Combining Short and Wide Baseline Stereo Cameras for Improved Maritime Target Tracking*

Nicholas Dalhaug†, Annette Stahl†, Rudolf Mester‡ and Edmund Førland Brekke†

***Abstract*—Target tracking is essential for autonomous vehicles to avoid collisions. Using a stereo camera for the target tracking gives a dense representation of the targets, contrary to the the sparser data on typical radars and lidars. With a wider baseline stereo camera the depth measurements are more accurate, but the stereo matching challenge is greater, especially in the maritime domain with reflections on the water. Earlier classical methods of tracking using stereo cameras have often tracked targets by first doing water surface estimation and then finding objects perturbing the plane. The challenge is then to get a good estimate of the water surface plane while still having precise measurements to the targets. We propose both a short baseline method and a multi-baseline method for target detection. The multi-baseline method uses a short baseline stereo camera to find the water plane and uses a wider baseline stereo camera to get accurate target measurements. The targets are consistently being tracked when using data collected during the summer of 2023 from an autonomous ferry prototype compared to ground truth GNSS tracks. The short baseline method achieves minimal error for a day cruiser boat 40 m away using a camera baseline of only 12 cm. The multi-baseline method further improves the accuracy of boat measurements, especially for a far-away small kayak.**

***Index Terms*—Multi-object tracking, Classic boat detection, Stereo cameras, Maritime situational awareness.**

I. Introduction

Multi-Object Tracking (MOT) enables autonomous vessels to manage its surrounding vessels [1]. Often this is done in either a tracking-by-detection framework or track-before-detect. Tracking-by-detection assumes that the input are detected objects and outputs the resulting tracks. Track-before-detect assumes the measurements can be the objects and estimates both what are objects and their tracks. In this way, tracking-by-detection can aid collision avoidance systems with improved situational awareness, if there is an accurate way to detect the surrounding objects.

Stereo cameras are sensors that can allow for the object detection and typically give denser measurements than sensors like radars and lidars. While there exists radars and lidars with different properties, they both typically give a sparse point cloud of the surroundings. The most important benefits of each are the long range of the radar and the accurate distance measurements of a lidar. While the stereo camera can have approximately the same number of points as a lidar, the field of view is decreased. Therefore, stereo camera object detections give more points on the target objects which can aid in the target tracking.

Fig. 1. A point cloud from disparity estimation colored by the rgb image values. The boat in the foreground is visible as well as the quay background. The red line is the estimated track from the complete scenario.

While stereo cameras are commonly used on autonomous cars, they are less prevalent in the maritime domain. One reason for this is the scarcity of available datasets with ground truth tracks, whereas the automotive domain has readily available datasets such as KITTI [2]. Another reason might be the distance requirements on the situational awareness. Far-away objects can become small in the images, and the depth accuracy at a specific distance depends on the stereo camera baseline, which is the distance between the two monocular cameras. For example, the SceneFlowFields [3] paper limits the depth measurements to $35\,\mathrm{m}$ on KITTI [2] where the baseline is about $50\,\mathrm{cm}$. In the maritime domain, relevant objects are often located farther away, suggesting that if a stereo cameras is to be used it should have a baseline of at least $50\,\mathrm{cm}$.

A third reason for the less prevalent use of stereo cameras in the maritime domain might be the greater challenge of stereo matching, which we discuss in section III. This is important since many classical methods for tracking in the maritime domain using stereo cameras estimate the water plane from the resulting point cloud [4]–[6]. To address these challenges related to stereo camera target tracking at sea, we propose a multi-baseline method analyzed using ground-truth tracks. It integrates both short and wide baselines to estimate the water plane and achieve accurate target tracking. To the best of our knowledge, such an approach has not been previously explored.

Our contributions can be summarized as follows: (i) We

* This work was supported by The Research Council of Norway (project number 333917).

† Nicholas Dalhaug, Annette Stahl and Edmund Førland Brekke are with the Department of Engineering Cybernetics, NTNU, 7034 Trondheim, Norway {nicholas.dalhaug, annette.stahl, edmund.brekke}@ntnu.no

‡ Rudolf Mester is with the Department of Computer Science, NTNU, 7034 Trondheim, Norway rudolf.mester@ntnu.no

present a method for short baseline stereo camera boat tracking. This method uses water plane detection and processes disparity images into clusters of points that can be used for tracking. Our approach utilizes a Joint Integrated Probabilistic Data Association (JIPDA) framework which is well-suited for handling low-quality detections and is able to model the boat in the world frame, unlike methods that track solely in the image frame. (ii) We present a multi-baseline method for improved target-background discrimination. A smaller baseline stereo camera setup is used to generate a disparity image (as shown in fig. 1), benefiting from the increased similarity in the images due to the more similar views. Since finding the water plane can be challenging for wide baseline cameras, the method uses the short baseline camera to detect the water plane in 3D and uses a wide baseline camera to get accurate depth measurements to the targets and extend the tracking range. (iii) We analyze the methods on the gathered data showing that although the short baseline method achieves impressive results, the multi-baseline method has improved accuracy.

