

EGO-SLAM: A Robust Monocular SLAM for Egocentric Videos

Suvam Patra Kartikeya Gupta Faran Ahmad Chetan Arora Subhashis Banerjee
Indian Institute of Technology Delhi

Abstract

Regardless of the tremendous progress, a truly general purpose pipeline for Simultaneous Localization and Mapping (SLAM) remains a challenge. We investigate the reported failure of state of the art (SOTA) SLAM techniques on egocentric videos [24, 40, 42]. We find that the dominant 3D rotations, low parallax between successive frames, and primarily forward motion in egocentric videos are the most common causes of failures. The incremental nature of SOTA SLAM, in the presence of unreliable pose and 3D estimates in egocentric videos, with no opportunities for global loop closures, generates drifts and leads to the eventual failures of such techniques. Taking inspiration from batch mode Structure from Motion (SFM) techniques [4, 55], we propose to solve SLAM as an SFM problem over the sliding temporal windows. This makes the problem well constrained. Further, as suggested in [4], we propose to initialize the camera poses using 2D rotation averaging, followed by translation averaging before structure estimation using bundle adjustment. This helps in stabilizing the camera poses when 3D estimates are not reliable. We show that the proposed SLAM technique, incorporating the two key ideas works successfully for long, shaky egocentric videos where other SOTA techniques have been reported to fail. Qualitative and quantitative comparisons on publicly available egocentric video datasets validate our results.

1. Introduction

Egocentric or first-person cameras [15, 16, 29] are wearable cameras, typically harnessed on a wearer’s head. First person perspective coupled with always-on nature have made these cameras popular in extreme sports, law enforcement, life-logging, home automation, and assistive vision applications [1, 2, 5, 11, 12, 22, 28, 36, 39].

Simultaneous Localization and Mapping (SLAM) has received a lot of attention from computer vision researchers over the years. While a variety of strategies including incremental [19, 9, 32, 7, 8, 26, 31, 47, 49, 51], hierarchical [50] and global [55] approaches have been proposed, the incremental ones remain popular because of their efficiency and

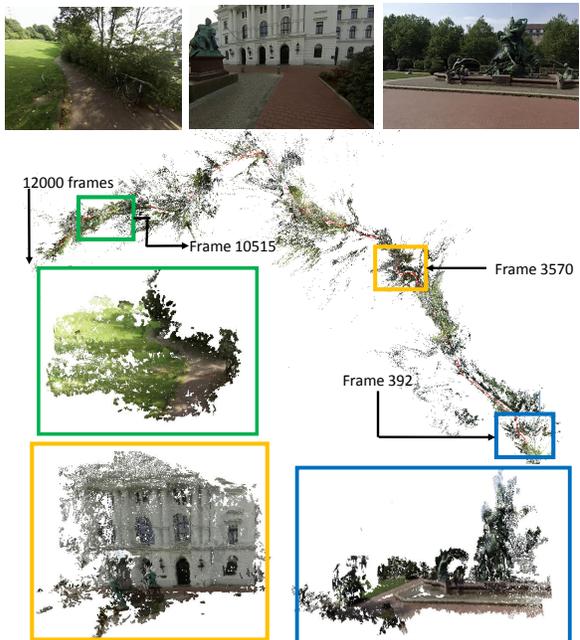


Figure 1: Incremental nature of state of the art SLAM [32, 9, 19] as well as SFM [56, 55, 50] techniques are unsuitable for extremely unstable egocentric video when the pairwise camera pose and 3D estimates are unreliable. We propose a robust SLAM (EGO-SLAM) which solves the SLAM as an SFM problem over sliding temporal windows. The SFM problem is solved globally over the window, by first stabilizing poses using rotation and translation averaging, before going for bundle adjustment. The figure shows 3D point clouds and trajectory estimated, using the proposed algorithm, over 12000 frames from a Hyperlapse [24] bike07 sequence where all the other SLAM and SFM techniques have been reported to fail.

scalability. Such approaches pick one frame at a time from a video stream and estimate the camera poses with respect to the 3D structure obtained so far, especially from the last few frames. Some techniques use an additional loop closure step to recognise the places visited earlier and refine the pose estimates over the loop so as to make them consistent at the intersection [54]. While the existing systems have advanced the state of the art tremendously, robustness and accuracy remain the key problems in incremental SLAM that prevent their use as general-purpose methods.

In this paper, we investigate the monocular SLAM problem with a special emphasis on **EGO**centric videos, and propose a new robust **SLAM** which we call **EGO-SLAM**. We observe that the incremental nature of the state of the art (SOTA) SLAM techniques are unsuitable for the egocentric videos. Sharp head rotations and primarily forward motion in egocentric videos cause quick changes in the camera view and result in short and noisy feature tracks. This combined with low parallax due to dominant 3D rotation because of natural head motion of the wearer causes triangulation errors in both feature-based [32] and direct techniques [9]. Relying on such 3D points causes drifts in the estimated camera trajectories leading to failure of the whole pipeline. The failure of the SOTA SLAM techniques on egocentric videos have been observed and reported by various research groups working in the area [24, 40, 42].

One of the key insights from the current work is to make the SLAM problem over egocentric videos better constrained by solving it as an SFM problem over a temporal window. Even without a careful analysis, Kopf et al. [24] also did something similar, when they carried out bundle adjustment [56] over large batches of 1400 frames, thereby making the problem well conditioned. However, merely doing SFM over batches is not sufficient. The idea behind batches is to make the problem better constrained, and incremental SFM techniques such as VisualSFM [56] over the batches defeat the core idea. In our experiments, and also reported by other research groups, [24, 40, 42], batch mode visual SFM works better for egocentric videos compared to incremental SLAM, but still does not solve the problem. The second key insight of this paper is to suggest solving the batch SFM problem globally. We adopt the technique proposed by Bhowmik et al. [4], which first computes pairwise estimates, followed by 2D based rotation averaging, and translation averaging. The pose estimates thus obtained are fed to the bundle adjustment for joint pose and structure estimation over a temporal window. This has two advantages. First, it avoids the noisy estimates of incremental SFM. Second, motion averaging does not require use of 3D estimates which may be erroneous at this stage. This helps to stabilize the pose estimates and obtain improved initialization for bundle adjustment (BA) which is crucial for obtaining good solutions using a BA algorithm.

