Robust Selective Stereo SLAM without Loop Closure and Bundle Adjustment

Fabio Bellavia, Marco Fanfani, 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 presents a novel stereo SLAM framework, where a robust loop chain matching scheme for tracking keypoints is combined with an effective frame selection strategy. The proposed approach, referred to as *selective SLAM* (SSLAM), relies on the observation that the error in the pose estimation propagates from the uncertainty of the three-dimensional points. This is higher for distant points, corresponding to matches with low temporal flow disparity in the images. Comparative results based on the reference KITTI evaluation framework show that SSLAM is effective and can be implemented efficiently, as it does not require any loop closure or bundle adjustment.

Keywords: SLAM, Structure from Motion, RANSAC, feature matching, frame selection

1 Introduction

The interest for visual Simultaneous Localization and Mapping (SLAM) has been increasingly growing in the computer vision community during the last few years. The main goal of SLAM is to simultaneously estimate both the camera positions and a geometrical 3D representation of the environment with real-time constraints [1]. Early SLAM implementations were based on probabilistic frameworks [2, 3], and employed Bayesian filtering techniques, such as the Extendend Kalman Filter (EKF), to couple together in the same process the 6 DoF camera positions and all the 3D points, incrementally updated. Later, alternative SLAM implementations were proposed [4], influenced by the Structure from Motion (SfM) paradigm. These approaches exploit the epipolar geometry constraints to first estimate the camera positions and the 3D map, in general by using the RANdom SAmple Consensus (RANSAC) paradigm [5, 6]. Successive refinement steps by iterative non-linear optimization techniques, such as bundle adjustment [7], over a selected sub-set of frames (keyframes) are used to minimize the global error. This allows separating pose estimation from 3D map computation, thus efficiently decoupling the process flows, as 3D structure needs not be optimized at each pose update but only when needed [8].

Both kinds of approaches have some drawbacks. In the Bayesian frameworks, points have to be added and discarded as the estimation proceeds, since the 3D map cannot grow excessively for computational limits. On the other hand, keyframe-based approaches, in order to achieve real-time operation, can perform local optimizations only occasionally. According to [9], keyframe based solutions outperform Bayesian approaches, due to their ability to maintain more 3D points in the estimation procedure.

Single camera (or mono) [1,3,8], stereo [10–12] or multiple cameras [13] setups can be used in SLAM systems, and different features and matching strategies can be employed to detect and track keypoints across image frames [1, 8, 13, 14]. In general, stereo or multiple camera configurations provide better solutions, since the rigid calibration of the cameras increases the accuracy in the 3D map computation and provides more robust matching correspondences. Further issues must be taken into account in mono SLAM design, such as the delayed 3D feature initialization [2] (i.e. when a point is seen for the first time) and the scale factor uncertainty [15].

Since SLAM system design is affected by the input scene, different implementation choices can be found for indoor [1, 8], outdoor [11, 12, 15] or even underwater [16] environments. Large scenarios present more challenging tasks since long tracks tend to accumulate an error drift. In order to alleviate this issue and to achieve finer and better estimates, loop closure detection techniques [17] have been developed to enforce pose constraints by recognizing already visited scenes, which obviously requires the camera to perform a looping path.

SLAM systems have to deal with errors mainly introduced during the extraction and the matching of 2D features. This paper implements a stereo SLAM system with a robust loop chain matching scheme [14], where the recent HarrisZ detector [18] and the sGLOH descriptor [19] are used to extract and match keypoints respectively, in order to provide more stable matches. Furthermore, only high temporal flow disparity frames are used to estimate the pose, since the error in the pose estimation is propagated from the uncertainty of the three-dimensional points, which is larger for distant points corresponding to low temporal flow disparity matches in the images. This strategy is effective at detecting and discarding frames with a similar visual content. The proposed system, referred to as *selective SLAM* (SSLAM), does not rely upon loop closure detection or bundle adjustment, and only considers the most reliable data measurements. This proves beneficial to both algorithmic accuracy and efficiency.

The paper is organized as follows. In Sect. 2 details of SSLAM are presented. Comparative results according to the reference KITTI evaluation framework [20] are discussed in Sect. 3. Conclusions and final remarks are offered in Sect. 4.

