# SAMSLAM:
# Simulated Annealing Monocular SLAM

Marco Fanfani, Fabio Bellavia, Fabio Pazzaglia, and Carlo Colombo

Computational Vision Group, University of Florence
Via Santa Marta, 3, 50139, Florence, Italy
{marco.fanfani,carlo.colombo}@unifi.it,
{bellavia.fabio,fabio.pazzaglia}@gmail.com

**Abstract.** This paper proposes a novel monocular SLAM approach. For a triplet of successive keyframes, the approach inteleaves the registration of the three 3D maps associated to each image pair in the triplet and the refinement of the corresponding poses, by progressively limiting the allowable reprojection error according to a simulated annealing scheme. This approach computes only local overlapping maps of almost constant size, thus avoiding problems of 3D map growth. It does not require global optimization, loop closure and back-correction of the poses.

**Keywords:** SLAM, Structure from Motion, RANSAC, Feature Matching, Disparity, Simulated Annealing, 3D Registration, Pose Estimation.

## 1 Introduction

Simultaneous Localization and Mapping (SLAM) approaches are designed to estimate both the camera positions and the 3D map of the environment in real-time. Early SLAM implementations were based on the Extendend Kalman Filter [1]. Alternative approaches inspired by Structure from Motion (SfM) techniques were proposed recently [2], and proved to outperform the former [3].

Single camera [1, 2], stereo or multiple camera [4] SLAM systems have appeared. While stereo or multiple camera configurations provide more reliable solutions, monocular SLAM leads to a more general and simple operative environment.

Different feature description and matching strategies [1,2,5,6] have been used to detect and track keypoints across the image frames: Robust to high degrees of blur [6], with hierarchical pose refinement [7], or exploiting the high computational power offered by modern GPUs through a dense approach [5].

SfM-based approaches typically exploit iterative non-linear optimization refinement schemes (such as Bundle Adjustment [8]) over sub-sequences of relevant frames (keyframes). While beneficial for accuracy improvement, such global optimization schemes require a good inizialization and are not tolerant to outliers. Moreover, optimization of long sub-sequences requires a large memory space, and refined estimates are obtained with some delay with respect to the current

camera position. Loop closure detection [9], enforcing pose constraints on already visited scenes, which obviously requires looping paths, is also frequently employed to reduce error accumulation over long tracks.

This paper proposes a novel monocular SLAM system, where a local, robust simulated annealing scheme replaces the global, SfM optimized approach for the purpose of obtaning both the 3D map and the camera pose. The proposed approach works locally on triplets of successive overlapping keyframes, thus guaranteeing scale and 3D structure consistency. Each update step uses RANSAC and alternates between the registration of the three 3D maps associated to each image pair in the triplet and the refinement of the corresponding poses, by progressively limiting the allowable reprojection error. Since the proposed method does not require neither global optimization nor loop closure, it doesn't perform any back-correction of the poses and does not suffer of 3D map growth. In addition, the method can be implemented in an efficient way through a multi-thread scheme.

The paper is organized as follows. Section 2 introduces the proposed SLAM system and the novel approach to the computation of the camera pose and the registration of the 3D points based on the simulated annealing process, while the experimental evaluation of the system is described in Sect. 3. Conclusions and final discussions are given in Sect. 4.

