**Noname manuscript No.**
(will be inserted by the editor)

# Accurate Keyframe Selection and Keypoint Tracking for Robust Visual Odometry

**Marco Fanfani · Fabio Bellavia · Carlo Colombo**

Received: date / Accepted: date

**Abstract** This paper presents a novel stereo Visual Odometry (VO) framework based on Structure from Motion (SfM), where a robust keypoint tracking and matching is combined with an effective keyframe selection strategy. In order to track and find correct feature correspondences a robust loop chain matching scheme on two consecutive stereo pairs is introduced. Keyframe selection is based on the proportion of features with high temporal disparity. This criterion relies on the observation that the error in the pose estimation propagates from the uncertainty of 3D points—higher for distant points, that have low 2D motion. Comparative results based on three VO datasets show that the proposed solution is remarkably effective and robust even for very long path lengths.

**Keywords** Visual Odometry · Structure from Motion · RANSAC · feature matching · keyframe selection

## 1 Introduction

The real-time estimation of the camera trajectory and the construction of a 3D map of the scene, based on images acquired in an unknown environment, has received an increasing interest in the computer vision community during the last few years. This task is usually referred to as *visual Simultaneous Localization And Mapping* (vSLAM) [7]. vSLAM systems typically include a *Visual Odometry* (VO) module [26,30], aimed at the incremental estimation of the camera path using local information. Optimization algorithms over the whole estimated path and map can also be present, so as to enforce global consistency when revisiting part of the scene [15].

Related work on camera path estimation—which is the main topic addressed in this paper—is discussed hereafter.

### 1.1 Related Work

Methods for real-time camera tracking are mainly based either on probabilistic frameworks [23,7] or on the SfM paradigm [26,24]. In the former case they employ Bayesian filtering techniques, such as the Extendend Kalman Filter (EKF), to couple together in the same process camera positions and 3D points, incrementally updated. On the other side, the latter approaches exploit the epipolar geometry constraints [14] to compute the camera positions and the 3D map through robust estimators, such as the RANdom SAmple Consensus (RANSAC) [9]. Successive refinement steps are usually applied by iterative non-linear optimization techniques—such as bundle adjustment [35]—over a selected sub-set of frames (keyframes).

Both kinds of approaches have their 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, thus resulting in a loss of estimation accuracy. On the other hand, in order to achieve real-time operation, keyframe-based approaches can perform local optimizations only occasionally. Nevertheless, according to [33], keyframe based solutions outperform Bayesian approaches, due to their ability to maintain more 3D points in the estimation procedure.

M. Fanfani · F. Bellavia · C. Colombo
Computational Vision Group (CVG)
University of Florence, Via S. Marta, 3, Florence, Italy
E-mail: {name}.{surname}@unifi.it
Tel: +39 055 275 8509 - Fax: +39 055 275 8570

Stereo configurations have been widely used [26,27, 20,22,11]. In general, stereo systems provide better solutions than single camera setups, since the rigid calibration of the cameras increases the accuracy of the 3D map and provides more robust matches. This avoids issues such as the delayed 3D feature initialization [23] (i.e. when a point is seen for the first time) and the scale factor uncertainty [34].

Camera tracking systems can also be characterized by the feature matching strategy adopted to detect and track keypoints across the image frames [6,17,25, 29,18,11]. In addition to the classical keypoint matching and tracking methods [17], solutions robust to high degrees of blur [29], relying on edges [18], with hierarchical pose refinement [32], or exploiting the high computational power offered by modern GPUs through a dense approach [25] have been proposed. Effective stereo matching strategies [11] and sequentially overlapping 3D maps [8] have also been employed.

### 1.2 Our Contribution

The main idea of the proposed system, named SSLAM, is to use only highly reliable data in the estimation process, as reflected mainly in the feature matching scheme and the choice of good frames.

The feature matching process is the main source of noise in a camera tracking system. Wrong matches can lead to erroneous estimates, which can be only partially corrected using robust outlier rejection strategies. To limit as much as possible the introduction of errors in early processing stages, we choose to employ an accurate and relatively slow matching strategy instead of less accurate solutions. In particular, a robust loop chain matching scheme is adopted, improving upon VISO2-S [11], for using a more robust detector-descriptor pair. The adopted robust matching strategy avoids upfront the introduction of strong noise and thus the need of further global optimization steps. In addition, this strategy can find correspondences also in images with high spatial and/or temporal disparity—a critical issue for any approach based on tracking [20].

The other aspect characterizing our system is the selection of the keyframes used as base references for the measurement of the 3D landmark positions for the camera trajectory computations. Keyframes are selected only if a strong feature temporal disparity is detected. This idea arises from the observation that errors may propagate also from the uncertainty of the 3D points, which is higher for distant points corresponding to low temporal flow matches in the images. The proposed strategy can be more stable and effective with respect to using a threshold on the average temporal disparity [19] or a constant keyframe interleaving [26]. Moreover, evaluating 2D measures such as the feature temporal flow leads to a more robust keyframe selection compared to approaches that evaluate the distance among frames in 3D space [11].

This paper significantly extends our previous work [2], by providing a detailed description of the proposed method in Sect. 2, followed by a comprehensive evaluation and comparison on the KITTI [10], New College [31] and New Tsukuba [21] datasets in Sect. 3. Conclusions and final remarks are given in Sect. 4.

## 2 Method overview

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 (see Fig. 1). The first step matches keypoints between the last keyframe $f_i$ and the current frame $f_j$, while the second 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. If the new pose is successfully estimated and sufficient temporal disparity is detected between $f_j$ and $f_i$, the frame $f_j$ is updated as the new keyframe.

Assuming that $\mathrm{R}_{i,i} = \mathrm{I}$ and $\mathbf{t}_{i,i} = \mathbf{0}$ (where I and $\mathbf{0}$ are respectively the identity matrix and the null vector) the *absolute pose* at time $n$ is defined as $\mathrm{P}_n = \mathrm{P}_{0,n}$. $\mathrm{P}_n$ 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 $0 < k < \ldots < i < j$ belong to accepted keyframes and $n > j$ is the current frame.

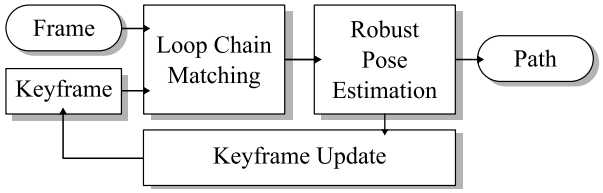


**Fig. 1:** Pipeline of the proposed method

### 2.1 Loop Chain Matching

The proposed loop chain matching draws inspiration from the *circle match* of VISO2-S [11], as the candidate correspondences should be consistent among the four image pairs $(I_i^l, I_i^r)$, $(I_i^l, I_j^r)$, $(I_i^r, I_j^r)$, $(I_j^l, I_j^r)$. However differently from [11], instead of a less accurate keypoint detector and descriptor based on simple image

filters, a robust detector and descriptor pair is used. This also avoids using the two step matching strategy employed by VISO2-S to further refine correspondences, and permits achieving longer and more stable keypoint tracks, crucial for the pose estimation, without re-initialization issues and keypoint losses occurring with tracking strategies such as KLT [20].

