

Comparative Study of a commercial tracking camera and ORB-SLAM2 for person localization

Safa Ouerghi, Nicolas Ragot, Rémi Boutteau, Xavier Savatier

▶ **To cite this version:**

Safa Ouerghi, Nicolas Ragot, Rémi Boutteau, Xavier Savatier. Comparative Study of a commercial tracking camera and ORB-SLAM2 for person localization. 15th International Conference on Computer Vision Theory and Applications, Feb 2020, Valletta, France. pp.357-364, 10.5220/0008980703570364 . hal-02567816

HAL Id: hal-02567816
https://hal.science/hal-02567816

Submitted on 8 May 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Comparative Study of a commercial tracking camera and ORB-SLAM2 for person localization

Safa Ouerghi[1], Nicolas Ragot[1], Remi Boutteau[1] and Xavier Savatier[1]
[1]*Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, 76000 Rouen, France*
email{Safa.Ouerghi, Nicolas.Ragot, Remi.Boutteau, Xavier.Savatier }@esigelec.fr

Keywords: Intel T265, ORB-SLAM2, benchmarking, person localization.

Abstract: Aiming at localizing persons in industrial sites is a major concern towards the development of the factory of the future. During the last years, developments have been made in several active research domains targeting the localization problem, among which the vision-based Simultaneous Localization and Mapping paradigm. This has led to the development of multiple algorithms in this field such as ORB-SLAM2 known to be the most complete method as it incorporates the majority of the state-of-the-art techniques. Recently, new commercial and low-cost systems have also emerged in the market that can estimate the 6-DOF motion. In particular, we refer here to the Intel Realsense T265, a standalone 6-DOF tracking sensor that runs a visual-inertial SLAM algorithm and that accurately estimates the 6-DOF motion as claimed by the Intel company. In this paper, we present an evaluation of the Intel T265 tracking camera by comparing its localization performances to the ORB-SLAM2 algorithm. This benchmarking fits within a specific use-case: the person localization in an industrial site. The experiments have been conducted in a platform equipped with a VICON motion capture system, which physical structure is similar to a one that we could find in an industrial site. The Vicon system is made of fifteen high-speedtracking cameras (100 Hz) which provides highly accurate poses that were used as ground truth reference. The sequences have been recorded using both an Intel RealSense D435 camera to use its stereo images with ORB-SLAM2 and the Intel RealSense T265. The two sets of timestamped poses (VICON and the ones provided by the cameras) were aligned then calibrated using the point set registration method. The Absolute Trajectory Error, the Relative Trajectory Error and the Euclidian Distance Error metrics were employed to benchmark the localization accuracy from ORB-SLAM2 and T265. The results show a competitive accuracy of both systems for a handheld camera in an indoor industrial environment with a better reliability with the T265 Tracking system.

1 INTRODUCTION

Improving the performance and safety conditions in industrial sites represent a major challenge that particularly requires people tracking to verify in real time their authorization to accomplish the task they are doing. To be able to fulfill such a high level task, the pose of humans in the industrial space has to be accurately known.

Within the context of localization and tracking, developments in several active research fields such as SLAM (Simultaneous Localization and Mapping), computer vision, Augmented Reality (AR), Virtual Reality (VR), indoor Geographic and Information Systems (GIS) have been made. Nowadays, visual SLAM (V-SLAM) for tracking is a systematic problem. The core of the algorithm development has become mature, but the success still relies on a complete and robust hardware-software solution that fits within the application.

For instance, the localization issue has primarily been tackled within traditional industrial applications and in autonomous vehicles that involve robots with limited mobility and a defined kinematic model. However, in most studies, the SLAM used by humans and humanoid robots doesn't make specific optimization for the motion characteristics. It rather directly carries experiments and evaluates the results of other SLAM modules. Hence, although the maturity of SLAM, new applications imply that additional experiments have to be carried out.

