Enhancing Accuracy in Visual SLAM by Tightly Coupling Sparse Ranging Measurements Between Two Rovers

Chen Zhu*, Gabriele Giorgi†,
*Institute for Communications and Navigation
Technische Universität München
Munich, Germany
Email: chen.zhu@tum.de, younghee.lee@tum.de

Young-Hee Lee*, Christoph Günther*†
†Institute of Communications and Navigation
German Aerospace Center (DLR)
Oberpfaffenhofen, Germany
Email: gabriele.giorgi@dlr.de, christoph.guenther@dlr.de

***Abstract*— Compared with stand-alone rovers, cooperative swarms of robots equipped with cameras enable a more efficient exploration of the environment, and are more robust against malfunctions of an individual platform. VSLAM (Visual Simultaneous Localization and Mapping) techniques have been developed in recent years to estimate the trajectory of vehicles and to simultaneously reconstruct the map of the surroundings using visual clues. This work proposes a tight coupling sensor fusion approach based on the combined use of stereo cameras and sparse ranging measurements between two dynamic rovers in planar motion. The Cramér-Rao lower bound (CRLB) of the rover pose estimator using the fusion algorithm is calculated. Both the lower bound and the simulation results show that to what extent the proposed fusion method outperforms the vision-only approach.**

I. Introduction

Autonomous robotic platforms can be utilized in the exploration of extreme environments, e.g., extraterrestrial exploration or in disaster areas. The autonomous navigation of the robots often relies on several sensors such as mobile radio receivers, Inertial Measurement Units (IMUs), laser scanners and cameras [1]. VSLAM (Visual Simultaneous Localization and Mapping) techniques using stereo camera rigs have been developed in recent years to estimate the trajectory of vehicles and to simultaneously reconstruct the map of the environment [2][3].

In order to increase the system robustness against hazards inherent to the missions (e.g., the rover being incapacitated due to wheel slippage in complicated terrains or blocks in the trajectory), and to improve the exploration efficiency, we propose to use a robotic swarm including multiple autonomous units [4]. For such scenario, several multi-agent cooperative VSLAM approaches have been devised [5] [6]. Estimating the relative pose between different rovers is a core problem in multi-robot SLAM. All the state-of-the-art methods are either based on the merging of images or maps, e.g., [7] and [8], which requires overlapping exploration areas and significant amounts of data transmission, or require to detect another rover in the camera field of view, such as the methods in [9] and [10]. By establishing a wireless radio link between two rovers, ranging measurements can be obtained using pilot signals and round-trip-delay (RTD) estimation methods [11]. The additional information can be used to improve the exploration based on VSLAM techniques. Using the methods proposed in [12], the relative pose between the two rovers can be estimated by using cameras and range measurements, without transmitting any image or feature point and without requiring another rover to appear in the field of view of the cameras. However, the method is based on loose coupling of the sensors and does not exploit the range measurements to improve the visual SLAM accuracy besides consistent scale estimation. Therefore, we propose in this work a tight coupling sensor fusion method that exploits both the ranging measurements and the stereo camera images, and shows to what extent the rover pose estimation can be improved.

The organisation of the paper is as follows: in Section II, we define the system model and give a brief introduction of stereo-camera-based VSLAM. In Section III, the Cramér-Rao lower bound is calculated for VSLAM in planar motion based on stereo cameras. Subsequently, a sensor fusion method is proposed in Section IV, which exploits a ranging link between two dynamic rovers. Simulation results are provided in Section V and conclusions are drawn from the analysis.

II. System Model and Visual SLAM using Stereo Camera Rigs

Fig. 1 illustrates the system, composed of two rovers arbitrarily moving in a plane. The rovers, each equipped with a stereo camera rig and a wireless radio receiver, execute SLAM tasks on the ground. The motion of both vehicles is constrained to be planar. Let $\vec{\beta}_{j,[k]}^{(W)} \in \mathbb{R}^2$ be the position of robot j in the world frame (W) at time k. In the remainder of this paper, we use a superscript with parentheses $^{(\cdot)}$ to denote the coordinate frame in which the vector is represented. Vectors such as $\vec{\beta} \in \mathbb{R}^2$ with geometric meanings are written with an arrow notation on top. Time, denoted with square brackets $[\cdot]$, refers to keyframes, i.e., the time reference instances in which both the range measurements and the trajectory estimation are available. We use (k) to express the local coordinate frame (i.e., the frame integral with the rovers' bodies) at keyframe k. We choose the initial position of the camera projection center of rover 2 as the coordinate reference system's origin, and the camera's principal axis

The project VaMEx-CoSMiC is supported by the Federal Ministry for Economic Affairs and Energy on the basis of a decision by the German Bundestag, grant 50NA1521 administered by DLR Space Administration.

