2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Visual Pitch and Roll Estimation For Inland Water Vessels

Dennis Griesser[1], Georg Umlauf[1] and Matthias O. Franz[1]

***Abstract*— Motion estimation is an essential element for autonomous vessels. It is used e.g. for lidar motion compensation as well as mapping and detection tasks in a maritime environment. Because the use of gyroscopes is not reliable and a high performance inertial measurement unit is quite expensive, we present an approach for visual pitch and roll estimation that utilizes a convolutional neural network for water segmentation, a stereo system for reconstruction and simple geometry to estimate pitch and roll. The algorithm is validated on a novel, publicly available dataset[2] recorded at Lake Constance. Our experiments show that the pitch and roll estimator provides accurate results in comparison to an Xsens IMU sensor. We can further improve the pitch and roll estimation by sensor fusion with a gyroscope. The algorithm is available in its implementation as a ROS node[3].**

I. INTRODUCTION

In the shipping industry there is currently an intense effort to increase safety through automation. In this context, research is being conducted on autonomous vessels, similarly to autonomous driving. However, as there is much stronger inherent movement in maritime vehicles due to wave action, wind and current, this must also be precisely measured. Calibrated high performance inertial measurement units (IMU's) with a high sampling rate are suitable for this task. They provide high-precision acceleration and orientation data even in situations where other sensors like a global positioning system (GPS) fail. Unfortunately, these high performance IMU's with low gyro bias and exact orthogonality between the axes are very expensive. On the other hand, cameras are affordable and deliver large amounts of information about the environment. In most autonomous vessels, cameras are available for perception tasks anyway. Therefore, our paper proposes an alternative approach based on a stereo camera system to estimate pitch and roll on inland waters without the need of an IMU. Furthermore, we suggest a sensor fusion with a gyroscope because it measures orientation very well. A gyroscope is very accurate over short time scales, but is not applicable on larger time scales due to gyroscope drift during dead reckoning and also the initial orientation is missing. However, the proposed visual pitch and roll estimation is more accurate over the long term similar to the fusion of acceleration and magnetometer measurements of an IMU.

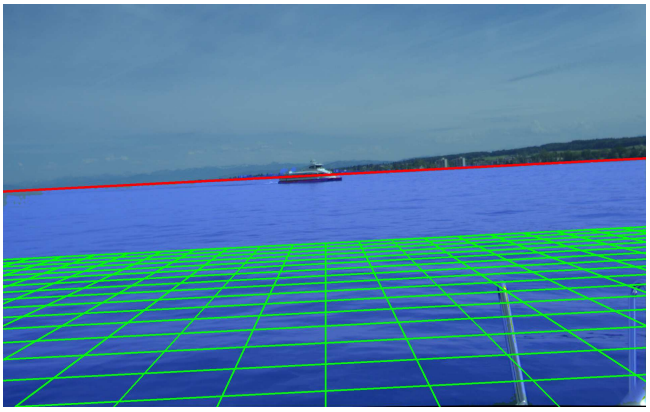

Fig. 1. Estimation of the artificial horizon (red), ground plane (green) and segmentation result as blue overlay.

The experiments show that we can improve the visual pitch and roll estimation by sensor fusion with a gyroscope. Thus, our approach offers an alternative measurement system to an accelerometer and a magnetometer. An advantage of our system compared to the magnetometer is that it does not need a hard and soft iron calibration to compensate magnetic disturbances. In addition to pitch and roll estimation, the sensor height can also be determined which can be used to estimate the artificial horizon.

The algorithm uses two calibrated cameras to acquire a synchronized pair of stereo images. To locate where water occurs a convolutional neural network (CNN) segments the pixels of the image into foreground (blue overlay in Fig. 1) or background. All pixels of the water surface are foreground, all others are background. The spatial position of all foreground pixels is reconstructed using stereo to obtain the 3d position of the water surface. The water surface is approximated with a plane (green grid in Fig. 1). The detection of this ground plane is also useful for other tasks such as object detection, mapping or path planning. The estimated ground plane in relation to the camera leads to pitch and roll estimates which can be plotted as an artificial horizon (red line in Fig. 1). The artificial horizon provides a simple tool to visualize changes in orientation. For the validation of the segmentation CNN, we introduced a new test dataset[2] with images recorded at Lake Constance. To evaluate the pitch and roll estimation a new dataset was created as well. The dataset was also recorded at Lake Constance and contains nine sequences with stereo images and IMU data. The results are compared to a high performance Xsens IMU where the accelerometer data and magnetometer data are fused with gyroscope data using a Kalman Filter.

