Journal of Machine Vision and Applications (MVAP) manuscript No.
(will be inserted by the editor)

Visual Odometry with a Single-Camera Stereo Omnidirectional System

Carlos Jaramillo, Liang Yang, J. Pablo Muñoz, Yuichi Taguchi, and Jizhong Xiao

Received: date / Accepted: date

Abstract This paper presents the advantages of a single-camera stereo omnidirectional system (SOS) in estimating egomotion in real-world environments. Dynamic conditions, deficient illumination, and poor textured surfaces result in the lack of features to track in the observable scene. This negatively affects the pose estimation of visual odometry systems, regardless of their field-of-view. We compare the tracking accuracy and stability of the single-camera SOS versus an RGB-D device under various real circumstances. Our quantitative evaluation is performed with respect to 3D ground truth data obtained from a motion capture system. The datasets and experimental results we provide are unique due to the nature of our catadioptric omnistereo rig, and the situations in which we captured these motion sequences. We have implemented a tracking system with deterministic rules for both synthetic and real scenes. Our implementation does not make any motion model assumptions, and it maintains a fixed configuration among the compared sensors. Our experimental outcomes confer the robustness in 3D metric visual odometry estimation that the single-camera SOS can achieve under normal and special conditions in which other perspective narrow view systems such as RGB-D cameras would fail.

C.J., L.Y., and J.X. were supported by U.S. Army Research Office grant No. W911NF-09-1-0565, US National Science Foundation grants No. IIS- 0644127 and No. CBET-1160046, Federal Highway Administration (FHWA) grants No. DTFH61-12-H-00002 and No. DTFH61-17-C-00007.

C. Jaramillo
Computer Science Department, The Graduate Center of The City University of New York (CUNY), 365 Fifth Avenue, New York, NY, USA
E-mail: omnistereo@gmail.com

L. Yang
Electrical Engineering Department, The City College, CUNY, Convent Ave & 140th Street, New York, NY, 10031, USA
E-mail: lyang@ccny.cuny.edu

J. P. Muñoz
Intel Corporation, 2200 Mission College Blvd, Santa Clara, CA, USA
E-mail: pablo.munoz@intel.com

Y. Taguchi
Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA, USA
E-mail: taguchi@merl.com

J. Xiao *(corresponding author)*
Electrical Engineering Department, The City College, CUNY, Convent Ave & 140th Street, New York, NY, 10031, USA
Tel.: +001-212-650-7268
E-mail: jxiao@ccny.cuny.edu

1 Introduction

Visual odometry (VO) is an important building block for a vast number of applications in the realms of robotic navigation and augmented reality. Several camera types, lenses, mirrors, and their combinations have been used to estimate egomotion in the past. Using a single camera has the main drawback of the unknown absolute scale factor for the solution in the scene. On the other hand, the scale problem can be solved by rigidly combining various cameras at the cost of price, energy, size, weight, computer I/O ports, and hardware synchronization issues. VO estimation on a small robot requires of a portable sensor providing 3D metric information. Therefore, we conceive a single-camera Stereo Omnidirectional System (SOS) as the essence of the VO solution presented in this work. Fig. 1 shows the single-camera SOS based on a vertical catadioptric configuration designed by Jaramillo et al. [1].

Depending on the situation, the choice of a wide field-of-view (FOV) camera for visual odometry can have advantages over the narrow viewing angle solutions, as it was observed by Zhang et al. [2]. However, they only employed monocular sensors without 3D metric scale information, and their comparisons concluded that the trade-off in pixel res-

olution for a wider FOV affects the VO accuracy in open outdoors spaces. Our work is different by employing the single-camera SOS described in §2.2 that is capable of providing 3D metric information, so we study its practical advantages. We emphasize in showing its robust operation under dynamic environments, and some exceptional circumstances that robotics navigation encounters under weakly illuminated or poorly textured areas.

Existing approaches for VO usually make assumptions about operating in mostly-static environments and that the observed scene is discriminative enough. However, when confronted with weakly textured, dynamic, partially occluded, or poorly illuminated environments, real challenges arise [3]. We have chosen to employ the feature-based method due to its versatile compatibility with different types of cameras, plus the ample availability of standard software libraries that guarantee the abstraction and portability of our VO implementation (§3). Our purpose is not to design a SLAM framework since plenty of alternatives exist already, i.e., [4,5,6]. Instead, our goal is to show the potential of the proposed single-camera SOS and projection model (§2.2) as a viable alternative to more traditional, affordable sensors such as the popular RGB-D device. The application of omnidirectional vision for robotic navigation has been discussed in previous works (§2.1), but this has not been demonstrated until now for a single-camera SOS. We believe to be the first to provide such concrete findings (§4) using both real and synthetic datasets created along this investigation and that we release publicly[1].

2 Related Work

2.1 Feature-based VO with Omnidirectional Cameras

Various approaches to omnidirectional VO that have employed sparse features from the scene are [7,8,9]. Lemaire and Lacroix [10] presented a solution to the bearing-only SLAM problem using a calibrated para-catadioptric ODVS on top of a rover taking long trajectories. In [10], the 3D metric information of their VO estimation was given by an external stereo camera mounted on the same rover. Gutierrez et al. [9] adapted a 1-Point RANSAC technique to achieve EKF SLAM with omnidirectional images whose projection rays were linearized via the unified sphere model [11]. Their solution tracked FAST features [12] sparsely detected on the omnidirectional image. Schönbein and Geiger [13] achieved impressive results for dense omnidirectional mapping, where the vehicle's motion was estimated by tracking FAST keypoints that got triangulated as 3D points. In a RANSAC perspective from *n* points fashion, the 3D points from the previous frame were reprojected onto the current frame, and the relative pose of the vehicle was obtained as a reprojection error minimization problem similar to our approach. They also constructed a plane-based 3D model through block stereo matching. In fact, [13] is the closest attempt to ours as for estimating VO with a catadioptric SOS. However, they employed a pair of omnidirectional cameras operating independently on top of the vehicle, so the large horizontal baseline allowed them to operate outdoors at the cost of some self-occlusion at the singularity points [14]. As discussed previously, there exist practical disadvantages in managing a multi-camera system, in particular, the increase in size and required computer resources limit their adoption to more powerful robots.

