Localization for Ships during Automated Docking using a Monocular Camera

Dinosshan Thiagarajah[1] Håkon Hagen Helgesen[1] Øivind Kåre Kjerstad [2,3] Tor Arne Johansen[1]

***Abstract*— Automating docking operations of ships requires at least two robust and precise localization systems for reaching the typically required safety and redundancy levels. Global navigation satellite systems are typically chosen as one of these, while the second needs to be independent and preferably use another measurement principle. Optical sensors are versatile, low-cost, and assumed to provide sufficient localization range. Used to simultaneously locate the vessel and map the harbor environment, this technology is believed to offer the necessary properties to complement navigation satellite systems in a resilient manner. In this paper, a monocular camera together with the state-of-the-art ORB-SLAM3 algorithm is used for localization. Umeyama's method is used to create an initialization procedure to determine the unknown scale factor encountered in monocular camera odometry and to find the transformation between the camera and a world-fixed coordinate frame. The proposed system is validated using data recorded on a commercial high-speed passenger ferry in nominal operation. The results indicate a localization range of more than $200\,\mathrm{m}$. The mean absolute position error is less than $0.5\,\mathrm{m}$ with an estimated heading error of 0.5° in favorable weather conditions.**

***Index Terms*— Localization, Visual Odometry, SLAM, Autonomous Vehicle Navigation**

I. Introduction

Ship docking and harbor maneuvering are complex tasks that rely on skilled and experienced operators, both with respect to the ship dynamics and the environmental forces. A docking maneuver is stressful for human operators since it requires simultaneous control of several thrusters in a dynamic environment. Human error is the most common reason for accidents in the maritime industry [1]. Therefore, it is of interest to investigate requirements for development of automated control systems in harbor maneuvering and docking. A multi-vehicle framework for guidance, navigation and control is introduced to automate maneuvers in harbor areas in [1]. Moreover, [2] investigates optimal maneuvering and control for multiple vessels within a harbor. [3] designs a trajectory planning and control framework for automatic docking with full-scale experiments.

The control algorithms calculating the forces and moments required to bring the ship into and maintaining contact with the quay are an integral part of an automated docking system. However, for modern ships with electric thrusters, their ability to fulfill the control objective with sufficient precision relies mostly on the precision of the localization system. Obtaining the planar position and heading from sensor data is, therefore, an important aspect for surface ships. Global navigation satellite systems (GNSS) with real-time kinematic (RTK) corrections is the state-of-the-art solution in ship navigation systems [4]. However, harbors are often located near tall buildings and large structures. This can degrade the GNSS performance or lead to outage. Moreover, GNSS have known vulnerabilities such as jamming and spoofing [5], [6]. Therefore, automated control systems require at least one additional and independent localization system that can guide the ship in case of GNSS failure.

Cost-effective localization systems without the need for additional infrastructure on the quay are desirable. It is also preferable that the system uses another measurement principle for redundancy. Passive optical sensors, radars, and LIDARs fulfill these requirements. LIDARs are precise, but are also expensive. Maritime radars are used to detect other structures, but are not suitable for high-precision localization. Consequently, passive optical sensors are more attractive as a low-cost localization system during docking. Electro-optical (EO) sensors are versatile, and have the required range and resolution in sufficient lighting and weather conditions [7]. Moreover, they are also commonly used as part of automatic safety-critical systems such as collision detection and avoidance systems [8], [9].

Camera-based localization is often based on visual odometry (VO), visual simultaneous localization and mapping (VSLAM) or specific camera markers. Monocular SLAM has been a widespread solution to localize drones, robots, and vehicles. Vision-based localization has also been a particularly important research topic for the automotive industry in recent years. VO for automotive applications was studied in [10] and a survey of VO and its applications was presented in [11]. Computer vision for automated parking systems was studied in [12]. A few works have also investigated vision-based localization in maritime environments. [13] presents a monitoring system for ship localization during docking based on artificial intelligence. However, the camera is mounted on the quay which requires wireless communication for feedback and provides no flexibility on the choice of docking position. [14] uses bearing measurements fused with

[1]Center for Autonomous Marine Operations and Systems (NTNU-AMOS), Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), O. S. Bragstads plass 2D, 7491 Trondheim, Norway hakon.helgesen@ntnu.no

[2] NTNU, Department of Ocean Operations and Civil Engineering, Aalesund

[3] Kongsberg Maritime AS, Aalesund, Norway

inertial measurements, but is only verified in an indoor lab environment. Underwater docking using a visual sensor is studied in [15]. [16] presents a vision-based localization for docking but requires physical markers in the scene. This limits the distance for when the algorithm works and requires a permanent installation on the quay. Consequently, there has not been much research in localization for surface ships in harbor environments using an on-board camera system without additional infrastructure on land.