[1]Dennis Griesser, Georg Umlauf and Matthias O. Franz are with the Institute for Optical Systems, University of Applied Sciences Konstanz, Germany. Email: {dgriesse, umlauf, mfranz}@htwg-konstanz.de.

This research has been financed by the Baden-Württemberg Stiftung gGmbH and BMBF (01IS19083A).

[2]https://git.ios.htwg-konstanz.de/dgriesse/constance_orientation_dataset/-/archive/main/constance_orientation_dataset-main.zip

[3]https://github.com/dionysos4/water_surface_detector

II. RELATED WORK

Stereo vision in maritime environments

In the last few years, a number of papers [1]–[13] showed that cameras are promising sensors for perception tasks to realize autonomous navigation for vessels. Larson et al. [1] presented already in 2007 a stereo vision system to detect objects above the water surface. Zhang et al. [2] focused on the development of a stereo camera system for unmanned surface vehicles (USV) for the reconstruction of objects. Their special interest was in stereo calibration and stereo matching for USV navigation. The stereo system was then used to create obstacle maps for obstacle avoidance [3]. Huntsberger et al. [4] used two stereo pairs to increase the field of view. They also created range images and computed the plane of the water surface, but they have not mentioned how they computed the plane. Furthermore, they computed 2d grid maps and classified the objects with a supervised learning algorithm. Additionally, they tracked the objects to obtain position and velocity. All of these methods use the stereo vision system for detection and tracking tasks, but there was no study that investigated if a stereo system is able to estimate pitch and roll accurately in maritime environments.

Water surface estimation

Wang and Wei [5] presented also a stereovison based detection and tracking approach. They mentioned that the sea surface can be considered as flat and used random sample consensus (RANSAC) [14] for plane fitting. Shin et al. [13] presented an autocalibration approach for the stereo system and obtained the sea surface also with RANSAC. Most similar to our application is the work of Bovcon et al. [8]. They also use a stereo camera system and conduct RANSAC plane fitting. Similar to our work, they use the plane normal to estimate pitch and roll, but their focus is also on object detection and object tracking. All the mentioned approaches approximated the water surface with a plane, but there was no experiment that validated how accurate pitch and roll is estimated with the help of the plane normal.

Segmentation

For save maritime navigation it is important to distinguish between water surface, objects and shoreline. Bovcon et al. [9] used a graphical model in combination with an IMU to segment the water surface. Liang et al. [6] developed a semi-supervised video water segmentation network especially for changes of color and texture of the water between consecutive frames. Steccanella et al. [7] used also a binary segmentation network to separate water and non-water regions to detect the waterline and obstacles in front of an USV. Bovcon and Kristan [10] proposed a deep encoder-decoder network to separate water and obstacles. The novel decoder fuses IMU information with features from the decoder to improve estimation of the water edge. Zhou et al. [11] improved the encoder-decoder architecture by incrementally fusing the encoded features into the decoder and Zhou et al. [12] presented a CNN which reliable segments images in inland waterways. There are a variety of segmentation networks for maritime navigation. However, our focus is not on subpixel accurate image segmentation. For our application it is only important to roughly segment the maximum number of pixels which belong to the water plane so that a large part of the water surface is reconstructed.

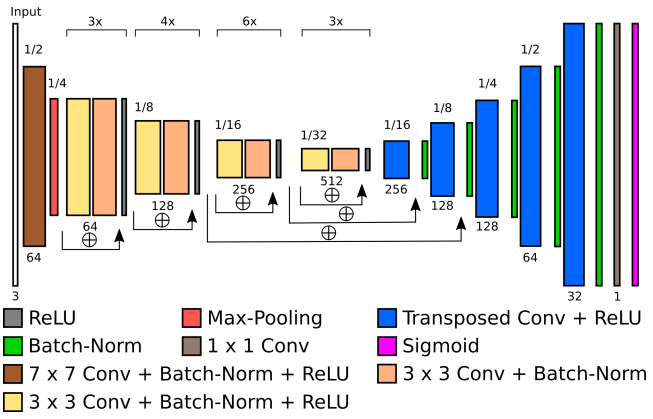

Fig. 2. Architecture of the segmentation network.