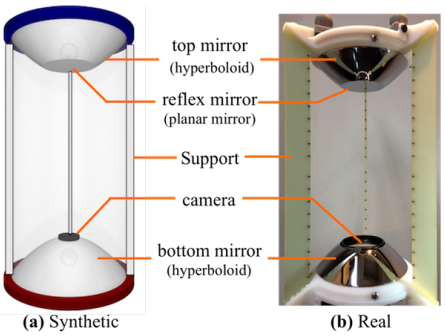


Fig. 1: Single-camera SOS prototypes designed in [1].

2.2 Single-Camera Stereo Omnidirectional System

The challenge of applying omnidirectional stereo vision via a single camera is what separates our work from others reviewed in §2.1. Inspired by the catadioptric single-camera SOS configuration presented by Jang et al. [15], we optimized and analyzed the geometric characteristics of this kind of sensor for its end-use on top of a micro aerial vehicle (MAV) [1]. As the example given in Fig. 3a, the spatial resolution of the SOS is sacrificed by combining two simultaneous omnidirectional views on a single image, which are conveniently rectified as a pair of registered panoramas, as illustrated in Fig. 7. We can perform the typical search for stereo correspondences between these panoramas, and triangulate 3D points as a result. Our single-camera SOS enables instantaneous 3D metric information, whose acquisition is modeled via a generalized unified model for stereo, GUMS, proposed in [16]. The parameters of GUMS are numerically estimated to minimize the total pixel reprojection error that a real sensor exhibits as it deviates from the theoretical central configuration. This optimization is done in a non-linear least-squares fashion with the Huber-norm as its loss function in a highly-coupled calibration procedure. The unit spheres onto which projections are normalized as bearing vectors are visualized in Fig. 2.

[1] http://ubuntuslave.github.io/publication/2018-vo_sos

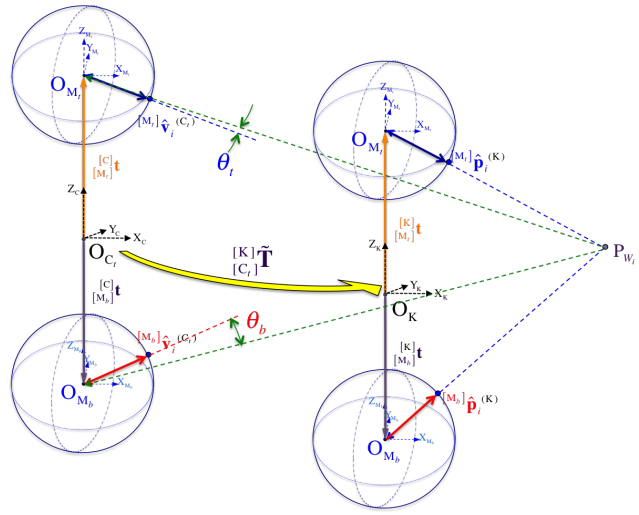


Fig. 2: Illustration of the registration error for a 3D point P_{W_i} due to a noncentral absolute pose ${}^{[K]}_{[C_t]}\tilde{\mathbf{T}}$ between SOS frames via GUMS. The error is considered with respect to the current tracking frame $[C_t]$ as the angle θ_k formed between the back-projecting vector ${}^{[M_k]}\hat{\mathbf{v}}_i^{(C_t)}$ and the forward projecting vector ${}^{[M_k]}\mathbf{p}_i^{(C_t)} \leftarrow {}^{[C]}_{[M_k]}\mathbf{T}^{-1} \cdot {}^{[K]}_{[C_t]}\tilde{\mathbf{T}}^{-1} \cdot {}^{[C]}_{[M_k]}\mathbf{T} \cdot {}^{[M_k]}\mathbf{p}_i^{(K)}$, with $k \in \{t,b\}$ for the top and bottom mirrors, respectively.

3 Tracking Algorithm for Single-Camera SOS

This paper intended to demonstrate the general VO robustness of our SOS in comparison with an RGB-D sensor under the same camera tracking framework. We employed an RGB-D sensor instead of a binocular stereo camera due to our empirical notion that RGB-D devices provide regarding pixel-depth registration reliability. We implemented a frame-to-frame pose estimation algorithm with a deterministic termination criterion. Since we did not make any motion assumptions, i.e., dominant direction and speed, all frames were considered for evaluation according to §4.2. Each frame $[C_t]$ was tracked with respect to its reference keyframe [K] created under the heuristics of exceeding a 1 cm change in translation or 1° in the relative rotation angle, and if and only if the number of tracking correspondences of the candidate frame was at least 10% of the cumulative moving average for the number of keypoints tracked against the current keyframe.