Furthermore, over the last few years, new sensors such as Time Of Flight (TOF) and RGB-D cameras have pushed the boundaries of robot perception significantly (Zollhfer et al., 2018). The maturity of V-SLAM has also contributed to the emergence of low-cost systems in the market such as the Intel RealSense T265 tracking camera that estimates the 6-DOF mo-

tion (Intel, 2019). Thus, it would be worth considering to investigate such available commercial visual sensors and discuss about their usability for a reliable tracking of a human as for instance in the context of industrial environments.

This paper aims at evaluating the tracking performance of the new imaging and tracking system, the Intel RealSense T265 released by Intel in 2019 by comparing its performances to ORB-SLAM2 (Mur-Artal and Tardos, 2016). ORB-SLAM2 has been particularly chosen as it is one of the most accurate open-source V-SLAM algorithms that integrates the majority of state-of-the-art techniques including multi-threading, loop-closure detection, relocalization, bundle adjustment and pose graph optimization. As has been previously stated, the context of the benchmarking involves a hand-held camera by a person moving in an industrial environment.

The paper is organized as follows: section 2 presents some works including V-SLAM and new imaging systems. Section 3 gives details about the used sensors in this study as well as the evaluation metrics used to assess their performance. Section 4 highlights the calibration method used between the camera estimation and the VICON's one to put both of them in the same reference frame. Finally, Section 5 presents a comparative study between the RealSense T265 tracking camera and the stereo ORB-SLAM2 followed by a discussion of the findings and conclusions.

2 RELATED WORK

Our work is related to the fundamental and heavily researched problem in computer vision: the visual SLAM, through the comparison of the performances of the new low-cost RealSense tracking sensor T265 and the RealSense D435 coupled with the ORB-SLAM2 algorithm running in the stereo mode. The history of the research on SLAM has been over 30 years, and the models for solving the SLAM problem can be divided into two main categories: filtering based methods and graph optimization based methods. The filtering based methods usually use the Extended Kalman Filter (EKF), Unscented Kalman Filter (UKF) or Particle Filter (PF). These methods first predict both the pose and the 3D features in the map and then update these latters when a measurement is acquired. The state of the art key-methods based on filtering are the MonoSLAM (Davison et al., 2007) that uses an EKF and FastSLAM (Montemerlo et al., 2002) that uses a PF. The methods based on graph optimization generally use bundle adjustment to simultaneously optimize the poses of the camera and the 3D points of the map which corresponds to an optimization problem. A key-method is PTAM proposed by Klein et al. (Klein and W. Murray, 2009) which introduced the separation of the localization and mapping tasks into different threads and performing bundle-adjustment on keyframes in order to be able to meet the real-time constraint. ORB-SLAM uses multi-threading and keyframes as well (Mur-Artal et al., 2015) and could be considered as an extension of PTAM. On top of these functionalities, ORB-SLAM performs loop-closing and the optimization of a pose-graph. ORB-SLAM was first introduced to work with monocular cameras and has subsequently been extended to stereo and RGB-D cameras in (Mur-Artal and Tardos, 2016). It therefore represents the most complete approach in the state-of-the-art-methods and has been used as a reference method in several works. Moreover, a popular research axis in SLAM is the visual-inertial SLAM based on the fusion of vision sensor measurements with an Inertial Measurement Unit (IMU). As well as visual SLAM, VI-SLAM methods can be divided into filtering-based and optimization-based. A review of the main VI-SLAM methods has been presented in (Chang et al., 2018).

