently, direct approaches were also applied to monocular cameras, in a dense [23], semi-dense [9] and sparse Direct Sparse Odometry (DSO) [5]. Wang et al. extended DSO to stereo vision for highly accurate real-time visual odometry [6].

There are two main categories VIO can be identified. Filtering-based approaches [24][25] operate on a probabilistic state representation, in a Kalman-filtering framework. Optimization approaches on the other hand operate on a minimum loss function based representation in a non-linear optimization framework.

IMU pre-integration technique was firstly proposed by Lupton and Sukkarieh [13], and Forster et al. extended it to Lie group [2], forming a set of elegant theoretical system. Forster realized the IMU pre-integration in the GTSAM 4.0 optimization toolbox [14], and completed the combination with SVO [15]. This theory has been widely applied in VIO based on Bundle Adjustment and feature-based optimization framework, including VI-ORB [4], VINS [16]. Usenko et al. combines IMU measurements with direct tracking of a semi-dense subset of points in the image [8] called Stereo-VI-LSD. Stumberg et al. presented monocular VI-DSO [10] which merges IMU measurements with monocular direct sparse odometry and initializes system with an arbitrary scale instead of having to delay the initialization until everything is observable. To the best of our knowledge, there isn't a mature method in stereo direct sparse odometry coupled with inertial sensor. One of the direct approaches [26] claims to combine IMU measurements with stereo direct image tracking, but does not provide a theoretical derivation, a systematic with inertial sensor evaluation and comparison to other state-of-the-art methods.

III. CONTRIBUTION

In this paper we present a stereo inertial extension of Direct Sparse Odometry. The main novelty of this paper is the formulation of tight IMU integration into stereo direct sparse

image alignment within sliding window optimization. We provide a detailed derivation in an additional supplementary material.

We evaluate our approach on different datasets and compare it to alternative stereo DSO systems, while running in real-time on a modern CPU. In addition, we introduce evaluation method based on an autonomous driving simulation platform and open the dataset made with it.

IV. NOTATION

Throughout the paper, we will write matrices as bold capital letters (\mathbf{R}) and vectors as bold lower case letters ($\boldsymbol{\xi}$), light lower-case letters to denote scalars (s), typewriter letters are used to represent functions (\mathbf{I}).

Rigid-body orientation directly is described as elements of $\mathfrak{so}(3)$ and poses as $\mathfrak{se}(3)$. We can identify every skew symmetric matrix with a vector in \mathbb{R}^3 using the hat (\wedge) operator [2, eq. (1)]. Exponential map associates Lie Algebra to a pose and logarithm is anti-mapping following [2, eq. (3)]: $\text{Exp} \triangleright \mathfrak{so}(3) \leftrightarrow SO(3), \mathfrak{se}(3) \leftrightarrow SE(3) \triangleleft \text{Log}$.

Unlike DSO uses left jacobian and perturbation, we use right method in line with [2]. The term $J_r(\phi), J_r(\boldsymbol{\xi})$ is the right Jacobian of $SO(3), SE(3)$. We write directly as vectors, i.e., $\phi \in \mathbb{R}^3$ and $\boldsymbol{\xi} \in \mathbb{R}^6$. we use the right perturbation retraction for $SO(3)$,

$$\mathbf{R}_2 = \mathbf{R}_1 \text{Exp}(\delta\phi), \delta\phi \in \mathbb{R}^3 \quad (1)$$

and for $SE(3)$, we perturb transformation on the right,

$$\mathbf{T}_2 = \mathbf{T}_1 \text{Exp}(\delta\boldsymbol{\xi}), \delta\boldsymbol{\xi} \in \mathbb{R}^6 \quad (2)$$

The input for our Stereo VI-DSO is a stream of IMU measurements and stereo camera frames. In IMU body frame (abbreviated as “B”), the gyroscope and accelerometer measurements at time k , namely ${}_{\text{B}}\tilde{\boldsymbol{\alpha}}_k$ and ${}_{\text{B}}\tilde{\boldsymbol{\omega}}_k$, are affected by additive white noise $\boldsymbol{\eta}$ and a slowly varying sensor bias \mathbf{b} . Δt is sampling interval. The state of IMU at time i is described by the orientation, position, velocity from “B” to the world frame “W” and biases:

$$\mathbf{x}_i = [{}_{\text{WB}}\mathbf{R}_i, {}_{\text{W}}\mathbf{p}_i, {}_{\text{W}}\mathbf{v}_i, \mathbf{b}_i] \quad (3)$$

Velocities live in a vector space, i.e., ${}_{\text{W}}\mathbf{v}_i \in \mathbb{R}^3$. IMU biases can be written as $\mathbf{b}_i = [{}_{\text{B}}\mathbf{g}_i^g, {}_{\text{B}}\mathbf{g}_i^a] \in \mathbb{R}^6$, where ${}_{\text{B}}\mathbf{g}_i^g, {}_{\text{B}}\mathbf{g}_i^a \in \mathbb{R}^3$ are the gyroscope and accelerometer bias. We model them with “Brownian motion” which is integrated white noise.