The geometric pose estimation of the sensor was performed by a noncentral 3D-to-2D Perspective from n Points solution (PnP), which requires a minimum of three feature-pair correspondences to model an $SE3$ pose hypothesis ${}^{[K]}_{[C_t]}\tilde{\mathbf{T}}$ as shown in Fig. 2. The model-independent projection metric employed according to the OpenGV framework created by Kneip and Furgale [17] was the angle θ_{vp} computed between the pair of correspondences $\left({}^{[C_t]}\hat{\mathbf{v}}_i, {}^{[C_t]}\hat{\mathbf{p}}_i\right)$ related to the back-projective and forward-projective bearing vectors, respectively. The 2D correspondence at the keyframe is associated to a 3D point ${}^{[K]}\mathbf{p}_i$ that gets transformed onto the current frame coordinates by ${}^{[C_t]}\mathbf{p}_i \leftarrow {}^{[K]}_{[C_t]}\mathbf{T}^{-1}\,{}^{[K]}\mathbf{p}_i$. The error score e_i for each feature match was computed based on their reprojection angles cosine function, $\cos(\theta_{vp})_i = \langle \hat{\mathbf{v}}_i, \hat{\mathbf{p}}_i \rangle$, as:

$$e_i := 1.0 - \cos(\theta_{vp})_i, \quad e_i \in [0,2] \tag{1}$$

The reprojection angle threshold for our RANSAC model fitting was set to 1° in our experiments.

Without the need for bootstrapping, we assume the existence of a set P of metric 3D points that could be registered into the global world frame coordinates, [W], conveniently set at the initial sensor's frame, $[C_0]$, also acting as the first keyframe. For a camera frame $[C_t]$ at a given time step $t > 1$, a set of 2D keypoints D_{I_t} was detected on the image I_t to be initially matched against all the keypoints in set D_K pertaining the keyframe, [K]. Note that for the SOS, we did the feature extraction on the panoramic images instead. We used the OpenCV[2] implementation of "Good Features to Track" [18] for detecting up to 1000 corners via the minimal eigenvalue gradient matrix method, which provided a good number of keypoints (bucketed at every 10° for the panoramic images). These keypoints were consequently described as ORB (Oriented Robust Binary) features [19] due to its relative speed and empirical performance for feature description generation and matching on the panoramic images as visualized in Fig. 7. This initial set of feature matches, $M_{f2f}^{(s_0)}$, was further refined via Kneip's Non-Perspective-three-Point (NP3P) algorithm [20] in a RANSAC fashion. The final set of inlier correspondences, $M_{f2f}^{(s_r)}$, after some r iterations from RANSAC, was used to solve for a final pose ${}^{[K]}_{[C_t]}\mathbf{T}^*$ in a local non-linear optimization with the objective of minimizing the sum of the bearing vector angle errors (1) via the Levenberg-Marquardt algorithm employed in OpenGV. Given the absolute pose ${}^{[W]}_{[K]}\mathbf{T}$ of the keyframe with respect to the world frame [W], the pose of a tracking frame was ultimately transformed into the world by ${}^{[W]}_{[C_t]}\mathbf{T} \leftarrow {}^{[W]}_{[K]}\mathbf{T} \cdot {}^{[K]}_{[C_t]}\mathbf{T}$.

The VO algorithm described above was implemented for both sensors. The RGB-D sensor obeyed the pinhole camera model. The single-camera SOS, however, was modelled by the Generalized Unified Model for Stereo (GUMS) proposed in [16], so we could deploy a "noncentral" absolute pose adapter according to OpenGV's design pattern. In this sense, we had two viewpoints in the camera system established via the fixed rigid transform ${}^{[C]}_{[M_k]}\mathbf{T}$, which was obtained during the GUMS coupled-calibration procedure for the top and bottom mirrors identified by subscript $k \in \{t,b\}$, respectively. In other words, GUMS allowed us to map any keypoint located at ${}^{[\Xi_k]}\boldsymbol{m}_i$ on panoramic image Ξ_k into its back-

[2] `http://opencv.org`

projecting vector ${}^{[\mathrm{M_k}]}\hat{\mathbf{v}}_i$ with respect to the current frame's viewpoint $[\mathrm{M_k}]$. Fig. 2 illustrates the noncentral configuration and the reprojection error angles, θ_k, due to an estimated pose ${}^{[\mathrm{K}]}_{[\mathrm{C}]}\tilde{\mathbf{T}}$ out of a set of 3D-to-2D point correspondences.

The tracking termination criterion was $\tau_{P3P} = 3$, indicating the minimum number of unique keypoint features needed for the generation of a $SE3$ pose via P3P. As opposed to other full-fledged SLAM / SfM frameworks like PTAM[21] and ORB-SLAM[6] that keep on ignoring the lost tracker until another thread relocalizes the system out of newly arriving images, we instead terminated the VO immediately when the tracking was lost. This unforgiving termination criterion allowed us to measure each sensor's susceptibility to feature quantity information that we analyze in §4.3.

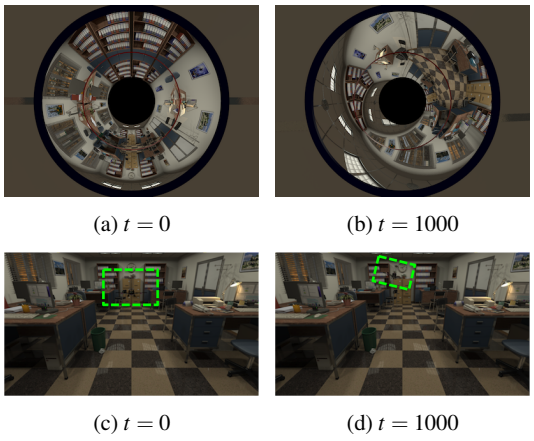
4 Experiments

For all our experiments, we kept consistent settings as stated in the implementation description (§3). We also limited the Euclidean distance of 3D point measurements according to our single-camera SOS capabilities: between 0.25 m and 20 m for disparities greater than 1 pixel. The range for the ASUS Xtion PRO LIVE used in the real experiments is between 0.8 m and 3.5 m as specified by the manufacturer, but it helped to increase it to up to 7 m in order to be compared with the SOS. The SOS was calibrated via a numerical method that approximated the GUMS through the reduction of projection error of control points as proposed in [16]. For the 100 mm-cell calibration chessboard located at ≈ 1.5 m, we obtained: a mean projection error of 1.5 ± 1 pixels, and a 3D triangulation accuracy of 0.018 ± 0.007 m based on the ground-truth information obtained via our motion capture system. On the other hand, for the RGB-D sensor, we used the factory-default settings for the depth-registered images without additional rectification.