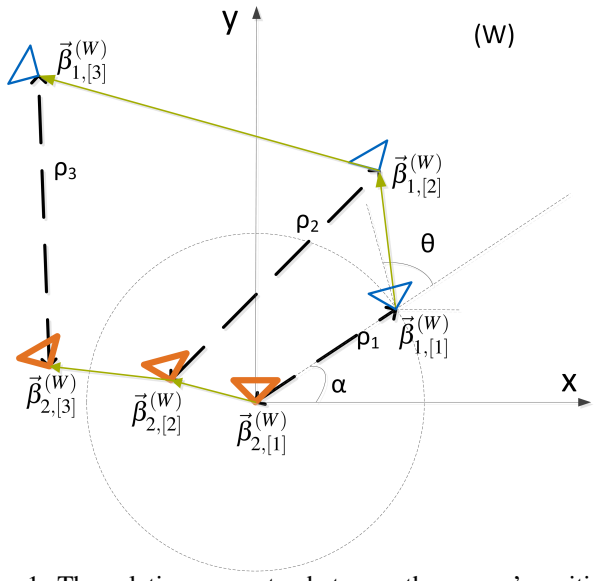

Fig. 1: The relative geometry between the rovers' positions in the global frame (W)

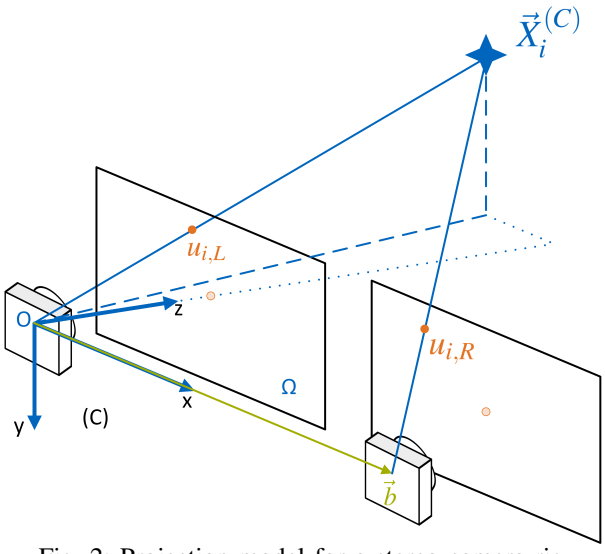

Fig. 2: Projection model for a stereo camera rig

as the y-axis. Generally, the transformation between two coordinate frames (P) and (Q) follows

$$\vec{X}^{(Q)} = R_{(P\to Q)}\vec{X}^{(P)} + \vec{t}_{(P\to Q)}, \tag{1}$$

where $\vec{X}^{(P)}$ and $\vec{X}^{(Q)}$ denote the coordinates of an arbitrary 3D point $\vec{X} \in \mathbb{R}^3$ expressed in the corresponding (P) and (Q) frames, $R_{(P\to Q)} \in \mathbf{SO}(3)$ denotes the rotation matrix, and $\vec{t}_{(P\to Q)}$ denotes the translation vector from the origin of (P) to the origin of (Q).

The origin of the body frame identifies the position of the ranging sensor. Since the relative pose between the stereo camera rig and the ranging sensor can be obtained by calibration, the body frame and camera frame are not distinguished. This assumption does not affect the validity of the algorithm if the body is assumed to be rigid.

Fig. 2 shows the projection model for the chosen stereo setup. The origin of the camera frame is defined at the

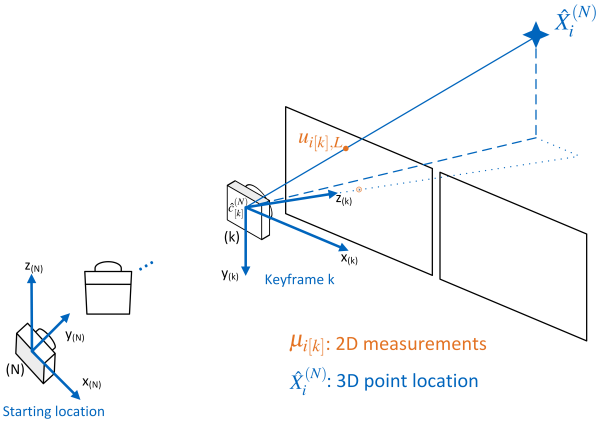

Fig. 3: Projection of a point in the navigation frame

projection center of the left camera. $\Omega \subset \mathbb{R}^2$ is the image plane. Applying the pinhole model, the perspective projection can be formulated as

$$\tilde{u}_{iL} = d_i[u_{iL}, 1]^T = K_L\vec{X}^{(C)}_i, \tag{2}$$

where $d_i = X^{(C)}_{i,z}$ is the depth of the point, and K_L is the camera intrinsic matrix. $u_{iL} \in \mathbb{R}^2$ denotes the Cartesian coordinates of the point's two-dimensional (2D) location in the image, and $\tilde{u}_{iL} \in \mathbb{P}^2$ is the corresponding homogeneous coordinates in the extended Euclidean space. Assuming the image planes of both cameras in the stereo rig to be coplanar (possibly after rectification) and the right camera to be set from the left one with a pure translation along the x-axis, the position of the right camera is $\vec{b}^{(C)} = [l, 0, 0]^T$. The projection of the same point on the right camera is

$$\tilde{u}_{iR} = d_i[u_{iR}, 1]^T = K_R(\vec{X}^{(C)}_i - \vec{b}^{(C)}). \tag{3}$$

Using the matched visual features at both image planes, the depth d_i can be retrieved and the three-dimensional (3D) location of the point can be estimated as $\hat{X}^{(C)}_i$.