Homogeneous camera calibration matrices are denoted by \mathbf{K} . ${}_{\text{BC}}\mathbf{T}$ is the pose of the camera frame “C” in the body frame, known from prior calibration. The “delta” pose from time j to time i is a homogeneous transformation consist by:

$$\Delta \mathbf{T}_{ij} = \mathbf{T}_i^{-1} \mathbf{T}_j \in \mathbb{R}^{4 \times 4} \quad (4)$$

where we dropped the coordinate frame subscripts for readability (the notation should be unambiguous from now on).

V. DIRECT SPARSE VISUAL-INERTIAL STEREO ODOMETRY

We tightly couple inertial integration with non-linear error terms arising from direct image alignment – minimization of the photometric error. To make the problem computationally feasible the optimization is performed on a window of recent frames. Our approach can be viewed as a direct formulation of [2]. In contrast to [2], we perform a full bundle-adjustment like optimization instead of including structure-less vision error. Compared to Stereo-VI-LSD [8], we upgrade visual front-end tracking and back-end bundle adjustment.

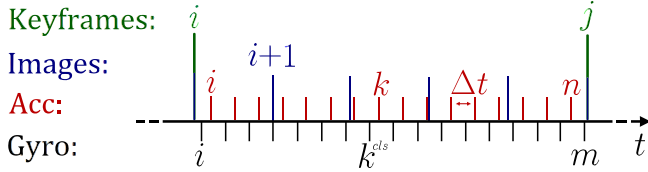


Figure 2. Different rates of IMU and camera: one IMU term uses all accelerometer and gyro readings between successive camera measurements. Furthermore, because of asynchronous but same frequency for accelerometer and gyro data, there will be different quantity samples of these two sensors. We take strategy that for accelerometer sample in time k , combining with time-closest k^{cls} gyro sensor sample.

Compared with VI-DSO[10], scale can be directly calculated from static stereo from the known baseline of the stereo camera. Static stereo can also provide initial depth estimation for multi-view stereo. Initialization coupled with IMU is faster and more robust than monocular.

The proposed approach estimates states by minimizing the energy function with a coupling factor α :

$$E_{total} = E_{IMU} + \alpha E_{CAM} \quad (5)$$

which consists of the an inertial error term E_{IMU} (section V-A) and photometric error E_{CAM} (section V-B).

The system contains two main parts running in parallel:

- The coarse pyramid tracking is executed for every frame and uses direct image alignment. IMU data isn't used in this part, but saved in data packs.
- When a new keyframe in time i is created, all accelerometer and gyroscope data between two consecutive keyframes i and j are pre-integrated followed [2]. We perform a visual-inertial bundle adjustment like optimization that estimates the state of all active keyframes.

In contrast to [2], IMU-measurements are not necessarily synchronized with the camera measurement. Furthermore, accelerometer and gyro-scope data is asynchronous but at the same frequency in our IMU sensors (Figure 2).

The new keyframe will be generated and added to the active window. For all keyframes in the active window, a joint optimization of IMU state, affine brightness parameters, as well as the inverse depths of all the observed 3D points is performed. To maintain the size of the active window, old keyframes and 3D points are marginalized out using the Schur complement. Suppose \mathcal{F} is the set of the consecutive keyframes in the current window.

A. IMU Error Factors

Given the pre-integrated measurement model in [2], We extend this model to asynchronous measurement. Constant large IMU biases part $\bar{\mathbf{b}}$ update by a small amount $\delta\mathbf{b}$ during optimization, i.e., $\mathbf{b} = \bar{\mathbf{b}} + \delta\mathbf{b}$. There are m gyro data and n acc data ($m \neq n$) between two consecutive keyframes i and j in Figure 2. The estimate of rigid "delta" rotation $\Delta\bar{\mathbf{R}}_{ij}$ is computed iteratively by gyro data in timestamp k which is independent to accelerometer, and first-order Taylor expansion is a large value $\Delta\bar{\mathbf{R}}_{im}(\bar{\mathbf{b}}_i^g) = \Delta\bar{\mathbf{R}}_{im}$ multiply a small perturbation $\Delta\bar{\mathbf{R}}_{im}(\delta\mathbf{b}_i^g) = \text{Exp}(\frac{\partial\Delta\bar{\mathbf{R}}_{im}}{\partial\mathbf{b}_i^g}\delta\mathbf{b}_i^g)$ on right:

$$\Delta\bar{\mathbf{R}}_{ik} = \begin{cases} \mathbf{I}_{3 \times 3}, & k = i \\ \Delta\bar{\mathbf{R}}_{i(k-1)}\text{Exp}((\tilde{\omega}_{k-1} - \bar{\mathbf{b}}_i^g)\Delta t), & k \in [i+1, m] \end{cases} \quad (6)$$

Jacobian $\frac{\partial\Delta\bar{\mathbf{R}}_{im}}{\partial\mathbf{b}_i^g}$ is recursively calculated:

$$\frac{\partial\Delta\bar{\mathbf{R}}_{ik}}{\partial\mathbf{b}_i^g} = \begin{cases} \mathbf{0}_{3 \times 3}, & k = i \\ \Delta\bar{\mathbf{R}}_{(k-1)k}^T \frac{\partial\Delta\bar{\mathbf{R}}_{i(k-1)}}{\partial\mathbf{b}_i^g} - \mathbf{J}_r^{k-1} \Delta t, & k \in [i+1, m] \end{cases} \quad (7)$$

where $\mathbf{J}_r^{k-1} = \mathbf{J}_r((\tilde{\omega}_{k-1} - \bar{\mathbf{b}}_i^g)\Delta t)$ and

$$\Delta\bar{\mathbf{R}}_{(k-1)k}^T = \text{Exp}((\tilde{\omega}_{k-1} - \bar{\mathbf{b}}_i^g)\Delta t).$$

