Redesigning SLAM for Arbitrary Multi-Camera Systems

Juichung Kuo, Manasi Muglikar, Zichao Zhang, Davide Scaramuzza

***Abstract*— Adding more cameras to SLAM systems improves robustness and accuracy but complicates the design of the visual front-end significantly. Thus, most systems in the literature are tailored for specific camera configurations. In this work, we aim at an adaptive SLAM system that works for arbitrary multi-camera setups. To this end, we revisit several common building blocks in visual SLAM. These techniques make little assumption about the actual camera setups and prefer theoretically grounded methods over heuristics. We adapt a state-of-the-art visual-inertial odometry with these modifications, and experimental results show that the modified pipeline can adapt to a wide range of camera setups (e.g., 2 to 6 cameras in one experiment) without the need of sensor-specific modifications or tuning.**

Supplementary Material

Video: https://youtu.be/JGL4H93BiNw

I. Introduction

As an important building block in robotics, visual(-inertial) odometry (VO/VIO), or more general, simultaneous localization and mapping (SLAM) has received high research interest. Modern SLAM systems are able to estimate the local motion accurately as well as build a consistent map for other applications.

One of the remaining challenges for vision-based systems is the lack of robustness in challenging environments, such as high dynamic range (HDR) and motion blur [1]. Among different approaches that have been explored for better robustness (e.g., [2] [3]), adding more cameras in SLAM systems proves to be effective and is already exploited in successful commercial products, such as Oculus Quest [4] and Skydio [5].

As the workhorse for modern (keyframe-based) SLAM systems, bundle adjustment (BA) like nonlinear optimization naturally generalizes to multiple sensors, including visual-inertial and multi-camera systems, as long as the measurement process is modeled correctly. The design of the so-called front-ends is much less theoretically grounded. Many details, such as initialization, keyframe selection, and map management, are designed heuristically. Moreover, such designs are often tailored to specific sensor setups, and it is not clear to what extent they can be applied to more general sensor configurations. For example, one popular method for selecting keyframes is to consider commonly visible features in the current frame with respect to the last keyframe. While this works well for monocular setups or stereo pairs with highly overlapping field-of-views (FoV), it quickly becomes complicated as more cameras are added, as different cameras may have drastically different view conditions (e.g., the number of features).

The authors are with the Robotics and Perception Group, Dep. of Informatics, University of Zurich, and Dep. of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland— http://rpg.ifi.uzh.ch.

Fig. 1: Multi-camera systems achieve superior performance in perception algorithms and are widely used in real-world applications, such as omni-directional mapping [6], autonomous drones [5], and VR headsets [4]. To facilitate the use of such systems in SLAM, our method adapts to arbitray multi-camera systems automatically.

To remove the dependence on sensor-specific assumptions and heuristics, we resort to adaptive and more principled solutions. First, instead of using hard-coded rules, we propose an adaptive initialization scheme that analyzes the geometric relation among all cameras and selects the most suitable initialization method online. Second, instead of engineering heuristics, we choose to characterize the uncertainty of the current pose estimate with respect to the local map using the information from all cameras, and use it as an indicator of the need for a new keyframe. Third, instead of relying on the covisiblity graph, we organize all the landmarks in a voxel grid and sample the camera frustums via an efficient voxel hashing algorithm, which directly gives the landmarks within the FoVs of the cameras. The contribution of this work is an adaptive design for general multi-camera VO/VIO/SLAM systems, including

- an adaptive initialization scheme,
- a sensor-agnostic, information-theoretic keyframe selection algorithm,
- a scalable, voxel-based map management method.

Since the proposed method is not limited to specific implementations or sensing modalities, we will use the term SLAM in general for the rest of the paper.

II. Adaptive Initialization

For any multi-camera setups with known intrinsic and extrinsic calibrations, our method is able to select the proper initialization method accordingly, without the need to change the algorithm settings manually. Specifically, it utilizes an overlapping check between the camera frustums to identify all the possible stereo camera pairs. If there exists stereo pairs, the initial 3D points are created from the stereo matching of these stereo pairs. Otherwise, the 5-point algorithm is run on every camera as in a standard monocular setup,

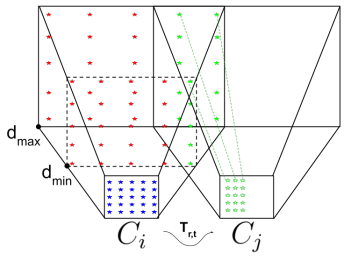

Fig. 2: An illustration of the stereo overlapping check between two cameras, C_i and C_j. The blue stars are the sampled points on the image plane of camera i. The green stars are the 3D points that are successfully projected to camera j, and the red ones are the points that fall out of the image plane.