In addition, new camera technologies have been investigated in the context of visual SLAM. RGB-D cameras have been extensively used in recent years and several works document their performance. In (Weng Kuan et al., 2019), a comparison of three RGB-D sensors that use near-infrared (NIR) light projection to obtain depth data is presented. The sensors are evaluated outdoors where there is a strong sunlight interference with the NIR light. Three kinds of sensors have been used namely a TOF RGB-D sensor, the Microsoft Kinect v2, a structured-light (SL) RGB-D sensor, the Asus Xtion Pro Live and an active stereo vision (ASV) sensor the Intel RealSense R200. These three sensors have been as well compared in the context of indoor 3D reconstruction and concluded that the Kinect v2 has better performance in returning less noisy points and denser depth data. In (Yao et al., 2017), a spatial resolution comparison has been presented between Asus Xtion Pro, Kinect v1, Kinect v2 and the R200. This comparison showed that the Kinect v2 performs better than both the Primesense sensors and the Intel R200 indoors. In (Halmetschlager-Funek et al., 2019), ten depth cameras have been evaluated. The experiments have been performed in terms of several evaluation metrics including bias, precision, lateral noise, different lighting conditions, materials and multiple sensor setups in indoor environments. Authors expressed

that the Microsoft Kinect v2 behaves significantly differently compared to the other sensors as it outperforms all sensors regarding, bias, lateral noise and precision for $d > 2m$ and is less precise for the range $0.7m < d < 2m$ than the structured light sensors.
Recently, Intel has released the Tracking sensor T265 (Intel, 2019) from the Intel RealSense line of products. The T265 is a standalone tracking camera that uses a proprietary visual inertial SLAM algorithm for accurate and low-latency tracking targeting multiple applications such as robotics, drones, augmented reality (AR), and virtual reality. Current literature does not seem to include any research work to directly compare the performance of the T265 tracking camera with an existing state-of-the-art algorithm. The investigation is therefore centered on comparing the T265 RealSense camera with the stereo ORB-SLAM2.

3 MATERIALS AND EVALUATION METRICS

This section briefly describes the operational principles of the used camera sensors namely the Intel RealSense D435 and the Intel RealSense T265. Then, it presents the evaluation metrics used to benchmark the performance of the T265 tracking system vs ORB-SLAM2.

3.1 Characteristics of the Used Sensors

3.1.1 The Intel D435 RealSense Depth camera

The Intel RealSense D400 depth camera series technology represent an important milestone as it introduces inexpensive, easy-to-use 3D cameras for both indoor and outdoor. The Intel D435 is the successor of the depth camera D415. Both are stereo cameras and have an Infra Red (IR) projector to obtain a good field rate and an RGB camera as well. The difference between them is that D435 has a wider field of view. The RealSense D-400 series support depth output and enable capturing a disparity between images up to a 1280×720 resolution, at up to 90 fps (Intel, 2017).

3.1.2 The Intel T265 tracking camera

The Intel RealSense Tracking Camera T265 is a standalone 6-DOF tracking sensor that runs a visual-inertial SLAM algorithm onboard. It can, additionally, integrate wheel odometry for greater robustness in robotics. The T265 uses inputs from dual fisheye cameras and an IMU along with processing capabilities from the Movidius MA215x ASIC allowing it to be a low-power, high performance device, adequate for embedded systems. The SLAM algorithm running onboard is a proprietary algorithm based on fusing images, inertial data, sparse depth and wheel odometry if available in an embedded system. It also uses a Sparse-Kalman filtering approach, poses at $200Hz$ and an appearence-based relocalization. Intel claims that the loop-closure error is below 1% of path length. Intel states that the T265 tracking camera is for use with drones, robots and AR/VR applications. In fact, the two fisheye cameras provide a large field of view for robust tracking even with fast motion. However, unlike previous Intel RealSense cameras such as the D400 series, the T265 is not a depth camera. Intel does note that it is possible to use the image feed from the two fisheye lenses and sensors to compute dense depth, but the results would be poor compared to other RealSense depth cameras, as the lenses are optimized for wide tracking field of view, rather than depth precision, and there is no texture projected onto the environment to aid in depth fill. However, the T265 can be paired with the RealSense D400 camera for increased capabilities where the tracking camera and the depth camera are used in combination as for instance for occupancy mapping and obstacle avoidance (Intel, 2019).

3.2 Evaluation metrics

For evaluating the trajectory accuracy, some evaluation metrics have been employed including the absolute trajectory error (ATE) and the Relative Pose Error (RPE) as presented by Sturm et al. (Sturm et al., 2012). We, additionally, use the Euclidean Error (EE) to benchmark the T265 tracking performance.