Because of $\Delta\tilde{\mathbf{v}}_{ij} = \Delta\tilde{\mathbf{v}}_{im} + \frac{\partial\Delta\tilde{\mathbf{v}}_{in}}{\partial\mathbf{b}_i^g}\delta\mathbf{b}_i^g + \frac{\partial\Delta\tilde{\mathbf{v}}_{in}}{\partial\mathbf{b}_i^a}\delta\mathbf{b}_i^a$ is related to both gyro and accelerometer. We choose time-closest $\Delta\bar{\mathbf{R}}_{ik}^{cls}$ and $\frac{\partial\Delta\bar{\mathbf{R}}_{ik}^{cls}}{\partial\mathbf{b}_i^g}$ for an accelerometer data $\tilde{\mathbf{a}}_k$ and recursively calculate:

$$\begin{aligned} \Delta\tilde{\mathbf{v}}_{ik} &= \Delta\tilde{\mathbf{v}}_{i(k-1)} + \Delta\bar{\mathbf{R}}_{ik}^{cls}((\tilde{\mathbf{a}}_{k-1} - \bar{\mathbf{b}}_i^a)\Delta t), \\ \frac{\partial\Delta\tilde{\mathbf{v}}_{ik}}{\partial\mathbf{b}_i^g} &= \\ \frac{\partial\Delta\tilde{\mathbf{v}}_{i(k-1)}}{\partial\mathbf{b}_i^g} - \Delta\bar{\mathbf{R}}_{i(k-1)}^{cls}(\tilde{\mathbf{a}}_{k-1} - \bar{\mathbf{b}}_i^a) \wedge \frac{\partial\Delta\bar{\mathbf{R}}_{i(k-1)}^{cls}}{\partial\mathbf{b}_i^g} \Delta t & \quad (8) \\ \frac{\partial\Delta\tilde{\mathbf{v}}_{ik}}{\partial\mathbf{b}_i^a} &= \frac{\partial\Delta\tilde{\mathbf{v}}_{i(k-1)}}{\partial\mathbf{b}_i^a} - \Delta\bar{\mathbf{R}}_{i(k-1)}^{cls} \Delta t, & k \in [i+1, n] \end{aligned}$$

Repeating the same process for $\Delta\tilde{\mathbf{p}}_{ij}$, now it is easy to write the residual errors $\mathbf{r}_{I_{ij}} = [\mathbf{r}_{\Delta\mathbf{R}_{ij}}^T, \mathbf{r}_{\Delta\mathbf{v}_{ij}}^T, \mathbf{r}_{\Delta\mathbf{p}_{ij}}^T] \in \mathbb{R}^9$, where:

$$\begin{aligned} \mathbf{r}_{\Delta\mathbf{R}_{ij}} &= \text{Log}(\Delta\bar{\mathbf{R}}_{im} \text{Exp}(\frac{\partial\Delta\bar{\mathbf{R}}_{im}}{\partial\mathbf{b}_i^g}\delta\mathbf{b}_i^g))^T \mathbf{R}_i^T \mathbf{R}_j \\ \mathbf{r}_{\Delta\mathbf{v}_{ij}} &= \mathbf{R}_i^T(\mathbf{v}_j - \mathbf{v}_i - \mathbf{g}\Delta t_{ij}) \\ &\quad - (\Delta\tilde{\mathbf{v}}_{im} + \frac{\partial\Delta\tilde{\mathbf{v}}_{in}}{\partial\mathbf{b}_i^g}\delta\mathbf{b}_i^g + \frac{\partial\Delta\tilde{\mathbf{v}}_{in}}{\partial\mathbf{b}_i^a}\delta\mathbf{b}_i^a) \\ \mathbf{r}_{\Delta\mathbf{p}_{ij}} &= \mathbf{R}_i^T(\mathbf{p}_j - \mathbf{p}_i - \mathbf{v}_i\Delta t_{ij} - \frac{1}{2}\mathbf{g}\Delta t_{ij}^2) \\ &\quad - (\Delta\tilde{\mathbf{p}}_{im} + \frac{\partial\Delta\tilde{\mathbf{p}}_{in}}{\partial\mathbf{b}_i^g}\delta\mathbf{b}_i^g + \frac{\partial\Delta\tilde{\mathbf{p}}_{in}}{\partial\mathbf{b}_i^a}\delta\mathbf{b}_i^a) \end{aligned} \quad (9)$$

The noise covariance Σ_{ij} depending on rotation and position, has a strong influence on the MAP estimator (the inverse noise covariance is used to weight the terms in the optimization). We can get covariance recursively like previous calculations of asynchronous data. Here comes IMU error term in current window:

$$E_{IMU} = \sum_{i,j \in \mathcal{F}} \mathbf{r}_{I_{ij}}^T \Sigma_{ij}^{-1} \mathbf{r}_{I_{ij}} \quad (10)$$

B. Photometric Error Factors

We apply visual tracking strategy of Stereo DSO[6]:

- We track the motion of the camera towards a reference keyframe in the map and create new keyframes according to DSO.
- We estimate the inverse depth of selected points in the current reference keyframe from static and dynamic stereo cues. For static stereo we exploit the fixed baseline between the pair of cameras in the stereo configuration. Dynamic stereo is estimated from pixel correspondences in the reference keyframe towards subsequent images based on the tracked motion.