III. METHOD

A. Segmentation

To distinguish between water and background pixels in the image I, a binary segmentation network is used. The architecture is a Fully Convolutional Network (FCN-8s) [15] as visualized in Fig. 2 and predicts

$$I(x,y) = \begin{cases} 1 & \text{if } x,y \text{ is water plane} \\ 0 & \text{if } x,y \text{ is background} \end{cases}.$$

The network has an encoder-decoder structure, using a ResNet-34 [16] as encoder and a series of transposed convolutions with ReLU activation and batch normalization as decoder with intermediate skip connections. As final activation a sigmoid is used to obtain a pixelwise class score. The encoder transforms the input with convolutional and max-pooling layers to a low dimensional feature representation to realize a deeper network. Subsequently, the decoder increases the spatial resolution with transposed convolutions to retrieve the input resolution. The skip connections by additions preserve fine-grained details. Since the network is fully convolutional, the only limitation is that the input image resolution must be divisible by 32. The loss function is the binary cross entropy. The overall network architecture is illustrated in Fig. 2 and the hyperparameters are described in Section IV-A.

B. Reconstruction

In this step we are only interested in the pixels which are part of the water class. Therefore, all pixels in the original image of the left camera which are predicted as background were set to intensity value zero. In order to obtain spatial information of the water plane, a calibrated stereo camera system is needed with rectified image pairs. For disparity estimation of the water plane we used a modified version

of stereo processing by semiglobal matching and mutual information [17] which uses block matching and utilizes the Birchfield-Tomasi dissimilarity [18] instead of mutual information. Further, the costs are aggregated in only five directions. With the estimated disparity d and the camera parameters (c_x, c_y: coordinates of the principal point, f: focal length, B: baseline) of the calibrated stereo system, the point cloud of the water plane is reconstructed with

$$\begin{aligned} X &= \frac{-B(x - c_x)}{d}, \\ Y &= \frac{-B(y - c_y)}{d}, \\ Z &= \frac{-fB}{d}, \end{aligned} \tag{1}$$

where x, y are image coordinates and X, Y, Z are coordinates of a point of the reconstructed point cloud.

C. Plane fitting

Since the system is designed for distances below 1km, the earth curvature can be neglected. Therefore, the water surface can be approximated by a plane. Because neither segmentation nor stereo matching is outlier-free, RANSAC [14] is used to obtain the plane. The algorithm iteratively estimates the plane parameters ($\mathbf{n}$: normal, $\mathbf{p}$: support point) by picking three random points $[X_i, Y_i, Z_i]^T$ of the reconstructed water plane and taking those for which the consensus set is largest.

D. Pitch and Roll estimation

Especially because the captured water surface can be described as a plane, it is possible to estimate the orientation of the camera with respect to it. However, it is impossible to estimate the yaw angle just from the geometry of the water plane. Therefore, we concentrate to estimate pitch and roll. For this purpose, we orientate the plane normal downwards. To calculate pitch θ and roll φ, the coordinates of the estimated normal $\mathbf{n} = [a, b, c]^T$ are used according to

$$\theta = \arctan\left(\frac{c}{b}\right), \qquad \varphi = \arctan\left(\frac{a}{b}\right). \tag{2}$$

In addition to the pitch and roll estimation, our system is able to determine the height of the camera above the water surface with $\mathbf{n} \cdot \mathbf{p}$, where $\|\mathbf{n}\| = 1$. For an intuitive representation of the estimated plane, the horizon line projected into the image is a good feature to validate the results quickly. Object detection can also be accelerated with the horizon line. Objects at a certain distance (e.g. catamaran in Fig. 1) are on the horizon line, and therefore objects must be searched only along the line. For the projection of the line, we set the translation to map a point from the camera coordinate system to the plane coordinate system by

$$\mathbf{t}_{cam}^{plane} = \begin{bmatrix} 0 \\ -\mathbf{n} \cdot \mathbf{p} \\ 0 \end{bmatrix}. \tag{3}$$

The rotation matrix which rotates the plane with respect to the camera is defined as

$$\mathbf{R}_{plane}^{cam} = \begin{bmatrix} \mathbf{x} & \mathbf{n} & \mathbf{z} \end{bmatrix}. \tag{4}$$