We define a navigation frame (N) as a fixed coordinate frame with its origin at the starting location of the rover. The navigation frame of each rover is related to the world reference frame by a specific transformation dependent on the initial position and attitude of the vehicles. The projection of a point in the navigation frame is shown in Fig. 3. For a dynamic stereo rig with position $\vec{c}^{(N)}_{[k]}$ and attitude $R_{(k\to N)}$ at time k, the projection is:

$$u_{i[k],L} = \frac{\begin{bmatrix}1,0,0\\0,1,0\end{bmatrix} K_L R_{(N\to k)}\left(\vec{X}^{(N)}_i - \vec{c}^{(N)}_{[k]}\right)}{[0,0,1]K_L R_{(N\to k)}\left(\vec{X}^{(N)}_i - \vec{c}^{(N)}_{[k]}\right)} \tag{4}$$

$$u_{i[k],R} = \frac{\begin{bmatrix}1,0,0\\0,1,0\end{bmatrix} K_R\left(R_{(N\to k)}\left(\vec{X}^{(N)}_i - \vec{c}^{(N)}_{[k]}\right) - \vec{b}^{(C)}\right)}{[0,0,1]K_R\left(R_{(N\to k)}\left(\vec{X}^{(N)}_i - \vec{c}^{(N)}_{[k]}\right) - \vec{b}^{(C)}\right)}, \tag{5}$$

with $u_{i[k],L}$ and $u_{i[k],R}$ the coordinates in the left and right image respectively. For a stereo rig mounted on a vehicle constrained to be moving in a plane, the pose can be

parameterized by three parameters $\xi_{[k]}^{(N)} = [c_{[k],x}^{(N)}, c_{[k],y}^{(N)}, \phi_{[k]}^{(N)}]^T$ as

$$\vec{c}_{[k]}^{(N)} = \begin{bmatrix} c_{[k],x}^{(N)} \\ c_{[k],y}^{(N)} \\ 0 \end{bmatrix}, R_{(N\to k)} = \begin{bmatrix} \cos(\phi_{[k]}^{(N)}) & -\sin(\phi_{[k]}^{(N)}) & 0 \\ 0 & 0 & -1 \\ \sin(\phi_{[k]}^{(N)}) & \cos(\phi_{[k]}^{(N)}) & 0 \end{bmatrix}. \tag{6}$$

The planar position is $\vec{\beta}_k^{(N)} = [c_{[k],x}^{(N)}, c_{[k],y}^{(N)}]^T$. The reason for not denoting the poses with a two-dimensional group $\mathbb{SE}(2)$ is that even though the motion is constrained to be planar, the VSLAM problem still needs to handle 3D map points. Also, this model allows for a future extension of the proposed methods to 3D SLAM.

By stacking the measurements into a vector $u_{ik} = [u_{i[k],L}; u_{i[k],R}] \in \mathbb{R}^4$, a projection function $u_{ik} = \pi(\vec{X}_i^{(N)}, \xi_{[k]}^{(N)})$ can be defined for the point i and the vehicle pose at time k. The model of the corresponding noisy projective measurements is

$$\mu_{ik} = u_{ik} + n_{u,ik} \in \mathbb{R}^4, \tag{7}$$

with $E\{n_{u,ik}\} = \mathbf{0}, E\{n_{u,ik} n_{u,ik}^T\} = \Sigma_{u,ik}$. $E\{\cdot\}$ denotes the expected value function.

Using feature detectors, several feature points can be matched between the stereo images and tracked over frames for a period of time. To start the motion estimation, given a set of measurements $\{\mu_{i,1} : i = 1, ..., N_1\}$ and the initial pose estimate $\xi_{[1]}^{(N)}$, the 3D position of the point i can be obtained by stereo triangulation as $\hat{X}_i^{(N)} = \pi^{-1}(\mu_{i,1}, \xi_{[1]}^{(N)})$. Using N_k tracked features, the pose of the vehicle at time $k+1$ can be estimated by minimizing the reprojection error

$$\hat{\xi}_{[k+1]}^{(N)} = \arg\min_{\xi_{[k+1]}^N} \sum_{i=1}^{N_k} \left\| \mu_{i,k+1} - \pi(\hat{X}_i^{(N)}, \xi_{[k+1]}^{(N)}) \right\|_{\Sigma_{u,ik+1}^{-1}}^2, \tag{8}$$

where $\|\cdot\|_{\Sigma^{-1}}$ denotes the Mahalanobis distance in the metric given by the covariance matrix Σ. Using the estimated pose, the 3D position of the new features detected in frame $k+1$ can be updated using $\pi^{-1}(\cdot)$. As a result, the tracking can be continued as long as sufficient features can be tracked across consecutive frames.