Once creating a new keyframe, a sparse set of points is selected from the image, which will be called candidate points, this keyframe will be hostframe of selected points in the rest of the paper. Points that have sufficient image gradient are selected across the image. To make sure selected points distribute sparsely and evenly, the image is divided into small blocks and for each block an adaptive threshold is adopted.

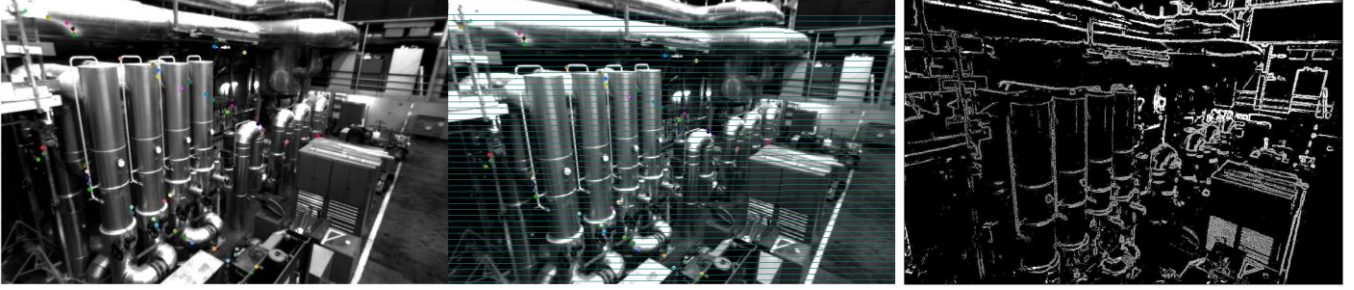


Figure 3. Visualization of typical data association on a European Robotics Challenge (EUROC) dataset: current stereo image pair (left) match points (colorful), Green line stands for epipolar lines. Depth map (right) of sufficient gradient points.

1) Static Stereo

Suppose a 2D image coordinate point \mathbf{p} is selected in keyframe i^L and observed by i^R , we search the corresponding pixel on the epipolar line, making the pixel with the highest similarity as the matching point in corresponding i^R . Considering the precision and robustness of match, NCC criterion was used in window match method. Once obtaining the pair of match points, a inverse depth initialization $d_{\mathbf{p}}^L$ can be calculated by using the method of triangular intersection illustrated in Figure 3. Frame i^L, i^R gray function is $\mathbb{I}_i^L, \mathbb{I}_i^R$. Static one-view stereo residuals are defined to:

$$\begin{aligned} E_{i^L i^R}^{\mathbf{p}} &= w_{\mathbf{p}} \|r_{\mathbf{p}}^s\|_{\gamma}, \\ r_{\mathbf{p}}^s &= \mathbb{I}_i^R(\mathbf{p}') - b_i^R - \frac{e^{a_i^R}}{e^{a_i^L}} ((I_i^L(\mathbf{p}) - b_i^L)) \end{aligned} \quad (11)$$

where $\|\cdot\|_{\gamma}$ is Huber norm, $a_i^L, a_i^R, b_i^L, b_i^R$ is affine brightness parameters to frame i^L and i^R , $w_{\mathbf{p}}$ is a gradient-dependent weighting parameter that down-weights high image gradients, c is a constant value.

$$w_{\mathbf{p}} = \frac{c^2}{c^2 + \|\nabla \mathbb{I}_i(\mathbf{p})\|_2^2}, \quad (12)$$

Relative transformation between the left and right cameras \mathbf{T}_{RL} is fixed. \mathbf{p} in frame I_i^L projected to I_i^R is \mathbf{p}' as:

$$\mathbf{p}' = d_{\mathbf{p}}^{i^R} \mathbf{K}(\mathbf{T}_{RL}((d_{\mathbf{p}}^{i^L})^{-1} \mathbf{K}^{-1} \mathbf{p})) \quad (13)$$

2) Dynamic Multi-View Stereo

Assume \mathbf{p} is observed and projected to another keyframe j^L denoted as \mathbf{p}' . Instead of using camera frame pose ${}^c\mathbf{T}_{ji}$ [8][10], we use IMU pose $\Delta \mathbf{T}_{ij}^{-1} = \mathbf{T}_j^{-1} \mathbf{T}_i$ and \mathbf{T}_{BC} maps a 3D landmark from i^L to j^L . We find this is more rigorous in theory to derive compared to assume that gyro and accelerometer of camera frame is same to IMU frame:

$$\mathbf{p}' = d_{\mathbf{p}}^{j^L} \mathbf{K}(\mathbf{T}_{BC}^{-1} \mathbf{T}_j^{-1} \mathbf{T}_i((d_{\mathbf{p}}^{i^L})^{-1} \mathbf{T}_{BC} \mathbf{K}^{-1} \mathbf{p})) \quad (14)$$