II. Related work

In the image tracking-by-detection framework, probabilistic methods that use as much information as possible from the detections were the focus before 2015 [7], [8]. For example, a probabilistic image tracker was used in [9], [10]. They still tend to be more popular for sensors such as radar and lidar, where there is a greater need to use all available data [11]. However, there has been a trend away from probabilistic methods such as the Joint Probabilistic Data Association (JPDA), which [8] explains by the increased singe-frame detections quality from neural networks. [12] did tracking-by-detection back in 2008 based on learned metrics, not using neural networks. State-of-the-art methods for tracking in images use deep learning for tracking-by-detection [1], [8]. Although better quality detections simplify data association, we argue that not all detection tasks have so much available training data that neural networks necessarily give accurate enough detections for the JPDA to be superfluous.

We have found that boat detections can be imperfect and that the JPDA is still relevant for filtering out clutter in the maritime domain. For pre-trained neural networks this is especially relevant for boats in inland waterways. The JIPDA [13] filter is a probabilistic tracking method that handles multiple targets, based on the JPDA. It uses a Kalman filter for each target, models the probabilities of different measurements being clutter or from each target, and then estimates with a weighted average of the association hypotheses. It also estimates the existence probabilities for each target and measurement, allowing for track initialization and false track discrimination. Our work uses a classical detection method with the Visual Joint Integrated Probabilistic Data Association (VJIPDA) [14] which further adds a visibility state in addition to the existence state, in order to better model an occlusion. In this way, the tracking method should be robust to detection quality, clutter and short occlusions.

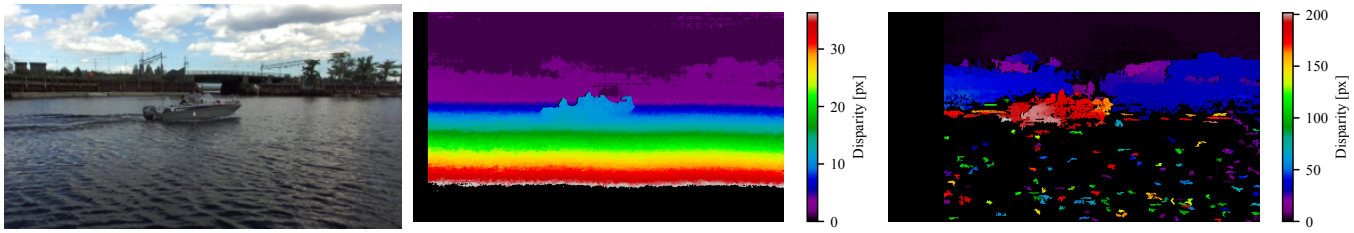
Although it has not been studied as much as in traffic scenes, tracking in the maritime environment has also been researched with the use of stereo cameras. [4] detected obstacles by estimating the sea surface and detecting what was perturbing it. The sea surface estimation was done using block matching [15] to generate the stereo point cloud, then using RANdom SAmple Consensus (RANSAC) to fit a plane. The baseline was 1.7m, and they argue that it had good accuracy in the range of 20-200m. [5] improved on it by using Semi-Global Block Matching (SGBM) for stereo matching and used the Fast Fourier Transform and template matching for the data association. [6] implemented a similar approach but enhanced it by incorporating an Inertial Measurement Unit (IMU) and semantic segmentation in their system. This integration allowed for improved performance, enabling accurate estimation even in calm water conditions. While not doing tracking, [16] used SGBM for stereo matching, found the water surface using RANSAC, then the artificial horizon and from this estimated roll and pitch. Classical boat tracking with stereo cameras typically employ water plane estimation.

The Hammerhead system [17] used a multi-baseline stereo camera setup in closed loop navigation. They placed on each side across a wide baseline two cameras that were oriented slightly away from each other. This increased the field of view of the cameras while keeping the same angular resolution. The paper did not explain much details of the stereo methods used, but they do state using a multi-scale approach for stereo matching in order to handle the detection of the water surface. In our paper we suggest using the short baseline stereo setup to detect the water surface. We also quantify the improvement of the target tracking compared to just using the short baseline.

III. The challenge of stereo matching in the maritime domain

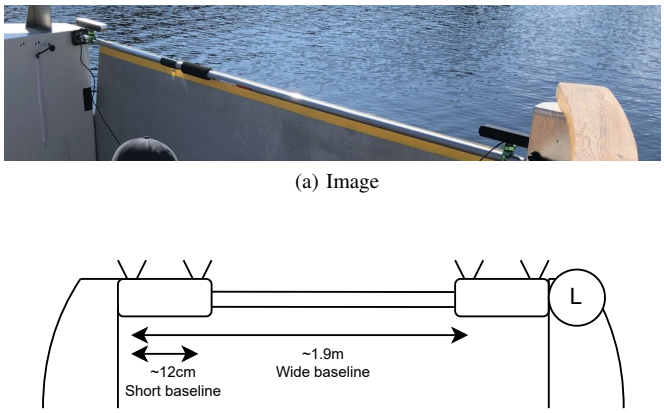
Stereo matching for disparity estimation in the maritime domain presents unique challenges due to the dynamic nature of water surfaces, including waves [18], reflections [19] and the lack of texture due to calm waters [6]. Researchers have explored various techniques to address these challenges, including block matching [4], [6], [15] and Semi-Global Matching (SGM) [5], [16], [20].