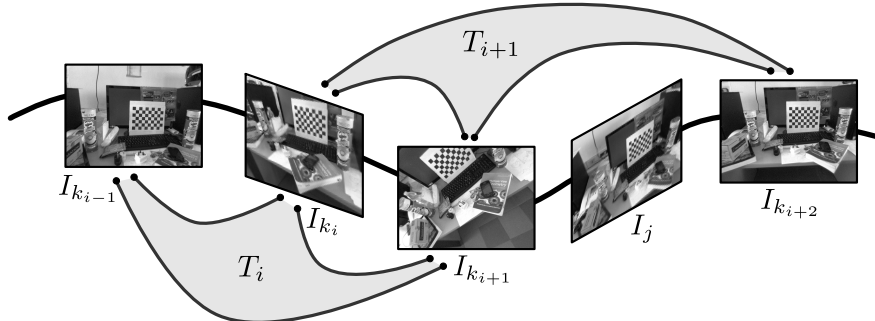## 2    The SAMSLAM Approach

### 2.1    Overview

Given a calibrated image sequence $S = \{I_t\}$, with radial distortion corrected, our SLAM approach proceeds by detecting successive triplets $T_i = \{I_{k_{i-1}}, I_{k_i}, I_{k_{i+1}}\}$ of image *keyframes* $\{I_{k_i}\} \subseteq S$, $k_0 = 0, k_i < k_{i+1}$ — see Fig. 1.

A local 3D map $\mathcal{M}_i$ is built upon the current keyframe triplet $T_i$ using the simulated annealing scheme described in Sect. 2.2, which also recovers the relative poses between keyframe pairs, $P_{k_{i-1},k_i}$ and $P_{k_{i-1},k_{i+1}}$. As the keyframe triplet is updated from $T_i$ to $T_{i+1}$, the first keyframe is dropped and a new one is queued, so that the 3D maps $\mathcal{M}_i$ and $\mathcal{M}_{i+1}$ overlap and the consistency of scale and 3D structure is guaranteed.

Image alignment for the generic pair $(I_{t_1}, I_{t_2})$ is based on keypoint matching. For each image, keypoints are extracted using the HarrisZ detector [10]. The sGLOH descriptor [11] with Nearest Neighbour matching is then used to obtain the candidate correspondences. These are then refined on a temporal constraint basis as follows. Let $\mathbf{x}_t = [x_t,\ y_t]^{\mathrm{T}} \in I_t$ be a generic keypoint of image $I_t$, a *match* $(\mathbf{x}_{t_1}, \mathbf{x}_{t_2})$ must satisfy the flow motion restriction $\| \mathbf{x}_{t_1} - \mathbf{x}_{t_2} \| < \delta_r$, where $\delta_r$ is the maximal flow displacement. Moreover, for a triplet $T_i$, after a further match refinement by normalized RANSAC [12], only matches which form a *loop chain*

$$\mathcal{C}_i = \left\{ (\mathbf{x}_{k_{i-1}}, \mathbf{x}_{k_i}), (\mathbf{x}_{k_i}, \mathbf{x}_{k_{i+1}}), (\mathbf{x}_{k_{i+1}}, \mathbf{x}_{k_{i-1}}) \right\}$$

through the corresponding keyframes are retained. The chain matches are used to estimate the 3D map $\mathcal{M}_i$ and the relative keyframe poses $P_{k_{i-1},k_i}$ and $P_{k_{i-1},k_{i+1}}$.

**Fig. 1.** Overview of the SAMSLAM approach. Keyframe triplets $T_i$ and $T_{i+1}$ are used to estimate successive overlapping local 3D point maps, which are then employed to retrieve the pose of a generic image frame $I_j$.

Note that, since outliers are dropped out by the simulated annealing scheme, only a fraction of the loop chain matches contribute to 3D points in the map $\mathcal{M}_i$.

The relative pose $P_{k_{i-1},j}$ for a generic image $I_j$, $k_{i+1} < j$, is estimated according to the 3D map $\mathcal{M}_i$ by employing a robust version of ePnP [13]. In order for $I_j$ to become the next keyframe $I_{k_{i+2}}$, a significant 2D motion with respect to $I_{k_{i+1}}$ has to be detected, in which case the current keyframe triplet $T_i$ is updated to $T_{i+1}$.