This paper investigates the performance of a camera-based VO system. The aim is to provide insight into how accurate a monocular camera can localize a ship using VO in harbor environments without additional sensors. This is important to investigate because many maneuvers cannot rely on loop closures, which is a key principle in VSLAM to avoid drift. Fusion of VO with data from an inertial measurement unit (IMU) is obviously an attractive solution, but it is desirable to assess the accuracy of a VO system before other sensors are added. Moreover, accurate heading information is crucial for ships and investigating the drift in heading using VO is an important research topic.

Maritime environments are often homogeneous with few distinguishable features, which poses a challenge for feature-based VSLAM and VO methods. The feature distribution is typically sparse in open waters but may be enriched when the ship moves closer to the quay. Moreover, features detected on waves and moving structures are problematic since a common assumption is static landmarks. Therefore, the feature distribution in harbor environments is studied in this paper. ORB-SLAM3 is used as a proof-of-concept VO method. The performance of ORB-SLAM3 is analyzed experimentally and compared with a conventional navigation filter based on RTK-GNSS and an IMU. The results presented in this paper are relevant for other VSLAM and VO architectures, since they share many similar challenges.

This paper is structured as follows. Section II describes preliminaries important for the rest of the paper. Section III describes ORB-SLAM3 and state-of-the-art methods. Section IV describes Umeyama's alignment method. Section V describes the experiments carried out to collect field data and the methods. Section VI presents experimental results before the paper is concluded in Section VII.

II. Preliminaries

The motion of a ship can be described in several coordinate frames. The 6-DoF generalized pose vector $\mathbf{x}$ consists of three parameters representing the position (x, y, and z) and three parameters representing the attitude (e.g., roll, pitch, and yaw). For ships in harbor maneuvering with a monocular camera, the relevant coordinate frames are typically a camera frame, the body (vehicle) frame, and the North-East-Down (NED) frame. The NED frame acts as a world frame and is a local tangent plane considered to be inertial locally [4]. The relevant coordinate systems are illustrated in fig. 1 and also defined in [4].

A position vector $\mathbf{x}$ given in one coordinate frame is transformed to another coordinate frame using homogeneous transformations. For example, the homogeneous transformation from the NED frame to the camera frame is

$$\tilde{\mathbf{x}}^c = \mathbf{T}_{cn}\tilde{\mathbf{x}}^n$$

where $\tilde{\mathbf{x}}^c$ and $\tilde{\mathbf{x}}^n$ are the position vectors represented using homogeneous coordinate vectors and decomposed in the camera frame and the NED frame, respectively. The homogeneous transformation matrix is defined as:

$$\begin{aligned}\mathbf{T}_{cn} &= \mathbf{T}_{cb}\mathbf{T}_{bn} \\ &= \begin{bmatrix}\mathbf{R}_{cb} & \mathbf{t}^c_{cb} \\ \mathbf{0}^T & 1\end{bmatrix}\begin{bmatrix}\mathbf{R}_{bn} & \mathbf{t}^b_{bn} \\ \mathbf{0}^T & 1\end{bmatrix}\end{aligned}$$

where $\mathbf{T}_{cb}$ and $\mathbf{T}_{bn}$ are the transformation matrices describing the transformation from the camera frame to the body frame, and the NED frame to the body frame, respectively. $\mathbf{R}_{cb}$ is the rotation matrix between the camera and body, and $\mathbf{t}^b_{cb}$ is the translation between the origins.

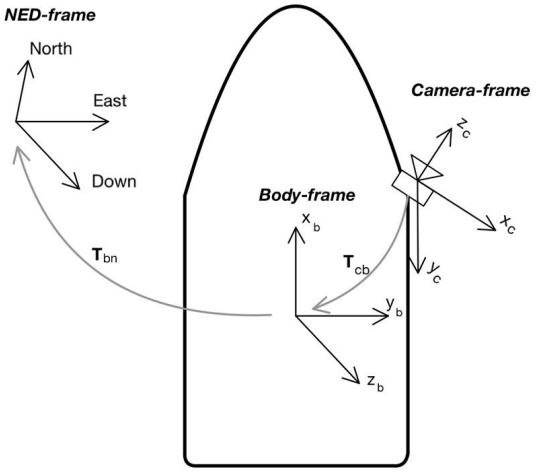

Fig. 1. Illustration of the coordinate frames of interest. $\mathbf{T}_{bn}$ and $\mathbf{T}_{cb}$ are homogeneous transformation matrices describing the relative orientation and translation between the three coordinate frames.

A Bayesian probabilistic model is typically used to formulate and solve VO and VSLAM problems. The true state vector X is unknown and includes the camera pose and position of detected map (feature) points:

$$X = \begin{bmatrix}\textit{camera pose} \\ \textit{map points}\end{bmatrix} \tag{1}$$

where *camera pose* is 6-DoF camera pose vector $\mathbf{x}$, and the *map points* is the position coordinates of all points in a map created by the VO or VSLAM algorithm decomposed in the camera frame. The objective in VO and VSLAM is to find an estimate $\hat{X}$, given a set of noisy sensor measurements Z and the initial states $X(0)$. The MAP estimator is most often used to maximize the posterior density $p(X|Z)$ by applying Bayes' theorem. Moreover, these methods apply Bayesian smoothing to improve the accuracy at the expense of computational complexity.