An additional challenge for the wide baseline stereo camera is the appearance difference due to a larger viewing angle. The difference between the disparity images for the two baselines is shown in fig. 2. The most prominent difference is the density of estimated disparities, where the wide baseline does not get as many valid disparities on the water surface. Assuming correct data association, the wider baseline gives a more precise distance measurement to the objects. The increased disparity amplifies the effect of parallax, making it easier to calculate the depth with higher accuracy. Essentially, as the baseline widens, even small changes in disparity become more detectable, allowing for finer distinctions in the measured distances to objects within the scene. This suggests that the short baseline stereo camera can be used to estimate the water



(a) Left image (b) Short baseline SGBM disparity image (c) Wide baseline SGBM disparity image

Fig. 2. Disparities for short and wide baseline. The colorbars show the disparity values in pixels. Pixels with zero disparity are invalid.



(a) Image

(b) Drawing

Fig. 3. The two ZED cameras mounted on a pole on the autonomous ferry prototype. Both an image of the setup and an illustrative drawing with approximate baseline lengths. The L is a lidar used for calibration.

surface, while the wider baseline camera can be used to get accurate depth estimates on targets.

In our work we use the OpenCV [21] implementation of SGBM, based on SGM, for both wide and short baseline stereo matching due to its user-friendliness and enhanced performance relative to other accessible methods.

IV. Experimental Setup

During the summer of 2023, a data acquisition campaign was conducted in order to gather data related to target tracking of boats using the multi-baseline stereo camera. The campaign took place in the canal between Brattøra and Trondheim City Centre in Trondheim, Norway. We had two target boats in addition to the autonomous ferry prototype *milliAmpere2*: the day cruiser Buster XL and a kayak. The ferry has Real-Time Kinematic (RTK) Global Navigation Satellite System (GNSS) and the targets both have GNSS receivers, with raw data, thereby allowing Post-Processing Kinematics (PPK) for increased accuracy.

Two stereo cameras were placed on the ferry prototype, a ZED 1 and a ZED 2 from Stereolabs, see fig. 3. They were placed with a distance of about $1.9\,\mathrm{m}$ between them. The wide baseline stereo camera uses both left hand cameras in the two ZEDs, while the short baseline uses just the left ZED, with a baseline of about $12\,\mathrm{cm}$. The stereo cameras came pre-calibrated with their respective intrinsic and extrinsic parameters. All that was missing with respect to calibration were the extrinsic parameters between the two stereo cameras, which we found using a lidar and manual selection of points in the images. The two cameras have different fields of view, resulting in lower resolution images on the wide field of view camera after stereo rectification. The settings for the two cameras were HD1080 resolution at 15Hz.

TABLE I
These are the parameters used for the SGBM using OpenCV [21]. The values for n_{P1} and n_{P2} were suggested in the OpenCV documentation. The rest were manually tuned.

	Short baseline	Wide baseline
$n_{\text{num disparities}}$	$0.04 \cdot 1920$	$0.15 \cdot 1920$
$n_{\text{block size}}$	4	9
$n_{\text{min disparity}}$	1	1
n_{P1}	$8 \cdot 3 \cdot n^2_{\text{block size}}$	$8 \cdot 3 \cdot n^2_{\text{block size}}$
n_{P2}	$32 \cdot 3 \cdot n^2_{\text{block size}}$	$32 \cdot 3 \cdot n^2_{\text{block size}}$
$n_{\text{disp12MaxDiff}}$	0	0
$n_{\text{uniquenessRatio}}$	0	6
$n_{\text{speckleWindowSize}}$	200	300
$n_{\text{speckleRange}}$	1	2

V. Short Baseline Stereo Camera Tracking

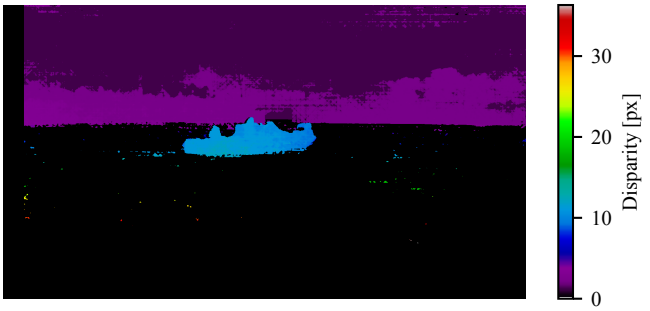
To explain the multi-baseline tracking, we will first describe the short baseline tracking approach. Their information flow is similar. Where the short baseline method uses only two images, the wide baseline method gets the target measurements using an additional image across the wide baseline. See the information flow illustrated in fig. 6.

A. Rectification, stereo matching and 3D projection

The left and right images are first rectified. The images are then matched using SGBM with the parameters shown in table I which results in disparity images similar to those in fig. 2b. From the disparity image, the corresponding point cloud is found using the regular stereo camera equations.

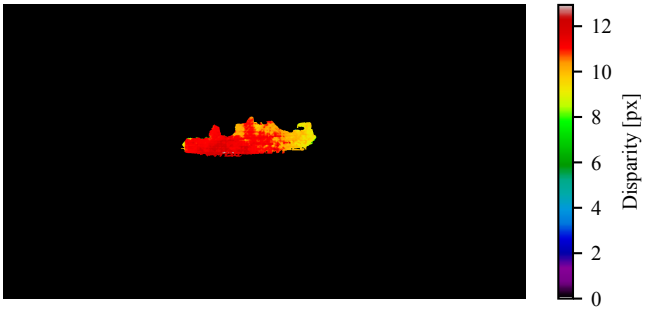
B. Water plane estimation

The points corresponding to water are removed to find potential objects to track. To find the water, RANSAC is used on the point cloud data to identify the predominant plane. This results in the plane parameters in the left rectified camera frame.