We note that the typical motion profiles of handheld videos, even if unstable, are not similar to that of egocentric videos. In videos obtained from handheld cameras, there are typically no dominant 3D rotations and a user often looks at the same scene from multiple viewpoints in a scanning motion. As a result, the generated feature tracks are longer. Multiple scans give enough opportunities for loop closures making such problems simpler in comparison.

Though, the focus of this paper is on egocentric videos, the resulting framework is also suitable for other scenarios

having low parallax (leading to unstable 3D estimate) and lack of loop closure opportunities, where SOTA incremental SLAM algorithms are unsuitable. One such example is videos are taken from a vehicle-mounted camera. We show that our algorithm, though not specifically designed for such situations, improves upon the SOTA on such videos as well.

The specific contributions of this paper are:

1. Our analysis on the failure of existing SLAM techniques for egocentric videos: We posit that computing geometry from unreliable pose estimation in an incremental fashion is the primary cause of such failures.
2. Our two key novel proposals: We suggest solving the first person SLAM as an SFM problem over temporal windows and solve each window/batch as global SFM with initialization based on 2D motion averaging. The motivation for the specific suggestions have been described above.
3. Though not specifically the focus of this paper, we also test EGO-SLAM for the vehicle-mounted cameras where similar situations exist. We show that EGO-SLAM improves the SOTA there as well.

We contrast the proposed system to Bundler [46, 45] and VisualSFM [56], which are general purpose SFM pipelines. Our EGO-SLAM is specific SLAM pipeline for egocentric videos. Our experiments over a large set of publicly available egocentric datasets show the success, where all other SOTA SLAM algorithms have been shown to fail (Fig. 1 shows one such result). Thus, the proposed technique closes a long standing, open problem in egocentric vision and will prove to be helpful to the community.

2. Related Work

Based on the method of feature selection for pose estimation, SLAM algorithms from a monocular camera can be classified as feature-based, dense, semi-dense or hybrid methods. Feature-based methods, both filtering based [25] and key frame based [19, 23, 32], use sparse features like SIFT [27], ORB [43], SURF [3], etc. for tracking. The sparse feature correspondences are then used to refine the pose using structure-from-motion techniques like bundle adjustment. Due to the incremental nature of all these approaches, a large number of points are often lost during the resectioning phase [52].

Dense methods initialize the entire or a significant portion of an image for tracking [33]. The camera poses are estimated in an expectation maximization framework, where in one iteration the tracking is improved through pose refinement by minimizing the photometric error, and, in alternate iterations, the 3D structure is refined using the improved tracking. To increase the accuracy of estimation,

semi-dense methods perform photometric error minimization only in regions of sufficient gradient [9, 10]. However, these methods do not fare well in cases of low parallax and wild camera motions mainly because structure estimation cannot be decoupled from pose refinement.

SLAM techniques also differ on the kinds of scene being tracked: road scenes captured from vehicle mounted cameras, indoor scans from a hand-held camera, and from head-mounted egocentric cameras usually accompanied by sharp head rotations of the wearer. Visual odometry algorithms have been quite successful for hand-held or vehicle-mounted cameras [9, 10, 21, 23, 32, 33], but their incremental nature does not fare well for egocentric videos because of instabilities in the computation due to unrestrained camera motion, wide variety of indoor and outdoor scenes and presence of moving objects [24, 40, 41, 42].

Just like SLAM, structure-from-motion (SFM) techniques can also be categorized into global, batch and incremental ones. As the names suggests, global approaches [55] solve the global problem jointly, whereas incremental approaches like Visual SFM [56] inserts one frame into the estimated structure at a time. Batch mode techniques [4] try to trade between the efficiency of incremental and robustness of global approaches. In recent years, the SFM techniques have seen a lot of progress, using the concepts of rotation averaging (RA) [6] and translation averaging (TA) [17, 20, 30, 55]. The computational cost being linear in the number of cameras, these techniques are fast, robust and well suited for small image sets. They provide good initial estimates for camera pose and structure using pairwise epipolar geometry, which can be refined further using standard SFM techniques.

3. Background

The pose of a camera I' w.r.t a reference image I is denoted by a 3×3 rotation matrix $\mathbf{R} \in \text{SO}(3)$ and a 3×1 translation direction vector \mathbf{t} . The pairwise pose can be estimated from the decomposition of the *essential matrix* E which binds two views using pairwise epipolar geometry such that: $E = [\mathbf{t}]_{\times} \mathbf{R}$ [18, 34]. Here $[\mathbf{t}]_{\times}$ is a skew-symmetric matrix corresponding to the vector \mathbf{t} . A view graph has the images as nodes and the pair-wise epipolar relationships as edges.

3.1. Motion Averaging

Given such a view graph, embedding of the camera poses into a global frame of reference can be done using motion averaging [6, 20, 30, 55]. The motion between a pair of cameras i and j can be expressed in terms of the pairwise rotation (\mathbf{R}_{ij}) and translation direction (\mathbf{t}_{ij}) as:

$\mathbf{M}_{ij} = \begin{bmatrix} \mathbf{R}_{ij} & s\mathbf{t}_{ij} \\ \mathbf{0} & 1 \end{bmatrix}$, where, s is the scale of the translation. If \mathbf{M}_i and \mathbf{M}_j are the motion parameters of cameras

i and j respectively in the global frame of reference, then we have the following relationship between pairwise and global camera motions: $\mathbf{M}_{ij} = \mathbf{M}_j \mathbf{M}_i^{-1}$.

