C-KLAM: Constrained Keyframe Localization and Mapping for Long-Term Navigation

Esha D. Nerurkar[†], Kejian J. Wu[‡], and Stergios I. Roumeliotis[†]

***Abstract*— In this paper, we present C-KLAM, a Maximum A Posteriori (MAP) estimator-based keyframe approach for SLAM. As opposed to many existing keyframe-based SLAM approaches, that discard information from non-keyframes in order to reduce the computational complexity, the proposed C-KLAM presents a novel and computationally-efficient technique for incorporating most of this information, resulting in improved estimation accuracy. Specifically, C-KLAM projects information from the non-keyframes to the keyframes, using marginalization, while *maintaining the sparse structure of the information matrix*, to generate fast and efficient solutions. The performance of C-KLAM has been tested in both simulations and experimentally, using visual and inertial measurements, to demonstrate that it achieves performance comparable to that of the computationally-intensive batch MAP-based 3D SLAM that uses all available measurement information.**

I. INTRODUCTION AND RELATED WORK

For mobile robots navigating in large environments over long time periods, one of the main challenges in designing an estimation algorithm for Simultaneous Localization and Mapping (SLAM) is its inherently high computational complexity. For example, the computational complexity of the Minimum Mean Squared Error (MMSE) estimator for SLAM, i.e., the Extended Kalman filter [1], is $O(N^2)$ at each time step, where N is the number of landmarks in the map. Similarly, for the batch Maximum A Posteriori (MAP) estimator-based SLAM (smoothing and mapping) [2], the worst-case computational complexity is $O([K+N]^3)$, where K is the number of robot poses in the trajectory. While existing batch MAP-based SLAM approaches such as the $\sqrt{SAM}$ [2], g^2o [3], and SPA [4] generate efficient solutions by exploiting the sparsity of the information matrix, for large-scale SLAM with frequent loop closures, this cost eventually prohibits real-time operation.

The approximate solutions developed to reduce MAP-based SLAM's computational complexity can be classified into three main categories. The first category of approaches such as iSAM [5] and iSAM2 [6] *incrementally* optimize over all robot poses and landmarks, using *all* available measurement information. However, for trajectories with frequent loop closures, (i) fill-ins are generated between periodic batch updates for iSAM, when the number of constraints is greater than five times the number of robot poses [5], and (ii) many nodes in the Bayes tree used by iSAM2 have to be relinearized, hence degrading the performance of these approaches.

The second category includes fixed-lag smoothing approaches such as [7], [8] that consider a constant-size, sliding-window of recent robot poses and landmarks, along with measurements only in that time window. Here, old robot poses and landmarks are *marginalized* and the corresponding measurements are discarded. However, marginalization destroys the sparsity of the information matrix, and the cost of this approach becomes $O(R^3)$, hence limiting the number of poses, R, in the sliding window. Moreover, this approach is unable to close loops for long trajectories.

The third category consists of *keyframe*-based approaches, such as PTAM [9], FrameSLAM [10], and view-based maps (pose graphs) [11], [12], [13] that process measurement information from only a *subset* of all available views/keyframes/robot poses. Here, information from non-keyframes is *discarded* (as opposed to marginalized) in order to retain the sparsity of the information matrix, hence trading estimation accuracy for reduced computational cost.

In this paper, we present the batch MAP-based Constrained Keyframe Localization and Mapping (C-KLAM), which estimates the keyframes along with the positions of landmarks observed from these keyframes. Importantly, information from non-keyframes, acquired between keyframes, is not discarded. Instead, this information is projected on to the keyframes, in order to generate tight constraints between them. Our main contributions are as follows:

- In contrast to existing keyframe methods, C-KLAM utilizes *all* available measurement information, both proprioceptive (e.g., IMU) and exteroceptive (e.g., camera), from non-keyframes to generate tight constraints between the keyframes. This is achieved by marginalizing the non-keyframes along with the landmarks observed from them.
- In contrast to sliding-window approaches, C-KLAM incorporates information from marginalized frames and landmarks *without* destroying the sparsity of the information matrix, and hence generates fast and efficient solutions.
- The cost of marginalization in C-KLAM is cubic, $O(M^3)$, only in the number of non-keyframes, M, between consecutive keyframes, and *linear* in the number of landmarks, F_M, observed exclusively from the M non-keyframes, where $M \ll F_M$.
- The keyframes' poses and the associated landmark-map are maintained over the entire robot trajectory, and thus C-KLAM enables efficient loop closures, necessary for