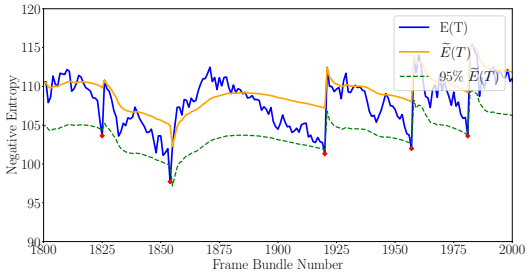

Fig. 3: Running average $\tilde{E}(\mathtt{T})$ and keyframe selection. The running average filter (yellow) tracks the localization quality since the last keyframe. When the negative entropy of the current frame (blue) falls below a certain percentage of the running average (green dash), a new keyframe is selected (red dots) and the running average filter is reset.

and the map is initialized whenever there exists a camera that triangulates the initial map successfully (i.e., enough parallax, and the camera is not undergoing strong rotation).

The core part of the aforementioned initialization scheme is the overlapping check. The overlapping checking algorithm checks all the possible pairs in a multi-camera configuration, denoted as C_{ij}, where $i, j \in 1 \dots n,\ i \neq j$, and n is the total number of cameras in the system, and finds all possible stereo pairs. For each pair C_{ij}, the algorithm is illustrated in Fig. 2. A pair of cameras is considered as a stereo pair if the overlapping ratio, $\frac{\text{\# of Successful Projection}}{\text{\# of Total Samples}}$, is above a user-defined threshold.

The proposed sampling-based method is generic. By using the camera projection/backprojection directly, we can find all stereo pairs across different types of camera models without the need to explicitly calculate the overlapping volume of possibly very different frustums (e.g., between pinhole and fisheye cameras), which can be non-trivial to compute analytically. Moreover, the check can be computed offline, and the valid stereo pairs be directly used at runtime.

III. Entropy-based Keyframe Selection

The concept of keyframe naturally generalizes to a keyframe bundle for a multi-camera setup, as in [7]. A keyframe bundle contains the frames from all the cameras at the same time. In the following, we will use the terms keyframe and keyframe bundle interchangeably. To determine when a keyframe should be added, we design an entropy-based mechanism. In particular, the local map contains 3D points (organized as keyframes or voxels as in Section IV) against which new frames can localize. Intuitively, a keyframe should be selected when the current map is not sufficient for tracking, since new points will be initialized at the insertion of a keyframe. We select keyframes based on the uncertainty of the keyframe bundle pose with respect to the current map. Compared with heuristics, our method is more principled, has less parameters (*only 1*) and generalizes to arbitrary camera configurations. In this section, we first provide necessary background on the uncertainties in estimation problems and then describe our keyframe selection method.

A. Uncertainties Estimation in Nonlinear Least Squares

For a parameter estimation problem of estimating $\mathbf{x}$ from measurement $\mathbf{z}$ with normally distributed noise, a common method is to cast the problem as a nonlinear least squares (NLLS) problem. In iterative algorithms of solving NLLS problems, such as Gauss-Newton, the uncertainties of the estimated parameters can be obtained as a side product in each iterative step. Specifically, the normal equation at step i is $(\mathtt{J}^\top \Sigma_{\mathbf{z}}^{-1} \mathtt{J}) \delta \mathbf{x}_i = \mathtt{J}^\top \mathbf{r}(\mathbf{x}_i)$, where $\mathbf{r}(\mathbf{x}_i)$ is the residual given the current estimate $\mathbf{x}_i$, $\delta \mathbf{x}_i$ the optimal update, and $\mathtt{J}$ the Jacobian of $\mathbf{z}$ with respect to $\mathbf{x}$. With first-order approximation, the covariance of the estimate can be obtained by backward propagating the measurement noise to the parameters, which is simply:

$$\Sigma_{\mathbf{x}} = (\mathtt{J}^\top \Sigma_{\mathbf{z}}^{-1} \mathtt{J})^{-1}, \tag{1}$$

which is an important tool to quantify the estimation quality of NLLS solutions [8, Chapter 5, App. 3]. $\mathtt{I}_{\mathbf{x}} = \mathtt{J}^\top \Sigma_{\mathbf{z}}^{-1} \mathtt{J}$ is also known as the Fisher information.