Rotation averaging : Using the above expression, the relationship between global rotations and pairwise rotations can be derived as: $\mathbf{R}_{ij} = \mathbf{R}_j \mathbf{R}_i^{-1}$, where \mathbf{R}_i and \mathbf{R}_j are the global rotations of cameras i and j . From a given set of pairwise rotation estimates \mathbf{R}_{ij} , we can estimate the absolute rotations of all the cameras by minimising a robust sum of discrepancies between the estimated relative rotations \mathbf{R}_{ij} and the relative rotations suggested by the terms $\mathbf{R}_j \mathbf{R}_i^{-1}$ [6]: $\{\mathbf{R}_1, \dots, \mathbf{R}_N\} = \underset{\{\mathbf{R}_1, \dots, \mathbf{R}_N\}}{\operatorname{argmin}} \sum_{(i,j)} \Phi(\mathbf{R}_j \mathbf{R}_i^{-1}, \mathbf{R}_{ij})$,

where $\Phi(\mathbf{R}_1, \mathbf{R}_2) = \frac{1}{\sqrt{2}} \|\log(\mathbf{R}_2 \mathbf{R}_1^{-1})\|_F$, which is the intrinsic bivariate distance measure defined in the manifold of 3D rotations on the $\text{SO}(3)$. Having outlier pairwise rotation estimates is a common problem for any rotation averaging technique. In our experiments, we have used the implementation of [6], which handles such outliers by iterative re-weighting of the constraints using the Huber loss function.

Translation averaging : The global translations \mathbf{T}_i and \mathbf{T}_j are related with pairwise translation directions \mathbf{T}_{ij} as: $\mathbf{t}_{ij} \times (\mathbf{T}_j - \mathbf{R}_{ij} \mathbf{T}_i) = \mathbf{0}$. Global camera positions ($\mathbf{C}_i = \mathbf{R}_i^T \mathbf{T}_i$) can be obtained as: $\{\mathbf{C}_i\} = \underset{\{\mathbf{C}_i\}}{\operatorname{argmin}} \sum_{(i,j)} d(\mathbf{R}_j^T \mathbf{t}_{ij}, \frac{\mathbf{C}_i - \mathbf{C}_j}{\|\mathbf{C}_i - \mathbf{C}_j\|})$, where the summation is over all camera-camera and camera-point constraints derived from feature tracks and scene 3D points. A common concern in translation averaging is to handle degeneracies arising out of linear motion. We handle the problem using additional camera-point constraints harnessed from feature tracks, and using 3D scene points as suggested in [55].

3.2. Bundle Adjustment

Triggs *et al.* [52] suggests using Structure-from-Motion (SFM) to recover both camera poses and 3D structure by minimizing the following reprojection error using bundle adjustment:

$$\min_{\mathbf{c}_j, \mathbf{b}_i} \sum_{i=1}^n \sum_{j=1}^m V_{ij} D(P(\mathbf{c}_j, \mathbf{b}_i), \mathbf{x}_{ij} \Psi(\mathbf{x}_{ij})) \quad (1)$$

where, $V_{ij} \in \{0, 1\}$ is the visibility of the i^{th} 3D point in the j^{th} camera, P is the function which projects a 3D point \mathbf{b}_i onto camera \mathbf{c}_j which is modelled using 7 parameters (1 for focal length, 3 for rotation, 3 for position), \mathbf{x}_{ij} is the actual projection of the i^{th} point on the j^{th} camera, $\Psi(\mathbf{x}_{ij}) = 1 + r \|\mathbf{x}_{ij}\|^2$ is the single parameter (r) distortion function and D is the Euclidean distance. Two kinds

of bundle adjustment methods are used in the literature to minimize Eq. (1). Incremental bundle adjustment technique traverses the graph sequentially starting from an image pair as a seed for the optimization and then keeps on adding images sequentially through resectioning of 3D-2D correspondences. The technique is used in the majority of SLAM algorithms [19, 23, 32]. On the other hand, Batch-mode bundle adjustment technique optimizes for all the camera poses at once by minimizing (1) globally. The approach is less susceptible to discontinuities in reconstruction or drift due to joint optimization of all cameras at once. Also, it requires an initialization for camera parameters and 3D structure which can be provided through motion averaging and linear triangulation [18].

4. Proposed Algorithm

Key Framing We start by processing each frame and designate a new key-frame whenever there is sufficient parallax between frames. The designation is made after every P frames, or whenever the average optical flow crosses m pixels, whichever happens first. This allows our method to adapt to wild motions, for example when turning. In our experiments, we typically set $P = 30$ and $m = 20$. We use “good features to track” (GFTT) for calculating optical flow. Our initial experiments with SIFT did not yield any significant benefit but increased computation time. We use bi-directional sparse iterative version of the Lucas-Kanade optical flow on intensity values of the images. Point correspondences obtained using only optical flow contain noise and therefore can create drift in the estimation of flow vectors. We filter out the correspondences which have higher bi-directional positional error.

Temporal Window Generation We do pose estimation in a temporal window of key-frames. Processing in temporal windows makes the camera motion and structure estimation problem well constrained when parallax between successive frames is small due to dominant 3D rotations, as is common in egocentric videos. We allocate a number of key-frames into a temporal window and then process each window independently. Typically each window contains around 10-30 key-frames with each key-frame separated by about 5-7 frames for the case when the wearer is walking. While lack of parallax justifies creating temporal windows, making too large a window is also problematic considering instability of motion averaging for large windows. A smaller window size also helps in controlling drifts and breaks in structure estimation.