III. Visual Simultaneous Localization and Mapping and Visual Odometry

The methods investigated in this research are based on monocular camera systems. In monocular SLAM, a single camera, which is freely moving through its environment, represents the sole sensory input to the system [17]. One major characteristic of monocular SLAM is the scale ambiguity, meaning that it cannot estimate the absolute scale of the scene, and thus the perceived scale will drift over time. The scale ambiguity is caused by the lack of depth information in monocular camera images. Adding more sensors, such as another camera and obtaining stereo vision, or fusing the measurements with an IMU are some of the strategies to recover the scale in real-time applications. However, this paper only concerns VO using a single camera, where an initialization procedure handles the scale ambiguity.

The most widely used VO and VSLAM algorithms are sparse and indirect. DSO [18] is sparse and direct, while LSD-SLAM [19] and DTAM [20] are dense and direct. The performance of some state-of-the-art algorithms is benchmarked in [21], where ORB-SLAM3 is shown to outperform the other methods in terms of accuracy and robustness, which is the main motivation for using ORB-SLAM3 in this paper. Without utilizing its loop closure feature, only the VO capability of ORB-SLAM3 is used. ORB-SLAM and its successors are described in the next section.

A. ORB-SLAM

ORB-SLAM is a feature-based, indirect, and sparse monocular SLAM algorithm with real-time capabilities. ORB-SLAM includes the following modules:

1) ORB feature detector: Oriented FAST and Rotated BRIEF (ORB) is the feature selector, which builds on FAST keypoint detector [22] and BRIEF [23] feature descriptor. ORB is both scale and rotation-invariant, computationally efficient, and is also invariant to the viewpoint. This allows for scenes captured from wide viewpoints to be matched, which is an advantage in urban areas.

2) Main Threads: There are three main threads that run simultaneously: *tracking*, *local mapping*, and *loop closing*. The *tracking* thread is responsible for localizing the camera for every new frame and deciding when to add a new keyframe. Feature matching is done with the previous keyframe and the camera pose is optimized using *motion-only bundle adjustment*. The *local mapping* thread is responsible for inserting keyframes into the map, creating new map points, removing map point outliers and redundant keyframes, and performing *local bundle adjustment*. *Loop closing* is responsible for searching for new loops on every new keyframe. Loop closure is not relevant for the experiments, and thus not described further.

B. ORB-SLAM3

ORB-SLAM3 [21] is the most recent version of ORB-SLAM [24] with improved accuracy and robustness. The method has some new features such as *full bundle adjustment* which optimizes all map points and camera poses to improve the accuracy and map consistency. It also includes tight integration of inertial measurements, which is, however, not exploited in this paper.

Using this algorithm in a harbor environment has its advantages and disadvantages. The method runs in real-time and is known to provide excellent localization accuracy in several environments. Since it is an indirect system, it is less sensitive to brightness variations. However, contrary to direct methods, it can be vulnerable to motion blur and a sparsely textured environment. The number of features can be a challenge in maritime environments and poses some limitations on the localization range of the system. It is further addressed in Section VI. Motion blur can occur, but it is not normally an issue in low-speed harbor maneuvering.

IV. Umeyama's Alignment Method

Umeyama's $SIM(3)$ alignment is a method used to find the similarity transformation parameters $\mathbf{s}$ (scaling), $\mathbf{R}$ (rotation), and $\mathbf{t}$ (translation) that give the least mean-squared error between two point patterns [25]. The method can also be used to align and compare a ground-truth pose and an estimated pose. To put it formally, given N estimated positions $\{\hat{\mathbf{p}}_i\}_{i=0}^{N-1}$ and the ground-truth positions $\{\mathbf{p}_i\}_{i=0}^{N-1}$ from the corresponding time steps, it is possible to find a similarity transformation $\mathbf{S}' = \{\mathbf{s}', \mathbf{R}', \mathbf{t}'\}$ that satisfies:

$$\mathbf{S}' = \underset{\mathbf{S}=\{\mathbf{s},\mathbf{R},\mathbf{t}\}}{\arg\min} \sum_{i=0}^{N-1} ||\mathbf{p}_i - \mathbf{s}\mathbf{R}\hat{\mathbf{p}}_i - \mathbf{t}||^2$$

To solve this least squares problem, the method in [25] is often used. If the ground-truth positions are given in a different coordinate frame compared to the estimated positions, then the similarity transformation can be used to obtain the relative pose of the two coordinate systems. In this paper, Umeyama's $SIM(3)$ alignment method is used to estimate and transform the VO pose given in the camera-frame to the NED-frame for localization and evaluation purposes. Moreover, the method is used to estimate the unknown scale factor present in monocular VO as part of an initialization procedure described in Section V-D.3.