In particular, the HarrisZ detector [3], 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 with the sCOr Nearest Neighbour matching [4] on the $L_1$ distance is used instead 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 (see hereafter).

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 $\delta_y$ is the error band allowed by epipolar rectification and $\delta_x$ is the maximum allowed disparity (i.e. the corresponding stereo point must lie inside a $2\delta_x \times 2\delta_y$ rectangular window). In the case of a *temporal* match $(\mathbf{x}_i^d, \mathbf{x}_j^d)$ between corresponding images at different frames, the flow restriction

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

is taken into account, where $\delta_r$ is the maximum flow displacement (i.e. the corresponding point in the next frame must lie inside a circular window of radius $\delta_r$). Only matches that 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. 2); however, some outliers can still be present. For this reason, each matching pair of the loop chain $\mathcal{C}$ is further filtered by RANSAC to refine the matches. These four RANSAC runs have an almost immediate convergence due to the high presence of inliers. Only loop chains whose all pair matches survive to the four RANSACs are finally collected into the set $C_{i,j} \subseteq \{\mathcal{C}\}$.

## 2.2 Robust Pose Estimation

The relative pose $\mathrm{P}_{i,j}$ between $f_i$ and $f_j$ is estimated in the second step of the SSLAM approach (see again Fig. 2). The 3D point $\mathbf{X}_{i,j}$ corresponding to the match pair $(\mathbf{x}_i^l, \mathbf{x}_i^r)$ in keyframe $f_i$ can be estimated by triangulation [14], since the intrinsic and extrinsic calibration parameters of the system are known—in particular, we use the iterative linear triangulation method described in [13].

Let $\widetilde{\mathbf{x}}_j^l$ and $\widetilde{\mathbf{x}}_j^r$ be the projections of $\mathbf{X}_{i,j}$ onto frame $f_j$, according to the estimated relative pose $\mathrm{P}_{i,j} = [\mathrm{R}_{i,j}|\mathbf{t}_{i,j}]$. The distance

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

among the matches of the chain set $C_{i,j}$ must be minimized, in order for the estimate 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, where the number $\mathcal{D}_R(\mathrm{P}_{i,j})$ of outliers chain matches over $C_{i,j}$ exceeding a threshold value $\delta_t$ is minimized so that pose $\mathrm{P}_{i,j}$ be consistent with data:

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

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

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

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 [9]. 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 [16] 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 $\mathrm{P}_{i,j}$.

With respect to the pose estimation method described above, SSLAM filters the frame sequence according to the observation that the image resolution provides a lower bound to the uncertainty of the position of the keypoints used in the matching process, although subpixel precision is used. Matches 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 disparity and the associated 3D point position has a

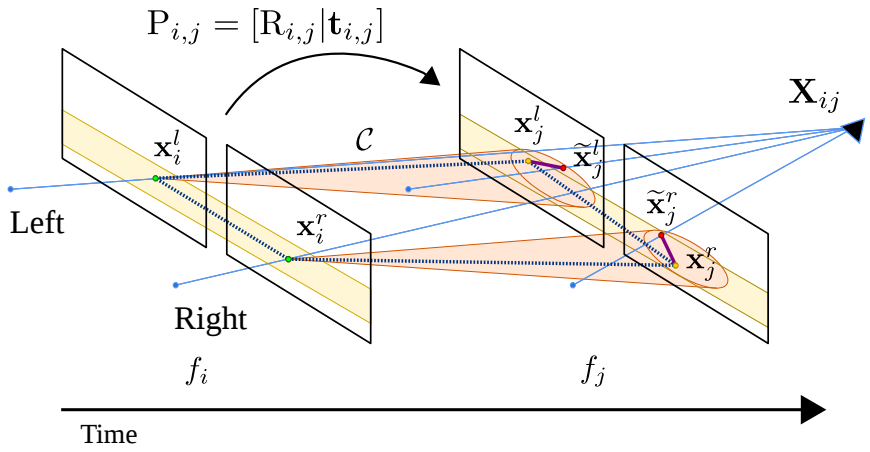


**Fig. 2:** (Best viewed in color) Keypoint matches between the keyframe $f_i$ and the new frame $f_j$ must satisfy the spatial constraint imposed by the epipolar rectification (yellow band) as well as the temporal flow restriction (orange cone). Furthermore, the four matching points must form a loop 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 $\mathbf{X}_{i,j}$ obtained by triangulation of $\mathbf{x}_i^l$, $\mathbf{x}_i^r$ in $f_i$ in order for the chain $\mathcal{C}$ to be consistent with the pose $\mathrm{P}_{i,j}$. However, due to data noise, in the real case it is required that the distances $\| \widetilde{\mathbf{x}}_j^l - \mathbf{x}_j^l \|$ and $\| \widetilde{\mathbf{x}}_j^r - \mathbf{x}_j^r \|$ are minimal

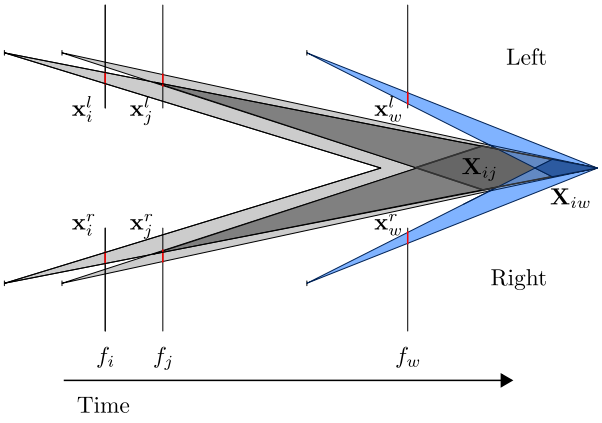


**Fig. 3:** (Best viewed in color) The uncertainty of matches in the image planes is lower bounded by the image resolution (red) and it is propagated to the 3D points. In order to estimate the 3D point $\mathbf{X}_{i,j}$, by using close frames $f_i$ and $f_j$, a low temporal disparity is present in the image planes, and the 3D point location $\mathbf{X}_{i,j}$ can assume a 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

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 both the translational and rotational motion, as shown in Fig. 3. This idea is a straight generalization of the well-known baseline length issues related to the trade-off between reliable correspondence matching and accurate point triangulation [14].

Exploiting this idea, SSLAM defines two subsets $F_{i,j}$ and $\bar{F}_{i,j}$ of the set of chain matches $C_{i,j}$ for $f_i$ and $f_j$ which respectively include *fixed* and *non-fixed* points with respect to the temporal flow, i.e.