Direct residuals $E_{ij}^{\mathbf{p}}$ are defined as:

$$\begin{aligned} E_{i^L j^L}^{\mathbf{p}} &= w_{\mathbf{p}} \|(r_{\mathbf{p}}^d)_{ij}\|_{\gamma}, \\ (r_{\mathbf{p}}^d)_{ij} &= \mathbb{I}_j^L(\mathbf{p}') - b_j^L - \frac{e^{a_j^L}}{e^{a_i^L}} (\mathbb{I}_i^L(\mathbf{p}) - b_i^L) \end{aligned} \quad (15)$$

where a_j^L, b_j^L are affine brightness parameters to frame j^L .

Assume a point set \mathcal{P} are selected, $obs(\mathbf{p})$ is the set of the keyframes in \mathcal{F} that can observe \mathbf{p} , a coupling factor λ . Total dynamic multi-view stereo and static one-view stereo residuals is listed as (16).

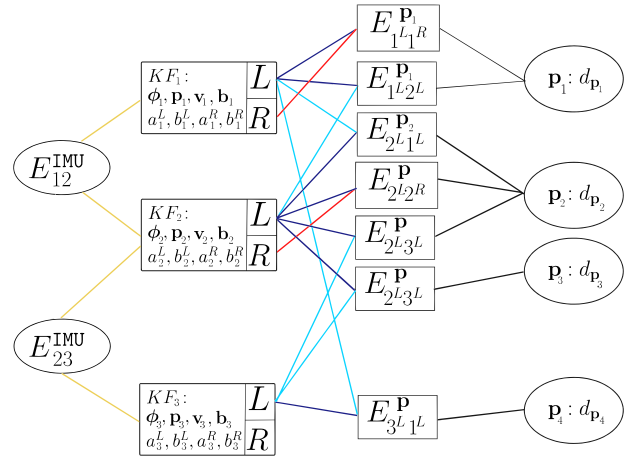


Figure 4. Factor graph of the energy function. In this example, 4 points are observed by 3 keyframes. Each energy factor is related to one point and two keyframes, thus depends on IMU pose, velocity, bias, inverse depth of the point, their affine brightness correction factors. Cameras constraints from host keyframes and static stereo are shown in dark blue and red respectively. Remaining constraints in light blue are the ones from the keyframes the points are observed. IMU constraints from two consecutive keyframes is yellow.

$$E_{CAM} = \sum_{i^L \in \mathcal{F}} \sum_{\mathbf{p} \in \mathcal{P}_i} \left(\sum_{j^L \in obs(\mathbf{p})} E_{i^L j^L}^{\mathbf{p}} + \lambda E_{i^L i^R}^{\mathbf{p}} \right) \quad (16)$$

C. Optimization

The total energy is optimized iteratively using Gauss-Newton algorithm with IMU pose, velocity, biases, affine brightness and inverse depth parameters to be optimized:

$$\begin{aligned} &(\mathbf{J}_I^T \mathbf{W}_I \mathbf{J}_I + \mathbf{J}_C^T \mathbf{W}_C \mathbf{J}_C) \delta \chi \\ &= -(\mathbf{J}_I^T \mathbf{W}_I \mathbf{r}_I + \mathbf{J}_C^T \mathbf{W}_C \mathbf{r}_C) \end{aligned} \quad (17)$$

$$\chi^{\text{new}} = \chi \otimes \delta \chi$$

where $\mathbf{r}_I, \mathbf{r}_C$ contains the stacked residuals of E_{IMU}, E_{CAM} , and $\mathbf{J}_I, \mathbf{J}_C$ is the Jacobian, $\mathbf{W}_I, \mathbf{W}_C$ is the weight matrix. The parameters we want to optimize are enclosed in (18). Where N_f and N_p are the numbers of keyframes and active points in the current window, respectively. The \otimes -operator is state space update and $\delta \chi$ is a right-multiplied increment to the current state. A factor graph is shown in Figure 4.

To keep the active window of bounded size, old keyframes are removed by marginalization using the Schur complement to marginalize a subset of variables [5][11].

$$\chi = \begin{pmatrix} (\phi_1, \dots, \phi_{N_f})^T \\ (\mathbf{p}_1^T, \dots, \mathbf{p}_{N_f}^T)^T \\ (\mathbf{v}_1^T, \dots, \mathbf{v}_{N_f}^T)^T \\ (\mathbf{b}_1^T, \dots, \mathbf{b}_{N_f}^T)^T \\ (d_{\mathbf{p}_1}, \dots, d_{\mathbf{p}_{N_f}})^T \\ (a_1^L, a_1^R, b_1^L, b_1^R)^T \\ \vdots \\ (a_{N_f}^L, a_{N_f}^R, b_{N_f}^L, b_{N_f}^R)^T \end{pmatrix} \in \mathbb{R}^{19N_f + N_p}, \quad (18)$$

$$\phi_i = \text{Log}(\mathbf{R}_i)$$

TABLE I. TRANSLATIONAL DRIFT EVALUATED OVER DIFFERENT SEGMENT LENGTHS

Length(m)	30	60	90	120	RMSE(m)
<i>Our</i>	0.369	0.703	0.991	1.309	0.8273
<i>S-DSO</i>	0.636	1.105	1.894	4.814	2.1179

TABLE II. ACCURACY OF THE ESTIMATED TRAJECTORY ON THE EUROC DATASET FOR SEVERAL METHODS. WE RUN AND CALCULATE RMSE OF VINS AND OKVIS IN OUR OWN LAPTOP.