3.2.1 Absolute trajectory error

This metric evaluates the global consistency of the estimated trajectory by comparing the absolute distances between the estimated and the ground truth ones. This metric was introduced in (Sturm et al., 2012) and consists first in aligning the two trajectories and then evaluating the root mean squared error over all time indices of the translational components. The alignment allows to find the rigid-body transformation $\mathbf{S}$ referring to the least-squares solution that maps the estimated trajectory $\mathbf{P}_{1:n}$ onto the ground truth trajectory $\mathbf{Q}_{1:n}$, where n is the number of poses. Hence, the absolute trajectory error $\mathbf{F}_i$ at time step i can be computed as

$$\mathbf{F}_i = \mathbf{Q}_i^{-1}\mathbf{S}\mathbf{P}_i. \quad (1)$$

The root mean squared error over all time indices of the translational components could, hence, be evaluated which refers to

$$\text{RMSE}(\mathbf{F}_{1:n}) = (\frac{1}{n}\sum_{i=1}^{n} \|trans(\mathbf{F}_i\|^2)^{1/2}. \quad (2)$$

3.2.2 Relative Pose Error

The RPE measures the local accuracy of a trajectory over a fixed time interval which refers to the drift in a trajectory suitable for evaluating visual odometry systems. While the ATE assesses only the translational errors, the RPE evaluates both: the translational and rotational errors. Therefore, the RPE is always greater than the ATE (or equal if there is no rotational error). The RPE metric gives, indeed, a away to combine rotational and translational errors into a single measure. The instant RPE is defined at time step i as $\mathbf{E}_i$

$$\mathbf{E}_i = (\mathbf{Q}_i^{-1}\mathbf{Q}_{i+\Delta})^{-1}(\mathbf{P}_i^{-1}\mathbf{P}_{i+\Delta}), \quad (3)$$

where Δ is a fixed time interval that needs to be chosen. For instance, for a sequence recorded at 30 Hz, $\Delta = 30$ gives the drift per second which is useful for visual odometry systems as previously stated. From a sequence of n camera poses, we obtain in this way $m = n - \Delta$ individual relative pose errors along the sequence. From these errors the RMSE over all time indices of the rotation component is computed.

$$\text{RMSE}(\mathbf{E}_{1:n}, \Delta) = (\frac{1}{m}\sum_{i=1}^{m} \|trans(\mathbf{E}_i\|^2)^{1/2}. \quad (4)$$

In fact, it has been reported in (Sturm et al., 2012) that the comparison by translational errors is sufficient as rotational errors show up as translational errors when the camera is moved. For SLAM systems, this metric is used by averaging over all possible time intervals by computing

$$\mathbf{E}_{1:n} = \frac{1}{n}\sum_{\Delta=1}^{n} \text{RMSE}(E_{1:n}, \Delta). \quad (5)$$

3.2.3 Euclidian Error

We report the use of the Euclidian Distance Error root-mean squared as an additional evaluation metric. As we are targeting the localization of a person, we use this metric to evaluate the pose error on the ground plane. We define the root-mean squared Euclidian Error (EE) as ε

$$\varepsilon = \text{RMSE}(\mathbf{T}_{1:n}) = (\frac{1}{n}\sum_{i=1}^{n} \mathbf{T}_i^2)^{1/2}, \quad (6)$$

where $\mathbf{T}_i$ is the magnitude of the Euclidean Distance Error along the horizontal plane between the estimated and the ground truth pose at instant i.

4 GEOMETRIC CALIBRATION BETWEEN THE CAMERA AND THE VICON

The extrinsic calibration consists in estimating the relative pose between the camera and the VICON motion capture system. The camera sensors D435, T265 and the markers tracked by the VICON system are rigidly attached to the same support. The knowledge of this rigid transformation between the camera's optical center and the VICON's reference is essential in order to express the camera estimate in the VICON's reference frame. This implies first the time alignment of the poses and then the estimation of the rigid body transformation.