2 Selective Stereo SLAM

Given a calibrated and rectified stereo sequence $S = \{f_t\}$, where the frame $f_t = (I_t^l, I_t^r)$ is composed by the left I_t^l and right I_t^r input images taken at time $t \in \mathbb{N}$, SSLAM alternates between two main steps. The former step matches

keypoints between the previous accepted SLAM frame f_i and the current frame f_j, while the latter estimates the relative camera pose $\mathrm{P}_{i,j} = [\mathrm{R}_{i,j}|\mathbf{t}_{i,j}] \in \mathbb{R}^{3\times4}$, where $\mathrm{R}_{i,j} \in \mathbb{R}^{3\times3}$ is the rotation matrix and $\mathbf{t}_{i,j} \in \mathbb{R}^3$ is the translation vector. Note that since $i < j$, frames that can potentially lead to a large error in pose estimation due to high uncertainty in the 3D data measurements can be discarded. In this way, the system is able to keep small errors also for long trajectories relative to the first frame f_0, taken as absolute reference, even without global optimization or loop closure techniques. The absolute pose at time n is defined as $\mathrm{P}_n = \mathrm{P}_{0,n}$. P_n can be computed by concatenating the poses $\mathrm{P}_{0,0}, \mathrm{P}_{0,k} \ldots, \mathrm{P}_{i,j}\mathrm{P}_{j,n}$, where time steps $0 < k < \ldots < j < i < n$ belong to accepted frames.

2.1 Keypoints Detection and Matching

In the matching step the HarrisZ detector [18], which provides results comparable to other state-of-the-art detectors, is used to extract robust and stable corner features in the affine scale-space on the images $I_i^l, I_i^r, I_j^l, I_j^r$. The sGLOH descriptor [19] with the Nearest Neighbour matching is used to obtain the candidate correspondences between image pairs $(I_i^l, I_i^r), (I_i^l, I_j^r), (I_i^r, I_j^r), (I_j^l, I_j^r)$ after spatial and temporal constraints have been imposed to refine the candidates matches.

Let $\mathbf{x}_s^d = [x_s^d,\, y_s^d]^\mathrm{T} \in \mathbb{R}^2$, $d \in \{l, r\}$, $s \in \{i, j\}$ be a point in the image I_s^d, a spatial match $(\mathbf{x}_s^l, \mathbf{x}_s^r)$ between the images on the same frame is computed by the stereo epipolar constraints imposed by the calibration

$$|x_s^l - x_s^r| < \delta_x \tag{1}$$
$$|y_s^l - y_s^r| < \delta_y \tag{2}$$

where δ_y is the error band allowed by epipolar rectification and δ_x is the maximal allowed disparity. In the case of a temporal match $(\mathbf{x}_i^d, \mathbf{x}_j^d)$ between corresponding images at different frames, the flow motion restrictions

$$\| \mathbf{x}_i^d - \mathbf{x}_j^d \| < \delta_r \tag{3}$$

are taken into account, where δ_r is the maximal flow displacement. Matches which form a *loop chain*

$$\mathcal{C} = \big((\mathbf{x}_i^l, \mathbf{x}_i^r), (\mathbf{x}_i^l, \mathbf{x}_j^l), (\mathbf{x}_j^l, \mathbf{x}_j^r), (\mathbf{x}_i^r, \mathbf{x}_j^r)\big) \tag{4}$$

are retained, see Fig. 1. In order to filter the candidate matches, four distinct RANSAC tests are finally run among the four image pairs to refine the epipolar geometry so that only a subset of chain matches $C_{i,j} \subseteq \{\mathcal{C}\}$ is selected. Note that the proposed matching scheme is similar to [14], but achieves longer and more stable keypoint tracks, which are crucial for the pose estimation—see the experimental evaluation.