Seq	Length (m)	Stereo-DSO		Stereo-VI-DSO		VINS		OKVIS
		Orien. (deg)	Pos. (m)	Orien. (deg)	Pos. (m)	Orien. (deg)	Pos. (m)	
<i>MH1</i>	78.7	17.808	1.282	10.798	0.754	4.536	0.444	0.597
<i>MH2</i>	70.1	14.652	1.133	8.461	0.659	3.939	0.327	0.698
<i>MH3</i>	132.4	12.224	3.913	8.352	0.701	8.612	0.335	0.551

VI. RESULTS

We evaluate our approach both qualitatively and quantitatively on different datasets, including a direct comparison to Stereo-DSO. We used the datasets captured by the hardware, EuRoC benchmark and datasets made of our autonomous driving simulation platform.

A. Real Experiments

The custom-built visual-inertial sensor consists of an BMI088 MEMS IMU and two embedded WVGA monochrome cameras with 12 cm baseline that are all rigidly connected. The data is streamed to a host computer via USB. The datasets used in this work were collected at an IMU rate of 200Hz, while the camera frame rate was set to 30 Hz.

We have taken the IMU noise parameters from the datasheet BMI088. We used the following IMU parameters: Gyroscope and accelerometer continuous-time noise density: $\sigma^g = 0.007[\text{rad}/(s\sqrt{\text{Hz}})]$, $\sigma^a = 0.019[\text{m}/(s^2\sqrt{\text{Hz}})]$, Gyro and accelerometer bias continuous-time noise density: $\sigma^{bg} = 0.02[\text{rad}/(s^2\sqrt{\text{Hz}})]$, $\sigma^{ba} = 0.06[\text{m}/(s^3\sqrt{\text{Hz}})]$.

The sensor was mounted onto an UGV for driving of a triangular region in our work zone and back to the starting point. 3D ground truth obtained using a high-precision BeiDou GNSS module. We furthermore neglect the offset between GPS antenna and IMU center. Using first keyframe IMU body to be world frame, we align GNSS coordinate systems to world

Some example images collected by the hardware are shown in Figure 5. In TABLE I., we show the RMSE error increases as the driving distance increases.

B. EuRoC Benchmark

We tested the proposed Stereo-VI-DSO on part of sequences in EuRoC dataset [27], in which a FireFly hex-rotor helicoptere quipped with VI-sensor (an IMU @ 200Hz and



Figure 5. Images from the real experiment (upper row: dynamic people, bottom row: sudden change of light) with depth estimates. Left is rgb image, then convert to grayscale, with color-coded depth estimates are shown on the right.

dual cameras 752×480 pixels @ 20Hz) was used for data collection.

In TABLE II., we present results of Stereo-DSO and Stereo-VI-DSO. The results of Stereo-DSO come from our approach removing the IMU constraint. VINS[16] and OKVIS[11] are open-source and the state of the art works. For comparison, we also provide accuracy RMSE results of VINS, OKVIS.

We also present all robot states estimation results in Figure 6. Figure 7. Figure 8. We can draw a conclusion that Stereo-VI-DSO have a significant improvement over Stereo-DSO in accuracy.

C. Simulation Experiments

We reconstructed the scene of our work zone on Unreal Engine 4 as shown in Figure 1. Benefit from Unreal real-time rendering, we also develop some simulation sensors which can output raw data of most sensors used in the field of autonomous driving. The simulated acceleration and gyroscope measurements are computed from the absolute vehicle pose in Engine and additionally corrupted by white noise and a slowly time-varying bias terms, according to the IMU model in [2] and parameters in real experiments. We evaluate the accuracy of our method with absolute ground-truth value base on simulation data shown in Figure 9.

VII. CONCLUSION

We have presented a novel approach to direct sparse, tightly integrated visual-inertial odometry. It combines a fully direct structure and motion approach – operating on per-pixel depth instead of individual keypoint observations – with tight, minimization-based IMU integration. Our method can outperform existing Stereo-DSO approaches in terms of tracking accuracy while running in real-time on a standard laptop CPU.

In future work, we will investigate tight direct visual IMU integration with Lidar. Lidar can correction depth of

landmarks, on the other hand, a better initial pose of VIO can help Lidar odometry optimization convergence quickly.