B. Negative Pose Entropy in SLAM

In keyframe-based SLAM, the pose of the current camera is usually obtained by solving a NLLS problem. For example, one common method is to solve a Perspective-n-Points (PnP) problem using the Gauss-Newton method. In this case, the Fisher information and the covariance of the camera pose can be directly obtained as

$$\mathtt{I}_{\mathtt{T}} = \mathtt{J}_{\mathtt{T}}^\top \Sigma_{\mathbf{u}}^{-1} \mathtt{J}_{\mathtt{T}},\ \Sigma_{\mathtt{T}} = (\mathtt{J}_{\mathtt{T}}^\top \Sigma_{\mathbf{u}}^{-1} \mathtt{J}_{\mathtt{T}})^{-1}, \tag{2}$$

where $\mathbf{u}$ is the observation, and $\mathtt{J}_{\mathtt{T}}$ is Jacobian of $\mathbf{u}$ with respect to the camera pose $\mathtt{T}$. [1] Note that in different NLLS problems, the Fisher information and covariance may be obtained differently (e.g., marginalization in a BA setup).

While (2) provides a principled tool, it is more desirable to have a scalar metric as keyframe selection criteria. Therefore, we utilize the concept of the differential entropy for a multivariate Gaussian distribution, which is $H(\mathbf{x}) = \frac{1}{2} m(1 + \ln(2\pi)) + \frac{1}{2} \ln(|\Sigma|)$ for a m-dimensional distribution with covariance Σ. Note that the magnitude of the entropy only depends on $\ln(|\Sigma|)$. Moreover, in the context of NLLS for pose estimation, from (2), we have $\ln(|\Sigma_{\mathtt{T}}|) = -\ln(|\mathtt{I}_{\mathtt{T}}|)$. Since that the Fisher information $\mathtt{I}_{\mathtt{T}}$ comes for free in the process of solving NLLS problems, we can actually avoid

[1] Technically, the Jacobian is with respect to a minimal parameterization of 6 DoF poses, which is omitted here for easy presentation.

Algorithm 1: Running average filter.

Input: newest entropy value $E(\mathtt{T})$
Result: Returns the current running average, $\widetilde{E}(T)$
initialization: $n = 0,\ \widetilde{E}(T) = 0$
for *each incoming* $E(T)$ **do**
  $n = n + 1$
  $\widetilde{E}(T) = \widetilde{E}(T) + (E(T) - \widetilde{E}(T))/n$
  $return\ \widetilde{E}(T)$

the matrix inversion and use

$$E(\mathtt{T}) \triangleq \ln(|\mathtt{I}_{\mathtt{T}}|) \tag{3}$$

to indicate how well the camera can localize in the current map. We refer to (3) as *negative entropy*. Since (2) is simply the sum of individual measurements, it is straightforward to incorporate the observations from all the cameras into one single scalar (3) in an arbitrary multi-camera setup.

C. Running Average Filter for Keyframe Selection

Using an absolute threshold for $E(\mathtt{T})$ as the keyframe selection criterion is not feasible as this absolute values may change for every run in the same environment.

Instead, we propose to track the negative entropy value using a running average filter (see Algo. 1) in the local map and selects a keyframe when $E(\mathtt{T})$ of a frame is below certain percentage of the tracked average $\tilde{E}(\mathtt{T})$. Since we localize the camera with respect to the latest map, and the map remains the same until a new keyframe is added, $\tilde{E}(\mathtt{T})$ essentially tracks the average pose estimation quality with respect to the local map up to the current time. Note that the running average filter is reinitialized every time the map is updated with a new keyframe, since the local map changes as a new keyframe is inserted. Moreover, we use a relative threshold with respect to the running average $\tilde{E}(\mathtt{T})$ so that the selection is adaptive to different environments. This threshold is the only parameter in our keyframe selection method, and it is intuitive to tune. A higher value means more frequent keyframe insertion, and vice versa (see Table II). An example of the running average filter is shown in Fig. 3.

IV. Voxel-Map Query

For new incoming images, the tracking process in SLAM is responsible to find the correspondences between the observations in the new images and the 3D points in the map. Traditionally this is done by searching for matches in the keyframes that overlap with the new frames. This introduces redundancy for general multi-camera setup. Therefore, we organize the map points in a voxel grid, and directly sample the camera frustums for possible 3D points to match, as proposed in [9].

Map query: To get the map points to match for a multi-camera system, we sample a fixed number of points in the camera frustums and find corresponding voxels in the voxel-map. The points inside these voxels are then used to match the observations in the new images. In this way, it is guaranteed that *all and only* the 3D points within the FoVs of all cameras are retrieved from the map. Moreover, we avoid the process of checking overlapping keyframes from different cameras, which may have many points in common and introduce redundant computation.