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

$$\bar{F}_{i,j} = C_{i,j} \setminus F_{i,j} \ , \tag{10}$$

for a given threshold $\delta_f$. In order for frame $f_j$ to be accepted as new keyframe, the number of non-fixed matches between frames $f_i$ and $f_j$ must be sufficient according to a threshold $\delta_m$:

$$1 - \frac{|F_{i,j}|}{|C_{i,j}|} > \delta_m \quad . \tag{11}$$

Indeed, if the estimation fails due to wrong matches or high noisy data, which practically leads to a final small RANSAC consensus set $G_{\overline{\mathrm{P}}_{i,j}}$, the frame $f_j$ is discarded and the next frame $f_{j+1}$ is tested. We also tried to verify if the use of only non-fixed matches as input to RANSAC pose estimation can lead to better results, but no improvements were found, so the proportion of fixed and non-fixed points is only used for keyframe selection. This means that, while all matches are used to estimate the camera position, in presence of enough non-fixed matches, a higher accuracy can be achieved by limiting bad solutions. Note that for determining fixed and non-fixed matches flow vectors are considered, so that in the case of strong rotations and weak translations, even if points are far, their higher

flow would lead to better measurements and accuracy with respect to the minimal measuring unit, i.e. a pixel.

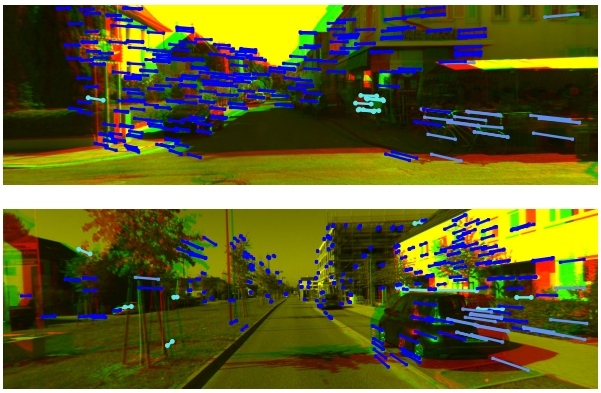

**Fig. 4:** (Best viewed in color) Examples of successive keyframes retained according to the temporal flow for two different sequences of the KITTI dataset. The two temporal keyframes involved are superimposed as for anaglyphs, only images for the left cameras are shown. Good fixed and unfixed matches are shown in blue and light blue, respectively, while wrong correspondences are reported in cyan

Examples of fixed point estimations are shown in Fig. 4. With respect to the average flow threshold commonly employed by other systems such as [19], our strategy is more stable and can handle better keyframe drops. As an example, referring to Fig. 4, the average flow in the top configuration is considerably higher than that of the bottom one. Lowering the threshold, to accept the bottom frame, would also include very low disparity frames (just consider to replace in the bottom frame the unfixed light blue matches by twice the matches with half disparity). In this sense, our measure is more robust, so that both the frames shown in the figure are retained as keyframes. In analogy, our frame selection resembles RANSAC while the average flow is close to the least-square approach.

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

$$\arccos\left(\mathbf{u}^{\mathrm{T}}\mathrm{R}_{i,j}^{\mathrm{T}}\mathrm{R}_{z,i}\mathbf{u}\right) < \delta_{\theta_1} \tag{12}$$

where $\mathbf{u} = \frac{1}{\sqrt{3}}[1\,1\,1]^{\mathrm{T}}$. Optionally, in the case of strong constrained movement, like that of a car, a further constraint on the corresponding translation directions $\mathbf{t}_{z,i}$ and $\mathbf{t}_{i,j}$ can be added

$$\arccos\left(\frac{\mathbf{t}_{i,j}{}^{\mathrm{T}}\mathbf{t}_{z,i}}{\| \mathbf{t}_{i,j} \| \| \mathbf{t}_{z,i} \|}\right) < \delta_{\theta_2} \tag{13}$$

This last constraint can also resolve issues in the case of no camera movement or when moving objects crossing the camera path cover the scene.

## 3 Experimental Evaluation

The KITTI vision benchmark suite [10], the New College sequence [31] and the New Tsukuba stereo dataset [21] were used to evaluate our SSLAM system.

Recently, the KITTI dataset has become a reference evaluation framework for VO systems. The dataset provides sequences recorded from car driving sessions on highways, rural areas and inside cities with vehicle speed up to 80 km/h. The benchmark consists of 22 rectified stereo sequences from 500 m to 5 km, taken at 10 fps with a resolution of $1241 \times 376$ pixels. Recorded scenes are not static, as moving vehicles in opposite direction or crossing the road are present. In order to train the parameters of the methods, ground truth trajectories are available only for the first 11 sequences. Results for the remaining sequences should be submitted online to get a final ranking. Translation and rotation errors normalized with respect to path length and speed are computed in order to rank the methods.

The New College dataset is made up of a very long sequence of 2.2 km for more than 50000 stereo rectified frames taken inside the Oxford New College campus using a Segway. Data were recorded at 20 fps with a resolution of $512 \times 384$ pixels. Although no reliable ground truth is available, the sequence consists of several different loops which can be used to qualitatively compare VO methods by visual inspection of estimated paths. Unlike the KITTI dataset, data are recorded at a lower speed and the camera movements are less constrained, i.e., strong camera shakes are present.

The New Tsukuba dataset is a virtual sequence that navigates into a laboratory reconstructed manually by computer graphics. Images with a resolution of $640 \times 480$ pixels are recorded at 30 fps for one minute while accurate ground truth positions are registered and provided to the users. The sequence is rendered with four different illuminations from the most classical *fluorescent* to the more challenging *flashlight* and *lamps*—see Fig. 5.

Unless otherwise specified, for SSLAM we set $\delta_f = 55$ px, $\delta_m = 5\%$, $\delta_{\theta_1} = 15°$ (see Sect. 2.2). About the spatial and temporal constraints, the triplet $(\delta_r, \delta_x, \delta_y)$ is set to $(500, 300, 12)$ px in the case of the KITTI