[†]E. D. Nerurkar, and S. I. Roumeliotis are with the Department of Computer Science and Engineering, Univ. of Minnesota, Minneapolis, USA {nerurkar,stergios}@cs.umn.edu

[‡]K. J. Wu is with the Department of Electrical and Computer Engineering, Univ. of Minnesota, Minneapolis, USA kejian@cs.umn.edu

This work was supported by the University of Minnesota (UMN) through the Digital Technology Center (DTC). E. D. Nerurkar was supported by the UMN Doctoral Dissertation Fellowship. The authors thank Chao X. Guo and Dimitrios G. Kottas for their help with the experimental dataset.

ensuring consistent long-term navigation.

II. ALGORITHM DESCRIPTION

We now present a brief overview of our proposed C-KLAM approach. Consider the current exploration epoch shown in Fig. 1(a). For $i \in \{K_1, M, K_2\}$, and $j \in \{K, B, M\}$, let $\mathbf{z}_{i,j}$ denote the set of exteroceptive measurements from poses in $\mathbf{x}_i$ to landmarks in $\mathbf{f}_j$, let $\mathbf{u}_i$ denote the proprioceptive measurements that relate poses in $\mathbf{x}_i$, and let $\mathbf{u}_{K_1,M}$, $\mathbf{u}_{M,K_2}$ (red dotted arrows in Fig. 1(a)) denote the proprioceptive measurements that relate the last pose in $\mathbf{x}_{K_1}$, $\mathbf{x}_M$ with the first pose in $\mathbf{x}_M$, $\mathbf{x}_{K_2}$, respectively. Lastly, let $\mathbf{x}_{K_{12}}$ denote the last pose in $\mathbf{x}_{K_1}$ and the first pose in $\mathbf{x}_{K_2}$. The batch MAP cost function, $\mathbb{C}$, associated with the current exploration epoch (see Fig. 1(a)) is given by:

$$\begin{aligned}&\mathbb{C}(\mathbf{x}_{K_1}, \mathbf{x}_M, \mathbf{x}_{K_2}, \mathbf{f}_K, \mathbf{f}_B, \mathbf{f}_M; \mathbf{z}_{K_1,K}, \mathbf{z}_{K_1,B}, \mathbf{z}_{K_2,K}, \mathbf{z}_{K_2,B},\\ &\quad \mathbf{z}_{M,B}, \mathbf{z}_{M,M}, \mathbf{u}_{K_1}, \mathbf{u}_M, \mathbf{u}_{K_2}, \mathbf{u}_{K_1,M}, \mathbf{u}_{M,K_2})\\ &= \mathbb{C}_1(\mathbf{x}_{K_1}, \mathbf{x}_{K_2}, \mathbf{f}_K, \mathbf{f}_B; \mathbf{z}_{K_1,K}, \mathbf{z}_{K_1,B}, \mathbf{z}_{K_2,K}, \mathbf{z}_{K_2,B}, \mathbf{u}_{K_1}, \mathbf{u}_{K_2})\\ &+ \mathbb{C}_2(\mathbf{x}_M, \mathbf{x}_{K_{12}}, \mathbf{f}_B, \mathbf{f}_M; \mathbf{z}_{M,B}, \mathbf{z}_{M,M}, \mathbf{u}_{K_1,M}, \mathbf{u}_{M,K_2}, \mathbf{u}_M) \quad (1)\end{aligned}$$

In (1), $\mathbb{C}_2$ is the part of the cost function corresponding to measurements that *involve* the non-key poses $\mathbf{x}_M$ and landmarks $\mathbf{f}_M$ (denoted in red in Fig. 1(a)), while $\mathbb{C}_1$ corresponds to measurements that *do not involve* $\mathbf{x}_M$ and $\mathbf{f}_M$ (depicted in green in Fig. 1(a)).