## 2.2 Simulated Annealing 3D Map and Pose Estimation

The keyframe triplet $T_i$ is related to the matches $(\mathbf{x}_{k_{i+v}}, \mathbf{x}_{k_{i+w}}) \in \mathcal{C}_i$, with $v, w \in \{-1, 0, 1\}$ and $v < w$. The simulated annealing approach starts by associating to each pair $(I_{k_{i+v}}, I_{k_{i+w}})$ an initial 3D map $\mathcal{M}_i^{v,w}$, obtained by triangulation on the matches $(\mathbf{x}_{k_{i+v}}, \mathbf{x}_{k_{i+w}})$. The relative pose $P_{k_{i+v}, k_{i+w}}$ is extracted from the essential matrix for the first triplet $T_1$, while for all triplets $T_i$, $i > 1$ relative poses are initialized with the estimates obtained at triplet time $i - 1$. After this initialization, the method registers the 3D maps $\mathcal{M}_i^{v,w}$ and refines the poses $P_{k_{i+v}, k_{i+w}}$ at each iteration $q$ (in all experiments, a maximum of 8 iterations were run). A block diagram of the proposed method is illustrated in Fig. 2.

3D map registration is done by the Horn method [14], fixing a 3D reference map $\mathcal{M}_i^{\mathrm{ref}}$ from one maps $\mathcal{M}_i^{v,w}$ for all iterations. Inconsistent 3D points with negative depths in any of the three associated stereo configurations $(I_{k_{i+v}}, I_{k_{i+w}})$ are removed as well as points far from any of the corresponding camera centres, since the uncertainty in point localization increases with distance. The proportion $p$ of points discarded by this latter constraint linearly decreases with the iteration $q$ since a more refined model is obtained as the iterations go on. In our experiments $p$ is made to decrease from 30% to 1%. Remaining points in the resulting submap $\widetilde{\mathcal{M}}_i^{v,w}$ are registered to the reference submap $\widetilde{\mathcal{M}}_i^{\mathrm{ref}}$ through the Horn method, made robust to outliers by RANSAC. Reference 3D points of $\widetilde{\mathcal{M}}_i^{\mathrm{ref}}$ are mapped to $\widetilde{\mathcal{M}}_i^{v,w}$ according to the transformation $H_i^{v,w}$ estimated by the Horn method and back-projected to the corresponding images $I_v$ and $I_w$.
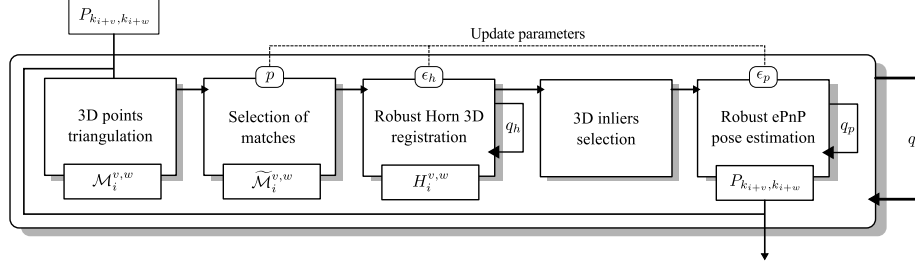
**Fig. 2.** Diagram of the simulated annealing 3D map and pose estimation executed for each keyframe triplet $T_i$

The distances between the back-projected points and the effective matches $\mathbf{x}_v$ and $\mathbf{x}_w$ are used to define inliers. The inlier threshold value $\epsilon_h$ linearly decreases with the iteration $q$, from 20 to 4 pixels in our experiments. At each RANSAC iteration $q_h$ the transformation $H_i^{v,w}$ is refined. The sampling set is a subset of the whole validation set and contains only the 25% of points in $\widetilde{\mathcal{M}}_i^{v,w}$ with maximal flow displacement. This is beneficial to map accuracy, since high disparity matches are characterized by a better localization in 3D space.

The pose refinement step also works on the reference 3D map $\widetilde{\mathcal{M}}_i^{\text{ref}}$. The ePnP with RANSAC is applied to points associated to the common inliers found in the Horn registration step between the maps $\widetilde{\mathcal{M}}_i^{v,w}$. The reprojection error threshold $\epsilon_p$ used to define inliers linearly decreases with the iteration $q$, from 5 to 3 pixels in the experiments. Similarly to 3D map registration, a constraint on the sampling set depending on the RANSAC iteration $q_p$ is used. The refined poses $\mathrm{P}_{k_{i+v},k_{i+w}}$ replace the previous ones for the next iteration $q$.