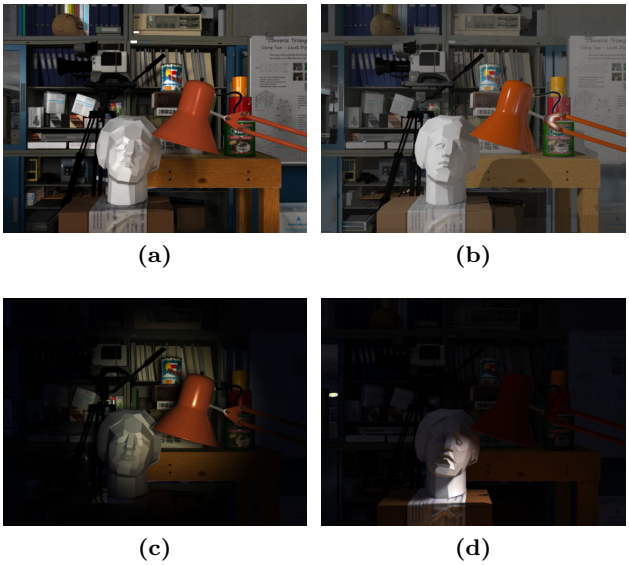

**Fig. 5:** Example frames of the New Tsukuba stereo dataset: (a) *fluorescent*, (b) *daylight*, (c) *flashlight*, (d) *lamps*

dataset and to $(100, 100, 12)$ px for the New College and New Tsukuba dataset, since these videos are taken at lower resolutions and baselines. In the rest of the evaluation we will mainly compare SSLAM against VISO2-S as it is the only method whose authors have kindly replied to our request to provide us their full code. Note that, for the sake of comparison, VISO2-S $(\delta_r, \delta_x, \delta_y)$ values are chosen as $(200, 200, 3)$ px (default values) for KITTI and $(100, 100, 3)$ in the New College dataset, where the latter values perform better than the default values. The translation constraint is $\delta_{\theta_2} = 10°$ for the KITTI dataset while it is not used for New College—due to high camera shakes—and for the New Tsukuba sequences.

Furthermore, we tested SSLAM using keypoints detected at full and half resolution videos; in the latter case, the notation SSLAM$^\dagger$ is used. In the case of SSLAM$^\dagger$ less accurate keypoints are found, with bigger (normalized) feature patches, more sensitive to fast camera movements. Note also that more keypoints are found in full resolution SSLAM implementation than with SSLAM$^\dagger$. Nevertheless, different image resolutions do not affect the other parameters of the methods since keypoint positions are rescaled at the full resolution before the constrained matching in both cases.

### 3.1 SSLAM Parameter Analysis

We compared different versions of our SSLAM system, corresponding to the successive improvements of the pipeline proposed in Sect. 2, in particular we analysed different versions of the more challenging SSLAM$^\dagger$. We indicated by SSLAM$^{\dagger\star}$ the first version which only includes the loop chain matching described in Sect. 2.1, while the adaptive keyframe selection is incorporated in the default SSLAM$^\dagger$.

In order to analyse the robustness and the effectiveness of the proposed method, the SSLAM$^\dagger$ system was tested with a different number of RANSAC iterations for the pose estimation. In particular, results of SSLAM$^\dagger$ with 500, 15 (set as default) and 3 RANSAC iterations, and SSLAM$^{\dagger\star}$ with 500 iterations are presented, indicated respectively by SSLAM$^\dagger$/500, SSLAM$^\dagger$/15, SSLAM$^\dagger$/3 and SSLAM$^{\dagger\star}$/500.

Figure 6 shows the average translation and rotation errors of the different SSLAM$^\dagger$ variants for increasing path length and speed, according to the first 11 sequences of the KITTI dataset [10]; we verified that similar results hold in the case of full resolution SSLAM.

The chain loop matching scheme together with the chosen keypoint detector and descriptor is robust even for long paths, without bundle adjustment or loop closure detection. SSLAM improves on the standard pose estimation without keyframes selection, allowing to track longer paths and confirming that the proposed keyframe selection strategy is effective.

Moreover, results for SSLAM$^\dagger$/15 and SSLAM$^\dagger$/500 are equivalent, while SSLAM$^\dagger$/3 obtains inferior results but similar to those obtained by SSLAM$^{\dagger\star}$/500, giving an evidence of the robustness of the proposed matching selection strategy and pose estimation.

A further test aiming at investigating the fixed point threshold $\delta_f$ used to accept a frame as keyframe was also done. This is the parameter that mainly affects the results, since selected keyframes decrease as $\delta_f$ increases, while we verified that the computation is stable with respect to the choice of the other parameters. Note that increasing $\delta_m$ can be considered similar to require a higher $\delta_f$ for a lower $\delta_m$ value. In particular, we run SSLAM for different values of $\delta_f = 30$, 50 (default), 80 px on KITTI and New College datasets. In the case of $\delta_f = 30$ slightly inaccurate paths are present with respect to $\delta_f = 50$ on both datasets, while for $\delta_f = 80$ higher pose errors are found.

Figure 7 shows the behaviour of SSLAM for the different values of $\delta_f$ on the New College sequence. Clearly the default set $\delta_f = 50$ px provides better results since even after a long path loops are correctly closed. This results confirm the observation that avoiding close keyframes improves the results, but this choice must be balanced with the tracking capability of the system. Moreover, if the system is unable to estimate the pose due to a low number of matches and tracking

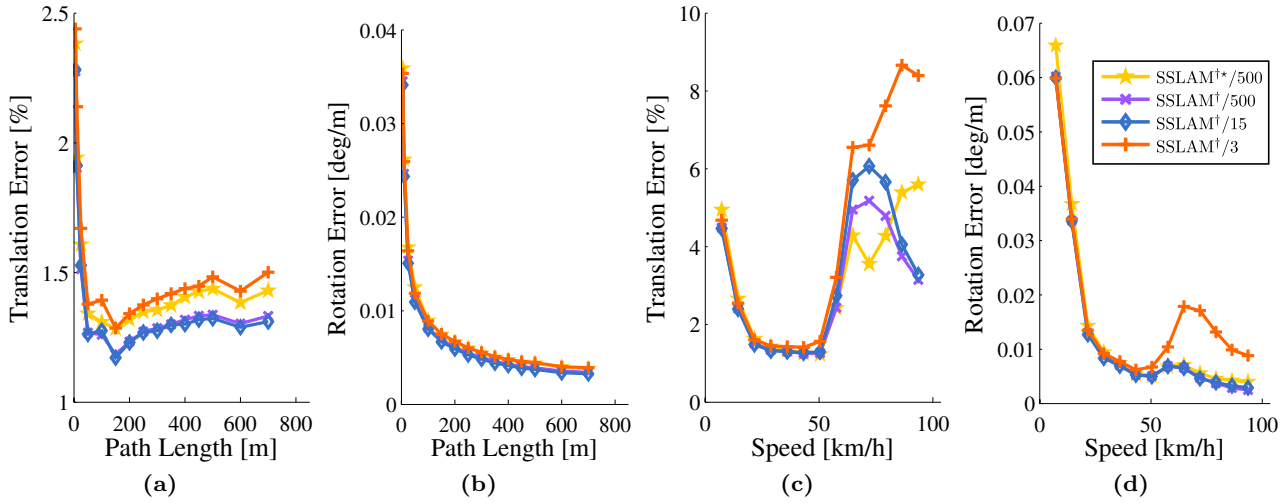


**Fig. 6:** (Best viewed in color) Average error on the first 11 sequences of the KITTI dataset. Plots (a-b) refer to the average translation and rotation error for increasing path length respectively, while plots (c-d) refer to increasing speed