(a) Disparity image after removing points close to or below the water plane estimate, compared to fig. 2b.

(b) Disparity image after removing points that are far away, compared to fig. 4a. Note some small specks of points remain in the image.



(c) Disparity image after removing specks from fig. 4b that have an area below a threshold.

Fig. 4. Point cloud and disparity filtering step from fig. 6, to identify objects relevant for tracking. The initial disparity image is shown in fig. 2b.

C. Point cloud and disparity filtering

Each point that is close to or below the water surface is then removed. This includes reflections of potential targets since those points can often be under the water surface, see fig. 4a. This does not mean that waves and wakes are fully removed since it is desirable not to remove too many points that are close to the water surface, ensuring that objects, especially low objects like kayaks, are detected.

The next step is to remove the background. We remove all points that have a distance to the camera that is over a certain threshold. This is a scenario specific way of removing the background. The result of this can look like fig. 4b. Note that the pixels that have small disparity in fig. 2b have zero disparity (are invalid) in fig. 4b.

The last step in the point cloud and disparity filtering is to remove specks of points. These are mostly due to waves and wakes in the water where the tops might not be regarded as water. The resulting disparity image is shown in fig. 4c. Note that the result is similar to a segmentation mask of relevant objects in the scene, but with added depth information.

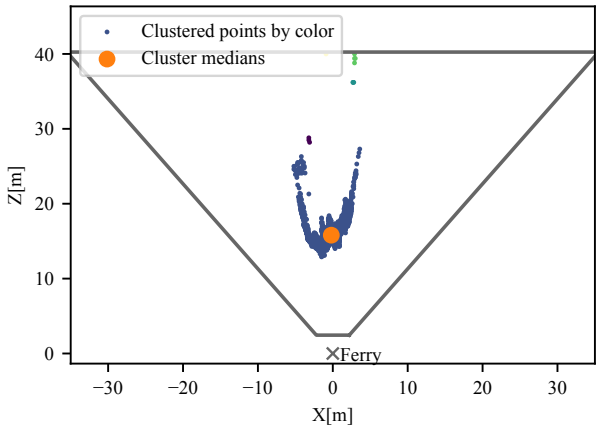


Fig. 5. The point cloud clusters after disparity filtering of the short baseline method. The point cloud is viewed from above.

D. Clustering

The point cloud can then have clusters of points belonging to different objects, see an example of this in fig. 5. Using hierarchical clustering in SciPy [22] these clusters are identified. For each cluster, the median is taken to get a single measurement for each object. The median was chosen in order to be more robust to outliers than the mean. Note that each cluster does not have points belonging to arbitrary points on the objects, but rather points belonging to the visually seen parts of the hull of the objects. This is important when results are shown in section VII.

E. VJIPDA

The points are then transformed to the world frame using the Inertial Navigation System (INS) of the autonomous ferry prototype and they are tracked using a VJIPDA tracker. It takes in point measurements in 2D, east and north position in this case, uses a constant velocity model for each points dynamics, and outputs a track for each identified target. The VJIPDA implementation used is from [14].

VI. Multi-baseline stereo camera tracking

The information flow in the multi-baseline method is shown in fig. 6. In addition to using the short baseline images to find the water plane, the method uses the right stereo camera to get more precise distance measurements.

In order to do the stereo matching across the wide baseline, the left and right camera images need to be rectified. This means that we want the intrinsic parameters of the two models to be the same and that we want epipolar lines to be horizontal. OpenCV can be used for the rectification given an accurate estimate of the transformation (translation and rotation) between the two cameras.

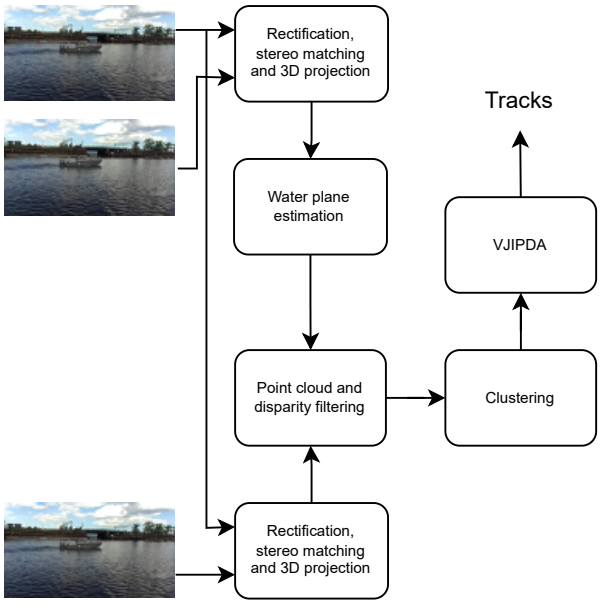


Fig. 6. Block diagram of the multi-baseline method. The inputs are the short and wide baseline images. The output is the tracks. The lower image is from the right stereo camera in fig. 3.

We use the same left camera for both the short and the wide baseline setup, as shown in fig. 3. However, they have different right cameras. The cameras are not perfectly aligned, and just a small difference in pose can give big differences in pixel measurements. Therefore, the rectification changes the image rotation slightly, which has to be accounted for when moving between short and wide baseline stereo setups.