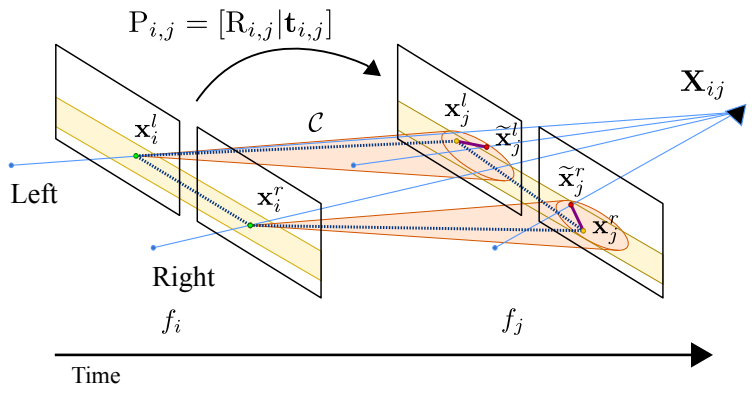


Fig. 1: (Best viewed in color) Keypoint matches between the frame f_i and f_j must satisfy the spatial constraint imposed by the epipolar rectification (yellow band) as well as the temporal flow motion restriction (orange cone). Furthermore, the four matching points must form the loop chain $\mathcal{C}$ (dotted line). In the ideal case, points $\mathbf{x}_j^l$, $\mathbf{x}_j^r$ in frame f_j must coincide with the projections $\widetilde{\mathbf{x}}_j^l$, $\widetilde{\mathbf{x}}_j^r$ of points $\mathbf{x}_i^l$, $\mathbf{x}_i^r$ in f_i obtained by triangulation of $\mathbf{X}_{i,j}$ in order for the chain $\mathcal{C}$ to be consistent with the pose $\mathrm{P}_{i,j}$. Due to data noise, in the real case the distances $\parallel \widetilde{\mathbf{x}}_j^l - \mathbf{x}_j^l \parallel$ and $\parallel \widetilde{\mathbf{x}}_j^r - \mathbf{x}_j^r \parallel$ must be minimal.

2.2 Pose Estimation Constrained by Temporal Flow

The relative pose $\mathrm{P}_{i,j}$ between frames f_i and f_j is estimated in the second step of our SSLAM approach—see again Fig. 1. The 3D point $\mathbf{X}_{i,j}$ corresponding to the match pair $(\mathbf{x}_i^l, \mathbf{x}_i^r)$ in frame f_i can be estimated by triangulation [5], since the intrinsic and extrinsic calibration parameters of the system are known. Let $\widetilde{\mathbf{x}}_j^l$ and $\widetilde{\mathbf{x}}_j^r$ be the projections of $\mathrm{X}_{i,j}$ onto frame f_j, according to the estimated relative pose $\mathrm{P}_{i,j}$. The distance

$$\mathcal{D}(\mathrm{P}_{i,j}) = \sum_{C_{i,j}, d \in \{l,r\}} \parallel \widetilde{\mathbf{x}}_j^d - \mathbf{x}_j^d \parallel \tag{5}$$

among the matches of the chain set $C_{i,j}$ must be minimized, in order for the estimated pose $\mathrm{P}_{i,j}$ to be consistent with the data. Due to the presence of outliers in $C_{i,j}$, a RANSAC test is run [14], where the number $\mathcal{D}_R(\mathrm{P}_{i,j})$ of outliers chain matches over $C_{i,j}$ exceeding a threshold value δ_t is minimized so that pose $\mathrm{P}_{i,j}$ be consistent with data:

$$\mathcal{D}_R(\mathrm{P}_{i,j}) = \sum_{C_{i,j}} T_d \left(\parallel \widetilde{\mathbf{x}}_j^d - \mathbf{x}_j^d \parallel > \delta_t \right) \quad . \tag{6}$$

In Eq. 6, $d \in \{l, r\}$, and the indicator function $T_x(P(x))$ is 1 if the predicate $P(x)$ is true for all the admissible values of x, and 0 otherwise. The final pose estimation $\overline{\mathrm{P}}_{i,j}$ between frames f_i and f_j is chosen as

$$\overline{\mathrm{P}}_{i,j} = \underset{\mathrm{P}_{i,j}}{\operatorname{argmin}}\, \mathcal{D}_R(\mathrm{P}_{i,j}) \quad . \tag{7}$$