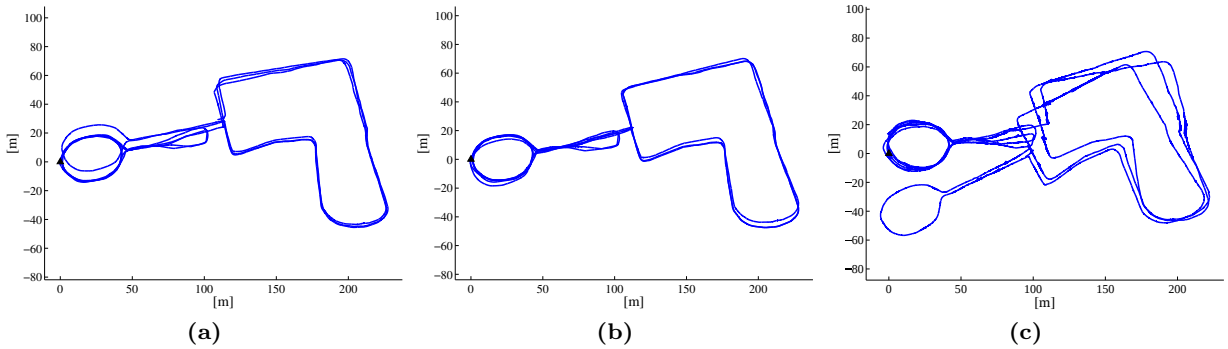


**Fig. 7:** SSLAM estimated paths for the New College video sequence with $\delta_f = 30$ px (a), $\delta_f = 50$ px (b) and $\delta_f = 80$ px (c)

loss, a recovery method must be implemented as for any other VO methods.

**Table 1:** Average number of frames between two consecutive keyframes and the corresponding standard deviations for different values of the threshold $\delta_f$

| $\delta_f$ | 35 | 55 | 85 | 35 | 55 | 85 |
|---|---|---|---|---|---|---|
| | Average | | | Std | | |
| KITTI | 1 | 2 | 3 | 1 | 1 | 2 |
| New College | 5 | 10 | 32 | 8 | 13 | 39 |

Table 1 shows the average number of frames between two consecutive keyframes and the corresponding standard deviations. Average keyframe rate depends upon $\delta_f$ and $\delta_m$ but also on the camera speed and the video frame rate. Slower camera speed and/or higher frame rate imply a lower keyframe rate, but on the other hand $\delta_f$ and $\delta_m$ depend also on the scene. According to Table 1, the average keyframe rate is lower for the New College dataset than for the KITTI dataset, due to their different camera speeds. Furthermore, as it can be noted in Fig. 8, the keyframe distribution is not uniform but it is denser near camera turns and accelerations.

### 3.2 Evaluation on the KITTI dataset

We report hereafter the results on the KITTI odometry benchmark for stereo methods only (more details are available online [10]) excluding methods that rely on laser data. Figure 9 shows the average translation and rotation errors of the different methods for increasing path length and speed. SSLAM and SSLAM$^\dagger$—ranked

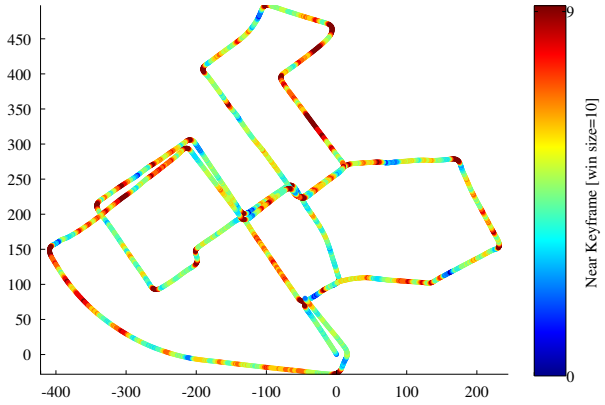


**Fig. 8:** (Best viewed in color) An example of keyframe distribution along the Sequence 00 of the KITTI dataset for SSLAM (default $\delta_f$=50 px). At each estimated camera position the number of keyframes that fall inside a window of 10 frames centred at the camera location is shown according to the colorbar gradation

among the first positions of the KITTI benchmark—obtain respectively a mean translation error of 1.57% and 2.14% w.r.t. the sequence length and a rotation error of 0.0044 and 0.0059 deg/m. These rank placements show the robustness of the proposed methodology. Note however that the benchmark provides partial results, since these error metrics cannot take into account all the properties of a VO system. In particular, referring to Fig. 10 where two sample tracks of the KITTI dataset are shown, it can be seen that while both MFI and VoBa (respectively ranked in $1^{st}$ and $4^{th}$ positions) provide slightly better results than SSLAM in term of KITTI metrics, on long paths SSLAM$^{\dagger}$ ($12^{th}$ ranked) clearly improves on the $7^{th}$ ranked eVO method. This can also be observed in the relative translation error for an increasing path length in Fig. 9(a), where SSLAM$^{\dagger}$ plot remains stable when compared to the increasing error of eVO. Additionally, evaluation on the KITTI sequence shows that our approach is robust in the case of non static scenes for common situations with other objects traveling in other directions. In particular, the $\delta_{\theta_2}$ constraint (see equation 13), allows to "remember" the recent past camera tracking data, thus avoiding to fall into wrong configurations in the case of Sequence 07 around frame 700, where a huge truck occupying nearly the whole scene crosses the road while the camera stands still.

Table 2 shows the input matches and the found inliers in the RANSAC pose estimation by SSLAM, SSLAM$^{\dagger}$ and VISO2-S. As it can be noted, while SSLAM$^{\dagger}$ outputs a comparable number of initial matches with VISO2-S, only 50% of these are inliers for VISO2-S: This implies that our matching strategy

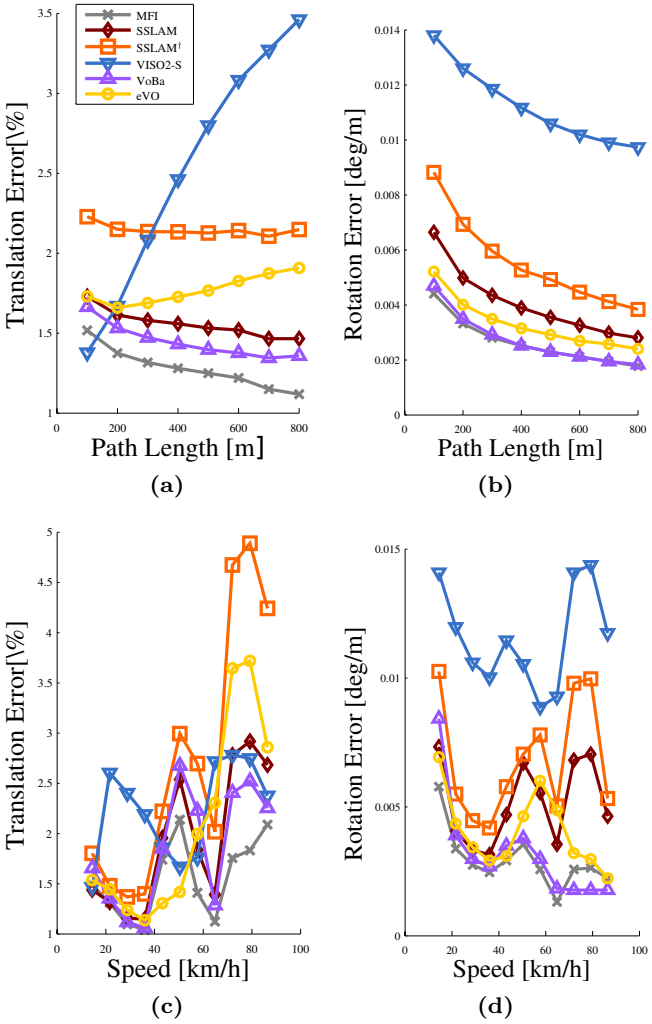


**Fig. 9:** (Best viewed in color) Average error on the KITTI benchmark. Plots (a-b) refer to the average translation and rotation error for increasing path length respectively, while plots (c-d) refer to increasing speed