Since the motion estimates are obtained with a dead-reckoning process, the estimation error will accumulate over time. In order to improve the accuracy of the estimation result, a global optimization for both 3D point position and the vehicle poses is performed using K keyframes and N_p map points:

$$\{\hat{\xi}_{[k]}^{(N)}\}, \{\hat{X}_i^{(N)}\} = \arg\min_{\{\xi_{[k]}^{(N)}, \vec{X}_i^{(N)}\}} \sum_{i=1}^{N_p} \sum_{k=1}^{K} F_{ik}(\xi_{[k]}^{(N)}, \vec{X}_i^{(N)}), \tag{9}$$

$$\text{with} \quad F_{ik} = v_{ik} \left\| \mu_{i,k} - \pi(X_i^{(N)}, \xi_{[k]}^{(N)}) \right\|_{\Sigma_{u,ik}^{-1}}^2, \tag{10}$$

where v_{ik} is a binary visibility mask, which assumes $v_{ik} = 1$ if feature i is visible to the camera at time instant k, otherwise $v_{ik} = 0$. This optimization is normally referred as bundle adjustment [13] in literatures.

Therefore, by executing the optimization in Eqn. (9), each rover obtains a set of egomotion estimates expressed in its own navigation frame, i.e., $\{\hat{\xi}_{[k]}^{(N_1)}\}$ and $\{\hat{\xi}_{[k]}^{(N_2)}\}$.

III. Cramér-Rao Bound for Planar Visual SLAM

Due to the presence of measurement noise, the accuracy of the estimated parameters is limited by a lower bound that depends on the noise level. The accuracy of an estimator can be evaluated by the Cramér-Rao lower bound (CRLB) [14]. It has been proved that for an unbiased estimator, the covariance of the estimated parameters is bounded by the inverse of its Fisher information matrix (FIM) I_ψ as

$$\text{cov}(\psi) \geq \text{CRLB}(\psi) = I_\psi^{-1}. \tag{11}$$

The Fisher information matrix is defined as

$$I_\psi = -E\left\{\nabla^2 \log\left(p(\mu|\psi)\right)\right\}, \tag{12}$$

where $\nabla^2 \log\left(p(\mu|\psi)\right)$ is the Hessian matrix of the function. μ and ψ are the measurements and the parameters to be estimated, respectively. In the stereo VSLAM problem outlined in Section II, the parameter vector is

$$\psi = \left[\vec{X}_1^{(N)}; ...; \vec{X}_{N_p}^{(N)}; \xi_{[1]}^{(N)}; ...; \xi_{[K]}^{(N)}\right] \in \mathbb{R}^{M\times 1}.$$

There are in total $M = 3N_p + 3K$ parameters in the vector ψ, with N_p the number of visual features used and K the total number of keyframes.

It is assumed that the outliers in feature tracking are already removed using outlier rejection schemes such as RANSAC [15], and the 2D feature location measurements of the inliers are multivariate Gaussian distributed variables.

Assuming all the 2D measurements are independent and identically distributed (i.i.d.), the log-likelihood function of all the measurements used to estimate the parameters is

$$\log\left(p(\mu|\psi)\right) = -\sum_{i=1}^{N_p}\sum_{k=1}^{K} \log(4\pi^2 \det(\Sigma_{u,ik})^{\frac{1}{2}}) - \frac{1}{2}\sum_{i=1}^{N_p}\sum_{k=1}^{K} v_{ik} \left\| \mu_{i,k} - \pi(X_i^{(N)}, \xi_{[k]}^{(N)}) \right\|_{\Sigma_{u,ik}^{-1}}^2. \tag{13}$$

with $\mu = \{\mu_{ik} | i = 1...N_p, k = 1...K\}$. As a result, for stereo VSLAM methods using maximum likelihood estimators, e.g., bundle adjustment, the parameter estimation accuracy is bounded by the diagonal terms of the inverse of the Fisher information matrix as

$$\text{var}(\psi_m) \geq (I_\psi^{-1})_{mm}.$$

IV. Tightly Coupled Cooperative Visual SLAM with a Ranging Link

The pose estimation in visual SLAM is purely based on dead reckoning methods, if the rovers do not revisit mapped places and detect loop closures. Consequently, the estimation error accumulates as the rover moves, and the obtained trajectory will drift away from the true one over time. By fusing the visual measurements with ranging measurements that are independently obtained, the drift can be mitigated since the ranging error does not accumulate

over time. Utilizing wireless radio, the range measurements can be obtained from pilot signals used for synchronization. Because a satisfactory clock synchronization between the two rovers cannot be achieved in most cases, round-trip-delay (RTD) techniques is a favorable choice to eliminate the impact of any clock offset. The details of ranging using RTD for slow-movement navigation purposes are discussed in [11]. For two cooperative rovers, a sparse set of noisy ranging measurements can be modeled as:

$$\rho_k = \left\| \vec{\beta}_{1,[k]}^{(W)} - \vec{\beta}_{2,[k]}^{(W)} \right\| + \eta_k. \tag{14}$$

As shown in Fig. 1, the initial position and attitude of the two rovers can be expressed in the reference frame as

$$\vec{\beta}_{1,[1]}^{(W)} = r_1 R(\alpha)[1,0]^T, \; R_{(N_1 \to W)} = R(\alpha + \theta - \frac{\pi}{2}). \tag{15}$$

$$\vec{\beta}_{2,[1]}^{(W)} = [0,0]^T, \quad R_{(N_2 \to W)} = I_2, \tag{16}$$

where r_1 is the true distance between the two rovers at time $k = 1$. I_2 denotes identity matrix, and $R(\cdot) \in \mathbf{SO}(2)$.