Figure 3(a) shows an example of the simulated annealing scheme on the first keyframe triplet $T_1$ of the *Monk* video sequence (see Sect. 3). Fig. 4 shows the corresponding 3D maps $\widetilde{\mathcal{M}}_i^{v,w}$ for different iterations $q$. The average reprojection error gradually decreases for each image pair $(I_{k_{i+v}}, I_{k_{i+w}})$ to less than 2 pixels, while the number of 3D point inliers increases and the 3D registration improves. Note that the first iteration $q = 1$ of the first keyframe triplet $T_1$ is the most time consuming in terms of RANSAC iterations with $q_h, q_p \simeq 500$, while in the other cases $q_h, q_p \simeq 50$ since only refinements are required. The RANSAC-based design can be useful to define efficient parallel and multi-threaded implementations of the simulated annealing scheme.

## 3   Results

In order to evaluate the performance of our monocular SLAM approach, two different experiments have been carried out: A quantitative direct measure of the odometry accuracy, and an indirect evaluation of the 3D reconstruction quality of an object acquired using a structured-light framework.

Three different indoor video sequences with a resolution of $640 \times 480$ pixels and about 800 frames have been used in the former case — see Table 1(a). The

(a)                                           (b)
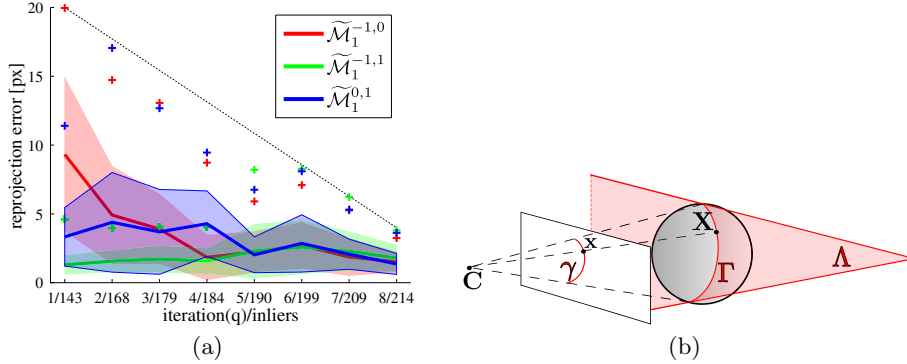
**Fig. 3.** (a) The reprojection errors as the iterations proceed for the first keyframe triplet $T_1$ of the *Monk* video sequence. The reference map is $\widetilde{\mathcal{M}}_1^{\mathrm{ref}} = \widetilde{\mathcal{M}}_1^{-1,1}$. Solid lines indicate the average reprojection errors, while bands show the behaviour of the standard deviation. Marks represents the maximal values and the dashed gray line is the RANSAC linear threshold bound $\epsilon_h$. (b) The laser scanner configuration for the evaluation of the *Monk* sequence. (Best viewed in color)