4.1 Time alignment

This step is essential in order to synchronize the timestamped data of the two sensors. An *opensource* method presented in (Furrer et al., 2017) has been used. This method relies on first resampling the poses at the lower frequency of the two pose signals, then, correlating the angular velocity norms of both of them.

4.2 Rigid transformation

The transformation is calculated using the corresponding point set registration. Considering two sets of 3D points, Set_{vicon} and Set_{camera} with Set_{vicon} given in the VICON's reference frame and Set_{camera} given in the camera's coordinate frame, solving for $\mathbf{R}$ and $\mathbf{t}$ from:

$$Set_{vicon} = \mathbf{R}.Set_{camera} + \mathbf{t}, \quad (7)$$

allows to find the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{t}$ that transform the points from the camera's frame to the VICON's frame. This consists in finding the optimal rigid transformation. First, the centroids of the two datasets are found using

$$centroid_{vicon} = \frac{1}{N}\sum_{i=1}^{N} P^i_{vicon}, \quad (8)$$

$$centroid_{camera} = \frac{1}{N}\sum_{i=1}^{N} P^i_{camera}, \quad (9)$$

where N is the number of corresponding points in the two datasets, P_{vicon} a 3D point in the VICON's frame and P_{camera} the corresponding point in the camera's frame with $P = [xyz]^T$.
The rotation matrix $\mathbf{R}$ is found by SVD where $\mathbf{H}$ is

first calculated

$$\mathbf{H} = \sum_{i=1}^{N} (P^i_{vicon} - centroid_{vicon})(P^i_{camera} - centroid_{camera}) \tag{10}$$

Then,

$$[\mathbf{U}, \mathbf{S}, \mathbf{V}] = \text{SVD}(\mathbf{H}) \tag{11}$$

allows to find the rotation matrix by performing

$$\mathbf{R} = \mathbf{V}.\mathbf{U}^T. \tag{12}$$

The translation vector is also found by using

$$\mathbf{t} = -\mathbf{R}.centroid_{camera} + centroid_{vicon}. \tag{13}$$

Using the rotation matrix and the translation vector, the coordinates expressed in the camera'S frame can be transformed to the VICON's frame.

5 EXPERIMENTS AND RESULTS

5.1 Experimental setup

We considered a real environment that is covered by the VICON system cameras at the number of 15. A snapshot of the hall, close to an industrial environment, is depicted in Figure 1. As previously stated, the VICON measurements serve as ground truth references as they are highly accurate (Merriaux et al., 2017). The experiments were conducted with ROS (Melodic version) on a Linux computer, with an Intel Core i7-4710 CPU. The *opensource* implementation of ORB-SLAM2 has been used with a moving stereo camera, the Intel D435. In order to emulate the human localization in an industrial environment, 6 sequences were recorded, described in Table 1. The first two have been recorded using the D435 stereo camera only, the next two using the T265 tracking camera only and the last two using both the D435 and T265 cameras rigidly fixed on the same support.

Table 1: Experimental Sequences.

Sequence	Camera	3D length(m)
Seq1	D435	34.99
Seq2	D435	13.15
Seq3	T265	24.95
Seq4	T265	31.99
Seq5	D435 and T265	58.10
Seq6	D435 and T265	43.16



Figure 1: Snapshot from the environment of the recorded sequences.

5.2 Experiments

The experiments considered a localization task of a hand-held moving camera in an industrial environment. The metrics presented in Section 3 have been employed to benchmark the Intel T265 tracking camera. Table 2 compares the Absolute Trajectory Error (ATE) for the different sequences recorded presented in Table 1. We also compare the Relative Pose Error (RPE) in Table 3 that is averaged over all possible time intervals and finally the Euclidian Error (EE) in the ground plane in Table 4. The values of ATE, RPE and EE are root mean squared. We also show the aligned trajectories over X, Y and Z directions for Sequence 6: ORB-SLAM2's estimation in Figure 2 and T265's one in Figure 3.