In order to reduce the computational complexity of MAP-based SLAM and ensure real-time operation over long time durations, our objective is to maintain only a few key poses, representative of the current epoch, and use information from non-key poses to provide constraints between the key poses (denoted by blue arrow in Fig. 1(b)). In C-KLAM, this is achieved by: (i) marginalizing the non-key poses, $\mathbf{x}_M$, and the landmarks, $\mathbf{f}_M$, observed exclusively from $\mathbf{x}_M$, (ii) projecting this information on the key poses, $\mathbf{x}_{K_{12}}$, and (iii) then discarding all measurements (denoted with red arrows in Fig. 1(a)) involving the marginalized states. In terms of the batch MAP estimation problem, this is equivalent to approximating the part $\mathbb{C}_2$ of the nonlinear least-squares batch MAP cost function, $\mathbb{C}$, in (1), using the second order Taylor series, as follows:

$$\begin{aligned}&\mathbb{C}_2(\mathbf{x}_M, \mathbf{x}_{K_{12}}, \mathbf{f}_B, \mathbf{f}_M; \mathbf{z}_{M,B}, \mathbf{z}_{M,M}, \mathbf{u}_{K_1,M}, \mathbf{u}_{M,K_2}, \mathbf{u}_M) \quad (2)\\ &\approx \alpha + \mathbf{b}_d^T(\mathbf{x}_{K_{12}} - \hat{\mathbf{x}}_{K_{12}}) + \frac{1}{2}(\mathbf{x}_{K_{12}} - \hat{\mathbf{x}}_{K_{12}})^T \mathbf{A}_d (\mathbf{x}_{K_{12}} - \hat{\mathbf{x}}_{K_{12}})\end{aligned}$$

Here, $\hat{\mathbf{x}}_{K_{12}}$ denote the best estimates available for $\mathbf{x}_{K_{12}}$ at the time of marginalization. α is independent of $\mathbf{x}_{K_{12}}$, and $\mathbf{b}_d$ and $\mathbf{A}_d$ are the Jacobian and Hessian, respectively. Importantly, we note that C-KLAM uses information from the discarded measurements to generate constraints *only* between consecutive key poses, $\mathbf{x}_{K_{12}}$, hence maintaining the sparsity of the information matrix. Moreover, note that even though the measurements from non-key poses, $\mathbf{x}_M$, to landmarks in $\mathbf{f}_B$ are discarded, after incorporating their information in (2), the landmarks $\mathbf{f}_B$ themselves are not marginalized, since they appear in $\mathbb{C}_1$ and remain in the state to provide a sparse representation of the environment.

Next, we show that the marginalization described above can be carried out with cost $O(|\mathbf{x}_M|^3)$, where $|\mathbf{x}_M|$ denotes the cardinality of $\mathbf{x}_M$. The structure of the Hessian, $\mathbf{H}$, corresponding to $\mathbb{C}_2$ in (2), is shown in Fig. 2. Based on $\mathbf{H}$, $\mathbf{A}_d$ is calculated as:

$$\begin{aligned}\mathbf{A}_d &= \mathbf{A}_K - \mathbf{B}_K(\mathbf{D} - \mathbf{B}_B^T \mathbf{A}_B^{-1} \mathbf{B}_B)^{-1} \mathbf{B}_K^T\\ \text{where} \quad \mathbf{D} &= \mathbf{A}_M - \mathbf{A}_{M f_M} \mathbf{A}_{f_M}^{-1} \mathbf{A}_{f_M M} \quad (3)\end{aligned}$$

In (3), note that both $\mathbf{A}_B$ and $\mathbf{A}_{f_M}$ are block diagonal and hence their inverses can be calculated with linear cost, $O(|\mathbf{f}_B|)$ and $O(|\mathbf{f}_M|)$, respectively. The most computationally-intensive calculation in (3) is that of $(\mathbf{D} - \mathbf{B}_B^T \mathbf{A}_B^{-1} \mathbf{B}_B)^{-1}$, which is cubic, $O(|\mathbf{x}_M|^3)$, in the number of non-key poses currently being marginalized. Since this size is bounded, the marginalization in C-KLAM can be carried out with minimal computational overhead. Note that the analysis for the Jacobian $\mathbf{b}_d$ is similar and is omitted due to space constraints.