is more robust. Note also that the spatial and temporal flow constraints of VISO2-S are stricter, which would lead theoretically to a higher number of matches since the probability to make an accidental wrong match is higher for SSLAM and SSLAM$^{\dagger}$ (except for the epipolar constraint $\delta_y$, the other thresholds are about equal to the minimal image size). Yet, as it can be seen from Table 2, the opposite holds, in favour of the robustness and stability of the proposed methodology.

### 3.3 Evaluation on the New College dataset

We tested SSLAM and SSLAM$^{\dagger}$ versus VISO2-S not only on the whole sequence but also on the two subsequences corresponding to the small and large loops

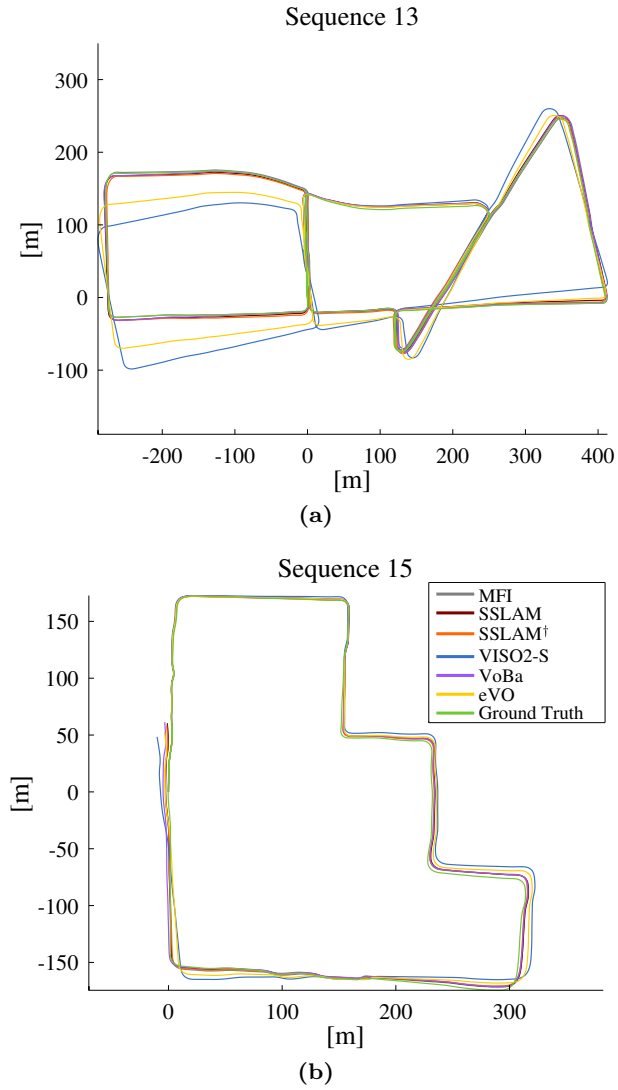


**Fig. 10:** (Best viewed in color) Trajectories on the sequences 13 (a) and 15 (b) of the KITTI dataset

**Table 2:** Average number of input matches before the RANSAC pose estimation and final inlier ratios

| | KITTI | | New College | |
|---|---|---|---|---|
| | pts | inl(%) | pts | inl(%) |
| SSLAM | 766 | 98 | 780 | 99 |
| SSLAM† | 222 | 96 | 201 | 97 |
| VISO2-S | 245 | 50 | 156 | 84 |

present in the sequence. This is done to analyse the behaviour of the methods at different starting points. Figure 11 shows the obtained tracks. While VISO2-S diverges as the sequence grows, both SSLAM and SSLAM† maintain the correct paths, closing the loops, without the need of bundle adjustment and loop closure techniques. In particular, full resolution SSLAM works slightly better than SSLAM†. This becomes noticeable only at the end of the last part of the video sequence.

The New College video sequence seems more reliable than the KITTI sequences, since as it can be seen from Table 2, all methods achieve a higher number of tracked keypoints but also inliers, maybe due to slower camera movements. Anyway, VISO2-S still obtains a lower number of matches and inliers with respect to SSLAM and SSLAM†. Note also that the absence of the optionally translation constraint $\delta_{\theta_2}$ in this sequence does not affect the quality of the results.

### 3.4 Evaluation on the New Tsukuba dataset

In order to investigate further into the robustness of our method, we tested SSLAM on the New Tsukuba sequence for all the available illuminations. In Fig. 12 estimated trajectories are reported together with the ground truth: Even if slight misalignments are present—especially in the first part of the *flashlight* sequence—SSLAM tracks well the camera movements for all illuminations. This is also clear by observing the translation and rotation errors in Fig. 13—computed using the KITTI metrics. For all illuminations, similar performance is obtained.

As already done for the KITTI and New College sequences, also for the New Tsukuba dataset we tested SSLAM on half-resolution images obtaining approximately the same errors reported for the full-resolution tests. In Fig. 14 are shown trajectories of both the full and half resolution input for the *fluorescent* sequence; Similar results are obtained for all the other illuminations.

It's worth noting that the *fluorescent* sequence of the New Tsukuba dataset was also used in [1] to evaluate the MFI method. Apart from a visual comparison between the trajectories estimated by MFI and SSLAM, from which no particular differences emerge, it is not possible to make a quantitative evaluation, since in [1] only relative accuracy improvements w.r.t. a base method are reported.