Using the images from the stereo camera rigs, the ego-motion of the two rovers in their navigation frames can be independently estimated as $\{\hat{\beta}_{1,[k]}^{(N_1)}\}$ and $\{\hat{\beta}_{2,[k]}^{(N_2)}\}$.

Using the method given in [12], the relative pose parameters $[\alpha, \theta, r_1]^T$ can be estimated by exploiting range measurements:

$$[\hat{\alpha}, \hat{\theta}, \hat{r}_1] = \arg\min_{\alpha,\theta,r_1} \|\rho - G(\alpha,\theta,r_1)\|^2_{Q^{-1}}, \quad \text{s.t. } r_1 > 0, \tag{17}$$

with vectors $\rho = [\rho_1, \rho_2, ..., \rho_K]^T$ and $G(\alpha,\theta,r_1) = [G_1, G_2, ..., G_K]^T$ with

$$G_k(\alpha,\theta,r_1) = \left\| R(\alpha + \theta - \frac{\pi}{2}) \vec{\beta}_{1,[k]}^{(N_1)} + r_1 R(\alpha)[1,0]^T - \vec{\beta}_{2,[k]}^{(N_2)} \right\|.$$

From the estimators in Eqn. (9) and Eqn. (17), we obtain $\{\hat{\xi}_{[k]}^{(N_1)}\}, \{\hat{\xi}_{[k]}^{(N_2)}\}$ and $[\hat{\alpha}, \hat{\theta}, \hat{r}_1]$, which can be regarded as initial coarse solutions of the rovers pose before the proper integration of both vision and ranging information.

Fig. 4 shows the Bayesian network of a tight coupling sensor fusion method exploiting both the visual and the ranging measurements. In order to optimize the overall pose graph, the two-rover system does not need to exchange any raw image or feature descriptor. As long as one of the rover can transmit the extracted 2D feature locations to the other, the poses of both rovers can be estimated in a tight coupling way using the visual features and ranging measurements from the radio link. Compared with algorithms based on map merging, this method requires much less data transmission in the communication. Applying the dependency among the random variables in the Bayesian network, the poses of both rovers can be obtained from the sensor fusion with the following maximum likelihood estimator:

$$\{\hat{\xi}_{1,[k]}^{(W)}, \hat{\xi}_{2,[k]}^{(W)}, \hat{X}_i^{(W)}\} = \arg\max \prod_{k=1}^{K} \prod_{i=1}^{N_p} p(\mu_{1i,k} | \pi(X_i^{(W)}, \xi_{1,[k]})$$
$$p(\mu_{2i,k} | \pi(X_i^{(W)}, \xi_{2,[k]}) p(\rho_k | \xi_{1,[k]}, \xi_{2,[k]}). \tag{18}$$

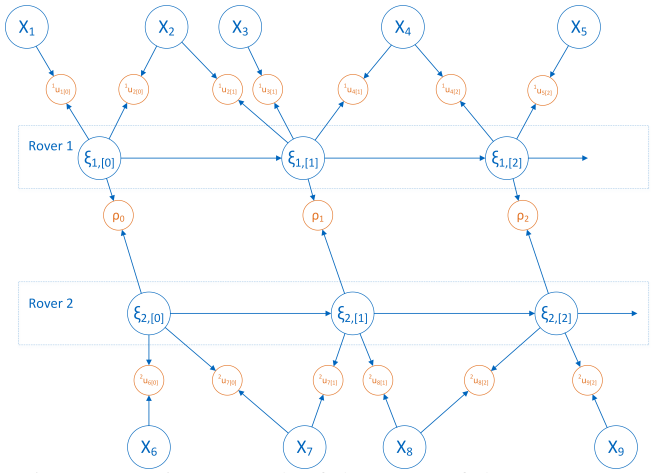

Fig. 4: Bayesian network of the states of the two rovers.

Under Gaussian noise assumption, the maximum linkelihood estimator can be transformed to an equivalent least squares (LS) estimator. Using the coarse estimates as initial value, the rovers' poses are obtained by solving the following LS estimation:

$$\{\hat{\xi}_{1,[k]}^{(W)}\}, \{\hat{\xi}_{2,[k]}^{(W)}\}, \{\hat{X}_i^{(W)}\} = \arg\min \sum_{k=1}^{K} \chi_k(\xi_{1,[k]}^{(W)}, \xi_{2,[k]}^{(W)})$$
$$+ \sum_{k=1}^{K} \sum_{i=1}^{N_p} (F_{ik}(\xi_{1,[k]}^{(W)}, \vec{X}_i^{(W)}) + F_{ik}(\xi_{2,[k]}^{(W)}, \vec{X}_i^{(W)})). \tag{19}$$

$F_{ik}(\cdot)$ is defined in Eqn. (9), and $\chi_k(\cdot)$ is defined as

$$\chi_k = w_k \left(\left\| \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \xi_{1,[k]}^{(W)} - \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \xi_{2,[k]}^{(W)} \right\| - \rho_k \right)^2, \tag{20}$$

where $w_k = (E\{\eta_k^2\})^{-1}$. The optimization problem can be solved using non-linear iterative solvers such as Levenberg-Marquart algorithm [16]. In practice, this process of batch optimization is burdened by a large computational complexity. As a feasible solution, advanced optimization algorihm such as iSAM2 [17] and [18] are used to reduce the complexity by exploiting the sparsity of the information matrix. Since the ranging measurements are sparse (the number of ranging measurements increases linearly with time), the computational complexity of the sensor fusion algorithm is almost the same as the vision-only optimization.