4.1 Datasets and Experimental Configurations

For our experiments, we produced synthetic and real-life datasets for both the RGB-D camera and the single-camera SOS. For the `synthetic` dataset, we rendered photo-realistic scenes with the open-source raytracer POV-Ray[3]. The four trajectories that correspond to a real moving camera are based on the ICL-NUIM dataset [22]. Fig. 3 shows a couple of instances from the first sequence rendered using the theoretical hyperbolic single-camera SOS.

For real experiments, we collected ground-truth data indoors within a volume of $6\,\mathrm{m} \times 3\,\mathrm{m} \times 2\,\mathrm{m}$ using our VICON mocap system. The single-camera SOS prototype shown in Fig. 1 was rigidly attached next to the RGB-D baseline sensor. This SOS employed a Pointgrey Blackfly USB3 camera capturing 1920×1200 pixels images (global shutter) at 30 FPS. Each stereo panoramic image measured 1412×140 pixels as the example given in Fig. 7. The RGB-D camera ran at 30 FPS with VGA resolution of 640×480 pixels. Each sensor operated independently, but via the recorded time-stamps, we associated the images and the respective VICON poses. In order to link the observable ground-truth frame of the rig and the sensor's camera frame, the necessary hand-eye transformations were estimated separately. When ground-truth data is available, the poses are given in the standard TUM-format [23].



(a) $t = 0$ (b) $t = 1000$ (c) $t = 0$ (d) $t = 1000$

Fig. 3: A few frames at time-stamp t from the first `synthetic` sequence employing the single-camera hyperbolic SOS on top of an MAV. We show the omnidirectional images **(a,b)** above the associated external views **(c,d)**.

We grouped the real-life sequences according to aspects they were meant to address. The `conventional` sequences were choreographed for trajectories such as spinning in-place, walking around a square path, going up-and-down, and some free-style motion. We also recorded a trajectory going in/out of the mocap room into a hallway (≈ 50 m long), but for this sequence the ground-truth information only exists for the minority of the path. In addition, we captured sequences for various `special` conditions such as moving into a low-textured surface: blank wall or darker room. Those results are discussed in §4.3. Last, we experimented with `static` sensors using the identity transformation as their ground-truth pose for some variations of dynamic environments with people moving as described in §4.4.

4.2 Quantitative Evaluation Criteria

The relative pose error (RPE) [23] and the absolute trajectory error (ATE) [23, 22] are common evaluation metrics for VO algorithms. In our experiments, the camera poses were

[3] `http://www.povray.org`

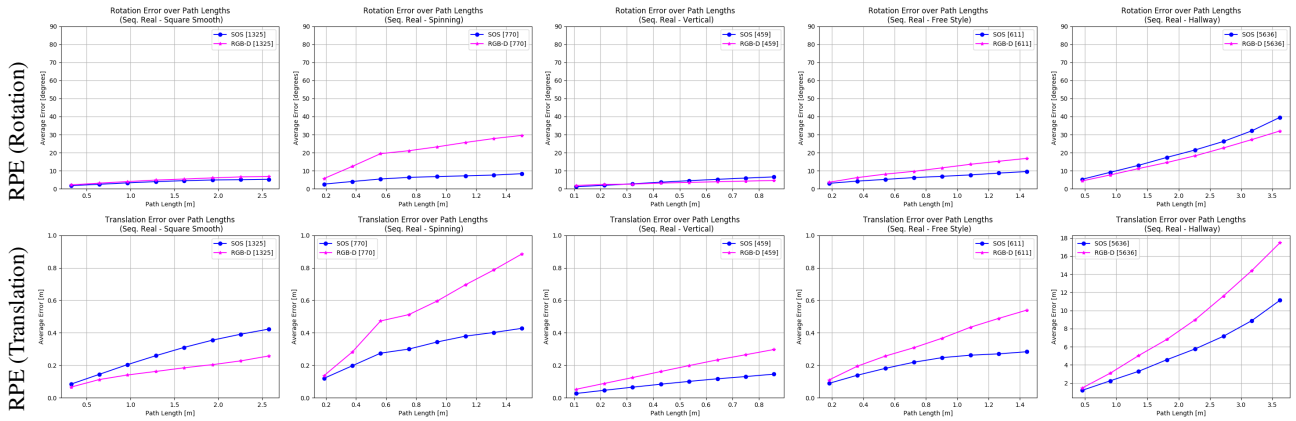
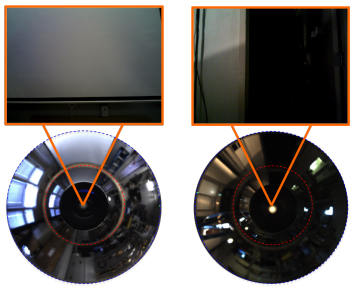


Fig. 4: RPE for some sequences moving in a `conventional` fashion. Here, the performance of both sensors is comparable.