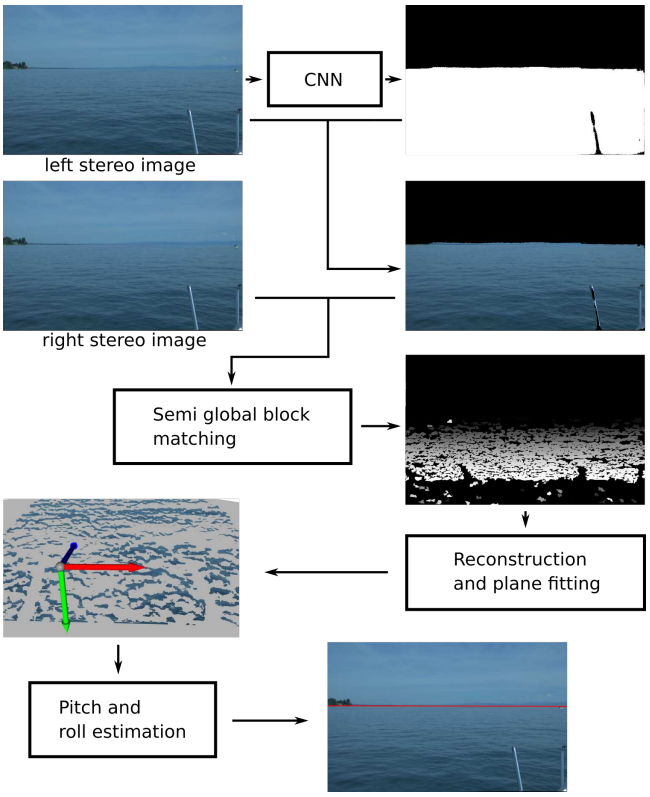

Fig. 3. System architecture of the pitch and roll estimation pipeline.

The columns of the rotation matrix are computed by the cross products

$$\begin{aligned} \mathbf{z} &= \mathbf{e}_x \times \mathbf{n}, \\ \mathbf{x} &= \mathbf{n} \times \mathbf{z}, \end{aligned} \tag{5}$$

where $\mathbf{e_x}$ is the unit vector $[1, 0, 0]^T$. This results in the final transformation to project a point defined in the plane coordinate system into the camera system

$$\mathbf{T}_{plane}^{cam} = \begin{bmatrix} \mathbf{R}_{plane}^{cam} & -\mathbf{R}_{plane}^{cam} \cdot \mathbf{t}_{cam}^{plane} \\ 0 & 1 \end{bmatrix}. \tag{6}$$

Starting from the mean earth radius r (6371km) and a sensor height h above the water surface ($\mathbf{n} \cdot \mathbf{p}$) of roughly 2m, we can compute the distance d to the horizon as

$$d = \sqrt{2rh + h^2}, \tag{7}$$

which is roughly 5000m. For visualization, we define homogenous coordinates $\mathbf{a}$ in the water plane coordinate system with x-range from -2500m to 2500m, set y to zero and z to 5000m. A visualization of this setup is shown in Fig. 4. Finally, we project the points of the artificial horizon into the image plane $[x/w,\ y/w]^T$ with the projection matrix $\mathbf{P}$ (estimated during camera calibration) with

$$\begin{bmatrix} x \\ y \\ w \end{bmatrix} = \mathbf{P} \cdot \mathbf{T}_{plane}^{cam} \cdot \mathbf{a}_{plane}. \tag{8}$$

The overall system architecture is visualized in Fig. 3.

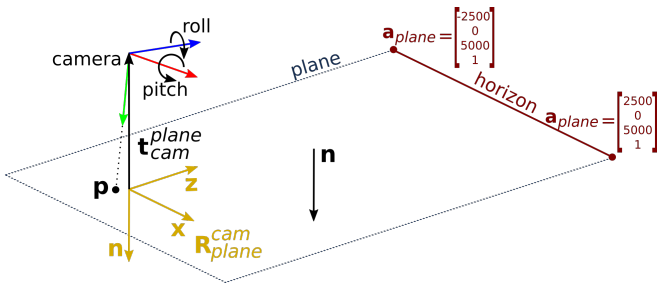

Fig. 4. Visualization of the configuration to estimate pitch and roll and the transformations to plot a 3d point into the image plane.

IV. EXPERIMENTS

A. Segmentation

We use three different datasets to train, validate and test the segmentation network.

1) The Tampere-WaterSeg dataset [19] contains 600 images from Lake Pyhäjärvi with corresponding pixelwise labels of water and background. There are sequences of three different situations where the boat is located in a channel, on the open lake and in a docking scenario. The training set consists of 400 open lake and docking scenes and the validation set of 200 channel images.
2) The WaterDataset [6] has 2188 labeled training images where 1888 images are from ADE20K [20] and 300 images from the River segmentation dataset [21]. The images from ADE20K [20] are recorded at seas, lakes, rivers, channels and swimming pools. On the other hand, the River segmentation dataset [21] contains only images of different rivers. The validation dataset has 212 labeled images and consists mostly of river, harbor and beach images.
3) The segmentation test dataset has been created within this work. We equipped our research boat with a camera and recorded at various scenarios at Lake Constance. We manually annotated 30 frames of different scenarios (docking, harbor, seerhein, open lake) and different weather conditions (sunny, cloudy, foggy). A subset of the dataset is shown in Fig. 5.