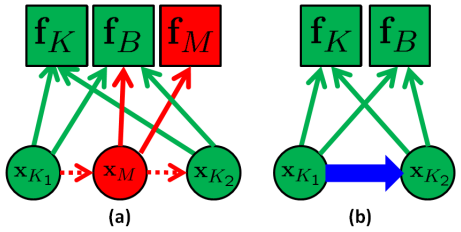

Fig. 1. (a) and (b) denote the structure of the current exploration epoch before and after the approximation employed in C-KLAM. $\mathbf{x}_M$ denotes the non-key poses between the two sets, $\mathbf{x}_{K_1}$ and $\mathbf{x}_{K_2}$, of key poses. $\mathbf{f}_K$, $\mathbf{f}_M$, and $\mathbf{f}_B$ denote the landmarks observed *only* from the key poses, *only* from the non-key poses, and *both* from the key and non-key poses, respectively. The arrows denote the measurements between different states.

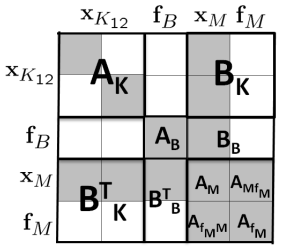

Fig. 2. Structure of the sparse information matrix corresponding to the cost function $\mathbb{C}_2$ in (2) (measurements shown with red arrows in Fig. 1(a)). The colored blocks denote non-zero elements. The sub-matrices $\mathbf{A}_B$, corresponding to landmarks observed from both key and non-key poses, and $\mathbf{A}_{f_M}$, corresponding to landmarks observed only from the non-key poses, are both block diagonal. The sub-matrix $\mathbf{A}_M$, corresponding to non-keyframes, is block tri-diagonal.

III. EXPERIMENTAL RESULTS

The experimental setup consists of a PointGrey Chameleon camera and a Navchip IMU, rigidly attached on a light-weight (100g) platform. The IMU signals were sampled at a frequency of 100 Hz while camera images were acquired at 7.5 Hz. The experiment was conducted in an indoor environment where the sensor platform performed a 3-D rectangular trajectory, with a total length of 144m and returned back to the initial position in order to provide an estimate of the final position error.

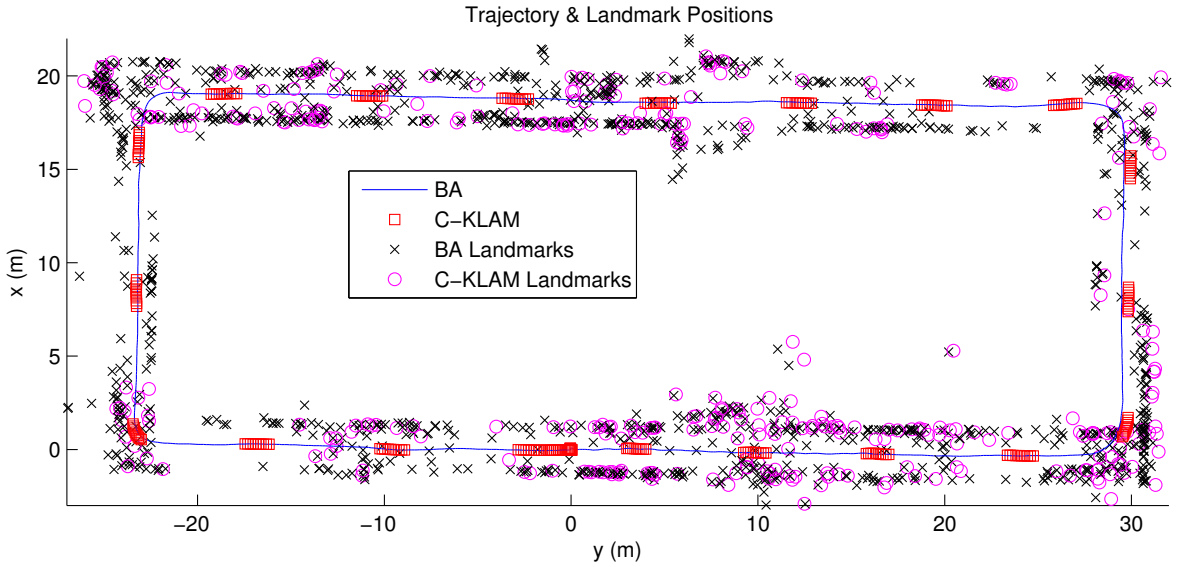

Fig. 3. Overhead $x-y$ view of the estimated 3-D trajectory and landmark positions. The C-KLAM estimates only keyframe poses (marked with red squares) and key features (marked with magenta circles).