We choose to rotate the water plane parameters from the short baseline reference frame to the wide baseline reference frame. One could also rotate the point cloud from the wide baseline frame to the short baseline frame and use the same plane parameters, but that would not give as many measurements in the resulting disparity image since the points are generated for each pixel and some points will after rotation end up in the same pixels in the short baseline reference image.

The plane equation for the short baseline case is

$$y_s = \alpha_s x_s + \beta_s z_s + \gamma_s, \tag{1}$$

where we know the parameters α_s, β_s and γ_s. We are looking for the parameters for the wide baseline equation

$$y_w = \alpha_w x_w + \beta_w z_w + \gamma_w, \tag{2}$$

where the plane parameters are α_w, β_w and γ_w. x_s, y_s and z_s are points in the point cloud relative to the short baseline reference frame, while the points in the wide baseline reference frame are

$$\begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = \mathbf{R}_s^w \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix},$$

Fig. 7. The resulting filtered disparity image from the wide baseline stereo camera images.

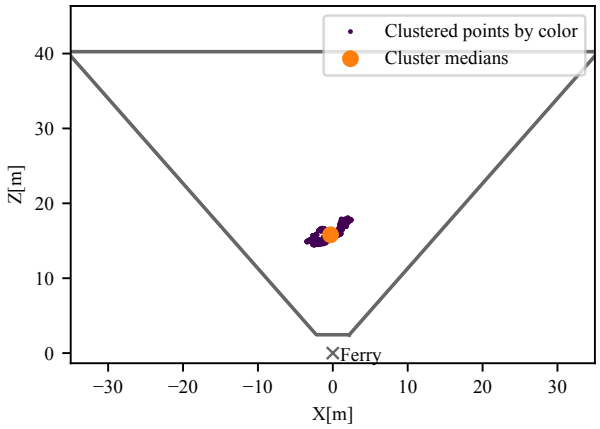


Fig. 8. A similar plot to fig. 5 but for the multi-baseline point cloud.

where $\mathbf{R}_s^w$ is the rotation matrix from the stereo camera rectification that rotates points from frame s to frame w. We rewrite eq. (1) with the wide baseline positions to get

$$\begin{bmatrix} -\alpha_s & 1 & -\beta_s \end{bmatrix} (\mathbf{R}_s^w)^\top \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} = \gamma_s.$$

The first vector and the rotation matrix are known and we call the resulting product $\begin{bmatrix} k_1 & k_2 & k_3 \end{bmatrix}$. This results in the equation

$$k_1 x_w + k_2 y_w + k_3 z_w = \gamma_s$$

which gives us the resulting parameters with respect to eq. (2):

$$\alpha_w = -\frac{k_1}{k_2}, \quad \beta_w = -\frac{k_3}{k_2}, \quad \gamma_w = \frac{\gamma_s}{k_2}.$$

When using such expressions we need to understand when they are valid. If $k_2 = 0$ there will be a problem. However, this could only happen when the distance to the plane is not dependent on y_w, meaning the ego-boat is either pitching or rolling with 90°, which should not happen.

The removal of points on the water surface happens in the point cloud from the wide baseline images using the plane parameters above with a plane estimated from the short baseline point cloud. It is important to use the plane here and not the segmentation mask from that plane estimate in the

TABLE II

DIFFERENT SORTS OF ROOT MEAN SQUARE ERROR (RMSE) ERRORS FOR THE DIFFERENT METHODS AND SCENARIOS. THE METHODS ARE THE MULTI-BASELINE (MB) AND THE SHORT BASELINE (SB). THE RMSE IS CALCULATED USING THE EUCLIDEAN DISTANCE BETWEEN THE ESTIMATED POINT AND THE GROUND TRUTH GNSS AS THE ERROR. THE SCENARIOS ARE THE MANEUVER (SM) WITH THE DAY CRUISER (DC), AND THE SCENARIO WITH CROSSING TARGETS (SC) WITH A KAYAK (K) AND A DAY CRUISER (DC). "WO SE" MEANS WE HAVE REMOVED THE START AND THE END OF THE SEQUENCE WHERE THE TARGET IS ONLY PARTLY VISIBLE. "FS" MEANS "FROM SHAPE" AND IS WHERE THE ERROR IS THE DISTANCE TO THE TARGET MODEL.

	JIPDA MB	**JIPDA SB**
SM DC	2.25	1.95
SM DC wo se	1.91	1.87
SM DC fs wo se	**0.08**	0.26
SC DC	3.37	2.43
SC DC wo se	2.29	2.11
SC DC fs wo se	**0.00**	**0.00**
SC K	**1.57**	2.84

short baseline disparity image. Far away objects are in the short baseline image smoothed into the background and water. If using the imperfect segmentation mask, objects would then not be detectable even though the wide baseline disparity better distinguishes those objects. An example of the object disparity image is shown in fig. 7 with the resulting point cloud shown in fig. 8.

The rest of the pipeline is similar to the short baseline method. In the multi-baseline method, the estimates have to be rotated back to the short baseline rectified left camera frame so that the estimates can be projected to the world coordinate frame using calibrated extrinsic parameters between sensors.

VII. RESULTS

Two scenarios have been analyzed and some results are shown in table II. The first scenario is the maneuvering scenario (SM) where a day cruiser (DC) moves counter-clockwise. The second scenario is the crossing scenario (SC) where the day cruiser moves in a straight line with a kayak (K) moving the other way, becoming occluded by the day cruiser.