To figure out on which training dataset the network generalizes best, we trained the segmentation network individually on the Tampere-WaterSeg dataset [19], on the WaterDataset [6] and finally on a combination of both to increase the data size. To avoid over- and underfitting we validated the network after each epoch with the associated validation data. The ResNet-34 encoder was initialized with pretrained ImageNet [22] weights. We rescaled the input images to a size of 960×640 (width, height) pixels. Each training was 60 epochs with a learning rate of 0.0001 and a mini-batch size of 4. For optimization RMSprop with a weight decay of 0.00001 was used. During training we optimized all network weights. To obtain the final performance of the network we tested on our novel segmentation test dataset. The results on the test dataset are shown in Table I. As performance measures

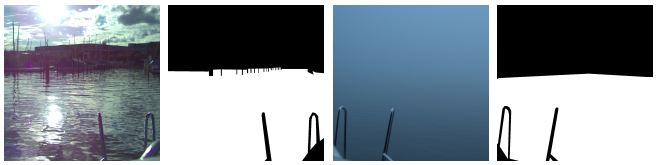

Fig. 5. Two samples of the segmentation test dataset which contain the RGB images and the associated labels as binary images.

for the segmentation results, we use intersection over union (IoU) [23], F_1-Score [24] and accuracy. However, we focus on IoU and the F_1-Score, because the accuracy leads to misleading results, if the classes of water and background are not balanced.

TABLE I
EVALUATION ON THE NOVEL SEGMENTATION TEST DATASET WITH NETWORKS TRAINED AND VALIDATED ON COMBINED, TAMPERE-WATERSEG AND WATERDATASET.

Train dataset	IoU	F_1	Accuracy
Combined	0.923	0.958	0.960
Tampere-WaterSeg	0.795	0.874	0.884
WaterDataset	**0.937**	**0.966**	**0.968**

The model trained only on the Tampere-WaterSeg dataset [19] does not generalize well (see Table I) to our segmentation test dataset, probably due to too small variations of the images. The combination of Tampere-WaterSeg dataset [19] and WaterDataset [6] achieves better segmentation results. However, the best generalization performance could be achieved when trained on the WaterDataset [6] only. We have also conducted the same experiments with a VGG16 FCN-8s [25], [15], a ResNet-34 FCN-32s [16], [15] and a U-net [26] architecture which all performed worse. For the final application, the test dataset was added to the WaterDataset training dataset and the model was retrained.

B. Dataset

To evaluate the entire system, we used a dataset which we also recorded at Lake Constance referred to as Lake Constance orientation dataset. This is a subset of the Lake Constance dataset [27]. For the recordings, the sensor system is equipped with a stereo system with a baseline of 1.6m and an image resolution of 1920×1200 pixels. The stereo system consists of two Basler acA1920-40gc cameras with a Kowa lens with 12.5mm focal length. Additionally, an IMU (Xsens-MTi-G-710) is mounted to record ground truth motion. The Xsens IMU uses a Kalman filter for sensor fusion of magnetometer, accelerometer and gyroscope. We use this data as ground truth because the fusion of these sensors determine an orientation with high accuracy (0.2 degrees root mean square in pitch/roll). In principle, the accuracy of a camera-driven orientation estimate should be even higher than the Xsens-estimate, but we currently have no experimental setup to validate this claim. The images of the stereo system were captured at a frame rate of 10Hz.

Fig. 6. Sequences of the Lake Constance orientation dataset.

The IMU data is recorded with 100Hz. Since we need synchronized data for the evaluation, we used spherical linear interpolation [28] between consecutive IMU messages to interpolate at the captured image timestamps. The IMU also delivers the delta angles of the gyroscope between two time steps. The dataset was recorded on two different days and is split into 9 different sequences (day 1: sequence 0-4, day 2: sequence 5-8). A thumbnail of each sequence is visualized in Fig. 6. The dataset contains 8321 stereo images with calibration data and the interpolated fused IMU data. We additionally provide a sequence with 596 stereo images and IMU measurements as calibration sequences for each day. We stored the data in the hierarchical data format (hdf5) to be able to store the stereo images and IMU data in combination with camera calibration parameters.