V. Experiments

A. Data Collection

The data used in this study was collected on a passenger ferry traveling between Trondheim and Vanvikan in Norway. Data from several crossings were captured, and two data sets are chosen for detailed analysis in this paper. The findings have been verified in the other data sets. Figure 2 shows the ferry docked at the terminal. The chosen portion of the data sets contain trajectories of about $190\,\mathrm{m}$ to $210\,\mathrm{m}$. These trajectories represent the part of the ferry mission with harbor maneuvering and docking. The trajectories cover the distance from where buildings and structures in the scenery begin to appear clearly in the camera images. The weather conditions were sunny with brief wind, which lead to images being captured with good brightness and minor camera motion due to waves and winds, as it can be seen in fig. 4 and fig. 5.

Fig. 2. Image of the ferry used to gather data at the dock

Fig. 3. Sensor placement on the ferry during data collection

B. Sensor Suite

The following sensor suite was used to capture data:

- Analog Devices Adis 16490 IMU providing measurements of specific force and angular rate at $250\,\mathrm{Hz}$.
- 2x uBlox Neo-M8T GNSS receivers paired with Harxon HX-GS288A-antennas. Dual antenna setup provides aiding measurements for heading estimation [4].
- Ueye UI-5260FA-C-HQ visual spectrum camera with a focal length of $12\,\mathrm{mm}$. A frame rate of $10\,\mathrm{Hz}$ was used, and the resolution was adjusted to 1936×1216.

To benchmark the VO system, a multiplicative extended Kalman filter (MEKF) was used to fuse data from the IMU and the GNSS receivers [26] with real-time kinematic (RTK) capability to obtain an accuracy of a few centimeters using the carrier phase observables [4]. The ground-truth reference is assumed to have an accuracy within $10\,\mathrm{cm}$ to $20\,\mathrm{cm}$ based on a floating-point solution for the GNSS integer ambiguity.

C. Mounting Pose

The camera was mounted on a metal railing, on the starboard side of the ferry, as shown in fig. 3. It was mounted in this manner to ensure that the quay area was within the field of view during docking. Figure 4 shows the docking area captured by the camera nearby the docking terminal. The metal railing is visible in the lower-left corner.

D. Methods

1) ORB-SLAM3: The open-source implementation of ORB-SLAM3 [21] was used as VO system in this research. The camera intrinsic matrix and distortion parameters were obtained through camera calibration [27].

2) EVO: To evaluate the localization accuracy of ORB-SLAM3, EVO [28] was used. It supports most of the well-known benchmarking data set file formats and provides algorithmic options for data association, alignment, and scale adjustment using Umeyama's method [28]. EVO can be used to estimate the camera mounting pose by aligning a reference trajectory with the SLAM trajectory during an initialization period. This is described next.

3) Initialization Procedure: The VO system must be anchored to NED initially, since only relative measurements are obtained in VO. Initialization using a single initial pose is not robust. Therefore, an initialization method is defined in this paper. Umeyama's method, described in Section IV, is used to find the transformation between the initial VO trajectory in the camera frame and a ground-truth trajectory in NED. The origin of NED is placed on the quay for convenience. The ground truth is based on a MEKF as mentioned in Section V-B and decomposed in the same NED frame. The initialization procedure is conducted to anchor the VO trajectory to NED and estimate the unknown scale factor of the camera. The length of the initialization period is tunable and chosen to be $25\,\mathrm{s}$ here. This choice is further explained in Section VI-B. The length of the initialization period affects the accuracy initially, but cannot prevent long-term drift. In practice, this strategy means that GNSS must be available initially before the camera acts independently in a dead-reckoning fashion using VO. The initialization procedure is typically carried out before the docking maneuver starts.

VI. Results

In this section, the results from the field tests are presented. First, the feature distribution and performance of the ORB feature detector are analyzed. This is followed by a case study that is conducted to benchmark the localization accuracy using a monocular camera and VO. All results are based on experimental data. The motivation behind the results is to emulate a likely situation in which RTK-GNSS position measurements are available initially but become unavailable, degraded or unreliable close to the dock. In open waters, the RTK-GNSS coverage is expected to be stable and can be

used as part of the initialization procedure. Only the camera is used for localization closer to the dock.

A. Feature Detection and Mapping

Feature detection is perhaps the most crucial part in VO, and the maximum number of features to detect is a typical tuning parameter. Choosing a large number leads to higher execution time. In the results, a maximum of 1000 features was allowed per frame, which makes the algorithm run at $1.4\,\mathrm{Hz}$ on a *9th Gen Intel Core vPro i7* processor. Figure 4 and fig. 5 show typical examples of how the ORB features are distributed across the images when the ferry is nearby the docking terminal and in open water during transit, respectively. Near the docking terminal, features are detected on the terminal structure as well as on the tall buildings in the background. In open waters, features are mainly detected in a smaller region of the camera frame, making it harder to estimate the pose of the ferry. Features detected in fig. 4 have short distances to the camera compared to fig. 5. In VO, map points detected at shorter distances with good distribution across the camera frame usually provide localization estimates with less uncertainty. This is due to epipolar geometry and its well-known depth uncertainty [29]. This uncertainty limits the localization range for surface vehicles in open sea.