The proposed fusion algorithm does not require any common field-of-view for the two stereo rigs, making the proposed approach more flexible and efficient in exploration tasks.

Stacking all the measurements $\{\mu_{1,ik}\}$, $\{\mu_{2,ik}\}$ and $\{\rho_k\}$ into a vector $\lambda \in \mathbb{R}^{(2N_p+1)K}$, and all the parameters $\{\xi_{1,[k]}^{(W)}\}$, $\{\xi_{2,[k]}^{(W)}\}$, and $\{X_i^{(W)}\}$ into $\Theta \in \mathbb{R}^{3(N_p+K)}$, the log-likelihood function $\log(p(\lambda|\Theta))$ can be calculated using $F_{ik}(\cdot)$ and $\chi_k(\cdot)$ in Eqn. (19). The CRLB of the estimated parameters using the tight coupling sensor fusion algorithm is

$$CRLB(\Theta) = I_\Theta^{-1} = -\left(E\left\{\nabla^2 \log(p(\lambda|\Theta))\right\}\right)^{-1}. \tag{21}$$

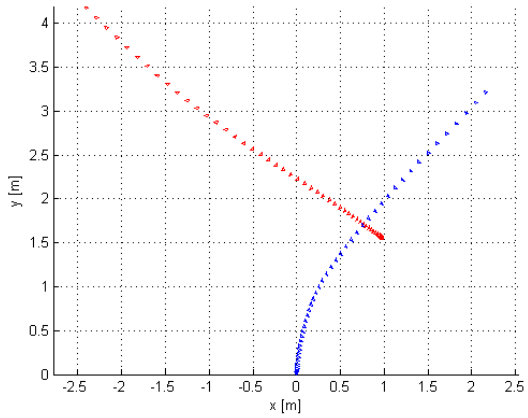

Fig. 5: First 50 keyframes of the trajectory

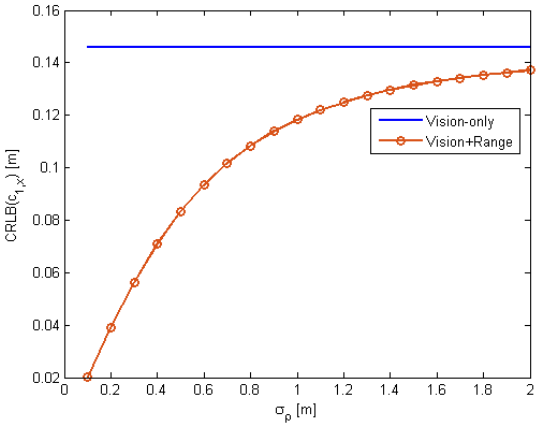

Fig. 6: Change of CRLB as function of σ_ρ, $\sigma_u = 0.1$ [pixel]

V. Simulation Results

The trajectories of two rovers, shown in Fig. 5, are generated to evaluate the proposed sensor fusion method in a simulated scenario. We set 10000 feature points distributed randomly in the 3D space. The stereo rigs' intrinsic parameters and sensor model are those of a real camera, a PointGrey Bumblebee2. The image sensor has a resolution of 1024*768 pixels, with pixel density ≈ 213.33 [pixels/mm]. The focal length of the lenses is 2.5 [mm]. The baseline length between the left and right camera is 12 [cm]. The 2D features are generated by using perspective projection as in Eqn. (2) and (3) with visibility check. White noise is added on both the 2D feature locations and the simulated range measurements.

Fig. 6 shows the CRLB as function of the ranging accuracy, represented by the standard deviation of the ranging noise. The y-axis is the CRLB for the x-component of the second rover's position. In the plot, the feature measurement noise is $\sigma_u = 0.1$ [pixel]. It can be inferred from the plot that when the ranging noise is small, the CRLB of the fusion-based method is much lower than the vision-only approach. When the ranging noise level is high, the accuracy of the fusion algorithm converges to the one of the vision-only method.

Fig. 7 illustrates the relation between the CRLB and the feature location accuracy. In this scenario, the ranging accuracy is fixed to 0.5 [m]. Since the baseline length of the stereo

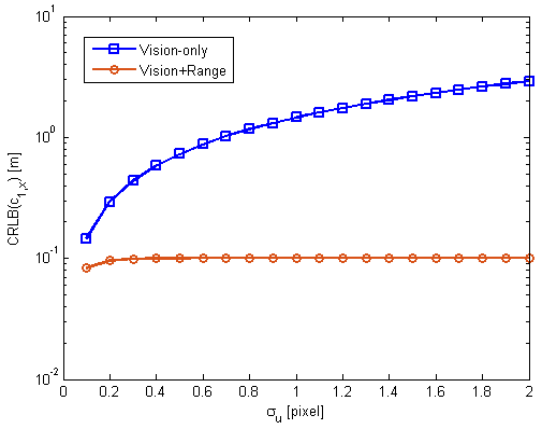

Fig. 7: Change of CRLB as function of σ_u, $\sigma_\rho = 0.5$ [m]

rig is only 12 [cm] and the resolution is not considerably high, the performance of the vision-only approach degrades significantly when the standard deviation characterizing the feature location inaccuracy exceeds 1 pixel. On the other hand, the bound for the fusion algorithm is much lower with the aid of the ranging measurements. Similar results are obtained for the other estimated parameters.