Given that the number of targets does not exceed two, these scenarios are overly simplistic for the purpose of comparing probabilistic multi-target trackers, such as the Probability Hypothesis Density (PHD) filter against the JIPDA. Nonetheless, these scenarios served as a basis for comparing the pre-processing of measurements into object detections using short baseline stereo cameras versus multi-baseline stereo cameras.

A. Choosing VJIPDA parameters

Some VJIPDA parameters ranges were found by comparing the stereo matching with lidar data. The parameters were afterward tuned empirically.

The termination threshold was set on the existence probability to $\tau_{term} = 0.1$. The detection probability was set to $P_D = 0.9$ while the visibility state handled the occlusion. A big difference was the number of allowed missing detections

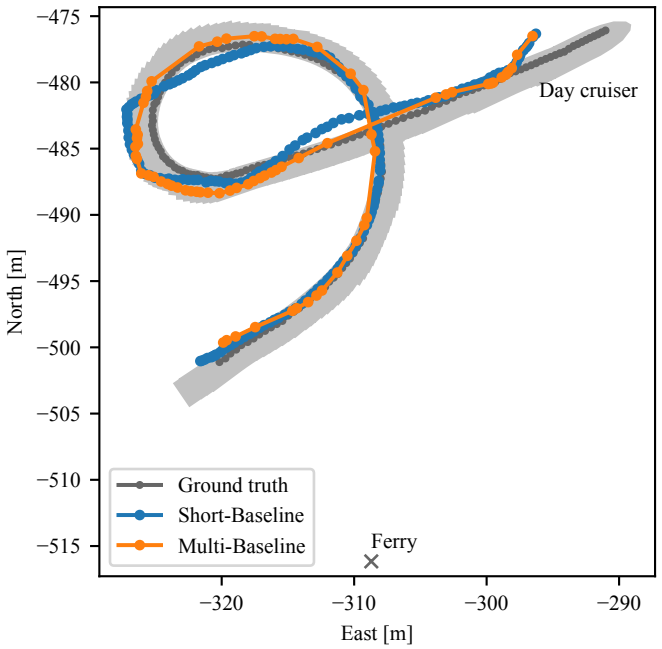


Fig. 9. The resulting tracks from the short baseline and multi-baseline tracking method on a single day cruiser target moving counter-clockwise. Ground truth is post processed GNSS data from an antenna right in front of the boat skipper. The ferry is the ego-vehicle. The day cruiser region is the union of the day cruiser shape for each timestep. See the shape in fig. 10.

before termination of the tracks, specifically for the occlusion, which was $\tau_{miss} = 6$ for the multi-baseline and $\tau_{miss} = 14$ for the short baseline.

Just using the JIPDA, the $P_D = 0.7$ had to be used to allow the tracker to not terminate at the threshold $\tau_{term} = 0.1$ for the existence probability when the kayak was occluded. This is where the VJIPDA allowed us to model the visibility and let the existence probability decrease slower than the visibility probability when we lacked correct detections.

The Cartesian standard deviation for the measurements was set to $\sigma_m = 0.4\,\mathrm{m}$ for the multi-baseline, but had to be increased to $\sigma_m = 0.8\,\mathrm{m}$ for the short baseline. The standard deviation for the process model in the constant velocity model were $\sigma_a = 0.2\,\mathrm{m\,s^{-2}}$. They were set as low as possible while still keeping the correct resulting tracks.

B. Scenario Maneuver (SM)

RMSE results are show in table II and the tracks are shown in fig. 9. The short baseline method exhibits impressive performance compared to the accuracy of the disparity image. With a baseline of just $12\,\mathrm{cm}$ the track did not deviate from the ground truth by more than just a few meters with a target at a distance of around $40\,\mathrm{m}$. This is in spite of the point cloud, as seen in fig. 5, having points on the target that deviate close to $10\,\mathrm{m}$ from the actual target. This is largely because the median is used as the measurement to the VJIPDA. Most points will be closer to the correct depth and taking the median makes the method partly robust to these outliers.

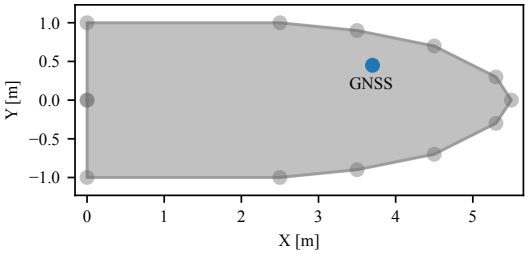


Fig. 10. The model of the boat used for the error metric, with the GNSS antenna location in the model. Based on measurements and images of the target boat.

The RMSE are similar in table II for the two methods when measuring the distance from the track position to the GNSS antenna, with a slightly lower RMSE for the short baseline method. One reason for this is that the tracks include the start and the end of the sequence where the target is only partially visible. In that case both methods are equally sensitive to outliers such as wakes and the results become more random.

Furthermore, we only want to measure the error as the distance to the hull of the boat. The methods did not track the antenna of the GNSS but rather the hull of the target. We do this by modelling where we places the antenna on the target, see the location in fig. 10. We have run a Kalman filter on the GNSS, with a constant velocity model, and aligned the heading of the boat with the velocity vector. We then take the error as the distance from the estimate to the boat shape, allowing all estimates inside the boat shape to have zero error. The result is shown in table II in the row "SM DC fs wo se" where the multi-baseline method performs better.