Fig. 4. ORB features detected in an image captured near the docking terminal.

Several features are detected in undesirable regions of the camera frame. This includes features on surface waves, the metal railing where the camera is mounted, and on moving targets such as the other ferry in fig. 4. The other ferry in fig. 4 was stationary during arrival in the experiments, but moving landmarks are undesirable. Moreover, waves are non-stationary unwanted features. Therefore, the maximum number of features has a limit where increasing the number of features can lead to worse accuracy since more undesired uncertain or non-stationary features will be detected. The metal railing is not part of the quay environment and is observed in the same locations in all images. Therefore, if the motion is estimated relative to the railing, then the camera seems to stand still, which contradicts the hypothesis

Fig. 5. ORB features detected in an image captured in the open water before arriving at the harbor area.

that the camera is moving relative to the environment. To provide robust estimates, ORB-SLAM3's culling policy handles anomalies like these, that do not describe the camera motion in a sensible manner with respect to the majority of the detected features [21].

B. Ferry Localization Using ORB-SLAM3

Data from two independent crossings are used to evaluate the localization accuracy of ORB-SLAM3. The initialization procedure described in Section V-D.3 was used to estimate the unknown scale factor, and anchor the VO estimates to NED in the initialization period. The trajectory of interest is when the ferry is within the harbor area. This corresponds to $95\,\mathrm{s}$ to $105\,\mathrm{s}$ with a trajectory length of about $185\,\mathrm{m}$ to $205\,\mathrm{m}$ in both data sets. Absolute position error (APE) is used to evaluate the accuracy of the VO trajectory against the true trajectory, and it is a frequently employed metric in the literature.

Length of Initialization Period

The VO system must be anchored to NED initially as described in Section V-D.3. An initialization period of $25\,\mathrm{s}$ was chosen. With the system running at $1.4\,\mathrm{Hz}$, the first $25\,\mathrm{s}$ corresponds to about 35 poses that are aligned between the VO and MEKF. The length of the initialization affects the overall accuracy and is thus interesting to analyze. Figure 6 shows the mean APE as a function of the length of the initialization procedure. The error converges if more than $50\,\mathrm{s}$ is used for initialization, but that corresponds to a large portion of the trajectory. An initialization period of $25\,\mathrm{s}$ was chosen since it gives the localization system a more relevant scenario with more than a minute without any other measurements. Navigating in a dead-reckoning fashion for more than a minute is challenging in applications which require precise localization. [30] investigates the dead reckoning capabilities of a state-of-the art inertial navigation system driven by MEMS IMU. The reported drift is about $1\,\mathrm{m}$ over $1\,\mathrm{min}$ for a vessel under dynamic positioning.

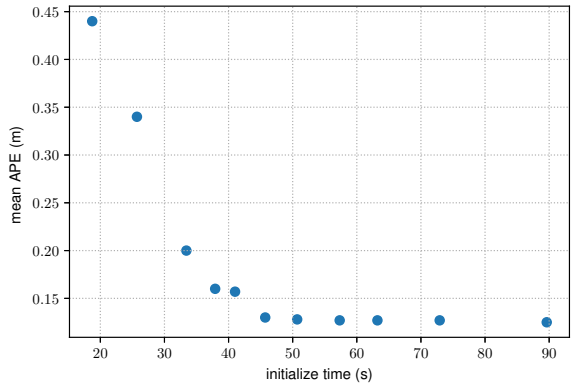

Fig. 6. Mean absolute position error as a function of the length of the initialization procedure.

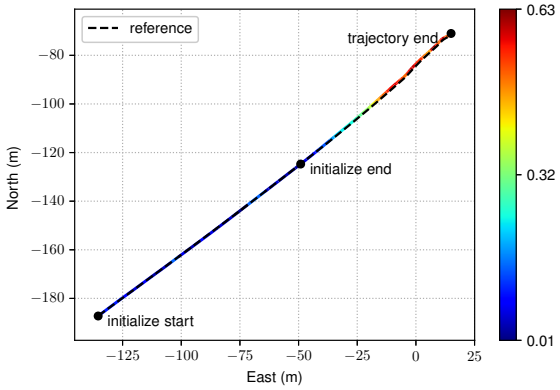

Fig. 7. North-east trajectory of the ferry with absolute position error mapped onto the trajectory with color coding. The maximum error is about 63 cm near the end.