computed without the scale ambiguity since both sensors operated at metric scale. ATE has a bias toward shorter path lengths, but RPE does not. Hence, RPE is capable of measuring the drift more indicatively than ATE, so we mainly interpret our results with regards to RPE computed among a set of 8 uniformly-divided path lengths for each trajectory. We arbitrarily set the longest path length to be $\frac{1}{3}$ of the complete path length, and we sampled the 8 segments at each frame in the trajectory. The plots in Fig. 4 and 6 demonstrate the accuracy measured via these RPE criteria for the available real-life sequences with ground-truth information. In Table 1, we provide the overall averages out of the existing sampled errors that have been normalized by the corresponding path lengths (in meters). Because we cannot compute the RPE or ATE when no ground-truth information exists, it was useful to measure the total path length in terms of number of frames successfully tracked. We expressed this quantity within square brackets in the plots' legends as an indicative of tracking loss. Since each experiment was non-numerically repeatable due to the use of RANSAC in the VO algorithm, we ran 3 trials and averaged their respective results for evaluation. Because noise in a real sensor is unavoidable, we validated our VO implementation by evaluating the pose estimation accuracy of the `synthetic` sequences (§4.1), where we noticed that the estimated 3D trajectories followed the ground-truth very closely. Table 1 also includes the results for these four sequences via a calibrated GUMS reflecting the theoretical centrality of the system. For the real-life sequences, the theoretical model did not produce any meaningful results, so the calibrated GUMS (§2.2) was necessary. In Fig. 4 and Table 1, we observe that the single-camera SOS produced slightly more accurate results than the RGB-D sensor for the majority of these sequences. This asserts its real-life utilization for VO estimation indoors (at the moment). In what follows, we experimentally evaluate the single-camera SOS performance under a series of



(a) Into a wall. (b) Into the dark.

Fig. 5: Illustrating the higher vulnerability that the RGB-D camera (top) has due to the lack of features to track in the scene. The associated omnidirectional images (bottom) have a higher probability to detect features under these special circumstances. Data with these issues are provided in the sequences from the `special` group of the dataset.

situations provoking dilemmas in real-world robotic navigation.

4.3 Feature-based Special Issues

It is evident that occlusions and the lack of distinguishable points in the scene are a problem for all feature-based VO methods. A very probable situation arises when the perception device gets too close to a surface covered by ambiguous patterns or not presenting recognizable corners for detection as illustrated in Fig. 5. The $SE3$ tracking process in a complete SLAM framework would usually get lost when this happens and try to re-localize with new frames. As with any omnidirectional camera, our SOS is less likely to be affected by the low presence of features in the environment due to its broader view of the scene [10]. The feature-availability likelihoods of the sensors depend mainly on their vertical and horizontal FOV angles. The wider FOV of the SOS helps

increase its likelihood of surviving the critical region threshold, τ_{P3P}. Experimentally, we showed the RGB-D sensor is more vulnerable to these feature-based issues due to its short range and FOV limitations. Both Fig. 6 and Table 1 support our claim of a better VO performance with the SOS in these `special` setups. In fact, VO with the RGB-D sensor failed most of the time, which is indicated by the fewer frames used for assessment.

4.4 Dynamic Environment Issues

The work by Xiong et al. [24] motivated us to investigate the capabilities of our catadioptric SOS in dynamic environments even though their work actually intended to detect moving objects, and not the system's egomotion. The undesirable pose estimation issues caused by dynamic features on the ground can be alleviated by directing the camera at the ceiling. However, our single-camera SOS configuration was designed to view around the equatorial region (§2.2), so the presence of dynamic features and occlusions holds with higher probability. An alternative is to combine a multitude of camera views to improve the VO's resiliency in dynamic environments, as what the Collaborative Visual SLAM by Zou and Tan [3] attempted. In our study, we elaborated various cases to measure the accuracy of the sensor's pose estimation (VO) when operating in a dynamic environment. We arbitrarily measured the effect of the proximity to the sensor as well as the density of moving subjects for the corresponding `dynamic` sequences collected in our dataset (§4.1). It is obvious that the VO of any camera is more susceptible to error in scenes with a higher dynamic density, so we evaluated this effect by arranging the number of people in the view as well as by controlling the distance to the sensors: about 1 m, 2 m, and 3 m away for a duration of ≈ 25 s. We also collected data from an uncontrolled setup in a public setup. Table 2 contains the resulting mean and standard deviation values for the translation and rotation components of the RPE computed for every tracked frame with the rig kept `static`. Here, we noticed that the SOS was less susceptible to the dynamic outliers due to the increased likelihood for sampling static features captured by its wider FOV. As expected, if the density of dynamic objects increases, both sensors are comparably affected. Due to perspective projection, the effect of the distance to the camera is the change in the image area occupied by the dynamic objects. In fact, when moving the rigs in a dynamic scene as detailed in the last two rows of Table 1, the RGB-D sensor was more affected when the dynamic objects moved relatively slower than the camera's true speed because feature outliers were tracked with higher confidence levels among frames. In Fig. 7d, we show an erroneous tracking instance for which the RGB-D camera got lost in the `stairs` sequence after a person passed by too close in-front of it. Although the camera was moving this time, this issue also relates to the results for the `static` rigs with moving people at different ranges given in Table 2. The public trajectories took place in Grand Central Terminal (GCT) in New York City. In this highly dynamic environment filled by both natural and artificial lighting, we were able to compare the qualitative VO performance of the systems while walking. The trajectories estimated via the SOS are visualized in Fig. 8. The corresponding video is available at `http://youtu.be/c5tyHqEkKQA`, where we can see the inconspicuous drift when returning to the starting points, and we can witness how the RGB-D camera gets lost when taking the `stairs`. Indeed, for the walk around the GCT `clock`, both sensors appeared to perform equivalently, but we had intentionally assisted the RGB-D camera's sensing range by directing it toward the information booth underneath the clock, where mostly static features existed during this experiment. The related issue due to lack of trackable features for the RGB-D sensor was presented and discussed in §4.3. Notice that we could not assume that only the foreground was dynamic and that the background was mostly static because both were dynamic. In fact, this dual effect was more pronounced for the SOS, but the RGB-D sensor was mainly affected by the dynamic foreground due to its shorter perception range.