C. IMU to stereo calibration

While the Xsens IMU directly outputs pitch and roll (θ_{imu}, φ_{imu}), we use our method from Section III to estimate pitch and roll (θ_{stereo}, φ_{stereo}) of the stereo system. Both sensors have different coordinate systems, which have to be aligned before testing. To align the two sensors we need to find the rotation $\mathbf{R}$ between them in a calibration step. In the calibration step θ_{imu}, φ_{imu} and θ_{stereo}, φ_{stereo} are converted into rotation matrices by rotating about axes y, x (intrinsic rotations) with

$$\begin{aligned} \mathbf{R}_{imu} &= \mathbf{R}_y(\theta_{imu})\mathbf{R}_x(\varphi_{imu}), \\ \mathbf{R}_{stereo} &= \mathbf{R}_y(\theta_{stereo})\mathbf{R}_x(\varphi_{stereo}). \end{aligned} \tag{9}$$

To calibrate the two sensors we compute the rotation between two measurements of the IMU (denoted by $\mathbf{A}$) and the camera (denoted by $\mathbf{B}$) at different time steps i, $i-1$ according to

$$\begin{aligned} \mathbf{A}_i &= \mathbf{R}_{imu,i-1}^T \mathbf{R}_{imu,i}, \\ \mathbf{B}_i &= \mathbf{R}_{stereo,i-1}^T \mathbf{R}_{stereo,i}. \end{aligned} \tag{10}$$

Since the camera and IMU are mounted rigidly next to each other on a fixed aluminum profile, we can solve the IMU camera calibration with the equation $\mathbf{AR} = \mathbf{RB}$. To obtain the unknown $\mathbf{R}$, we used the closed-form least squares solution suggested in [29].

D. Evaluation of the pitch and roll estimation

During the experiments we set the following parameters: the segmentation mask is predicted on to 960×640 resized camera images. The input images are zero centered by subtracting the mean in each channel and scaled by the standard deviation. The prediction threshold which specify the class for each pixel is 0.5. If a pixel of the CNN output is greater or equal than the threshold the pixel contains water otherwise background. For semiglobal matching the minimal disparity is 0, the number of disparities is 160 and the block size is set to 9. The disparity computation and reconstruction is performed on to 960×600 resized images. To fit the water plane, only points in range of 40m are considered. The maximum number of trials for RANSAC plane fitting is set to 1000 and the residual threshold which specify inlier points to 20cm. The experiments are executed on a workstation with Intel Core i7-4770 CPU and an NVIDIA Geforce GTX 1080 Ti GPU. During the experiments, the average prediction time of the neural network was 5.1ms (GPU). The average disparity computation time was 259ms (CPU). The ROS node processes images with 7Hz because the disparity is calculated on significantly smaller images. We expect an even higher frame rate if the disparity is also computed on the GPU. For evaluation of the visual pitch and roll estimation, we use our novel Lake Constance orientation dataset. To verify how well the vision estimates are, compared to the IMU measurements, we first calibrated the sensors with the calibration sequences for each day. $\mathbf{R}$ is estimated with the method described in Section IV-C. The stereo coordinate system is aligned with the IMU coordinate system according to

$$\mathbf{R}_{stereo}^{imu} = \mathbf{R}_{stereo}\mathbf{R}. \tag{11}$$

To check the calibration, the difference rotation matrix is calculated as

$$\mathbf{\Delta} = \mathbf{R}_{imu}\mathbf{R}_{stereo}^{imuT}. \tag{12}$$

If the estimated pitch and roll are equal to the ground truth, $\mathbf{\Delta}$ is the identity. However, in practice we have estimation errors due to noise, reconstruction errors and errors during plane fitting. To measure the spread of these estimates, we use the average misorientation angle and compute the confidence interval by bootstrapping [30]. First, the sample mean is computed over $\mathbf{\Delta}$ at each time step i

$$\widetilde{\mathbf{R}} = \frac{1}{n}\sum_{i=0}^{n} \mathbf{\Delta}_i. \tag{13}$$

Second, the mean rotation $\mathbf{M}$ is estimated by $\mathbf{M} = \mathbf{VW}$ [30] from the singular value decomposition of $\widetilde{\mathbf{R}}$

$$\widetilde{\mathbf{R}} = \mathbf{V\Sigma W}. \tag{14}$$

The misorientation angle which is the smallest rotation angle between $\mathbf{\Delta}_i$ and $\mathbf{M}$ [30] is then defined as

$$\mathrm{mis}(\mathbf{\Delta}_i, \mathbf{M}) = \arccos\left(\frac{\mathrm{tr}(\mathbf{\Delta}_i^T\mathbf{M}) - 1}{2}\right), \tag{15}$$

where the overall spread in the dataset is measured by the average misorientation angle

$$AMA = \frac{1}{n}\sum_{i=0}^{n} \mathrm{mis}(\mathbf{\Delta}_i, \mathbf{M}). \tag{16}$$

To obtain confidence intervals we use the same bootstrapping strategy as proposed in [30]. We sample 1000 times, where the sample size is equal to the dataset size. We also use the 2.5 and 97.5 percentiles as confidence bounds. The AMA and associated confidence intervals results of the 9 sequences of the Lake Constance orientation dataset are shown in Table II.