First experiment

Figure 7 shows the estimated trajectory of the ferry in the first experiment using VO with the APE mapped onto the trajectory. The GNSS-aided MEKF is used as ground truth. The initialization period only lasted for 25 s of the total trajectory length of 95 s but cover half of the trajectory length because the ferry speed was much higher initially. The speed decreased when the ferry moved closer to the docking station. The precision of the localization system is particularly important in the final part of the trajectory so that the propulsion system is able to steer the ferry towards the dock. Figure 8 shows the absolute positioning error. The mean APE is 6 cm during the initialization period and 44 cm after the initialization period. The maximum APE is 15 cm and 63 cm during and after the initialization period, respectively. The drift during VO is not significant and the positioning error only grows to 63 cm as shown in fig. 8. This corresponds to 0.8% of the trajectory length after the initialization procedure ended. Moreover, the error does not increase systematically with time. This is a promising result since the localization system worked in a dead-reckoning fashion for more than a minute.

Figure 9 shows the estimated heading angle compared with the reference. The maximum error is less than a degree without the drift increasing notably during the final part of the trajectory. Accurate heading estimation is key during docking, and the results show that a monocular camera can be a promising solution for localization without a gyrocompass or a GNSS compass. The long-term drift has not been investigated due to the lack of features further out at sea.

Second experiment

The second set of results originates from an independent data set captured with the same ferry. Figure 10 shows the estimated trajectory. The initialization period lasted for 25 s. The trajectory length after the initialization procedure is coincidentally somewhat longer in this experiment with a length of about 115 m.

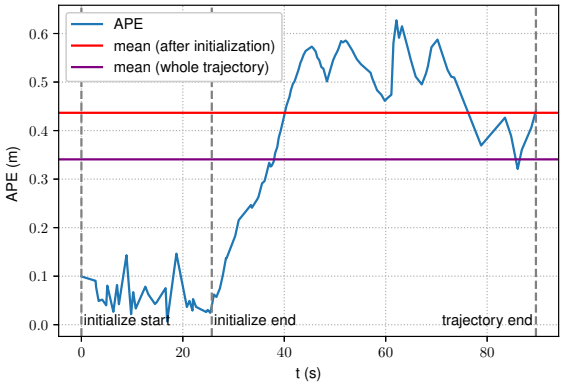

Fig. 8. Absolute position error during first experiment.

Figure 11 shows the absolute positioning error. The mean APE is 6 cm during the initialization period and 25 cm after the initialization period. The maximum APE is 14 cm and 70 cm during and after the initialization period, respectively. The drift during VO is not significant and the positioning error only grows to 70 cm as shown in fig. 11. This drift is about 0.6% of the trajectory length after the initialization procedure ended. The maximum APE occurs in a time period with significant variations in the heading as observed in fig. 12. This is expected since monocular VO is more prone to positioning errors during rotations. Potential time-synchronization inaccuracies are also more influential during rotations. Overall, the mean APE is smaller in this experiment and the accuracy better.

The estimated heading angle is shown in fig. 12. The estimates are more accurate for this experiment and follows the ground truth precisely, also during heading changes. This is promising and shows that a camera is a reliable sensor for heading estimation during harbor maneuvering.

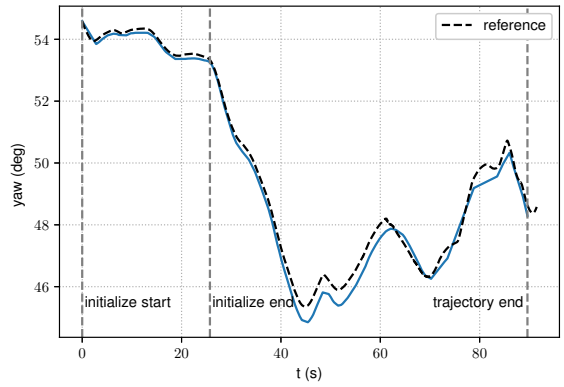

Fig. 9. Estimated heading. The maximum error is less than a degree and is not increasing noticeably after the initialization period.

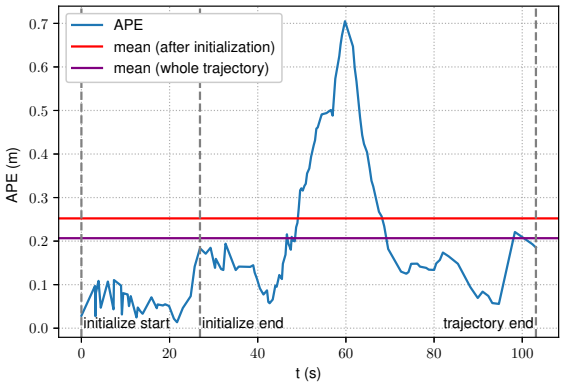

Fig. 11. Absolute position error during localization in the second experiment.