In the traditional approach used in [14], at each iteration RANSAC estimates a candidate pose $\mathrm{P}_{i,j}$ using a minimal set of matches, i.e., 3 matches, in order to be robust to outliers [6]. The candidate matches used to build the pose model $\mathrm{P}_{i,j}$ are sampled from the set of candidate matches $C_{i,j}$. The pose $\mathrm{P}_{i,j}$ is validated against the whole set of candidate matches $C_{i,j}$ according to (6) and the best model found so far is retained. The process stops when the probability to get a better model is below some user-defined threshold value, and the final pose $\overline{\mathrm{P}}_{i,j}$ is refined [21] on the set $G_{\overline{\mathrm{P}}_{i,j}}$ of inlier matches where

$$G_{\mathrm{P}_{i,j}} = \left\{ \mathcal{C} \in C_{i,j} \,|\, T_d \left(\| \, \widetilde{\mathbf{x}}_j^d - \mathbf{x}_j^d \, \| < \delta_t \right) \right\} \tag{8}$$

for a generic pose $\mathrm{P}_{i,j}$.

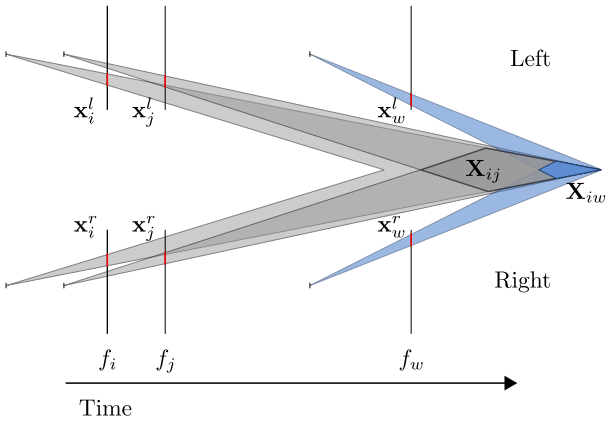


Fig. 2: (Best viewed in color) The uncertainty of matches in the image planes is lower bounded by the image resolution (red) and it is propagated to the 3D points. In order to estimate the 3D point $\mathbf{X}_{i,j}$, by using close frames f_i and f_j, a low temporal disparity flow is present in the image planes, and the 3D point location $\mathbf{X}_{i,j}$ can assume an higher range $\mathbf{X}_{i,j}$ of values (dark gray quadrilateral). In the case of distant frames f_i and f_w, the possible locations $\mathbf{X}_{i,w}$ are more circumscribed (blue quadrilateral), for the same resolution limits.

In our SSLAM approach, pose is estimated in a slightly different way, by taking advantage of the following observation. The image resolution provides a lower bound to the uncertainty in the keypoint match locations, which are triangulated to get the corresponding 3D point, and eventually estimate the relative pose between two temporal frames. Close frame matches have a low temporal flow disparity and the associated 3D point position has a high uncertainty with

respect to distant frames, due to the error propagation from the matches on the image planes. Only points with sufficient displacement can give information about the translational and rotational motion, as shown in Fig. 2. According to this observation, SSLAM singles out from the set of chain matches $C_{i,j}$ for frames f_i and f_j the set $F_{i,j}$ containing the fixed points, i.e., points with low temporal flow disparity:

$$F_{i,j} = \{\mathcal{C} \in C_{i,j}, T_d(\parallel \mathbf{x}_i^d - \mathbf{x}_j^d \parallel \leq \delta_f)\} \ , \tag{9}$$

for a given threshold δ_f. In order for frame f_j to be accepted, the number of fixed matches between frames f_i and f_j must be low:

$$\frac{|F_{i,j}|}{|C_{i,j}|} < \delta_m \quad . \tag{10}$$

Indeed, if the estimation is severely corrupted by noise and can lead to a very bad estimation, the frame f_j is discarded and the next frame f_{j+1} is tested. This provides an adaptive threshold to discard bad frames containing low motion information—examples are shown in Fig. 3. A similar approach has been recently employed in [22].