TABLE II

AVERAGE MISORIENTATION ANGLE IN DEGREES, CONFIDENCE INTERVAL WITH THE 2.5 PERCENTILE IN DEGREES AND THE 97.5 PERCENTILE IN DEGREES.

Sequence	AMA	CI
0	0.138	(0.133, 0.143)
1	0.197	(0.188, 0.206)
2	0.165	(0.158, 0.172)
3	0.180	(0.175, 0.184)
4	0.155	(0.146, 0.163)
5	0.189	(0.179, 0.198)
6	0.194	(0.184, 0.203)
7	0.634	(0.587, 0.686)
8	0.459	(0.436, 0.484)
0-8	0.180	(0.174, 0.186)

The AMA indicates that the visual pitch and roll estimations are very accurate in sequences $0-6$. However, in sequence 7 there is a strong deviation from the mean rotation. This arises from the glossy and cloudy water surface in this sequence. There the water surface is almost textureless so that the pattern matching fails. This leads to a very sparse point cloud. Subsequently, the RANSAC plane estimator returns a plane which lies in space arbitrarily and pitch and roll are inaccurate. Similarly, sequence 8 exhibits an increased value due to a sparse point cloud.

E. Gyroscope fusion

Our system estimates pitch and roll at each time step. However, there can be outliers in between. In this respect, our system is similar to the sensor fusion of magnetometer and accelerometer stable over the long term. In this section we combine the long term stability of our approach with the short term stability of a gyroscope. First we used the initial orientation of the IMU output and multiplied consecutive delta angles of the gyroscope (only the data from the gyroscope without sensor fusion) to find the orientation change over a sequence. The AMA compared with the IMU data (with sensor fusion) is shown in column 2 of Table III. Due to uncorrected bias or high frequency noise there is a gyroscope drift and the AMA is in all sequences over 5 degrees. Therefore, a gyroscope alone is useless. To combine the measurements of our system and a gyroscope we used a complementary filter [31]. The input to the complementary filter is the high frequency noise of our estimation and the low frequency noise of the gyroscope. The idea is to use a low pass filter to filter out high frequency noise and using its complement, a high pass filter to filter out low frequency noise. We implemented the discrete complementary filter as suggested in [31]. The complementary filter uses the visual pitch and roll estimation to slightly update the gyroscope into the correct orientation.

TABLE III

AVERAGE MISORIENTATION ANGLE IN DEGREES WITH THE INTEGRATED GYROSCOPE DATA. ADDITIONALLY, THE AVERAGE MISORIENTATION ANGLE IN DEGREES AFTER THE COMPLEMENTARY FILTER, CONFIDENCE INTERVAL WITH THE 2.5 PERCENTILE IN DEGREES AND THE 97.5 PERCENTILE IN DEGREES.

Sequence	AMA (gyroscope)	AMA (fused)	CI
0	11.672	0.129	(0.122, 0.137)
1	26.347	0.192	(0.181, 0.204)
2	8.646	0.129	(0.122, 0.137)
3	23.865	0.128	(0.122, 0.135)
4	8.180	0.141	(0.132, 0.153)
5	35.802	0.145	(0.135, 0.157)
6	5.717	0.165	(0.142, 0.193)
7	11.997	0.339	(0.322, 0.356)
8	13.018	0.250	(0.235, 0.266)
0-8	40.618	0.174	(0.166, 0.182)

There is a clear trend if we compare the AMA in Table II with the fused AMA in Table III. The sensor fusion improves the AMA in all sequences.

V. CONCLUSIONS

Our method constitutes a new approach to estimate pitch and roll on inland waters using a stereo system. In addition, a simple fusion with a gyroscope has been suggested. The water surface segmentation which was trained on open access datasets deliver reliable results on the novel segmentation test dataset. The overall system achieves accurate results in comparison to an IMU. However, the approach has limitations. If the water surface is foggy or cloudy the pattern matching fails. Another limitation is that our system can currently only be used during daytime as long as no additional night vision capabilities are provided. To make the system more robust against environmental influences such as foggy and cloudy water surfaces, more attention should be paid to the development of suitable pattern matching algorithms for stereo reconstruction. In scenarios where pattern matching fail, there is still some texture for example by reflection. A CNN which is optimized for such cases may help. The experiments showed that the visual system is able to supplement an IMU. It provides an additional pitch and roll estimate when wave action, wind and current affect the vessel. The fusion with the gyroscope can bridge the time when magnetometer measurements of an IMU are distorted due to external influences.