5 Conclusion

We presented the application of a single-camera stereo omnidirectional system (SOS) for egomotion. We are aware of stereo vision limitations such as the fixed baseline not being able to accommodate to the varying distances of objects under different circumstances. Our SOS has a fixed baseline, which is comparable to that of a traditional RGB-D sensor, so we directly compared it against. We performed experiments indoors within the practical sensing ranges of both devices. We evaluated the error in the estimated 3D visual odometry (VO) with the sequences that we have collected with associated ground-truth information where possible. We are making our datasets (§4.1) and implementation (§3) publicly availablefor the reproducibility of our findings (§4). Our experiments showed that both sensors are capable to achieve a comparable VO performance under conventional circumstances. Moreover, we have demonstrated the apparent advantages of the SOS under dynamic environments and where the number of distinguishable features decays. Under those circumstances, the wide-viewing angle (FOV) of the singe-camera SOS allows for the detection of features on the 360° panoramic images with more reliability than other 3D cameras with narrower FOV. We kept our VO implementation as consistent as possible in order to evaluate the frame-based tracking functionality of this unique SOS against the RGB-D camera. Without applying

Table 1: Average absolute trajectory error (ATE) and relative pose error (RPE, normalized) for the moving experiments.

	Sequence	ATE [m]		RPE (Translation) [%]		RPE (Rotation) [°/m]		Frames [#]	
		SOS	RGB-D	SOS	RGB-D	SOS	RGB-D	SOS	RGB-D
`Conventional`	Square Small	**0.12** ± 0.05	0.70 ± 0.23	**25.94** ± 8.18	41.76 ± 78.21	**3.46** ± 2.76	22.37 ± 45.52	619	619
	Square Smooth	**0.12** ± 0.06	0.14 ± 0.11	20.51 ± 8.54	**13.92** ± 7.89	**3.29** ± 1.97	4.11 ± 2.51	1325	1325
	Spinning	**0.30** ± 0.11	0.35 ± 0.08	**42.60** ± 19.59	68.30 ± 80.08	**8.64** ± 4.78	27.08 ± 36.18	770	770
	Vertical	**0.04** ± 0.02	0.14 ± 0.06	**19.86** ± 5.31	39.06 ± 16.12	**8.68** ± 3.55	8.76 ± 6.05	459	459
	Free Style	**0.14** ± 0.05	0.41 ± 0.14	**31.34** ± 13.57	45.75 ± 54.99	**9.65** ± 5.21	14.53 ± 20.35	611	611
	Hallway	0.95 ± 0.58	**0.81** ± 0.56	**262.53** ± 546.43	391.20 ± 763.28	10.14 ± 18.98	**8.54** ± 12.61	5636	5636
`Special`	Into Wall - Regular	**0.13** ± 0.04	0.22 ± 0.08	**37.70** ± 11.00	130.41 ± 79.68	**6.70** ± 1.86	57.27 ± 42.25	**1041**	315
	Into Wall - Slow	**0.09** ± 0.03	0.19 ± 0.09	**37.21** ± 10.64	165.29 ± 106.22	**6.72** ± 1.99	76.63 ± 65.46	**1400**	391
	Into Wall - Fast	**0.09** ± 0.04	0.18 ± 0.09	**36.02** ± 9.61	115.06 ± 79.37	**5.83** ± 1.91	47.08 ± 36.90	**896**	251
	Into Wall - Curvy	0.28 ± 0.08	**0.22** ± 0.11	**35.45** ± 14.35	136.35 ± 123.50	**6.39** ± 2.82	77.65 ± 81.61	**838**	309
	Into Dark - Straight	**0.06** ± 0.03	0.53 ± 0.24	**16.87** ± 6.31	213.43 ± 218.62	**4.54** ± 1.99	19.38 ± 14.02	**998**	554
	Into Dark - Turning	**0.13** ± 0.06	0.73 ± 0.23	**14.92** ± 6.07	141.20 ± 177.83	**5.50** ± 2.36	32.95 ± 37.47	1260	1260
`Dyn.`	Slow Dynamic	**0.02** ± 0.01	0.24 ± 0.15	**25.42** ± 9.65	483.48 ± 808.29	**7.53** ± 3.87	84.34 ± 153.37	**390**	278
	Fast Dynamic	**0.03** ± 0.01	0.17 ± 0.07	**23.31** ± 13.51	129.47 ± 104.12	**6.60** ± 3.75	21.79 ± 18.26	518	518
`Synth.`	Office-0	**0.03** ± 0.01	0.12 ± 0.06	**4.56** ± 2.24	14.45 ± 8.91	**0.83** ± 0.49	3.16 ± 2.24	1508	1508
	Office-1	**0.02** ± 0.01	0.13 ± 0.06	**3.13** ± 1.49	17.19 ± 16.08	**0.71** ± 0.37	5.99 ± 3.74	965	965
	Office-2	**0.04** ± 0.02	0.16 ± 0.06	**3.23** ± 1.60	8.93 ± 4.97	**0.75** ± 0.43	2.58 ± 1.37	880	880
	Office-3	**0.03** ± 0.02	0.05 ± 0.03	**3.57** ± 2.12	7.89 ± 7.14	**0.68** ± 0.40	2.98 ± 2.22	1240	1240