As another scenario with different geometries, Fig. 8 shows the trajectories of two stereo camera rigs mounted on rovers during a planar motion. The egomotion of the cameras can be estimated using frame-by-frame visual odometry. To improve the visual odometry coarse estimates, the rover poses and map point locations can be refined using global optimization, either with VSLAM-only approach, i.e., bundle adjustment, or with the proposed sensor fusion approach exploiting the ranging measurements. The performance of the methods are shown in Fig. 9 and Fig. 10. In these two plots, the uncertainty of the feature location is 1 pixel, and the standard deviation of the ranging noise is 0.9 [m]. The two figures shows the trajectory of rover 1. Fig. 9 is a zoomed-in plot for a few representative keyframes in the trajectory. The red triangles denote the ground truth of the camera poses. The magenta poses are the outcomes of the visual odometry, which are used as the initial values in the optimization. Due to the error accumulation, the magenta trajectory drifts gradually away from the true one. The green trajectory shows the estimation result of the camera-only bundle adjustment, while the blue one shows the sensor fusion outcome when using both visual and ranging measurements.

It can be seen from the plots that the drifts in visual odometry can be mitigated by both global optimization methods, but the sensor fusion algorithm outperforms the vision-only approach in accuracy. Similar conclusions can be drawn for larger ranging noise. Fig. 11 illustrates the estimated trajectories from both approaches with $\sigma_\rho = 1.7[m]$. The accuracy of the sensor fusion method is still slightly better.

Fig. 12 plots the root mean square error (RMSE) of the camera poses as function of the change of the ranging noise level. Since the latter does not affect the VSLAM algorithm, the error of the bundle adjustment approach remains mostly unchanged.

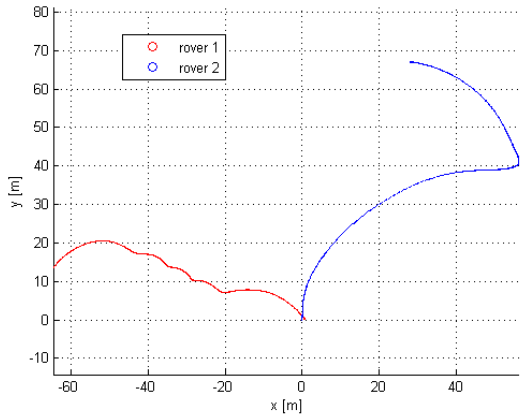

Fig. 8: The trajectories of the two rovers

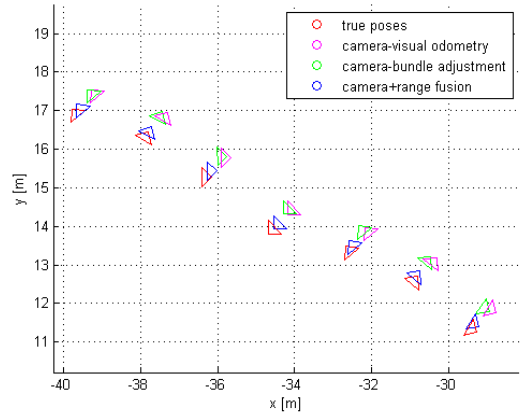

Fig. 9: A segment of the trajectory of rover 1 estimated using different methods, $\sigma_\rho = 0.9[m]$

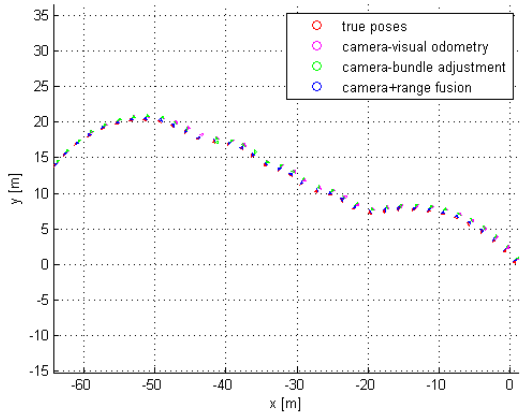

Fig. 10: The trajectory of rover 1 estimated using different methods, $\sigma_\rho = 0.9[m]$

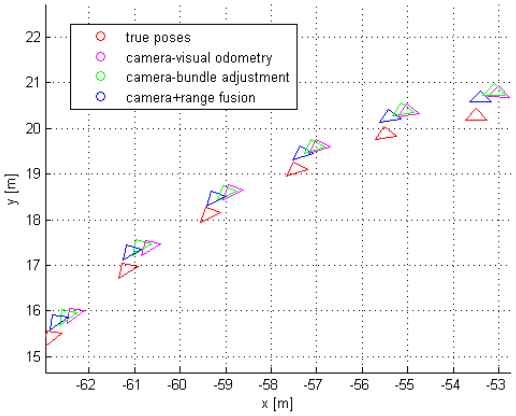

Fig. 11: The zoomed in trajectory of rover 1 estimated using different methods, $\sigma_\rho = 1.7[m]$

