Improving Computer Vision Based Perception for Collaborative Indoor Navigation

Laura Ruotsalainen, Aiden Morrison, Maija Mäkelä, Jesperi Rantanen and Nadezda Sokolova

***Abstract*— Collaborative navigation is the most promising technique for infrastructure-free indoor navigation for a group of pedestrians, such as rescue personnel. Infrastructure-free navigation means using a system that is able to localize itself independent of any equipment pre-installed to the building using various sensors monitoring the motion of the user. The most feasible navigation sensors are inertial sensors and a camera providing motion information when a computer vision method called visual odometry is used. Collaborative indoor navigation sets challenges to the use of computer vision; navigation environment is often poor of tracked features, other pedestrians in front of the camera interfere with motion detection, and the size and cost constraints prevent the use of best quality cameras resulting in measurement errors. We have developed an improved computer vision based collaborative navigation method addressing these challenges using a depth (RGB-D) camera, a deep learning based detector to avoid using features found from other pedestrians and for controlling the inconsistency of object depth detection, which would degrade the accuracy of the visual odometry solution if not controlled. Our analysis show that our method improves the visual odometry solution using a low-cost RGB-D camera. Finally, we show the result for computing the solution using visual odometry and inertial sensor fusion for the individual and UWB ranging for collaborative navigation.**

***Index Terms*— Collaborative navigation, computer vision, depth cameras, indoor navigation, Kalman filtering, object detection**

I. Introduction

At present, there is no solution that would provide required accuracy and reliability for indoor navigation at rescue operations. Indoors Global Navigation Satellite Systems (GNSS) are unavailable. Time and safety critical operations, in unknown environments, prevent relying on equipment placed into the area such as WiFi transmitters used widely indoor navigation. Therefore, navigation must be implemented using sensors carried by the users. In this case, the users are pedestrians, and therefore the equipment must also be lightweight and low-cost. Such requirements set fundamental challenge on the long term navigation performance. The application area requires real-time functioning with mobile devices, which excludes several computation methods.

Infrastructure-free navigation, namely using a system that is able to localize itself independent of any equipment pre-installed to the building, builds upon using various sensors monitoring the motion of the user. Fusion of the sensor measurements for propagating an intial position solution, provides continuous relative positioning [1]. Infrastructure-free indoor navigation is still an unsolved problem when the indoor time is not tightly bounded. Cost and size requirements of pedestrian navigation system demands use of Micro-Electro-Mechanical (MEMS) sensors, which results in measurement errors and rapidly drifting position solution [2]. Therefore, effort has been put into developing data fusion algorithms [3] for mitigating the effect of errors and extending the eligible navigation time indoors. The goal of our research is to provide an accurate navigation solution for a group of rescue personnel collaborating.

The most promising solution for such a group in rescue operations is collaborative navigation [4]. In collaborative navigation, ranges between group members are measured using radio signals such as Ultra-Wide Band (UWB) round trip timing (RTT) signals. Then, range information is shared between the users for improving the position estimates for each group member. The position estimates are significantly improved [5] especially when at least one of the group members is outdoors where GNSS is available, or experiencing a zero velocity update (ZUPT) [6]. The most feasible method for computing the individual infrastructure-free position estimates is visual inertial odometry [7], fusing a camera and an inertial measurement unit (IMU).

Cameras and computer vision, namely algorithms for ob-

Submitted for review on March 31 2021. This work was supported in part by the NATO Science for Peace and Security (SPS) funding, Grant for project G5406