It may be noted that if we choose a window size of 1, our method essentially becomes incremental strategy as is common in state of the art methods like VisualSFM [56]. In Figure 3, we analyze the effect of window size on trajectory

estimation for a window size of 1, 30 and 500. The estimation process fails for 1 and 500 but works best for 30 in this particular case.

In global SFM, densely connected entire view graph allows larger window sizes to yield more tightly bound constraints for motion averaging and hence better estimates. However, in our case feature correspondences are generated transitively through sequential tracking, which leads to drift in larger windows due to poor estimation of pairwise epipolar geometry for redundant paths (longer edges).

Though, most of the results in this paper have been produced with non-overlapping batches, in case a lower latency is required, one can use sliding windows with significant overlaps as well.

Local Loop Closures We use local loop closures to handle large rotations in our input videos. This gives extra constraints for stabilising the camera estimates. Global loop closure is an important step in traditional SLAM to fix the accumulated errors over pose estimation. However, in case of the egocentric videos, where the motion of the wearer is linear forward, a user may not revisit a particular scene point for long, making global loop closure sometimes impossible. Also, given the wild nature of egocentric videos, the camera poses and trajectories tend to drift quickly unless fixed by loop closures immediately. We observe that in a natural walking style, a wearer’s head typically scans the scene left to right and back. The camera looks at the same scene multiple times, thus providing opportunities for a series of short local loop closures. We take advantage of this phenomenon by using local loop closures Patra et al. [38], to improve the accuracy of the estimated camera poses.

We maintain a set of last few key-frames and when considering a new key-frame, we estimate its pairwise pose with these existing key-frames to discover redundant paths. The additional edges added in the view graph during this stage helps in local loop closures through motion averaging as described in the following step. Figure 4 shows the effect of loop closures on the estimation. In the absence of these extra edges and thus the local loop closures, the structure around the staircase gets deformed in the scale and also shifts above the ground, thus causing breaks.

Camera Pose Estimation After adding extra edges in the view graph for local loop closure, we use the five-point algorithm [34] for estimating the pair-wise epipolar geometry for a batch. This provides sufficient constraints for motion averaging. We first use rotation averaging for finding global rotation estimates followed by translation averaging. Note that providing a good initial estimate of camera pose, without using any 3D structural information, is essential for any SLAM approach to work successfully on egocentric videos.

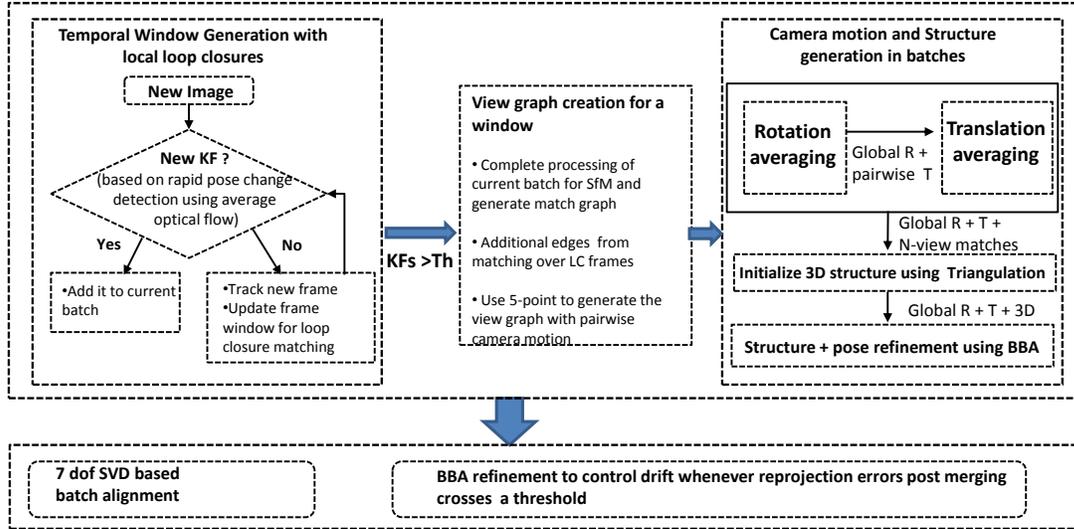


Figure 2: Flow chart of the proposed EGO-SLAM

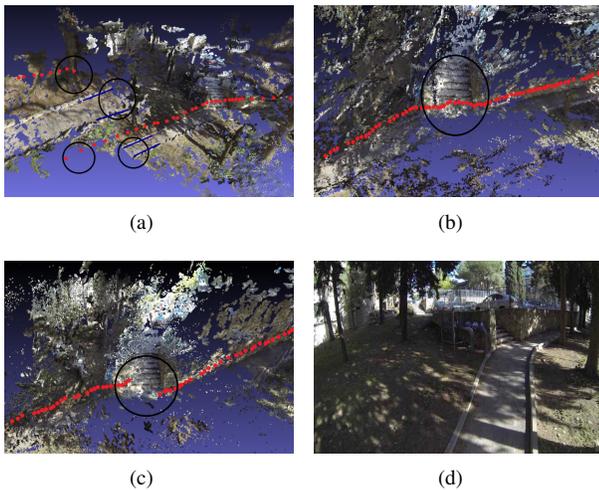


Figure 3: Incremental SLAM is problematic for egocentric videos due to lack of parallax between successive frames. We propose batch mode processing to stabilize the trajectory estimation first. (a), (b) and (c) show output with a batch size of 1, 30 and 500 respectively. (d) is the reference image. Large batch size may also cause problems in motion averaging convergence and breaks in SfM causing trajectory break highlighted in (c), structure error as well as trajectory break, highlighted in (a) but corrected by small batch size in (b). The sequence is taken from HUJI dataset [40].

This is where the proposed approach is critically different from state of the art approaches.