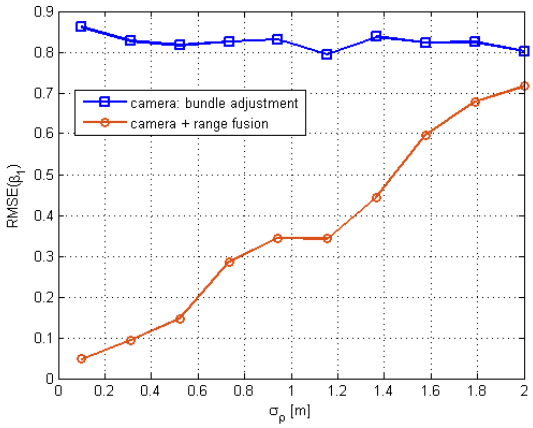

Fig. 12: The RMSE of the rover poses with respect to the ranging noise level

In conclusion, the sensor fusion approach significantly outperforms the vision-only method when the ranging noise is low. The performance of the proposed fusion method reduces to the one of classic VSLAM when the ranging measurement noise becomes very large (above meter level). These conclusions are further supported by the CRLB values shown in Fig. 6.

VI. CONCLUSION

In VSLAM-based exploration applications, using multiple cooperative rovers can improve the efficiency and robustness. We propose a tight coupling fusion algorithm to improve the SLAM accuracy by exploiting sparse range measurements between two rovers. The CRLB of the fusion approach is calculated and it is shown to outperform the vision-only method both theoretically (using the CRLB) and practically in various simulated scenarios.

REFERENCES

[1] M. Maimone, Y. Cheng, and L. Matthies, "Two years of visual odometry on the mars exploration rovers," *Journal of Field Robotics*, vol. 24, no. 3, pp. 169–186, 2007.

[2] R. Mur-Artal and J. D. Tardos, "Orb-slam2: an open-source slam system for monocular, stereo and rgb-d cameras," *arXiv preprint arXiv:1610.06475*, 2016.

[3] J. Engel, J. Stückler, and D. Cremers, "Large-scale direct slam with stereo cameras," in *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015, pp. 1935–1942.

[4] S. Sand, S. Zhang, M. Mühlegg, G. Falconi, C. Zhu, T. Krüger, and S. Nowak, "Swarm exploration and navigation on Mars," in *International Conference on Localization and GNSS, Torino, Italy*, 2013.

[5] D. Zou and P. Tan, "Coslam: Collaborative visual slam in dynamic environments," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 354–366, 2013.

[6] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-robot simultaneous localization and mapping: A review," *Journal of Field Robotics*, vol. 33, no. 1, pp. 3–46, 2016.

[7] R. Vincent, D. Fox, J. Ko, K. Konolige, B. Limketkai, B. Morisset, C. Ortiz, D. Schulz, and B. Stewart, "Distributed multirobot exploration, mapping, and task allocation," *Annals of Mathematics and Artificial Intelligence*, vol. 52, no. 2-4, pp. 229–255, 2008.

[8] C. Forster, S. Lynen, L. Kneip, and D. Scaramuzza, "Collaborative Monocular SLAM with Multiple Micro Aerial Vehicles," *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, vol. 143607, no. 200021, pp. 3963–3970, 2013.

[9] L. Carlone, M. K. Ng, J. Du, B. Bona, and M. Indri, "Rao-blackwellized particle filters multi robot SLAM with unknown initial correspondences and limited communication," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 243–249, 2010.

[10] O. De Silva, G. K. I. Mann, and R. G. Gosine, "Development of a relative localization scheme for ground-aerial multi-robot systems," *IEEE International Conference on Intelligent Robots and Systems*, pp. 870–875, 2012.

[11] E. Staudinger, S. Zhang, A. Dammann, and C. Zhu, "Towards a radio-based swarm navigation system on mars - key technologies and performance assessment," in *Wireless for Space and Extreme Environments (WiSEE), 2014 IEEE International Conference on*, Oct 2014, pp. 1–7.

[12] C. Zhu, G. Giorgi, and C. Günther, "Scale and 2d relative pose estimation of two rovers using monocular cameras and range measurements," in *Proceedings of the 29th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2016), Portland, Oregon*. Institute of Navigation, 2016, pp. 794–800.

[13] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustmenta modern synthesis," in *Vision algorithms: theory and practice*. Springer, 1999, pp. 298–372.

[14] C. R. Rao, "Advanced statistical methods in biometric research." 1952.

[15] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[16] J. J. Moré, "The levenberg-marquardt algorithm: implementation and theory," in *Numerical analysis*. Springer, 1978, pp. 105–116.

[17] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the Bayes tree," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 216–235, 2012. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0278364911430419

[18] R. Kümmerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "G2o: A general framework for graph optimization," in *Proceedings - IEEE International Conference on Robotics and Automation*, no. June, 2011, pp. 3607–3613.