Note that we only use voxel-map for querying visible landmarks. Keyframes are still selected for triangulation and potentially bundle adjustment.

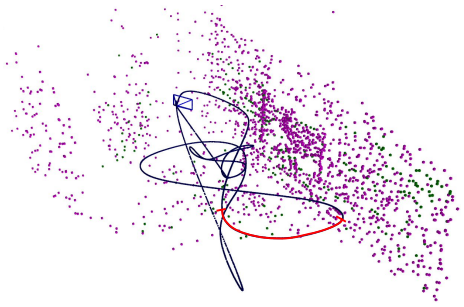

Fig. 4: Simulated figure 8 trajectory in the simulation environment. The trajectory was estimated by running the adapted VIO pipeline with 5 cameras. The segment where the monocular setup lost track is marked in red. The magenta dots are the tracked landmarks by SLAM systems.

V. Experiments

To validate the proposed method, we applied the aforementioned adaptations to a state-of-the-art keyframe-based visual-inertial odometry pipeline that consists of an efficient visual front-end [10] and an optimization-based backend similar to [11]. For real-world data, we tested the stereo setup with the EuRoC dataset [12] to show that the proposed method performs on par with standard methods but is much easier to tune. For quantitative evaluation of accuracy, we follow the evaluation protocol in [13]. We repeated the experiment on each sequence for 5 runs using the same setting unless specified otherwise. In each of the experiment, we kept the parameters *the same* for different camera configurations.

A. Real-world Experiment

1) EuRoC Dataset: We tested the multi-camera VIO pipeline on EuRoC dataset for the stereo setup. The number of keyframes in the sliding window was set to 10. To show the effect of the relative negative entropy for keyframe selection, we also experimented with different relative entropy thresholds.

The median values of the absolute trajectory error in 5 runs are shown in Table I. While there is no definite winner, the adapted pipeline in general performed similar or better than the default pipeline. This can also be confirmed from the odometry errors in Fig. 5 (we select three sequences only due to the limit of space). The adapted pipeline has lower estimate error in 10 out of 11 sequences and the entropy ratio of 98% has the most.

To summarize, as a generic pipeline, our method performed at least similarly good compared with a carefully tuned stereo pipeline, and our method was able to achieve similar accuracy with fewer keyframes. More importantly, we would like to emphasize that our method has *only one* parameter for keyframe selection, which makes the task of parameter tuning much easier.

Sensor Agnostic We also performed an experiment comparing the number of selected keyframes between monocular

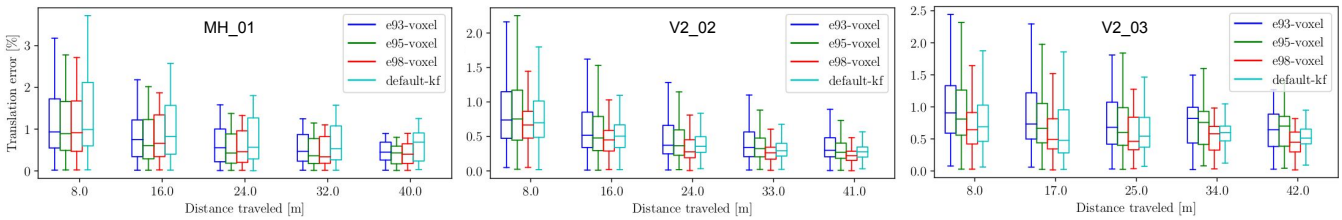

Fig. 5: Relative translation error percentages from the EuRoC dataset with BA.

TABLE I: Median RMSE (meter) on EuRoC dataset over 5 runs. Lowest error highlighted in **bold**.

Algorithm	MH_01	MH_02	MH_03	MH_04	MH_05	V1_01	V1_02	V1_03	V2_01	V2_02	V2_03
default-kf	0.140	0.078	0.091	**0.119**	0.330	0.042	0.070	0.047	0.056	0.066	0.127
e93-voxel	0.104	0.390	0.107	0.177	0.262	**0.038**	**0.036**	0.043	0.080	0.103	0.169
e95-voxel	**0.078**	0.084	0.093	0.182	0.237	0.040	0.047	0.049	0.056	0.087	0.171
e98-voxel	0.095	**0.074**	**0.088**	0.128	**0.180**	0.039	0.053	**0.041**	**0.046**	**0.057**	**0.111**

TABLE II: Average number of keyframes for 5 runs in EuRoC sequences.