After robustly estimating rotations, we use a mixture of two different methods for averaging the translations. To initialize the global translations we generate an initial guess using global convex optimization technique specified in [37] and subsequently refine the solution using the approach of [55]. This provides an excellent initial estimate for the

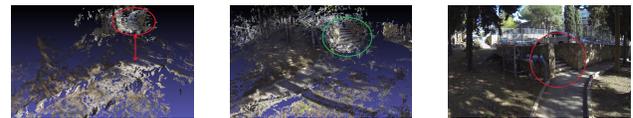


Figure 4: Loop closures are an important step in a SLAM algorithm but may never be applied in an egocentric video because of usual forward motion of the wearer. In this paper, we have suggested local loop closures for egocentric videos. First and second images show structure estimation without and with local loop closures respectively. Third image is the reference view. Note the ‘hanging’ stairs in the first image without loop closure.

camera poses. During this phase, the camera intrinsics remain constant.

3D Structure Estimation Once the camera poses are robustly initialized, the 3D structure is setup using linear triangulation as specified in [18]. We further refine the initial structure and camera poses using a final run of Window mode Bundle Adjustment (WBA). It estimates all the camera poses and 3D points simultaneously using Bundle adjustment [52]. The convergence of bundle adjustment is very fast due to the good initialization as described above. Also, this phase allows us to refine camera intrinsics through WBA.

Merging and WBA Refinement with Resectioning As the last step we merge the structure obtained from successive temporal windows using 7 dof alignment based on SVD as suggested in [53]. During the merging step, new points which were not used previously due to not being visible in most cameras, are added back, as these points get stable at this stage with more cameras viewing them now. A



Figure 5: Comparison of the estimated structure on a challenging Hyperlapse *climbing03* sequence [24]. State of the art SLAM fails here and authors of hyperlapse have reported using SFM algorithm by manually dividing the sequence into batches of 1400 frames. EGO-SLAM works without fail on the complete sequence. Left: Dense depth map generated by [24] using CMVS [13]. Middle: Corresponding dense depth map generated by EGO-SLAM. Right: A reference view

final round of global BBA based refinement is run in the background whenever the cross batch reprojection errors get high. This leads to a non-linear refinement in the scale of the estimated structure and poses. We describe the complete algorithm of EGO-SLAM in Figure 2.

5. Experiments and Results

In this section, we validate the robustness of EGO-SLAM on various publicly available video datasets captured from egocentric, handheld and vehicle mounted cameras. We have implemented portions of our algorithm in C++ and MATLAB. All the experiments have been carried out on a regular desktop with Core i7 2.3 GHz processor (containing 4 cores) and 32 GB RAM, running Ubuntu 14.04.

Our algorithm requires the intrinsic parameters of the cameras for SFM. For sequences taken from public sources, we use the calibration information based on the make and the version of the cameras provided on their websites.

Note that various egocentric research groups have reported the failure of various SLAM methods on egocentric videos. Therefore, we have restricted our attention to comparison with latest SLAM techniques which have been published after those reports: mainly ORB-SLAM, but also LSD-SLAM and PTAM for indicative purposes.

For visual clarity, we show the dense 3D map in all our examples by carrying out the dense reconstruction of some portions using CMVS [13]. We provide to CMVS the camera poses and the sparse structure computed using our algorithm. Note that CMVS can produce high quality output only if the pose and the initial structure estimates are correct, and this also serves as a test for our results. We present results with views of the point clouds from a single vantage point w.r.t. the reference image. For more views from better vantage points please refer to our supplementary video.

Please note that in our implementation, we use optical flow for image matching because of simplicity and speed. However, our pipeline does not preclude the use of feature descriptor based matching for relocalization and mapping applications (see Section 5.4 for a discussion). For a video of 1280×800 at 60fps with a batch size of 20 Key Frames



Figure 6: Our result on another challenging sequence from the HUJI data set [40]. Here the wearer is walking in a narrow alley and even makes a sharp 360 degree turn. Left: Estimated trajectory on superimposed on Google map. Middle: Dense depth map of a portion obtained using CMVS [13]. Right: Reference View

Seq. Name	#Frames	#Breaks in the Seq.		
		EGO	ORB[32]	LSD[9]
<i>Bike07</i> [24]	12000	0	14	17
<i>Climbing03</i> [24]	3866	0	4	20
<i>Yair_5</i> [40]	3634	0	2	2
<i>Yair_1_p2</i> [40]	3601	0	1	2
<i>Yair_6</i> [40]	3357	0	1	1

Table 1: Number of breaks suffered by various methods on 5 videos from the Hyperlapse [24] and the HUJI egoseg [40] datasets

(KF), on an average a batch lasts for 0.8 - 2 sec based on the type of video (usually shorter for an egocentric and longer for a car mounted video). Some indicative timings for a set of 49 frames from which 20 frames were chosen as KFs are: Relative Pose: 10.71 sec; Motion Averaging: 1.34 sec; Triangulation: 1.66 sec; BBA: 0.13 sec; and Batch Merging: 3 sec. Note that our code is unoptimized. E.g., Finding relative pose have been shown to work in real time by others.

5.1. Egocentric Videos

We have tested EGO-SLAM on various Hyperlapse sequences [24]. The *bike07* [24] video in the dataset is a very challenging sequence with wild head motions, fast forward movements, and sharp turns. Both [19, 32] break on this sequence. We have already shown the computed trajectory for the sequence in Figure 1. In the same figure, we have shown the 3D map by carrying out dense reconstruction of some portions using CMVS [13] based on the camera poses and the sparse structure computed using our algorithm. In Figure 5 we compare the dense 3D structure of a portion computed using EGO-SLAM with the one given in Hyperlapse. It is to be noted that in [24], pose and 3D structure are computed using SFM over batches of 1400 frames.