Table 2: ATE with OrbSLAM2 and T265.

Seq	System	ATE[m]	% of Seq length
Seq1	ORB-SLAM2	0.2597	0.74
Seq2	ORB-SLAM2	0.2511	1.9
Seq3	T265	0.4007	1.6
Seq4	T265	0.5217	1.63
Seq5	ORB-SLAM2	0.4591	0.79
	T265	0.4262	0.73
Seq6	ORB-SLAM2	0.3762	0.87
	T265	0.4303	0.99

Based on Table 2, the ORB-SLAM2 algorithm and the T265 Tracking camera perform almost equivalently in terms of accuracy. The rotational error calculated in degrees per second presented in Table 3 corroborates this observation. The rotational error is more important than the translational one as shown by the RPE values that encode both translational and rotational errors compared to the ATE values reflecting only the translational errors. We analyse the sequences 5 and 6 more thoroughly in Figure 4 where

Table 3: RPE with ORB-SLAM2 and T265.

Sequence	System	RPE[m]
Seq1	ORB-SLAM2	2.8047
Seq2	ORB-SLAM2	1.7381
Seq3	T265	3.0803
Seq4	T265	3.8270
Seq5	ORB-SLAM2	3.9459
	T265	3.8996
Seq6	ORB-SLAM2	2.6213
	T265	3.1742

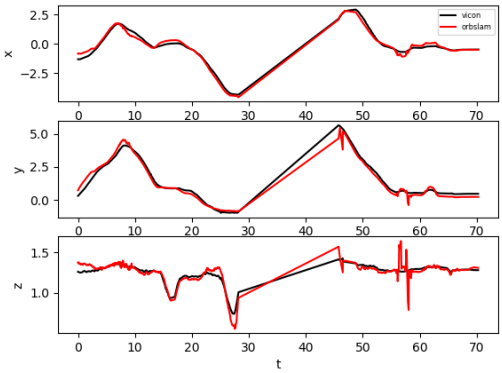


Figure 2: OrbSLAM2 vs Vicon over X, Y and Z axes.

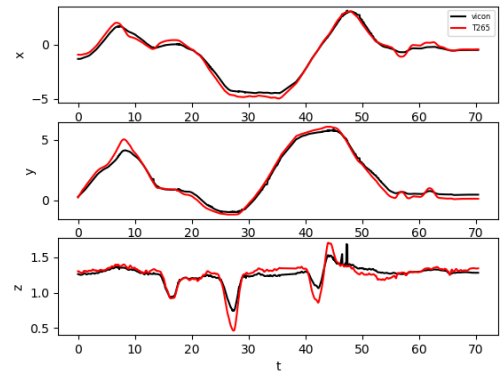


Figure 3: T265 vs Vicon over X, Y and Z axes.

some statistical parameters are compared namely the RMSE, the mean, the median, the standard deviation(std), the maximum value (max) and the minimum value (min). We denote that for the estimation of these parameters, 1065 poses have been used for ORB-SLAM2 against 4093 poses for T265 for Sequence 5 depicted in Figure 4a. For Sequence 6, 817 aligned poses for ORB-SLAM2 against 4754 poses for T265 have been used (Figure 4b). In fact, the frequency of the VICON system used is these experiments is of 100 Hz, the ORB-SLAM2 algorithm outputs estimations at a frequency around 20 Hz while the T265 outputs estimations at a much higher frequency around 200 Hz which justifies the varying number of camera poses aligned between the camera and the VICON despite the use of the same sequences. Thus, in order to have a better comparison of the results obtained from sequences 5 and 6, we rely on normalizing the RMSE. Various methods of RMSE normalization have been reported in literature including but not limited to the normalization by the mean, the standard deviation (std) and the difference between the maximum and minimum values as follows

Table 4: EE root-mean squared with ORB-SLAM2 and T265.