ACKNOWLEDGMENTS

We would like to thank Konstantin Christ for his help with initial experiments. We would also like to thank Matthias Albrecht and Tim Baur who helped us with the data acquisition.

REFERENCES

[1] J. Larson, M. Bruch, R. Halterman, J. Rogers, and R. Webster, "Advances in autonomous obstacle avoidance for unmanned surface vehicles," in *SPAWAR San Diego CA.*, 2007.

[2] W. Zhang, P. Zhuang, L. Elkins, R. Simon, D. Gore, J. Cogar, K. Hildebrand, S. Crawford, and J. Fuller, "A stereo camera system for autonomous maritime navigation (amn) vehicles," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 2009.

[3] L. Elkins, D. Sellers, and W. R. Monach, "The autonomous maritime navigation (amn) project: Field tests, autonomous and cooperative behaviors, data fusion, sensors, and vehicles," in *Journal of Field Robotics*, 2010.

[4] T. Huntsberger, H. Aghazarian, A. Howard, and D. Trotz, "Stereo visionbased navigation for autonomous surface vessels," in *Journal of Field Robotics*, 2011.

[5] H. Wang and Z. Wei, "Stereovision based obstacle detection system for unmanned surface vehicle," in *IEEE International Conference on Robotics and Biomimetics*, 2013.

[6] Y. Liang, N. Jafari, X. Luo, Q. Chen, Y. Cao, and X. Li, "Waternet: An adaptive matching pipeline for segmenting water with volatile appearance," in *Computational Visual Media*, 2020.

[7] L. Steccanella, D. Bloisi, A. Castellini, and A. Farinelli, "Waterline and obstacle detection in images from low-cost autonomous boats for environmental monitoring," in *Robotics and Autonomous Systems*, 2020.

[8] J. Muhovič, R. Mandeljc, B. Bovcon, M. Kristan, and J. Perš, "Obstacle tracking for unmanned surface vessels using 3d point cloud," in *IEEE Journal of Oceanic Engineering*, 2019.

[9] B. Bovcon, R. Mandeljc, J. Perš, and M. Kristan, "Stereo obstacle detection for unmanned surface vehicles by imu-assisted semantic segmentation," in *Robotics and Autonomous Systems*, 2018.

[10] B. Bovcon and M. Kristan, "A water-obstacle separation and refinement network for unmanned surface vehicles," in *IEEE International Conference on Robotics and Automation*, 2020.

[11] W. Zhou, X. Huang, and X. Zeng, "Obstacle detection for unmanned surface vehicles by fusion refinement network," in *IEICE Transactions on Information and Systems*, 2022.

[12] R. Zhou, Y. Gao, P. Wu, X. Zhao, W. Dou, C. Sun, Y. Zhong, and Y. Wang, "Collision-free waterway segmentation for inland unmanned surface vehicles," in *IEEE Transactions on Instrumentation and Measurement*, 2022.

[13] B. S. Shin, X. Mou, W. Mou, and H. Wang, "Vision-based navigation of an unmanned surface vehicle with object detection and tracking abilities," in *Machine Vision and Applications*, 2018.

[14] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," in *Association for Computing Machinery*, 1981.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[17] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[18] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998.

[19] J. Taipalmaa, N. Passalis, H. Zhang, M. Gabbouj, and J. Raitoharju, "High-resolution water segmentation for autonomous unmanned surface vehicles: a novel dataset and evaluation," in *IEEE 29th International Workshop on Machine Learning for Signal Processing*, 2019.

[20] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[21] L. Lopez-Fuentes, C. Rossi, and H. Skinnemoen, "River segmentation for flood monitoring," in *IEEE International Conference on Big Data*, 2017.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," in *International Journal of Computer Vision*, 2015.

[23] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," in *International Journal of Computer Vision*, 2010.

[24] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," in *BMC Medical Imaging*, 2015.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[27] D. Griesser, M. Albrecht, and T. Baur, "Lake constance dataset," In development.

[28] K. Shoemaker, "Animating rotation with quaternion curves," in *Association for Computing Machinery*, 1985.

[29] F. C. Park and B. J. Martin, "Robot sensor calibration: solving ax=xb on the euclidean group," in *IEEE Transactions on Robotics and Automation*, 1994.

[30] M. A. Bingham, "Quantifying spread in three-dimensional rotation data: comparison of nonparametric and parametric techniques," in *Journal of Statistical Distributions and Applications*, 2015.

[31] W. T. Higgins, "A comparison of complementary and kalman filtering," in *IEEE Transactions on Aerospace and Electronic Systems*, 1975.