The only part of the track where the multi-baseline method is outside of the boat model is in the leftmost part of fig. 9. This is because the constant velocity model is too simple and could not account for the target having to swing its aft in order to generate force to go east. The orientation of the "ground truth" model is in that region wrong and the tracking estimates are not as wrong as they seem. On the other hand, the short baseline has missed in depth in two additional regions where the target moved more or less straight forward.

The collected data experienced numerous frame drops, primarily due to the demanding setup of running four cameras simultaneously at full HD resolution, each capturing frames at a rate of 15 Hz, all on a single computer core. As a result, each camera managed to capture only approximately 38% of the expected number of frames. Particularly, the multi-baseline stereo method required both stereo cameras to avoid simultaneous frame drops. In that case, the frame count was further reduced to around 15% of the expected frame count. Consequently, the short baseline method could operate with more measurements than the multi-baseline method, as shown in fig. 9, as it only relies on a single stereo camera.

C. Scenario Crossing (SC)

This scenario is interesting since it has a small kayak at a far distance, it has an occlusion of that kayak with the day cruiser in front and it is therefore also a multi-target scenario. The results are seen in table II and fig. 11.

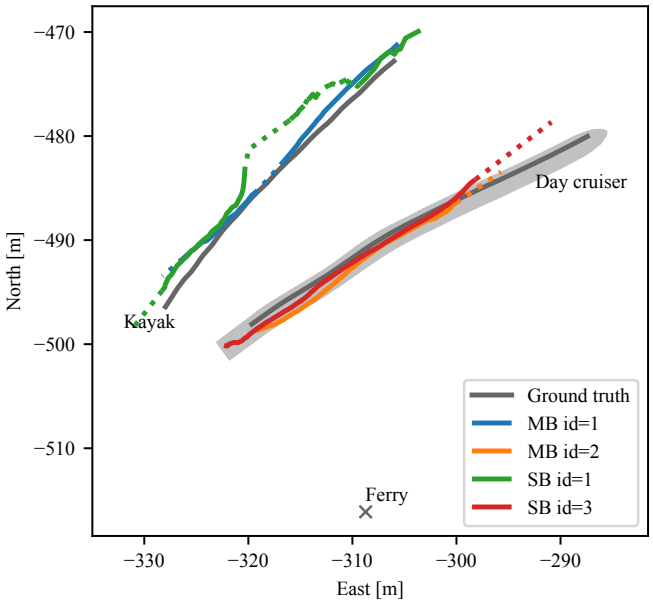


Fig. 11. The resulting tracks from the short baseline (SB) and multi-baseline (MB) tracking method in the crossing scenario. An important addition to the plot is the region that is the union of the day cruiser shape for each timestep. See the shape in fig. 10. The day cruiser moved north-east. The kayak moved south-west. The kayak is occluded by the day cruiser in the middle of its track. The dotted parts are where the visibility is below a threshold.

The cameras mostly see the starboard side of the day cruiser which is over a meter and up to four meters away from the GNSS antenna. This results in the high RMSE values for "SC DC wo se" in table II.

Similarly to the maneuver scenario, we have used the model of the day cruiser to quantify how good the tracking is. Both methods seem to track the day cruiser well with all estimates falling inside the model of the target, see "SC DC fs wo se" in table II.

An important difference between the two methods is seen on the kayak tracks. The short baseline method has a bigger difficulty with the far-away small target. Both methods are able to keep track of the kayak when it is occluded by the day cruiser, see the dotted part of the track, but with different parameters for the number of allowed missed detections, τ_{miss}. The position measurements far away were so noisy for the short baseline that we had to increase the measurement uncertainties in the VJIPDA parameters to associate the kayak after the occlusion to the earlier track that had diverged. This is because the disparity image is in some timesteps smoothing the kayak more into the background than at other timesteps, and the kayak detection is missing in many early frames. The spread of the point cloud is also big for the short baseline.

We have not found it necessary to use a geometric model for the kayak in the performance evaluation as the GNSS was in the middle of the kayak and the kayak was narrow compared to the day cruiser, and it was farther away.

D. Limitations

There are a few limitation to the way the tracking has been done in this paper.

Firstly, the runtime has not been optimized with respect to real-time performance. In particular, the water plane estimation using RANSAC does not run in real time in the current implementation. We believe that this can be achieved by reducing the number of points or by using deep learning, together with code optimization.

Secondly, the background has been removed in a naive way by distance filtering. For tracking close to shore with a movable platform, static land could be found in a better way, e.g., using sea maps.

Third, the methods have assumed the water being a plane. This is a fairly good assumption in the inshore scenario, but the waves and wakes could either be estimated for improved situational awareness or filtered out for more precise tracking. There are of course more advanced ways of segmenting out water, not just using the point cloud but also using statistical models or even deep learning.

Fourth, the methods do not guarantee tracking the same point on the targets over time, as they only took the median of the point clouds. This could be solved by Extended Object Tracking (EOT), see the tutorial paper [11] or the lidar tracking in [23].

VIII. Conclusion

A method of multi-baseline stereo camera boat tracking has been presented. The method alleviates some challenges related to stereo matching due to the properties of water.

Compared to a short baseline method with adequate performance, the multi baseline method gave more accurate tracks as well as more accurate point clouds on the targets. Furthermore, the results indicated that the multi-baseline method achieved stronger track continuity.

For future work, if a dense point cloud is not required, we suggest doing sparse stereo matching instead of dense stereo matching. The sparse stereo matching will ensure that only well-placed features are used for stereo matching and fewer miss-associations are used.