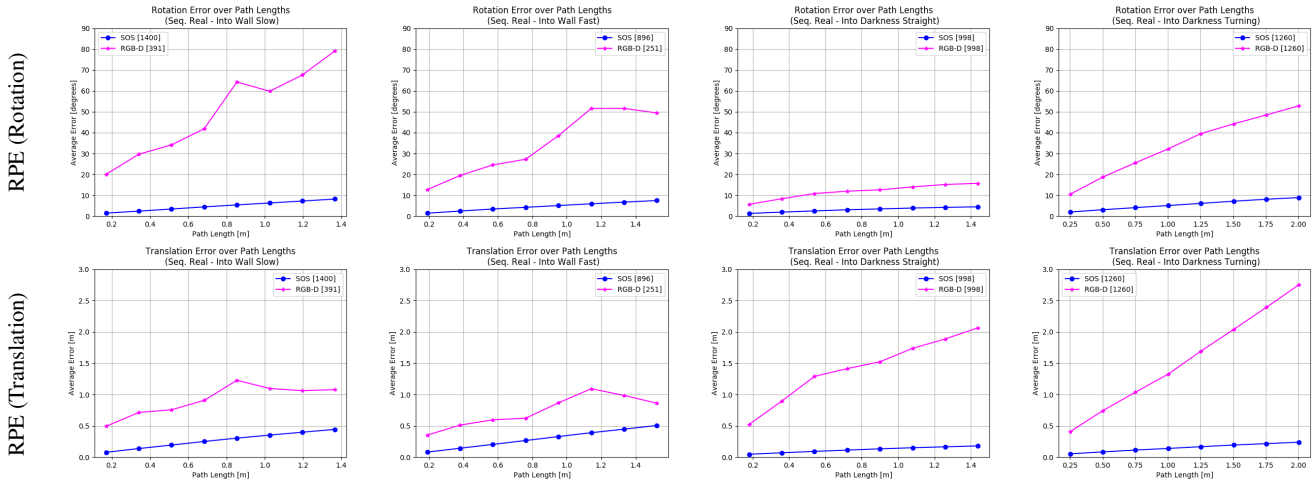
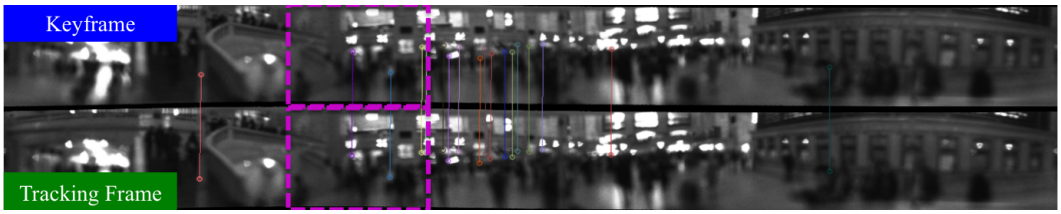


Fig. 6: RPE results of visual odometry estimation for some `special` sequences moving into poorly textured scenes, i.e., surface lacking features or too dark to find them. The value inside the brackets specify the number of frames that were successfully tracked before getting lost due to the lack of point features needed to solve PnP according to §3.

Table 2: `Static` rigs in dynamic environments

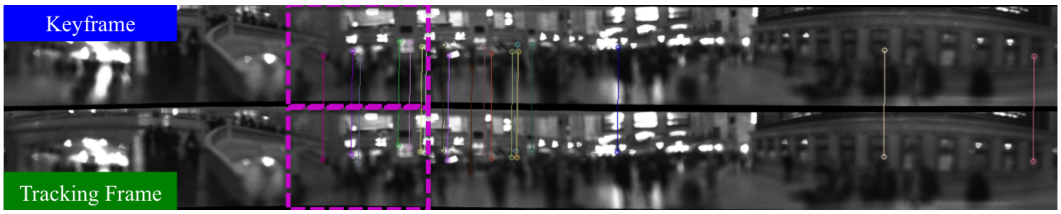
Prox [m]	Peop [#]	Translation Error [m] SOS	Translation Error [m] RGB-D	Rotation Error [°] SOS	Rotation Error [°] RGB-D
1	1	**0.021** ± 0.008	0.279 ± 0.207	**0.310** ± 0.120	5.230 ± 3.160
1	2	**0.021** ± 0.006	0.600 ± 0.348	**0.380** ± 0.080	8.720 ± 6.000
1	4	**0.047** ± 0.017	1.614 ± 0.768	**0.610** ± 0.240	21.320 ± 10.820
2	1	**0.015** ± 0.005	0.586 ± 0.330	**0.230** ± 0.070	7.940 ± 4.220
2	2	**0.017** ± 0.007	1.049 ± 0.404	**0.250** ± 0.120	17.300 ± 9.310
2	4	**0.030** ± 0.008	2.247 ± 0.982	**0.570** ± 0.270	13.400 ± 6.370
3	1	**0.013** ± 0.004	1.029 ± 0.632	**0.140** ± 0.040	11.940 ± 8.900
3	2	**0.022** ± 0.005	1.728 ± 0.598	**0.340** ± 0.090	24.240 ± 13.100
3	4	**0.028** ± 0.007	1.854 ± 0.681	**0.460** ± 0.110	13.710 ± 12.780
Var	2	**0.049** ± 0.021	0.481 ± 0.260	**0.900** ± 0.310	17.980 ± 10.900
Var	Var	**0.374** ± 0.193	4.487 ± 1.261	**5.630** ± 3.600	39.390 ± 11.120

any graph optimization or loop-closure techniques for localization and mapping, we concentrated in the actual VO performance of the sensors. This allowed us to plan future solutions to mitigate difficulties such as the eminent drift of the pose estimation without obscuring the root of the problem. In the presented work, our goal was to demonstrate the egomotion capabilities of the single-camera SOS for practical issues that real-world navigation may encounter.