Sequence	Camera	EE[m]
Seq1	ORB-SLAM2	0.4761
Seq2	ORB-SLAM2	0.4456
Seq3	T265	0.5939
Seq4	T265	0.6997
Seq5	ORB-SLAM2	0.5818
	T265	0.5614
Seq6	ORB-SLAM2	0.5616
	T265	0.6262

$$\text{NRMSE}_1 = \frac{\text{RMSE}}{\text{mean}} \tag{14}$$

$$\text{NRMSE}_2 = \frac{\text{RMSE}}{\text{std}} \tag{15}$$

$$\text{NRMSE}_3 = \frac{\text{RMSE}}{\text{max} - \text{min}} \tag{16}$$

The normalized RMSE values (NRMSE) according to the stated normalization methods for the Sequences 5 and 6 are presented in Table 5 and Table 6.

Table 5: Normalized RMSE for sequences 5.

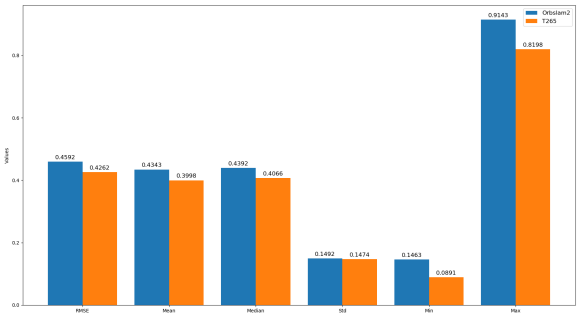
Method	NRMSE_1	NRMSE_1	NRMSE_1
ORB-SLAM2	1,057	3,0777	0,5979
T265	1,0660	2,8914	0,5832

Table 6: Normalized RMSE for sequences 6.

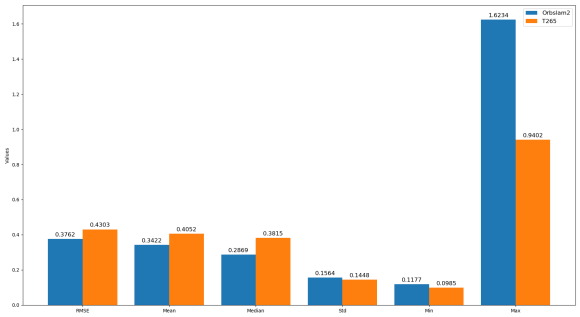
Method	NRMSE_1	NRMSE_1	NRMSE_1
ORB-SLAM2	1,0995	2,4056	0,2499
T265	1,0619	2,9713	0,5112

5.3 Discussion

From these results, it can be deduced that ORB-SLAM2 and T265 tracking camera give competitive accuracy. However, the T265 provides pose estimations at a 10× higher frequency. It should be noted



(a) Sequence 5.



(b) Sequence 6.

Figure 4: Benchmark evaluation over two sequences using different parameters.

that we relied on a simple stereo system for ORB-SLAM2 as industrial environments may contain either indoor or outdoor sites and the use of depth (RGB-D) cameras to reconstruct large outdoor environments is not feasible due to lighting conditions and low depth range. In fact, although the competitive accuracy, we evaluate the localization provided by T265 as more reliable for two main reasons. On the one hand, the localization provided by the T265 tracking camera is at a much higher frequency (200 Hz vs an average of 17 Hz for ORB-SLAM2). On the other hand, the statistical parameters in Figure 4 showed maximum error values higher with ORB-SLAM2 and closer mean and median errors for T265 than for ORB-SLAM2 which fits better a gaussian error and assesses more reliability against outliers. However, the T265 camera is only dedicated for the localization task while using the stereo output of the D435 with ORB-SLAM2 allows the person localization as well as other functionalities such as mapping and localizing objects using the depth information.