We present similar results on a similarly challenging *Huji_Yair_5* sequence from the HUJI EgoSeg dataset in Figure 6. All the state-of-the-art SLAM techniques have been reported to fail on these datasets [40, 41, 42].

One of our claims in this paper is on the robustness of proposed method over the SOTA. To validate this claim, we count the number of breaks/crashes suffered by various methods while processing egocentric videos from Hyperlapse [24] and HUJI EgoSeg [40] datasets. Table 1 shows

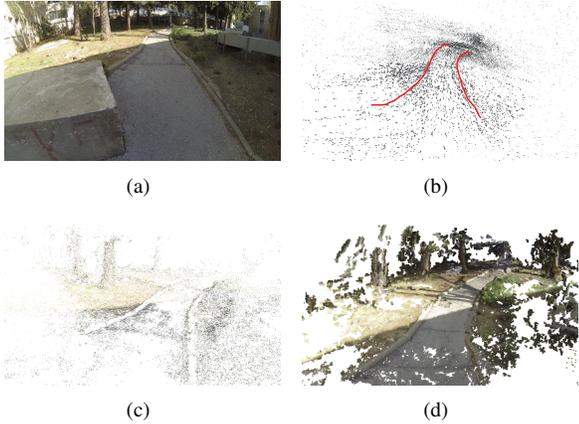


Figure 7: Poor 3D estimation by SOTA is one of the major reasons of breaks. (a) shows a reference view from a HUJI sequence [40], (b) poor structure estimation by ORB-SLAM (road highlighted) just before the break (c) shows the correct structure estimation by EGO-SLAM at the same point which is made dense by CMVS [13] in (d).

Motion profile	Traj. len. (m)	RMSE (cm) of 3D		
		EGO	ORB[32]	LSD[9]
frontal	2.0	20.6	30.5	39.8
left – right	4.0	24.9	26.6	35.0
egomotion	3.7	22.7	25.4	47.9

Table 2: Accuracy analysis of estimated structure and comparison with SOTA using a synthetic scene and different motion profiles.

the number of such breaks.

There are no benchmark datasets with ground truth trajectory or structure for egocentric videos. Therefore, we created a synthetic setup for quantitative error estimation. We created a synthetic scene with different planes of various sizes (max $5.12m \times 5.12m$) at different depths. Since the depths are known we now use the projected images under different motion profiles: frontal, left-right, and egocentric to estimate the 3D using LSD-SLAM, ORB-SLAM and EGO-SLAM, and then compare the estimated depths from each of these methods against the ground truth. We present this analysis in Table 2. More details and visualization can be found in supplementary.

Note that the breaks suffered by the SOTA are due to 3D tracking failures due to poor localization and inferred structure. We confirm this claim and show in Figure 7 the incorrect structure computed before the break by ORB-SLAM. We also show the corresponding correct structure computed by our method in the same figure for comparison.

We also compare our method with Hierarchical SFM (HSFM) [50]. We ran the free version of their commercial software (Zephyr Lite [57]) on bike07 sequence, and found it to break first time at around frame no 1927 and after that it

Seq.	Dimension $m \times m$	Trajectory RMSE (m)	
		EGO	ORB[32]
KITTI 00	564×496	3.34	6.68
KITTI 01	1157×1827	67.09	X
KITTI 02	599×946	7.75	21.75
KITTI 03	471×199	0.44	1.59
KITTI 04	0.5×394	1.75	1.79
KITTI 05	479×426	3.85	8.23
KITTI 06	23×457	11.63	14.68
KITTI 07	191×209	2.41	3.36
KITTI 08	808×391	5.87	46.58
KITTI 09	465×568	6.97	7.62
KITTI 10	671×177	0.85	8.68

Table 3: Our results on videos taken from vehicle mounted cameras on the KITTI dataset [14]. RMS error of computed trajectories (in meters) with respect to the ground truth trajectory show that we improve upon the SOTA on such videos as well. “X” denotes failure in estimation. Visualization of some of the computed trajectories is in the supplementary.

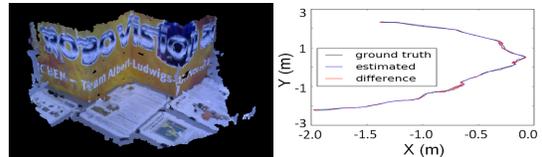


Figure 8: Left: Dense depth map computed using EGO-SLAM + CMVS [13] on *fr3_str.tex_far* seq. (TUM dataset [48]). Right: Comparison with ground truth trajectory after 7 dof alignment

suffered from breaks at multiple places due to resectioning (5 times within the first 7000 frames) whenever there were sharp turns. The structure estimated was also deformed.

5.2. Vehicle Mounted Cameras

Though, the focus of this paper is on egocentric videos, our algorithm is equally applicable for other capture scenarios where there is low parallax between consecutive frames and lack of global loop closure opportunities. One such case arises from vehicle mounted forward looking cameras. We have experimented with one such challenging dataset [14].

Table 3 shows the RMS error of the computed trajectory with respect to the ground truth. Comparison with a state of the art method, ORB-SLAM [32], indicates that we perform better on such videos as well. Note that LSD-SLAM [9] does not work on the KITTI videos.

5.3. Handheld Cameras

We emphasize that our technique is specially geared for egocentric videos. For hand held videos our technique holds no special advantage and only works as well as traditional SLAM algorithms. However, to confirm the applicability of it

Seq.	RMSE (cm) of Trajectory			
	EGO	ORB[32]	PTAM[23]	LSD[9]
f1_fl	1.60	2.99	X	38.07
f1_d	1.34	1.69	X	10.65
f3_l_off	1.03	3.45	X	38.53
f3_s_tf	1.06	0.77	0.93	7.95
f3_ns_tf	13.89	X	4.9 / 34.7	18.31
f3_s_tn	1.03	1.58	1.04	X
f3_ns_tn	1.31	1.39	2.74	7.54

Table 4: Comparison of RMS error with respect to ground truth trajectory on a few sequences from the TUM dataset [48] of hand-held video. Our error is better than LSD-SLAM on these sequences and also better than ORB-SLAM and PTAM in most cases. ‘‘X’’ denotes failure in estimation. Detailed information on the sequences can be found in the supplementary material.