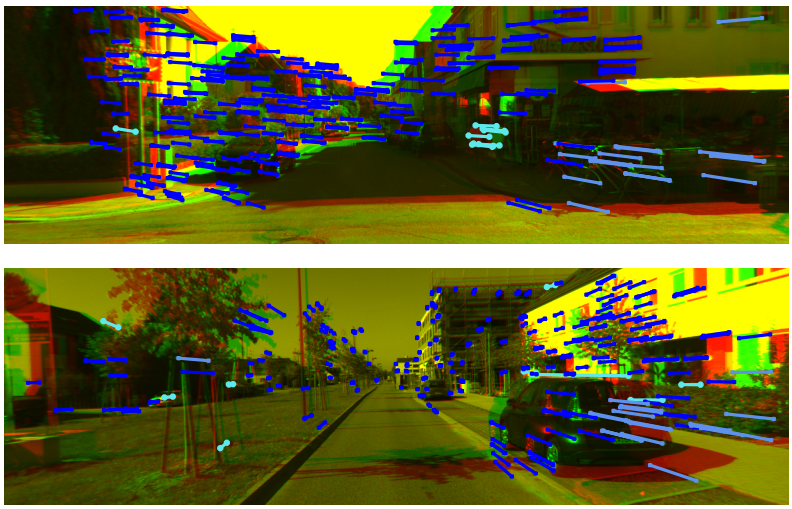


Fig. 3: (Best viewed in color) Examples of successive keyframes retained according to the temporal flow for two different sequences of the KITTI dataset. The two temporal keyframes involved are superimposed as for anaglyphs, only images for the left cameras are shown. Good fixed and not fixed matches are shown in blue and light violet, respectively, while wrong correspondences are reported in cyan.

Finally, we add a pose smoothing constraint between frames, so that the current relative pose estimation $\mathrm{P}_{i,j}$ cannot abruptly vary from the previous $\mathrm{P}_{z,i}$, $z < i < j$. This is achieved by imposing that the relative rotation around the origin between the two incremental rotations $\mathrm{R}_{z,i}$ and $\mathrm{R}_{i,j}$ is bounded, as well as for the corresponding translation directions $\mathbf{t}_{z,i}$ and $\mathbf{t}_{i,j}$:

$$|{\mathbf{r}^k_{i,j}}^{\mathrm{T}} \mathbf{r}^k_{z,i}| < \delta_{\theta_1} \tag{11}$$

$$\frac{|{\mathbf{t}_{i,j}}^{\mathrm{T}} \mathbf{t}_{z,i}|}{\| \mathbf{t}_{i,j} \| \| \mathbf{t}_{z,i} \|} < \delta_{\theta_2} \ , \tag{12}$$

where $\mathbf{r}^k_{a,b}$ is any k-th column of the rotation matrix $\mathrm{R}_{a,b}$. This last constraint can resolve issues in the case of no camera movement or when moving objects crossing the camera path cover the scene.

3 Evaluation

In order to evaluate the proposed system, the odometry dataset and the evaluation protocol of the KITTI vision benchmark suite has been used, which has becoming a reference evaluation framework for SLAM systems in recent years [20]. The dataset provides sequences recorded from car driving sessions on highways and inside cities. In particular we used the first 11 stereo sequences of the dataset, for which ground-truth data obtained by laser and GPS sensors are provided.

We compared different versions of our SSLAM system, corresponding to the successive improvements of the pipeline proposed in Sect. 2, both implemented as non-optimized Matlab code. In particular we indicate by SSLAM* the first version which only includes the loop chain matching described in Sect. 2.1, while the adaptive frame discarding strategy is incorporated in SSLAM.

The freely available SLAM library VISO2-S [14], one of the best performing SLAM in the KITTI ranking, is added in the evaluation as reference, since it uses a loop chain matching scheme similar to ours and a standard RANSAC pose estimation. The default parameter settings provided by the authors have been used for VISO2-S, while additionally for our systems we set $\delta_r = 500$ px, $\delta_x = 300$ px, $\delta_y = 14$ px, $\delta_f = 55$ px, $\delta_m = 0.05$, $\delta_{\theta_1} = 15^\circ$ and $\delta_{\theta_2} = 10^\circ$—see Sect. 2.2.