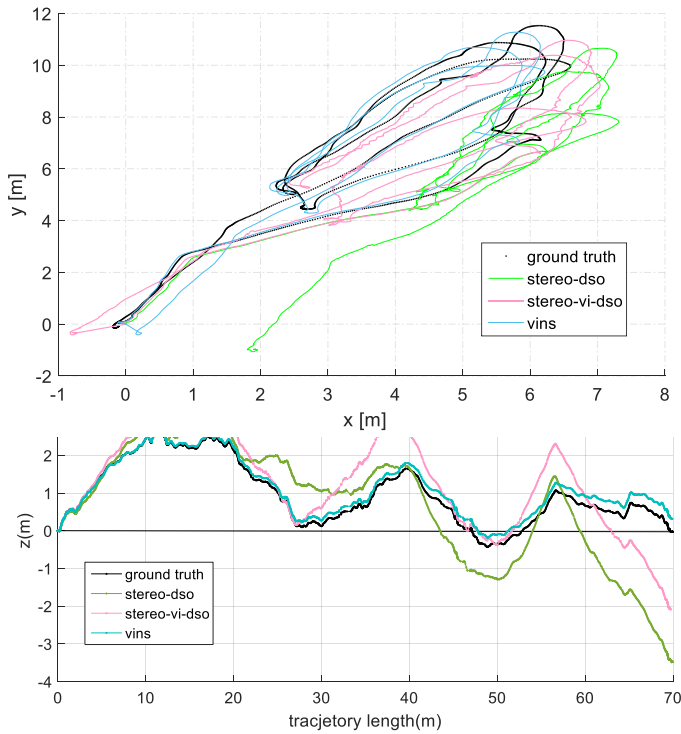


Figure 6. Trajectory and height estimates in MH2.

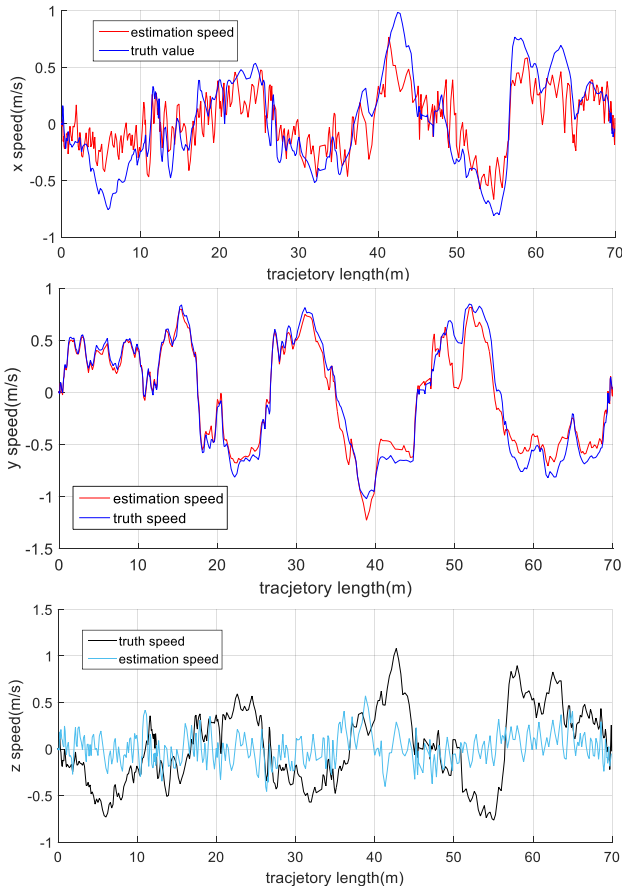


Figure 7. Velocity estimates in MH2.

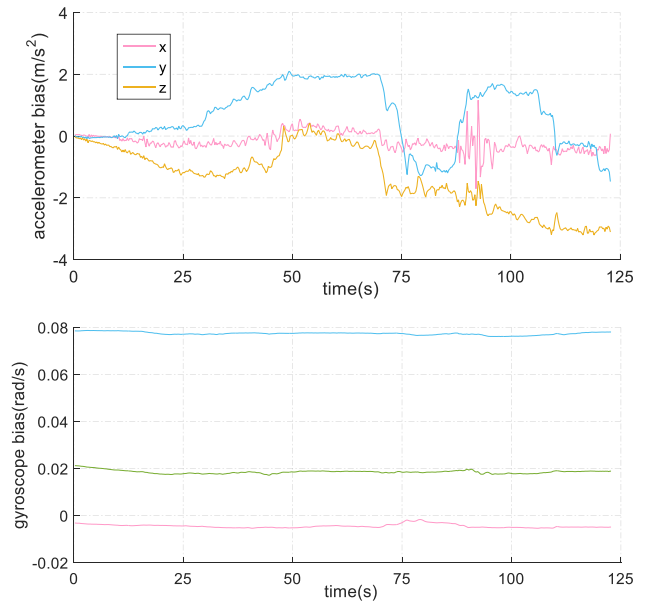


Figure 8. IMU bias estimates in MH2.

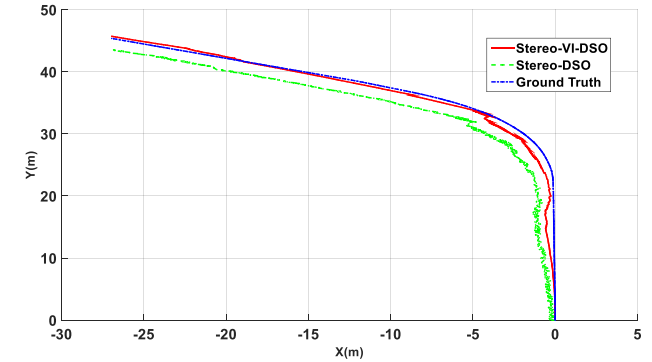


Figure 9. Trajectory estimates in simulation experiments. We install a stereo camera and an IMU on a Cadillac CT6. Driving manually for some distance, we record stereo images, IMU data and ground truth.