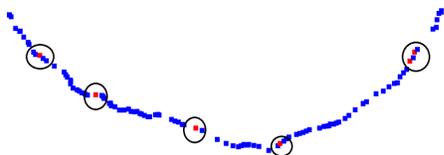


Figure 9: Our pipeline can also use standard feature descriptors for relocalization. The figure shows localized novel cameras on the precomputed trajectory using our method (see paper text for details). The estimated locations (red dots) near the trajectory indicate successful localization in TUM fr3_str_tex_far sequence.

the proposed technique as a generic SLAM, we provide experimental results on a few hand held benchmarks as well.

We have used the TUM Visual odometry dataset [48] for the analysis of videos captured from handheld cameras. Figure 8 shows the dense reconstruction and the trajectory estimated by the proposed method. Note that the graph shown in the figure also contains the ground truth trajectory, but the estimated trajectory is completely aligned with the ground truth and hides it completely.

The TUM dataset also allows us to compute the RMS error of the computed trajectory with respect to the ground truth trajectory. Table 4 shows the error for EGO-SLAM as well as the ones reported by the other SOTA techniques on the same sequence. We match and often improve the state of the art even for regular hand-held videos as well. Note that, for the fr3_nostructure_texture_far sequence ORB-SLAM fails due to planar ambiguity and PTAM produces ambiguous results due to different initializations every time. Hence for this case PTAM also produces unreliable results.

5.4. Relocalization

Relocalization error is a popular metric to measure the accuracy of estimated 3D structure. In EGO-SLAM, we use optical flow for image matching for the sake of simplicity and speed. Since optical flow vectors do not have associated feature descriptors, they cannot be used for relocalization and mapping. However, our pipeline does not preclude use

Method	Rotation (deg.)		Position (cm)	
	Mean	Median	Mean	Median
Without BA	0.0198	0.0216	1.004	1.040
With BA	0.0062	0.0051	0.975	0.977

Table 5: Quantitative analysis of relocalization error. We perform relocalization as shown in Figure 9 and compute error in camera rotation (degrees) and absolute position (cm) after relocalization for novel frames. Smaller error indicates successful localization.

of such feature descriptors for relocalization.

To demonstrate relocalization using our framework, we train a vocabulary tree [35] using the SIFT [27] features computed from the key-frames in the TUM fr3_str_tex_far sequence [48]. We then use a set of frames which are not key-frames to calculate the relocalization error. We carry out feature matching with the key-frames using vocabulary tree, reject outliers using the pre-computed trajectory of the key-frames, and estimate the pose of the unknown frames using 3D-2D correspondences.

In Figure 9, we plot the relocalized unknown frames on the computed trajectory. Location of the frames on the trajectory indicate the correctness of relocalization. In Table 5 we show the accuracy of relocalization with respect to the ground-truth both with and without a final BA refinement.

It may be noted that relocalization also facilitates global loop closures in our original match graph.

6. Conclusion

Despite tremendous progress made in recent SLAM techniques, running such algorithms for many categories of videos still remain a challenge. We believe that careful case by case analysis of such challenging videos may provide crucial insights into improving the SOTA. Egocentric a.k.a first person videos are one such category we focus on in this paper. We observe that incremental estimation employed in most current SLAM techniques often cause unreliable 3D estimates to be used for trajectory estimation. We suggest to first stabilize the trajectory using 2D techniques and then go for structure estimation. We also exploit domain specific heuristics such as local loop closures. Interestingly, we observe that the proposed technique improves the SOTA for videos captured from vehicle mounted cameras also. Finally, many applications like hyperlapse [24], and first person action recognition [39, 44] could have been solved by principled camera pose and structure estimation. But the authors of such works were forced to take other approaches because of inability of SOTA SLAM techniques on egocentric videos. We believe many such current and future researchers will benefit by the use of proposed technique.

Acknowledgement: This work was supported by a research grant from Continental Automotive Components (India) Pvt. Ltd.