In our C-KLAM algorithm implementation, the batch MAP-based SLAM problem is solved every 20 incoming camera poses. The exploration epoch is set to 60 camera poses, from which 10 consecutive camera poses are retained as keyframes, while the rest are marginalized. We compared the performance of C-KLAM to that of the full batch MAP-based SLAM (bundle adjustment [BA]), which optimizes over all robot poses and landmarks, using all available measurements. Thus, BA is computationally intensive, but provides high estimation accuracy. In our implementation of BA, the full batch MAP-based SLAM is solved every 20 camera poses.

Fig. 3 shows the $x-y$ view of the estimated trajectory and landmark positions. From the figure, we see that the estimates of the robot trajectory and landmark positions generated by C-KLAM coincide with those generated by the BA. Loop closure was performed in C-KLAM, and the final position error was 7 cm, only 5% more than that of the BA.

In terms of speed, the C-KLAM algorithm took only 4% of the time required for the entire BA. At the end of this experiment, the C-KLAM retained 240 keyframe poses and 350 key landmarks, while BA had 1040 camera poses and 1300 landmarks. Thus, the significant reduction in the number of estimated robot poses and landmarks in C-KLAM led to substantial improvement in efficiency. Moreover, by using information from non-key poses to constrain the key poses, we achieved estimation performance comparable to that of the BA. Lastly, we note that we tested the algorithm extensively in simulations (results not presented here due to space constraints) and our simulation results corroborate our experimental results.

REFERENCES

[1] R. Smith and P. Cheeseman, "On the representation and estimation of spatial uncertainty," *International Journal of Robotics Research*, vol. 5, no. 4, pp. 56–68, 1987.

[2] F. Dellaert and M. Kaess, "Square root SAM: Simultaneous Localization and Mapping via square root information smoothing," *International Journal of Robotics Research*, vol. 25, no. 12, pp. 1181–1203, Dec. 2006.

[3] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard, "g2o: A general framework for graph optimization," in *Proc. of the IEEE International Conference on Robotics and Automation*, Shanghai, China, May 9–13 2011, pp. 3607–3613.

[4] K. Konolige, G. Grisetti, R. Kummerle, W. Burgard, B. Limketkai, and R. Vincent, "Efficient Sparse Pose Adjustment for 2D mapping," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, Taiwan, Oct. 18–22 2010, pp. 22–29.

[5] M. Kaess, A. Ranganathan, and F. Dellaert, "iSAM: Incremental smoothing and mapping," *IEEE Transactions on Robotics*, vol. 24, no. 6, pp. 1365–1378, Dec. 2008.

[6] M. Kaess, H. Johannsson, R. Roberts, V. Ila, J. Leonard, and F. Dellaert, "iSAM2: Incremental smoothing and mapping using the bayes tree," *International Journal of Robotics Research*, vol. 21, pp. 217–236, Feb. 2012.

[7] G. Sibley, L. Matthies, and G. Sukhatme, "Sliding window filter with application to planetary landing," *Journal of Field Robotics*, vol. 27, no. 5, pp. 587–608, Sep./Oct. 2010.

[8] A. I. Mourikis, N. Trawny, S. Roumeliotis, A. Johnson, A. Ansar, and L. Matthies, "Vision-aided inertial navigation for spacecraft entry, descent, and landing," *IEEE Transactions on Robotics*, vol. 25, no. 2, pp. 264–280, Apr. 2009.

[9] G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. of the IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nara, Japan, Nov. 13–16 2007, pp. 225–234.

[10] K. Konolige and M. Agrawal, "FrameSLAM: From bundle adjustment to real-time visual mapping," *IEEE Transactions on Robotics*, vol. 24, no. 5, pp. 1066–1077, Oct. 2008.

[11] K. Konolige, J. Bowman, J. D. Chen, P. Mihelich, M. Calonder, V. Lepetit, and P. Fua, "View-based maps," *International Journal of Robotics Research*, vol. 29, no. 29, pp. 941–957, Jul. 2010.

[12] R. M. Eustice, H. Singh, and J. J. Leonard, "Exactly sparse delayed-state filters for view-based SLAM," *IEEE Transactions on Robotics*, vol. 22, no. 6, pp. 1100–1114, Dec. 2006.

[13] H. Johannsson, M. Kaess, M. Fallon, and J. Leonard, "Temporally scalable visual SLAM using a reduced pose graph," in *Proc. of the IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, May 6–10 2013, To appear.