### 3.5 Running Times

The SSLAM approach is implemented in C/C++ non-optimized multithreaded code, for which the download is available[1]. As it can be seen in Table 3, where the average running times for a single frame are reported, SSLAM scales with the resolution. By taking into account that only keyframes are required by SSLAM,

[1] https://drive.google.com/open?id=0B_3Nh0OK9BclM0I5VC1jNndTSTA

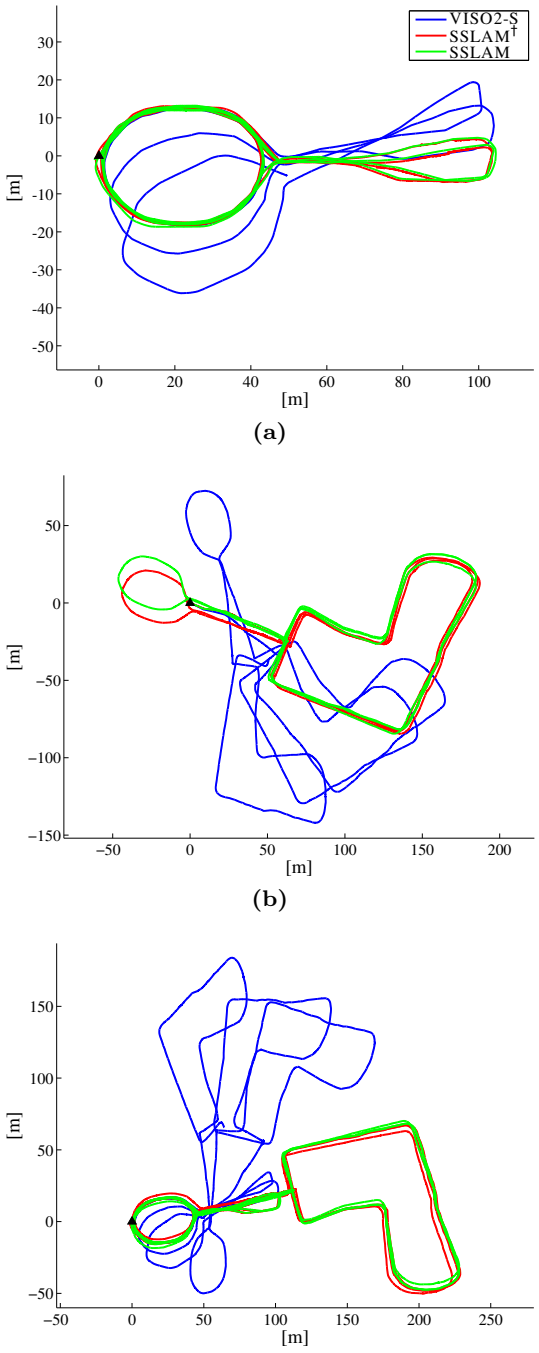


**Fig. 11:** (Best viewed in color) Estimated paths for the New College video sequence. The plots (a), (b) and (c) refer respectively to first subsequence (from frame 0 to frame 18400), to the last subsequence (from frame 18400 to frame 52479) and to the whole sequence. Note that to achieve the best top view, each sequence was rotated so that the displayed axes correspond to the major directions of the autocorrelation matrix of the point positions, i.e., to the two greatest eigenvectors

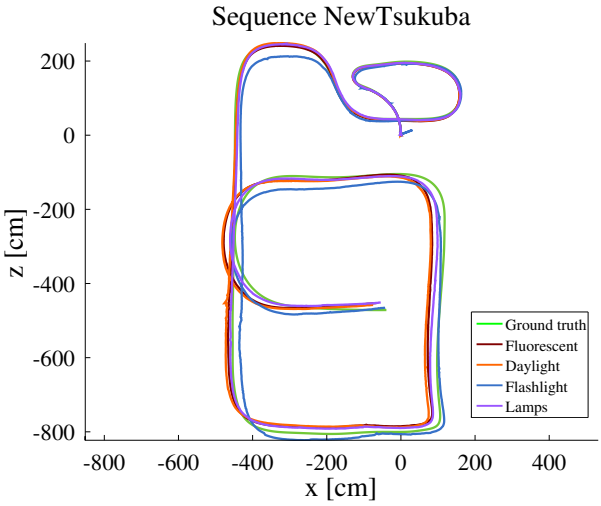


**Fig. 12:** (Best viewed in color) Trajectories estimated by SSLAM on the New Tsukuba sequence for all available illuminations

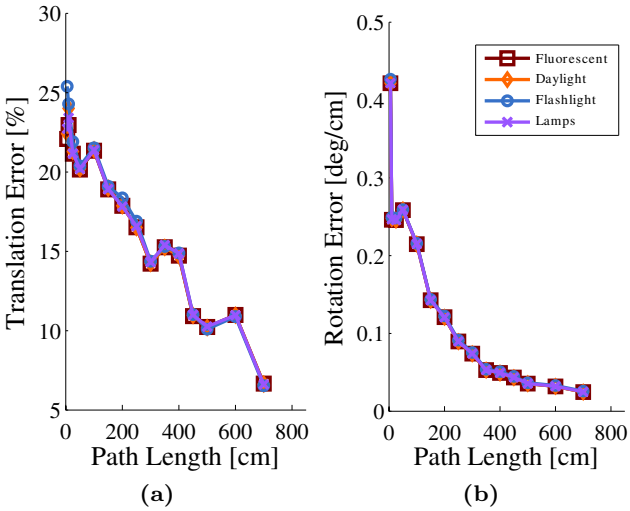


**Fig. 13:** (Best viewed in color) Average translation (a) and rotation (b) error for increasing path length for all New Tsukuba sequence illuminations

real-time performance is achieved when the keyframe computational time is less than $f_k/f_v$, where $f_k$ is the keyframe rate and $f_v$ is the frame rate of the video sequences ($[f_k, f_v]$ are respectively for the KITTI, New College and New Tsukuba datasets equal to $[2, 10]$, $[10, 20]$ and $[5, 30]$). This implies that the time to estimate a single keyframe must not exceed 0.20 s, 0.50 s and 0.17 s respectively for the KITTI, New College and New Tsukuba sequences. Although only SSLAM† can run almost in real-time, code optimization using GPU acceleration is planned to improve the running times. Furthermore, we found that the main bottleneck of the

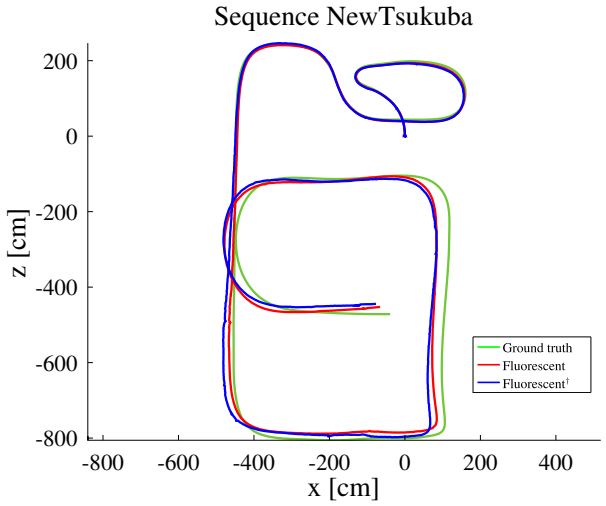


**Fig. 14:** (Best viewed in color) Trajectories estimated using full (Fluorescent, red track) and half (Fluorescent†, blue track) resolution images of the New Tsukuba *fluorescent* sequence

method is represented by the large size kernel convolutions employed by the accurate feature detector. Under this observation, further speed improvements should be achieved from fast and approximate convolution algorithms [12].