APPENDIX

A detailed derivation in an additional supplementary material below, if the article can't display online, you can click Download button for reading offline.

https://github.com/ArmstrongWall/Sup_S_VI_DSO/blob/master/Sup_S_VI_DSO.pdf.

Our autonomous driving simulation platform dataset:

https://github.com/ArmstrongWall/unreal_autonomous_driving_dataset.

We present evaluation results on EuRoC dataset and real experiments video.

<https://youtu.be/cEBNrhgElk4>.

REFERENCES

- [1] A. J. Davison, I. D. Reid, N. D. Molton and O. Stasse, "MonoSLAM: Real-Time Single Camera SLAM," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1052-1067, June 2007.

- [2] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-Manifold Preintegration for Real-Time Visual-Inertial Odometry," in *IEEE Trans. Robot.*, 2016.
- [3] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255-1262, (2017)
- [4] R. Mur-Artal and J. D. Tardos, "Visual-inertial monocular slam with map reuse," *IEEE Robot. and Autom. Lett.*, vol. 2, no. 2, 2017.
- [5] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *TPAMI*, vol. 40, 2018.
- [6] R. Wang, M. Schwörer, D. Cremers, "Stereo DSO: Large-Scale Direct Sparse Visual Odometry with Stereo Cameras," In *International Conference on Computer Vision (ICCV)*, (2017)
- [7] J. Engel, J. Stueckler, D. Cremers, "Large-Scale Direct SLAM with Stereo Cameras," In *International Conference on Intelligent Robots and Systems (IROS)*, (2015)
- [8] V. Usenko, J. Engel, J. Stueckler, and D. Cremers, "Direct visual-inertial odometry with stereo cameras," in *IEEE ICRA*, May 2016.
- [9] J. Engel, T. Schops, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. of ECCV*, 2014.
- [10] L. von Stumberg, V. Usenko and D. Cremers, "Direct Sparse Visual-Inertial Odometry using Dynamic Marginalization", In *International Conference on Robotics and Automation (ICRA)*, 2018.
- [11] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *The International Journal of Robotics Research*, 34(3):314–334, 2015.
- [12] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *IEEE ICRA*, 2013.
- [13] T. Lupton and S. Sukkarieh, "Visual-inertial-aided navigation for high-dynamic motion in built environments without initial conditions," *IEEE Trans. on Robotics*, vol. 28, no. 1, pp. 61–76, 2012.
- [14] F. Dellaert. "Factor graphs and GTSAM: A hands-on introduction," Technical Report GT-RIM-CP&R-2012-002, Georgia Institute of Technology, 2012.
- [15] C. Forster, M. Pizzoli, and D. Scaramuzza. "SVO: Fast SemiDirect Monocular Visual Odometry," In *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2014. doi: 10.1109/ICRA.2014.6906584.
- [16] Tong, Qin, L. Peiliang, and S. Shaojie. "VINS-Mono: A Robust and Versatile Monocular Visual-Inertial State Estimator." *IEEE Transactions on Robotics* (2018):1-17.
- [17] C. Kerl, J. Sturm, and D. Cremers, "Robust odometry estimation for RGB-D cameras," in *IEEE ICRA*, 2013.
- [18] J. Zhang and S. Singh. "LOAM: Lidar Odometry and Mapping in Real-time." Conference Paper, *Robotics: Science and Systems Conference*, July, 2014
- [19] J. Zhang, and S. Singh. "Low-drift and real-time lidar odometry and mapping." *Autonomous Robots* 41.2(2017):401-416.
- [20] D. Nister, O. Naroditsky, and J. Bergen, "Visual odometry," in *IEEE CVPR*, vol. 1, June 2004, pp. I-652–I-659 Vol.1.
- [21] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. of ISMAR*, 2007.
- [22] A. Comport, E. Malis, and P. Rives, "Accurate quadri-focal tracking for robust 3D visual odometry," in *IEEE ICRA*, 2007.
- [23] R. Newcombe, S. Lovegrove, and A. Davison, "DTAM: Dense tracking and mapping in real-time," in *IEEE ICCV*, 2011.
- [24] M. Li and A. Mourikis. High-precision, consistent EKF-based visual inertial odometry. *Int. Journal of Robotics Research (IJRR)*, 32:690–711, 2013.
- [25] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart. Robust visual inertial odometry using a direct EKF-based approach. In *Proc. of the IEEE/RSJ Int. Conf. on Intelligent Robot Systems (IROS)*, 2015.
- [26] S. Mann, D. Göhring, and R. Rojas. "Extension of Direct Sparse Odometry with Stereo Camera and Inertial Sensor,". URL <https://www.mi.fu-berlin.de/inf/groups/ag-ki/Theses/Completed-theses/Master-Diploma-theses/2017/Mann/index.html>
- [27] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The euroc micro aerial vehicle datasets,"