References

- [1] O. Aghazadeh, J. Sullivan, and S. Carlsson. Novelty detection from an ego-centric perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3297–3304, 2011. 1
- [2] L. Baraldi, F. Paci, G. Serra, L. Benini, and R. Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Proceedings of IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014. 1
- [3] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Distinctive image features from scale-invariant keypoints. *Computer Vision and Image Understanding (CVIU)*, 110(3):346–359, 2008. 2
- [4] B. Bhowmick, S. Patra, A. Chatterjee, V. M. Govindu, and S. Banerjee. Divide and conquer: A hierarchical approach to large-scale structure-from-motion. *Computer Vision and Image Understanding*, 157(Supplement C):190–205, 2017. 1, 2, 3
- [5] D. Castro, S. Hickson, V. Bettadapura, E. Thomaz, G. Abowd, H. Christensen, and I. Essa. Predicting daily activities from egocentric images using deep learning. In *Proceedings of the ACM International Symposium on Wearable Computers (ISWC)*, pages 75–82, 2015. 1
- [6] A. Chatterjee and V. M. Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 521–528, 2013. 3
- [7] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1052–1067, 2007. 1
- [8] E. Eade and T. Drummond. Scalable monocular slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–476, 2006. 1
- [9] J. Engel, T. Schops, and D. Cremers. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 834–849, 2014. 1, 2, 3, 6, 7, 8
- [10] J. Engel, J. Sturm, and D. Cremers. Semi-dense visual odometry for a monocular camera. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1449–1456, 2013. 3
- [11] A. Fathi, Y. Li, and J. M. Rehg. Learning to recognize daily actions using gaze. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 314–327, 2012. 1
- [12] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3281–3288, 2011. 1
- [13] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski. Towards internet-scale multi-view stereo. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1434–1441, 2010. 6, 7
- [14] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 7
- [15] Google. Glass. <https://www.google.com/glass/start>. 1
- [16] GoPro. Hero3. <https://gopro.com/>. 1
- [17] V. Govindu. Combining two-view constraints for motion estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 218–225, 2001. 3
- [18] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, 2nd edition, 2004. 3, 4, 5
- [19] C. D. Herrera, K. Kim, J. Kannala, K. Pulli, and J. Heikkil. DT-SLAM: deferred triangulation for robust slam. In *Proceedings of International Conference on 3D Vision (3DV)*, pages 609–616, 2014. 1, 2, 4, 6
- [20] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 481–488, 2013. 3
- [21] C. Kerl, J. Sturm, and D. Cremers. Robust odometry estimation for RGB-D cameras. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3748–3754, 2013. 3
- [22] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3241–3248, 2011. 1
- [23] G. Klein and D. Murray. Parallel tracking and mapping on a camera phone. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 83–86, 2009. 2, 3, 4, 8
- [24] J. Kopf, M. F. Cohen, and R. Szeliski. First-person hyperlapse videos. *ACM Transactions on Graphics (TOG)*, 33(4):78:1–78:10, 2014. 1, 2, 3, 6, 8
- [25] M. Li and A. I. Mourikis. High-precision, consistent ekf-based visual-inertial odometry. *International Journal of Robotics Research*, 32(6):690–711, 2013. 2
- [26] H. Lim, J. Lim, and H. J. Kim. Real-time 6-dof monocular visual slam in a large-scale environment. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 1532–1539, 2014. 1
- [27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004. 2, 8
- [28] W. Lucia and E. Ferrari. Egocentric: Ego networks for knowledge-based short text classification. In *Proceedings of ACM International Conference on Conference on Information and Knowledge Management (CIKM)*, pages 1079–1088, 2014. 1
- [29] Microsoft Research. SenseCam. <https://research.microsoft.com/en-us/um/cambridge/projects/sensecam/>. 1
- [30] P. Moulon, P. Monasse, and R. Marlet. Global fusion of relative motions for robust, accurate and scalable structure from motion. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pages 3248–3255, 2013. 3

- [31] E. . Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3d reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 363–370, 2006. **1**
- [32] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015. **1, 2, 3, 4, 6, 7, 8**
- [33] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison. DTAM: Dense tracking and mapping in real-time. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2320–2327, 2011. **2, 3**
- [34] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(6):756–777, 2004. **3, 4**
- [35] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, 2006. **8**
- [36] K. Ogaki, K. M. Kitani, Y. Sugano, and Y. Sato. Coupling eye-motion and ego-motion features for first-person activity recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–7, 2012. **1**
- [37] O. Ozyesil and A. Singer. Robust camera location estimation by convex programming. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2674–2683, 2015. **5**
- [38] S. Patra, H. Aggarwal, H. Arora, C. Arora, and S. Banerjee. Computing egomotion with local loop closures for egocentric videos. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 454–463, 2017. **4**
- [39] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2847–2854, 2012. **1, 8**
- [40] Y. Poley, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2537–2544, 2014. **1, 2, 3, 5, 6, 7**
- [41] Y. Poley, A. Ephrat, S. Peleg, and C. Arora. Compact CNN for indexing egocentric videos. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9, 2016. **3, 6**
- [42] Y. Poley, T. Halperin, C. Arora, and S. Peleg. EgoSampling: Fast-forward and stereo for egocentric videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4768–4776, 2015. **1, 2, 3, 6**
- [43] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2564–2571, 2011. **2**
- [44] S. Singh, C. Arora, and C. V. Jawahar. First person action recognition using deep learned descriptors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2620–2628, 2016. **8**
- [45] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3d. In *Proceedings of ACM SIGGRAPH*, pages 835–846, 2006. **2**
- [46] N. Snavely, S. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008. **2**
- [47] H. Strasdat, A. J. Davison, J. M. M. Montiel, and K. Konolige. Double window optimisation for constant time visual slam. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2352–2359, 2011. **1**
- [48] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers. A benchmark for the evaluation of rgb-d slam systems. In *Proceedings of the International Conference on Intelligent Robot Systems (IROS)*, 2012. **7, 8**
- [49] W. Tan, H. Liu, Z. Dong, G. Zhang, and H. Bao. Robust monocular slam in dynamic environments. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 209–218, 2013. **1**
- [50] R. Toldo, R. Gherardi, M. Farenzena, and A. Fusiello. Hierarchical structure-and-motion recovery from uncalibrated images. *Computer Vision and Image Understanding*, 140:127–143, 2015. **1, 7**
- [51] P. H. Torr and A. Zisserman. Feature based methods for structure and motion estimation. In *Proceedings of Vision Algorithms: Theory and Practice. Springer*, pages 278–294, 1999. **1**
- [52] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment a modern synthesis. In *Vision Algorithms: Theory and Practice, LNCS*, pages 298–372, 2000. **2, 3, 5**
- [53] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 13(4):376–380, 1991. **5**
- [54] B. Williams, M. Cummins, J. Neira, P. Newman, I. Reid, and J. Tardós. A comparison of loop closing techniques in monocular SLAM. *Robotics and Autonomous Systems*, 57(12):1188–1197, 2009. **1**
- [55] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 61–75, 2014. **1, 3, 5**
- [56] C. Wu. Towards linear-time incremental structure from motion. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 127–134, 2013. **1, 2, 3, 4**
- [57] Zephyr. <http://www.3dflow.net/3df-zephyr-pro-3d-models-from-photos/>. **7**