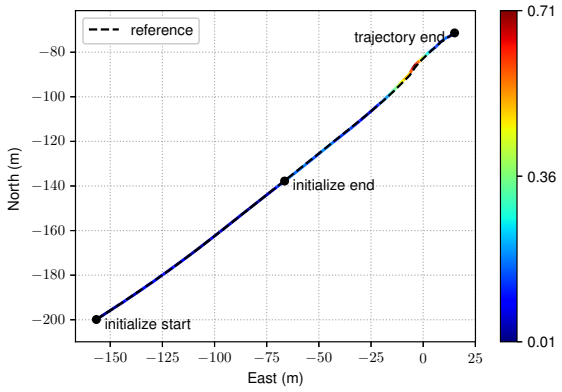

Fig. 10. North-east trajectory of the ferry with APE mapped onto the trajectory with color coding in the second experiment. The maximum error is about 70 cm near the end.

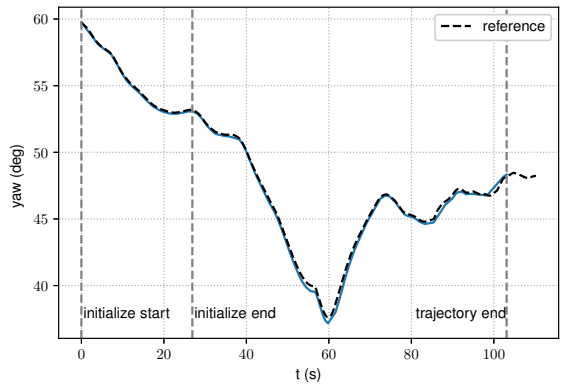

Fig. 12. Estimated heading in second experiment. The maximum error is less than a degree.

Discussion

The results indicate that a camera-based localization system using VO is a promising alternative in harbors with restricted GNSS coverage if properly initialized. An accuracy of 0.5 m in horizontal position and 0.5° in heading is considered to be sufficient for most ships during docking. This obviously depends on the type of ship and how the docking maneuver is conducted. In small, confined areas with obstacles, it may be necessary with even better accuracy. However, for docking at a longer quay without obstacles, an accuracy of 1 m might be sufficient. It is also worth highlighting that obstacles could enrich the feature distribution and improve the accuracy of the VO system. Also note that for longer ships, the heading accuracy is often even more important than the position accuracy since it is necessary to know where the bow and stern are. Finally, increasing the length of the initialization procedure will improve the accuracy since dead reckoning using VO will be conducted over a shorter time period. Therefore, it is obviously beneficial to use GNSS for as long as it is considered to be reliable.

This research does not fully cover the generalizability of the method for other docking areas, since both data sets were captured in the same harbor. For a complete feasibility assessment, the results must be validated in other densely and sparsely textured environments, with larger variations in light and environmental conditions. Moreover, challenging elements in the scene, such as moving objects, is a field that needs more research.

VII. Conclusion

In this study, the VO capabilities of ORB-SLAM3 is used to create a localization system for ferries in harbor maneuvering. Visual features are observed during maneuvering and used to estimate the navigation states of the ferry. By using detected features, the algorithm can simultaneously localize and map the harbor area in real-time, proving that the ORB-SLAM3-based localization system can act as a backup for state-of-the art inertial navigation systems aided by dual-antenna GNSS without a gyrocompass. The VO

system provides robust and satisfactory planar positioning and heading estimates up to a distance of $200\,\mathrm{m}$ if properly initialized. For shorter localization ranges, the accuracy is better since the drift in the estimates increases with the localization distance. As shown in the results, a mean APE of $0.5\,\mathrm{m}$ can be achieved when an increasingly larger portion of the RTK-GNSS data is used for initialization. This is not worse than the typical achievable accuracy for a dead-reckoning system based on a state-of-the- art MEMS-based inertial navigation system. Future work should therefore study fusion of VO and IMU data in harbor environments. Moreover, investigating the performance in poor lighting and weather conditions, and in other harbors are also interesting topics for further work.

ACKNOWLEDGMENT

This work was supported by Kongsberg Maritime through the University Technology Centre (UTC) at NTNU. It was also supported by the Research Council of Norway through the Centres of Excellence funding scheme, project number 223254. The authors are grateful for the cooperation with Kongsberg Maritime and the feedback received from them during this work, in particular, Bjørnar Vik and Sverre Torben. We are also grateful for the cooperation with FosenNamsos that has given us access to acquire data on their vessel.

REFERENCES

[1] M. Kurowski, S. Roy, J.-J. Gehrt, R. Damerius, C. Büskens, D. Abel, and T. Jeinsch, "Multi-vehicle guidance, navigation and control towards autonomous ship maneuvering in confined waters," in *European Control Conference (ECC)*, 2019, pp. 2559–2564.

[2] R. Zweigel, J. Gehrt, S. Liu, S. Roy, C. Büskens, M. Kurowski, T. Jeinsch, A. Schubert, M. Gluch, O. Simanski, E. Pairet-Garcia, F. Siemer, and D. Abel, "Optimal maneuvering and control of cooperative vehicles as case study for maritime applications within harbors," in *European Control Conference (ECC)*, 2019, pp. 3022–3027.