**Table 3:** Average computational time for a single frame on a Intel-i7 3.50GHz CPU, 8 cores are used

| | SSLAM | SSLAM† |
|---|---|---|
| KITTI | 3.85 s | 0.55 s |
| New College | 0.95 s | 0.20 s |
| New Tsukuba | 0.87 s | 0.24 s |

## 4 Conclusion

In this paper a new stereo VO system was presented. The approach achieves a low drift error even for long paths, is local and it does not rely 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 disparity frames for discarding highly uncertain models is an effective strategy to reduce error propagation from matches, but it must be balanced to avoid the loss of keypoint tracks across the video sequence. Results validated on the KITTI, New College and New Tsukuba datasets show the effectiveness of the system, which is robust even with an extremely small number of RANSAC iterations and able to work in various scenarios under different illuminations.

Future work will include an efficient optimized code to improve real-time performance and possible integrations with information from other sensors to improve the accuracy of the localization. Our work mainly focuses on strengthening the data retrieving and filtering phases, and relies on a well-know and simple pose estimation method [11]. Further future work will be addressed to extend and improve our VO pipeline by including novel numerical optimization techniques exploiting long tracks based on Local Bundle Adjustment [5,28] and tracking recovery mechanism to increase the SSLAM reliability.

## 5 Acknowledgments

This work was supported by the SUONO project (Safe Underwater Operations iN Oceans), SCN_00306, ranked first in the challenge on "Sea Technologies" of the competitive call named "Smart Cities and Communities" issued by the Italian Ministry of Education and Research.

## References

1. Badino, H., Yamamoto, A., Kanade, T.: Visual odometry by multi-frame feature integration. In: Proc. of the International Workshop on Computer Vision for Autonomous Driving at ICCV (2013)
2. Bellavia, F., Fanfani, M., Pazzaglia, F., Colombo, C.: Robust selective stereo SLAM without loop closure and bundle adjustment. In: Proc. of 17th International Conference on Image Analysis and Processing, ICIAP 2013, pp. 462–471 (2013)
3. Bellavia, F., Tegolo, D., Valenti, C.: Improving Harris corner selection strategy. IET Computer Vision **5**(2) (2011)
4. Bellavia, F., Tegolo, D., Valenti, C.: Keypoint descriptor matching with context-based orientation estimation. Image and Vision Computing (2014)
5. Cvisic, I., Petrovic, I.: Stereo odometry based on careful feature selection and tracking. In: Proc. of the European Conference on Mobile Robots ECMR, pp. 1–6 (2015)
6. Davison, A.: Real-time simultaneous localization and mapping with a single camera. In: Proc. of the 9th IEEE International Conference on Computer Vision, pp. 1403–1410 (2003)
7. Davison, A., Reid, I., Molton, N., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE Trans. on Pattern Analysis and Machine Intelligence **29**(6), 1052–1067 (2007)
8. Fanfani, M., Bellavia, F., Pazzaglia, F., Colombo, C.: SAMSLAM: Simulated annealing monocular SLAM. In: Proc. of 15th International Conference on Computer Analysis of Images and Patterns, CAIP 2013, pp. 515–522 (2013)

9. 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), 381–395 (1981)
10. 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). URL `http://www.cvlibs.net/datasets/kitti/eval_odometry.php`
11. Geiger, A., Ziegler, J., Stiller, C.: StereoScan: Dense 3D reconstruction in real-time. In: IEEE Intelligent Vehicles Symposium (2011)
12. Getreuer, P.: A survey of gaussian convolution algorithms. Image Processing On Line **3**, 286–310 (2013)
13. Hartley, R., Sturm, P.: Triangulation. Computer Vision and Image Understanding **68**(2), 146–157 (1997)
14. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004)
15. Ho, K., Newman, P.: Detecting loop closure with scene sequences. International Journal Computer Vision **74**(3), 261–286 (2007)
16. 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)
17. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Proc. of the IEEE/ACM International Symposium on Mixed and Augmented Reality, pp. 225–234 (2007)
18. Klein, G., Murray, D.: Improving the agility of keyframe-based SLAM. In: Proc. of the 10th European Conference on Computer Vision, pp. 802–815 (2008)
19. 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, pp. 1655–1660 (2011)
20. Lim, J., Pollefeys, M., Frahm, J.M.: Online environment mapping. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (2011)
21. Martull, S., Martorell, M.P., Fukui, K.: Realistic CG Stereo Image Dataset with Ground Truth Disparity Maps. In: Proc. of the ICPR2012 workshop TrakMark2012, pp. 40–42 (2012). URL `http://www.cvlab.cs.tsukuba.ac.jp/dataset/tsukubastereo.php`
22. 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)
23. Montiel, J., Civera, J., Davison, A.: Unified inverse depth parametrization for monocular SLAM. In: Proc. of Robotics: Science and Systems. IEEE Press (2006)
24. Mouragnon, E., Lhuillier, M., Dhome, M., Dekeyser, F., Sayd, P.: Real time localization and 3D reconstruction. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 363–370 (2006)
25. Newcombe, R., Lovegrove, S., Davison, A.: DTAM: Dense tracking and mapping in real-time. In: Proc. of the 13th International Conference on Computer Vision (2011)
26. Nistér, D., Naroditsky, O., Bergen, J.R.: Visual odometry. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–659 (2004)
27. Paz, L., Piniés, P., Tardós, J., Neira, J.: Large-scale 6-DoF SLAM with stereo-in-hand. IEEE Trans. Robotics **24**(5), 946–957 (2008)
28. Persson, M., Piccini, T., Felsberg, M., Mester, R.: Robust stereo visual odometry from monocular techniques. In: Proc of the IEEE Intelligent Vehicles Symposium IV2015, pp. 686–691 (2015)
29. Pretto, A., Menegatti, E., Bennewitz, M., Burgard, W.: A visual odometry framework robust to motion blur. In: Proc. of the IEEE International Conference on Robotics and Automation (2009)
30. Scaramuzza, D., Fraundorfer, F.: Visual Odometry: Part I - The First 30 Years and Fundamentals. IEEE Robotics and Automation Magazine **18**(4) (2011)
31. Smith, M., Baldwin, I., Churchill, W., Paul, R., Newman, P.: The new college vision and laser data set. The International Journal of Robotics Research **28**(5), 595–599 (2009). URL `http://www.robots.ox.ac.uk/NewCollegeData/`
32. 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)
33. Strasdat, H., Montiel, J., Davison, A.: Visual SLAM: Why filter? Image and Vision Computing **30**, 65–77 (2012)
34. Strasdat, H., Montiel, J.M.M., Davison, A.J.: Scale drift-aware large scale monocular SLAM. In: Proc. of Robotics: Science and Systems (2010)
35. 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)