In order to analyze the robustness and the effectiveness of the proposed method, the SSLAM system was tested with a different number of RANSAC iterations for the pose estimation. In particular, results of SSLAM with 500, 15 and 3 RANSAC iterations, and SSLAM* with 500 iterations are presented, indicated respectively by SSLAM/500, SSLAM/15, SSLAM/3 and SSLAM*/500. The VISO2-S system uses 200 RANSAC iterations by default.

Fig. 4 shows the average translation and rotation errors of the different SLAM methodologies for increasing path length and speed, according to the KITTI evaluation framework [20]. Only frames common to all the methods, i.e. not discarded during the process by any of the proposed implementations, are used.

This does not affect the results, since the computed error measures rely on the SLAM absolute positions, which remain the same. Both the different versions of the proposed method provide results better than the VISO2-S system.

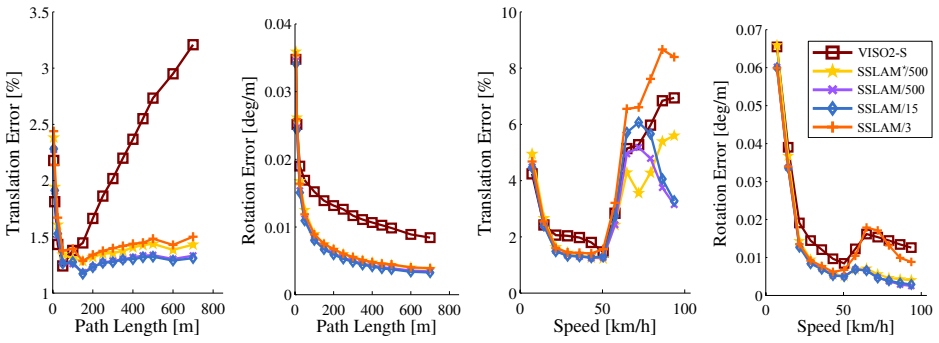


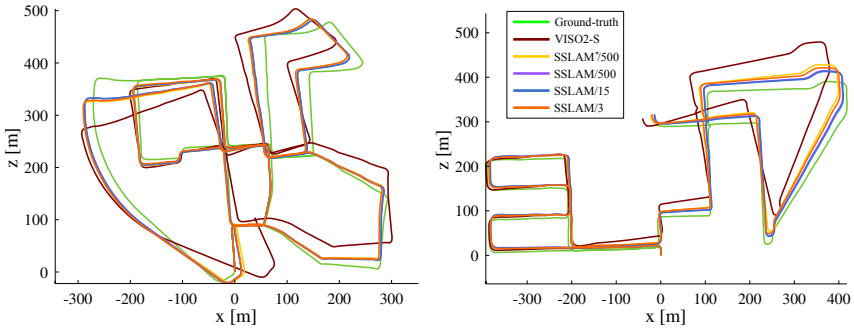
Fig. 4: (Best viewed in color) Average error on the first 11 sequences of the KITTI dataset. Plots (a-b) refer to the average translation and rotation error for increasing path length respectively, while plots (c-d) refer to increasing speed.

The chain loop matching scheme together with the chosen keypoint detector and descriptor is robust even for long paths, without relying on bundle adjustment or loop closure detection. Furthermore, dropping low temporal flow disparity frames in SSLAM improves on the standard pose estimation used in SSLAM*, allowing the tracking of longer paths. Note that SSLAM drops for each sequence from 35% to 70% of the total frames, which mainly occur on straight paths covered at low and medium speeds.

Moreover, results for SSLAM/15 and SSLAM/500 are equivalent, while SSLAM/3 obtains inferior results but similar to those obtained by SSLAM*/500, giving an evidence of the robustness of the proposed matching selection strategy and pose estimation, which can also improve the final running time.

Finally, on average the SSLAM and SSLAM* systems extract about 300 matches and VISO2-S around 250. The proposed SSLAM methodologies retain about 95% of extracted matches as inliers after the pose estimation, while only about 50% are instead preserved by VISO2-S, giving a further evidence on the effectiveness of the proposed system.

Fig. 5 shows the estimated paths for two sequences of the KITTI dataset. By inspecting the tracks it can be clearly seen that both SSLAM and SSLAM* (in this order) paths are closer to the ground-truth with respect to VISO2-S. In particular, rotations are the major source of incremental error, but our approach succeeds to better solve this issue.