6 CONCLUSION

In this paper, we proposed a benchmarking of the RealSense T265 tracking camera for person localization in an industrial environment. The presented work is based on a comparative study between the T265 camera and ORB-SLAM2 known to be the most complete up-to-date visual SLAM method as it includes the majority of state-of-the-art techniques. The study consisted in an experimental evaluation based on comparing the localization performances of both systems with the very accurate motion capture system VICON used as ground-truth. The estimated and ground-truth trajectories were first time-synchronized then compared using literature metrics such as the Absolute Trajectory Error and the Relative Pose Error as well as the Euclidian distance Error (EE) used to evaluate error on the ground plane. The experimental evaluation showed that both vision-based localization systems provide competitive accuracy, but the localization provided by the Intel RealSense T265 is more reliable. Furthermore, it has been noted that the Intel RealSense Tracking Camera T265 complements Intels RealSense D400 series cameras, and the data from both devices can be coupled for advanced applications like occupancy mapping, advanced 3D scanning and improved navigation and crash avoidance in indoor environments.

ACKNOWLEDGEMENTS

This work was carried out as part of the COPTER research project, and is co-funded by the European Union and the Region Normandie. Europe is involved in Normandy with the European Regional Development Fund (ERDF).

REFERENCES

Chang, C., Zhu, H., Li, M., and You, S. (2018). A review of visual-inertial simultaneous localization and mapping from filtering-based and optimization-based perspectives. *Robotics*, 7:45.

Davison, A. J., Reid, I. D., Molton, N. D., and Stasse, O. (2007). MonoSLAM: Real-Time Single Camera SLAM. *IEEE TPAMI*, 29(6):1052–1067.

Furrer, F., Fehr, M., Novkovic, T., Sommer, H., Gilitschenski, I., and Siegwart, R. (2017). *Evaluation of Combined Time-Offset Estimation and Hand-Eye Calibration on Robotic Datasets*. Springer International Publishing.

Halmetschlager-Funek, G., Suchi, M., Kampel, M., and Vincze, M. (2019). An empirical evaluation of ten depth cameras: Bias, precision, lateral noise, different lighting conditions and materials, and multiple sensor setups in indoor environments. *IEEE Robotics Automation Magazine*, 26(1):67–77.

Intel (2017). *Intel RealSense Depth Camera D400-Series*. 0.7 edition.

Intel (2019). *Intel RealSense Tracking Camera T265 datasheet*. 001 edition.

Klein, G. and W. Murray, D. (2009). Parallel tracking and mapping on a camera phone. pages 83–86.

Merriaux, P., Dupuis, Y., Boutteau, R., Vasseur, P., and Savatier, X. (2017). A study of vicon system positioning performance. *Sensors*, 17:1591.

Montemerlo, M., Thrun, S., Koller, D., and Wegbreit, B. (2002). Fastslam: A factored solution to the simultaneous localization and mapping problem. In *In Proceedings of the AAAI National Conference on Artificial Intelligence*, pages 593–598. AAAI.

Mur-Artal, R., Montiel, J. M. M., and Tards, J. D. (2015). Orb-slam: a versatile and accurate monocular slam system. *CoRR*, abs/1502.00956.

Mur-Artal, R. and Tardos, J. (2016). Orb-slam2: an open-source slam system for monocular, stereo and rgb-d cameras. *IEEE Transactions on Robotics*, PP.

Sturm, J., Engelhard, N., Endres, F., Burgard, W., and Cremers, D. (2012). A benchmark for the evaluation of rgb-d slam systems. pages 573–580.

Weng Kuan, Y., Oon Ee, N., and Sze Wei, L. (2019). Comparative study of intel r200, kinect v2, and primesense rgb-d sensors performance outdoors. *IEEE Sensors Journal*, PP:1–1.

Yao, H., Ge, C., Xue, J., and Zheng, N. (2017). A high spatial resolution depth sensing method based on binocular structured light. *Sensors (Switzerland)*, 17.

Zollhfer, M., Stotko, P., Grlitz, A., Theobalt, C., Niener, M., Klein, R., and Kolb, A. (2018). State of the art on 3d reconstruction with rgbd cameras. *Computer Graphics Forum*, 37:625–652.