(a)                              (b)                              (c)
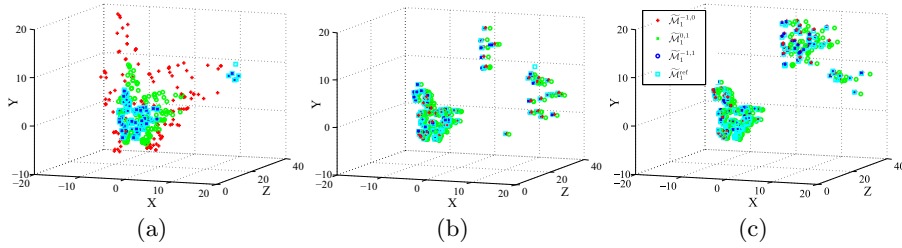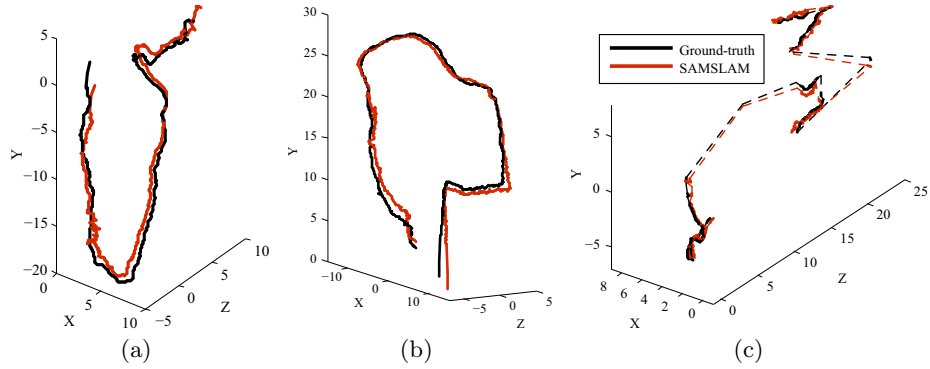
**Fig. 4.** The 3D maps $\widetilde{\mathcal{M}}_i^{v,w}$ and the reference map $\widetilde{\mathcal{M}}_1^{\mathrm{ref}}$ at iterations $q = 1, 5, 8$ for the keyframe triplet $T_1$ of the *Monk* video sequence. (Best viewed in color)

first two sequences (*Desk1* and *Desk2*) explore the same desktop environment as the camera undergoes two different motions, while the last sequence (*Monk*) contains an object scanned by a laser fan projector. This last sequence is also used for the indirect evaluation through 3D reconstruction. A known planar pattern is included in the background of all test sequences to recover accurate ground-truth poses using the approach described in [15].

Table 1(a) shows the Euclidean distance error of the camera centres normalized to the ground-truth path length, while corresponding tracks are shown in Fig. 5. Since the scale information is lost, the camera centres have been registered to the known ground-truth metric scale using the Horn method. SAMSLAM error is about 1% on average, i.e. less than 1 cm for a track length of 100 cm, and tracks are well aligned.

**Table 1.** (a) Distance error of the camera centres with respect to the ground-truth length. (b) 3D reconstruction error for the *Monk* sequence.

| | (a) | | | | (b) | | |
|---|---|---|---|---|---|---|---|



| | Desk1 | Desk2 | Monk |
|---|---|---|---|
| Mean(%) | 1.29 | 0.93 | 0.48 |
| Std(%) | 0.63 | 0.30 | 0.23 |
| Max(%) | 3.05 | 2.39 | 1.21 |
| Min(%) | 0.24 | 0.29 | 0.15 |
| Length(cm) | 71.31 | 100.35 | 74.90 |

Monk     Ground-truth SAMSLAM

3D reconstruction error ($10^{-3}$ cm)

| Mean | Std | Max |
|---|---|---|
| 0.105 | 0.112 | 1.616 |



| (a) | (b) | (c) |
|---|---|---|

**Fig. 5.** Track comparison for the video sequences *Desk1* (a), *Desk2* (b) and *Monk* (c). Dashed lines for the *Monk* sequence indicate that no ground-truth has been provided. (Best viewed in color)

For the 3D laser-scanned reconstruction test on the *Monk* sequence, a device equipped with a camera and a laser fan projector kept in fixed relative position is used in order to get an accurate 3D model. As depicted in Fig. 3(b), $C$ is the camera centre, $\Lambda$ the laser fan plane, $\Gamma$ the 3D laser trace and $\gamma$ the 2D laser image.