We are in the process of creating a dataset for 3D maritime tracking, which will ease the use of deep learning methods and make systematic benchmarking possible. It should have more scenarios and include heavier visual artifacts, e.g., heavier reflections.

Acknowledgements

We would like to thank the following people for help with the experiment: Trym Anthonsen Nygård, Øystein Kaarstad Helgesen, Petter Hangerhagen, Ingunn Helene Gjendemsjø and Kjetil Vasstein.

References

[1] L. Leal-Taixé, A. Milan, K. Schindler, D. Cremers, I. Reid, and S. Roth, "Tracking the Trackers: An Analysis of the State of the Art in Multiple Object Tracking," Apr. 2017, ArXiv:1704.02781 [cs] type: article.

[2] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *CVPR*, June 2012, pp. 3354–3361, ISSN: 1063-6919.

[3] R. Schuster, O. Wasenmuller, G. Kuschk, C. Bailer, and D. Stricker, "SceneFlowFields: Dense Interpolation of Sparse Scene Flow Correspondences," in *WACV*, Mar. 2018, pp. 1056–1065.

[4] H. Wang and Z. Wei, "Stereovision based obstacle detection system for unmanned surface vehicle," in *ROBIO*, Dec. 2013, pp. 917–921.

[5] B.-S. Shin, X. Mou, W. Mou, and H. Wang, "Vision-based navigation of an unmanned surface vehicle with object detection and tracking abilities," *Machine Vision and Applications*, Vol. 29, No. 1, pp. 95–112, Jan. 2018.

[6] J. Muhovič, R. Mandeljc, B. Bovcon, M. Kristan, and J. Perš, "Obstacle Tracking for Unmanned Surface Vessels Using 3-D Point Cloud," *IEEE Journal of Oceanic Engineering*, Vol. 45, No. 3, pp. 786–798, July 2020.

[7] P. Dendorfer, *et al.*, "MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking," *International Journal of Computer Vision*, Vol. 129, No. 4, pp. 845–881, Apr. 2021.

[8] S. Guo, *et al.*, "A Review of Deep Learning-Based Visual Multi-Object Tracking Algorithms for Autonomous Driving," *Applied Sciences*, Vol. 12, No. 21, p. 10741, Jan. 2022.

[9] S. Xu, A. Savvaris, S. He, H.-s. Shin, and A. Tsourdos, "Real-time Implementation of YOLO+JPDA for Small Scale UAV Multiple Object Tracking," in *ICUAS*, June 2018, pp. 1336–1341, ISSN: 2575-7296.

[10] H. Shertukde and Y. Bar-Shalom, "Tracking of crossing targets with imaging sensors," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 27, No. 4, pp. 582–592, July 1991.

[11] K. Granström and M. Baum, "A Tutorial on Multiple Extended Object Tracking," TechRxiv, Feb. 2022, Unpublished.

[12] B. Leibe, K. Schindler, N. Cornelis, and L. Van Gool, "Coupled Object Detection and Tracking from Static Cameras and Moving Vehicles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 30, No. 10, pp. 1683–1698, Oct. 2008.

[13] D. Musicki and R. Evans, "Joint integrated probabilistic data association: JIPDA," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 40, No. 3, pp. 1093–1099, July 2004.

[14] E. F. Brekke, A. G. Hem, and L.-C. N. Tokle, "Multitarget Tracking With Multiple Models and Visibility: Derivation and Verification on Maritime Radar Data," *IEEE Journal of Oceanic Engineering*, Vol. 46, No. 4, pp. 1272–1287, Oct. 2021.

[15] K. Konolige, "Small Vision Systems: Hardware and Implementation," in *Robotics Research*, Y. Shirai and S. Hirose, Eds. London: Springer, 1998, pp. 203–212.

[16] D. Griesser, G. Umlauf, and M. O. Franz, "Visual Pitch and Roll Estimation For Inland Water Vessels," in *ICRA*, May 2023, pp. 1961–1967.

[17] T. Huntsberger, H. Aghazarian, A. Howard, and D. C. Trotz, "Stereo vision–based navigation for autonomous surface vessels," *Journal of Field Robotics*, Vol. 28, No. 1, pp. 3–18, 2011.

[18] L. Su, Y. Chen, H. Song, and W. Li, "A survey of maritime vision datasets," *Multimedia Tools and Applications*, Vol. 82, p. 28873–28893, Mar. 2023.

[19] Y. Cheng, M. Jiang, J. Zhu, and Y. Liu, "Are We Ready for Unmanned Surface Vehicles in Inland Waterways? The USVInland Multisensor Dataset and Benchmark," *IEEE Robotics and Automation Letters*, Vol. 6, No. 2, pp. 3964–3970, Apr. 2021.

[20] H. Hirschmuller, "Accurate and efficient stereo processing by semi-global matching and mutual information," in *CVPR*, Vol. 2, June 2005, pp. 807–814 vol. 2, ISSN: 1063-6919.

[21] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[22] P. Virtanen, *et al.*, "SciPy 1.0: fundamental algorithms for scientific computing in Python," *Nature Methods*, Vol. 17, No. 3, pp. 261–272, Mar. 2020.

[23] M. Baerveldt, M. E. López, and E. F. Brekke, "Extended target PMBM tracker with a Gaussian Process target model on LiDAR data," in *FUSION*, June 2023, pp. 1–8.