[3] G. Bitar, A. B. Martinsen, A. M. Lekkas, and M. Breivik, "Trajectory planning and control for automatic docking of asvs with full-scale experiments," *21st IFAC World Congress*, vol. 53, no. 2, pp. 14 488–14 494, 2020.

[4] T. I. Fossen, *Handbook of marine craft hydrodynamics and motion control, second edition*. John Wiley & Sons, 2021.

[5] M. L. Psiaki and T. E. Humphreys, "Gnss spoofing and detection," *Proceedings of the IEEE*, vol. 104, p. 1258–1270, 2016.

[6] J. C. Grabowski, "Personal privacy jammers," *GPS World*, pp. 28–37, April 2012.

[7] M. O. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1, 2016.

[8] E. Dagan, O. Mano, G. P. Stein, and A. Shashua, "Forward collision warning with a single camera," in *IEEE Intelligent Vehicles Symposium, 2004*, 2004, pp. 37–42.

[9] H. Alvarez, L. M. Paz, J. Sturm, and D. Cremers, "Collision avoidance for quadrotors with a monocular camera," *Experimental Robotics: The 14th International Symposium on Experimental Robotics*, pp. 195–209, 2016.

[10] M. Persson, T. Piccini, M. Felsberg, and R. Mester, "Robust stereo visual odometry from monocular techniques," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015, pp. 686–691.

[11] M. O. Aqel, M. H. Marhaban, M. I. Saripan, and N. B. Ismail, "Review of visual odometry: types, approaches, challenges, and applications," *SpringerPlus*, vol. 5, no. 1, pp. 1–26, 2016.

[12] M. Heimberger, J. Horgan, C. Hughes, J. McDonald, and S. Yogamani, "Computer vision in automated parking systems: Design, implementation and challenges," *Image and Vision Computing*, vol. 68, pp. 88–101, 2017.

[13] H. Kim, D. Kim, B. Park, and S. M. Lee, "Artificial intelligence vision-based monitoring system for ship berthing," *IEEE Access*, vol. 8, pp. 227 014–227 023, 2020.

[14] S. de Marco, M.-D. Hua, T. Hamel, and C. Samson, "Position, velocity, attitude and accelerometer-bias estimation from imu and bearing measurements," in *European Control Conference (ECC)*, 2020, pp. 1003–1008.

[15] M. C. Nielsen, T. A. Johansen, and M. Blanke, "Cooperative rendezvous and docking for underwater robots using model predictive control and dual decomposition," in *European Control Conference (ECC)*, 2018, pp. 14–19.

[16] Ø. Volden, A. Stahl, and T. I. Fossen, "Vision-based positioning system for auto-docking of unmanned surface vehicles (usvs)," *International Journal of Intelligent Robotics and Applications*, vol. 6, no. 1, pp. 86–103, 2022.

[17] R. Munguía and A. Grau, "Monocular SLAM for visual odometry: A full approach to the delayed inverse-depth feature initialization method," *Mathematical Problems in Engineering*, vol. 2012, 2012.

[18] J. Engel, V. Koltun, and D. Cremers, "Direct Sparse Odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[19] J. Engel, J. Sturm, and D. Cremers, "LSD-SLAM: Large-Scale Direct Monocular SLAM," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449–1456, 2013.

[20] R. A. Newcombe, S. J. Lovegrove, and A. J. Davison, "Dtam: Dense tracking and mapping in real-time," in *International Conference on Computer Vision*, 2011, pp. 2320–2327.

[21] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multimap slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[22] E. Rosten and T. Drummond, "Machine Learning for High-Speed Corner Detection," in *Computer Vision – ECCV 2006*, 2006, pp. 430–443.

[23] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "BRIEF: Binary Robust Independent Elementary Features," in *Computer Vision – ECCV 2010*, 2010, pp. 778–792.

[24] R. Mur-Artal, J. M. Montiel, and J. D. Tardos, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, 2015.

[25] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 376–380, 1991.

[26] J. Sola, "Quaternion kinematics for the error-state kf," 2017, last accessed 2020-02-02. [Online]. Available: http://www.iri.upc.edu/people/jsola/JoanSola/objectes/notes/kinematics.pdf

[27] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, no. c, pp. 666–673, 1999.

[28] M. Grupp, "evo: Python package for the evaluation of odometry and slam." https://github.com/MichaelGrupp/evo, 2017.

[29] Z. Zhang, "Determining the Epipolar Geometry and its Uncertainty: A Review," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 161–195, 1998.

[30] R. H. Rogne, T. H. Bryne, T. I. Fossen, and T. A. Johansen, "On the usage of low-cost mems sensors, strapdown inertial navigation, and nonlinear estimation techniques in dynamic positioning," *IEEE Journal of Oceanic Engineering*, vol. 46, no. 1, pp. 24–39, 2021.