In basic projection geometry of laser profile $\Gamma$ onto the image, each point $\mathbf{x}$ of the imaged laser profile $\gamma$ can be backprojected onto the laser plane $\Lambda$, obtaining its pre-image $\mathbf{X} \in \Gamma$. The backprojection equation can be expressed as $\mathbf{X} = \alpha \mathtt{K}^{-1}\mathbf{x}$, where $\alpha = d[\mathbf{n}^{\top}\mathtt{K}^{-1}\mathbf{x}]^{-1}$, $\mathbf{n}^{\top}\mathbf{X} - d = 0$ is the equation of the laser plane $\Lambda$ in inhomogeneous camera-centred coordinates, $\mathbf{x}$ is a homogeneous 3-vector, and $\mathtt{K}$ is the camera calibration matrix. A 3D profile is obtained in the camera framework for each frame moving the scanning device. Knowing the estimated motion, it is possible to collate all the 3D profiles in a unique model.

Table 1(b) shows the 3D Euclidean reconstruction errors with respect to the ground-truth obtained with the estimated motion and the reconstructed model. Even in this case the error is low while the 3D reconstruction is almost identical to the ground-truth.

## 4    Conclusions

This paper presents a mono SLAM approach, relying on a local keyframe optimization, based on simulated annealing, which iteratively refines both the motion estimates and the 3D structure. Direct evaluation of track error and indirect validation through structured-light 3D reconstruction show good performance of the approach, that does not require neither global optimization nor loop closure techniques.

Future work will include solutions for a better 3D registration and pose handling in the case of noisy correspondences, due for example to motion blur, and to enforce the system for long tracks, also adaptively correcting the reference 3D map in the case of faults. Furthermore, efficient and optimized implementation of the system will be developed.

## References

1. Davison, A.: Real-time simultaneous localization and mapping with a single camera. In: Proc. 9th IEEE International Conference on Computer Vision, pp. 1403–1410 (2003)
2. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 225–234 (2007)
3. Strasdat, H., Montiel, J., Davison, A.: Visual SLAM: Why filter? Image and Vision Computing 30, 65–77 (2012)
4. Mei, C., Sibley, G., Cummins, M., Newman, P., Reid, I.: RSLAM: A system for large-scale mapping in constant-time using stereo. International Journal of Computer Vision 94, 198–214 (2011)
5. Newcombe, R., Lovegrove, S., Davison, A.: DTAM: Dense Tracking and Mapping in Real-Time. In: Proc. 13th International Conference on Computer Vision (2011)
6. Pretto, A., Menegatti, E., Bennewitz, M., Burgard, W.: A visual odometry framework robust to motion blur. In: Proc. IEEE International Conference on Robotics and Automation (2009)
7. Strasdat, H., Davison, A.J., Montiel, J.M.M., Konolige, K.: Double window optimisation for constant time visual SLAM. In: Proc. of the International Conference on Computer Vision, pp. 2352–2359 (2011)
8. Triggs, B., McLauchlan, P., Hartley, R., Fitzgibbon, A.: Bundle adjustment - a modern synthesis. In: Proc. of the International Workshop on Vision Algorithms: Theory and Practice, pp. 298–372 (2000)

 9. Ho, K., Newman, P.: Detecting loop closure with scene sequences. International Journal Computer Vision 74(3), 261–286 (2007)
10. Bellavia, F., Tegolo, D., Valenti, C.: Improving Harris corner selection strategy. IET Computer Vision 5(2) (2011)
11. Bellavia, F., Tegolo, D., Trucco, E.: Improving SIFT-based descriptors stability to rotations. In: Proc. of International Conference on Pattern Recognition (2010)
12. Bellavia, F., Tegolo, D.: noRANSAC for fundamental matrix estimation. In: Proc. of he British Machine Vision Conference, pp. 1–98 (2011)
13. Moreno-Noguer, F., Lepetit, V., Fua, P.: Accurate Non-Iterative O(n) Solution to the PnP Problem. In: Proc. of IEEE International Conference on Computer Vision (2007)
14. Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A 4(4), 629–642 (1987)
15. Fanfani, M., Colombo, C.: LaserGun: A tool for hybrid 3D reconstruction. In: Chen, M., Leibe, B., Neumann, B. (eds.) ICVS 2013. LNCS, vol. 7963, pp. 274–283. Springer, Heidelberg (2013)