(a) SOS frame successfully tracked at time t. Showing correspondences between top panoramic images.

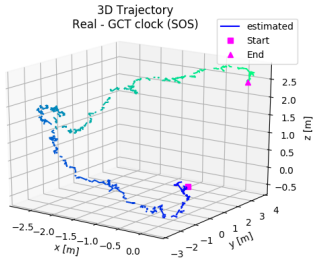
(b) RGB-D okay



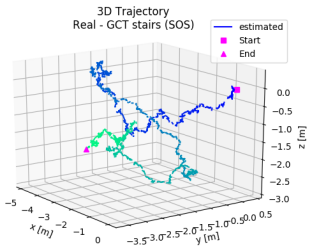
(c) SOS frame successfully tracked at time $t+1$. Showing correspondences between top panoramic images.

(d) RGB-D lost

Fig. 7: Example from the `stairs` sequence at GCT: the single-camera SOS succeeded while the RGB-D sensor failed due to the busy highly dynamic scene diminishing the number of trackable keypoint features when occluded by a sudden passerby in the view. The dotted boxes in the panoramas corresponds to the view in common with the RGB-D.



(a) Around the GCT `clock`.



(b) Taking the GCT `stairs`.

Fig. 8: Estimated 3D trajectory of the SOS in real-life environments: (a) walk around the `clock` in GCT; and (b) taking the `stairs` down and back up to the main lobby at GCT. Note: these trajectories are not fit to the ground plane.

Acknowledgment

We thank the MTA Metro-North Railroad for letting us collect video sequences at the main lobby of the Grand Central Terminal in NYC.

References

1. C. Jaramillo, R. G. Valenti, L. Guo, and J. Xiao, "Design and Analysis of a Single-Camera Omnistereo Sensor for Quadrotor Micro Aerial Vehicles (MAVs)," *Sensors*, vol. 16, no. 2, p. 217, 1 2016.
2. Z. Zhang, H. Rebecq, C. Forster, and D. Scaramuzza, "Benefit of Large Field-of-View Cameras for Visual Odometry," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA)*, 2016.
3. D. Zou and P. Tan, "Coslam: Collaborative visual slam in dynamic environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 354–366, 2013.
4. R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "DTAM: Dense tracking and mapping in real-time," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, Nov. 2011, pp. 2320–2327.
5. J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular SLAM," in *Proc. European Conf. Computer Vision (ECCV)*, Sept. 2014.
6. R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras," *IEEE Trans. Robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.
7. J.-P. Tardif, Y. Pavlidis, and K. Daniilidis, "Monocular visual odometry in urban environments using an omnidirectional camera," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS)*, 2008.
8. A. Rituerto, L. Puig, and J. J. Guerrero, "Visual SLAM with an Omnidirectional Camera," in *Proc. IEEE International Conf. Pattern Recognition (ICPR)*, 8 2010, pp. 348–351.
9. D. Gutierrez, A. Rituerto, J. M. M. Montiel, and J. J. Guerrero, "Adapting a Real-Time Monocular Visual SLAM from Conventional to Omnidirectional Cameras," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV) Workshops*, 2011.
10. T. Lemaire and S. Lacroix, "SLAM with panoramic vision," *Journal of Field Robotics*, vol. 24, no. 1-2, pp. 91–111, 2007.
11. C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems and practical implications," *Proc. European Conf. Computer Vision (ECCV)*, pp. 445–461, 2000.
12. E. Rosten, R. Porter, and T. Drummond, "Faster and better: a machine learning approach to corner detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 105–119, 1 2010.
13. M. Schönbein and A. Geiger, "Omnidirectional 3D Reconstruction in Augmented Manhattan Worlds," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS)*, 2014.
14. Z. Zhu, "Omnidirectional stereo vision," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA)*, 2001.
15. G. Jang, S. Kim, and I. Kweon, "Single camera catadioptric stereo system," in *OMNIVIS Workshop*, 2005.
16. C. Jaramillo, R. G. Valenti, and J. Xiao, "GUMS: A Generalized Unified Model for Stereo Omnidirectional Vision (Demonstrated Via a Folded Catadioptric System)," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS)*, 2016, pp. 2528–2533.
17. L. Kneip and P. Furgale, "OpenGV: A unified and generalized approach to real-time calibrated geometric vision," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA)*, 2014.

18. J. Shi and C. Tomasi, "Good Features to Track," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 1994, pp. 593–600.
19. E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An Efficient Alternative to SIFT or SURF," in *Proc. IEEE Int'l Conf. Computer Vision (ICCV)*, 2011.
20. L. Kneip, P. Furgale, and R. Siegwart, "Using multi-camera systems in robotics: Efficient solutions to the NPnP problem," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA)*, no. 2004, 2013.
21. G. Klein and D. Murray, "Parallel tracking and mapping for small AR workspaces," in *Proc. IEEE Int'l Symp. Mixed and Augmented Reality (ISMAR)*, Nov. 2007, pp. 1–10.
22. A. Handa, T. Whelan, J. B. McDonald, and A. J. Davison, "A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM," in *Proc. IEEE Int'l Conf. Robotics and Automation (ICRA)*, 2014.
23. J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of RGB-D SLAM systems," in *Proc. IEEE/RSJ Int'l Conf. Intelligent Robots and Systems (IROS)*, Oct. 2012, pp. 573–580.
24. Z. Xiong, W. Chen, and M. Zhang, "Catadioptric Omni-directional Stereo Vision and Its Applications in Moving Objects Detection," in *In Tech: Computer Vision*, 2008, no. November, ch. 26, pp. 493–538.