(a) (b)

Fig. 5: (Best viewed in color) An example of the final paths computed for Sequence 00 (a) and Sequence 08 (b) of the KITTI dataset.

4 Conclusion

In this paper a new stereo SLAM system was presented. The approach achieves a low drift error even for long paths, without relying on loop closure or bundle adjustment. A robust loop chain matching scheme for tracking keypoints is provided, sided by a frame discarding system to improve pose estimation. According to the experimental results, dropping low temporal flow disparity frames for discarding highly uncertain models is an effective strategy to reduce error propagation from matches. Results validated on the KITTI dataset showed the effectiveness of the system, which is robust even for an extremely small number of RANSAC iterations.

Future work will include the implementation of an efficient optimized code of the system, experimental results on a wider range of sequences and the development of a more advanced adaptive sampling scheme for model estimation.

Acknowledgements

This work has been carried out during the THESAURUS project, founded by Regione Toscana (Italy) in the framework of the "FAS" program 2007-2013 under Deliberation CIPE (Italian government) 166/2007.

References

1. Davison, A.: Real-time simultaneous localization and mapping with a single camera. In: Proc. 9th IEEE International Conference on Computer Vision. (2003) 1403–1410
2. Montiel, J., Civera, J., Davison, A.: Unified inverse depth parametrization for monocular SLAM. In: Proc. of Robotics: Science and Systems, IEEE Press (2006)

3. Davison, A., Reid, I., Molton, N., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Trans. on Pattern Analysis and Machine Intelligence **29**(6) (2007) 1052–1067
4. Mouragnon, E., Lhuillier, M., Dhomeand, M., Dekeyserand, F., Sayd, P.: Real time localization and 3D reconstruction. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. (2006) 363–370
5. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. 2nd edn. Cambridge University Press (2004)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6) (1981) 381–395
7. 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. (2000) 298–372
8. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. IEEE/ACM International Symposium on Mixed and Augmented Reality. (2007) 225–234
9. Strasdat, H., Montiel, J., Davison, A.: Visual SLAM: Why filter? Image and Vision Computing **30** (2012) 65–77
10. Paz, L., Piniés, P., Tardós, J., Neira, J.: Large-scale 6-DoF SLAM with stereo-in-hand. IEEE Trans. Robotics **24**(5) (2008) 946–957
11. Lim, J., Pollefeys, M., Frahm, J.M.: Online environment mapping. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition. (2011)
12. 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** (2011) 198–214
13. Zou, D., Tan, P.: CoSLAM: Collaborative visual SLAM in dynamic environments. IEEE Trans. on Pattern Analysis and Machine Intelligence **35**(2) (Feb. 2013) 354–366
14. Geiger, A., Ziegler, J., Stiller, C.: StereoScan: Dense 3D reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium. (2011)
15. Strasdat, H., Montiel, J.M.M., Davison, A.J.: Scale drift-aware large scale monocular SLAM. In: Proc. of Robotics: Science and Systems. (2010)
16. Mahon, I., Williams, S., Pizarro, O., Johnson-Roberson, M.: Efficient view-based SLAM using visual loop closures. IEEE Trans. on Robotics **24** (2008) 1002–1014
17. Ho, K., Newman, P.: Detecting loop closure with scene sequences. International Journal Computer Vision **74**(3) (2007) 261–286
18. Bellavia, F., Tegolo, D., Valenti, C.: Improving Harris corner selection strategy. IET Computer Vision **5**(2) (2011)
19. Bellavia, F., Tegolo, D., Trucco, E.: Improving SIFT-based descriptors stability to rotations. In: Proc. of International Conference on Pattern Recognition. (2010)
20. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. of Computer Vision and Pattern Recognition. (2012)
21. Horn, B.K.P.: Closed-form solution of absolute orientation using unit quaternions. Journal of the Optical Society of America A **4**(4) (1987) 629–642
22. Lee, G.H., Fraundorfer, F., Pollefeys, M.: RS-SLAM: RANSAC sampling for visual FastSLAM. In: Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems. (2011) 1655–1660