Algorithm	MH_01	MH_02	MH_03	MH_04	MH_05	V1_01	V1_02	V1_03	V2_01	V2_02	V2_03
default-kf	64.00	57.80	91.40	76.00	70.20	70.60	119.60	238.80	74.80	172.00	281.40
e93-voxel	46.00	46.30	67.80	58.80	61.80	52.80	56.20	120.40	30.80	63.40	86.80
e95-voxel	71.20	66.00	87.00	74.40	75.80	76.40	86.80	160.00	39.80	85.00	107.80
e98-voxel	154.20	137.20	181.20	138.60	143.60	176.80	177.80	305.20	84.40	169.00	203.60

TABLE III: The average number of keyframes by different keyframe selection criteria for monocular and stereo setups.

Algorithm	MH_01	MH_02	V2_01	V2_02
heuristic, mono	202.75	190.75	150.75	379.75
heuristic, stereo	90.00	117.25	84.75	204.5
entropy, mono	129.5	128.25	100.00	193.5
entropy, stereo	122.25	125.25	98.5	195

and stereo configurations. We only ran the visual front-end in this case to remove the influence of the optimization back-end, which caused the different keyframe numbers between Table II and III. The average number of keyframes on some sequences in EuRoC is shown in Table III. The heuristic method selected drastically different numbers of keyframes between monocular and stereo configurations because they had to be tuned differently for these configurations. In contrast, our entropy-based method selected very similar numbers of keyframes. This is due to fact that our method essentially summarizes the information in the map instead of relying on camera-dependent quantities. In particular, the stereo pair in EuRoC dataset has largely overlapping FoVs, and thus the visible areas of the environment were similar for monocular and stereo setups, leading to similar information for our keyframe selection method.

VI. Conclusion

In this work, we introduced several novel designs for common building blocks in SLAM to make an adaptive system for arbitrary camera configurations. In particular, we proposed an adaptive initialization scheme that is able to automatically find the suitable initialization method, an information-theoretic keyframe selection method that incorporates the information from all cameras elegantly and a voxel-map representation from which we can directly retrieve the landmarks in the camera FoVs. We applied these techniques to a state-of-the-art VIO pipeline, and extensive experimental results showed that the resulting pipeline was able to adapt to various camera configurations with minimum parameter tuning.

References

[1] C. Cadena, L. Carlone, H. Carrillo, Y. Latif, D. Scaramuzza, J. Neira, I. D. Reid, and J. J. Leonard, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, 2016. 1

[2] Z. Zhang, C. Forster, and D. Scaramuzza, "Active exposure control for robust visual odometry in hdr environments," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2017. 1

[3] A. Rosinol Vidal, H. Rebecq, T. Horstschaefer, and D. Scaramuzza, "Ultimate SLAM? combining events, images, and IMU for robust visual SLAM in HDR and high speed scenarios," *IEEE Robot. Autom. Lett.*, vol. 3, no. 2, pp. 994–1001, Apr. 2018. 1

[4] "Oculus Quest," https://www.oculus.com/quest/. 1

[5] "Skydio R1," https://robots.ieee.org/robots/skydior1/. 1

[6] D. Anguelov, C. Dulong, D. Filip, C. Frueh, S. Lafon, R. Lyon, A. Ogale, L. Vincent, and J. Weaver, "Google street view: Capturing the world at street level," *Computer*, 2010. 1

[7] A. Harmat, I. Sharf, and M. Trentini, "Parallel tracking and mapping with multiple cameras on an unmanned aerial vehicle," in *Intelligent Robotics and Applications*, 2012. 2

[8] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003, 2nd Edition. 2

[9] M. Muglikar, Z. Zhang, and D. Scaramuzza, "Voxel map for visual slam," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2020. 3

[10] C. Forster, Z. Zhang, M. Gassner, M. Werlberger, and D. Scaramuzza, "SVO: Semidirect visual odometry for monocular and multicamera systems," *IEEE Trans. Robot.*, vol. 33, no. 2, pp. 249–265, 2017. 3

[11] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial SLAM using nonlinear optimization," *Int. J. Robot. Research*, 2015. 3

[12] M. Burri, J. Nikolic, P. Gohl, T. Schneider, J. Rehder, S. Omari, M. W. Achtelik, and R. Siegwart, "The EuRoC micro aerial vehicle datasets," *Int. J. Robot. Research*, vol. 35, pp. 1157–1163, 2015. 3

[13] Z. Zhang and D. Scaramuzza, "A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2018. 3