L. Ruotsalainen is with the Department of Computer Science, University of Helsinki, Finland (e-mail: author@helsinki.fi)

A. Morrison and N. Sokolova are with SINTEF, Norway (e-mail: author@sintef.no)

M. Mäkelä and J. Rantanen are with the Finnish Geospatial Research Institute, Finland (e-mail: author@nls.fi).

taining understanding from images, have been used for improving the navigation solution for years. The methods have been developed first for robots [8], vehicles [9] and lately for pedestrians [1] and drones [11]. Computer vision provides speed and attitude measurements and results in improved navigation solution when fused with other sensors [10]. Speed and attitude of the pedestrian carrying a camera is computed by tracking the motion of features in consecutive images projected from static objects in the scene. Our challenging application complicates the use of computer vision in three ways. First of all, tracking requires many well visible objects in the environment, which is not usually the case indoors where the method suffers from poor lighting and areas with very uniform surfaces. This, in addition to the real-time processing requirement, advocates the use of visual odometry instead of a Simultaneous Localization and Mapping (SLAM) solution despite its good performance in navigation in more feasible situations [12]. Secondly, pedestrians need cost-effective and small sized cameras, traditionally the only option has been the use of a monocular camera. Monocular camera perception suffers from scale ambiguity, which degrades the visual odometry solution. Thirdly, the setup involving multiple pedestrians results in many dynamic objects around the camera, also degrading the visual odometry solution. In order not to degrade the visual inertial odometry and the resulting collaborative navigation solution, we have to solve these challenges before entering the visual result into the fusion.

In this paper, we will first discuss the computer vision challenges and our solution for the infrastructure-free collaborative indoor pedestrian navigation application. Our solution is based on visual odometry computed by tracking Speeded Up Robust Features (SURF) [25] matched between consecutive images taken with Intel RealSense depth camera [14]. To avoid tracking dynamic objects we have trained an object detector to detect the pedestrians and remove their features from tracked objects. Convolutional Neural Network (CNN) [18] based pedestrian detectors have gained good performance during the recent years [19]. The accelerator in the research has most likely been the emergence of solutions for autonomous vehicles and their requirement for complete situational awareness. However, our application differs from the vehicle based ones as the pedestrians are only partially observed from the images due to being at very close vicinity of the camera. Therefore, we have trained the model using images captured during two collaborative navigation scenarios. Then, we will present analysis of the performance improvement in the visual odometry solution alone arising from our solution and then an Extended Kalman filtering based collaborative navigation solution fusing loosely-coupled measurements from visual odometry, IMU and a barometer for height measurements. The performance of the solution is tested in real life experimentation. Finally, we conclude the paper and discuss our future work.

II. Computer vision solution for collaborative pedestrian indoor navigation

Cameras are used in pedestrian navigation for providing velocity and heading information, and thus complementing the conventional inertial sensors. While the inertial sensor measurements suffer from drift, computer vision methods are degraded by dynamic objects in the environment and lack of features when the scene consists of surfaces with very uniform material. Unfortunately, the scenes in our collaborative navigation consist of both; other members of our collaborative team as dynamic objects and outdoor areas covered with snow and feature poor office spaces indoors.

Traditionally monocular cameras have been used in pedestrian navigation due to their good performance combined with small size. However, speed measurements suffer from scale ambiguity due to the inability of monocular cameras to measure the distance between objects and themselves. Research has been active in solving the problem, but solutions have often been suitable only for limited use cases [1]. Recent improvement in the quality of small RGB-D cameras provides good opportunities for computer vision based navigation applications. RGB-D cameras, such as the RealSense product family manufactured by Intel [14], provide the depth of tracked objects and therefore solve the scale problem.

Depth camera measures the distance d between the camera's optical center and an object point $\mathbf{X} = (X, Y, Z)^T$. When the camera is modeled as a pinhole camera we can get the real 3D coordinates of $\mathbf{X}$ using

$$\mathbf{X} = \frac{d}{||\mathbf{K}^{-1}\mathbf{x}||}\mathbf{K}^{-1}\mathbf{x} \tag{1}$$

where $\mathbf{K}$ is the calibration matrix including the camera intrinsics and $\mathbf{x}$ homogeneous image pixel coordinates $\mathbf{x} = (x, y, 1)^T$ [20]. Rigid transformation of the point $\mathbf{X}$ from the camera centered coordinate system to the arbitrary world coordinate frame $\mathbf{X}_w$ is defined by the rotation $\mathbf{R}$ and translation $\mathbf{T}$ of the camera with respect to the world frame as

$$\begin{bmatrix}\mathbf{X}\\ 1\end{bmatrix} = \begin{bmatrix}\mathbf{R} & \mathbf{T}\\ \mathbf{0} & 1\end{bmatrix}\begin{bmatrix}\mathbf{X}_w\\ 1\end{bmatrix}. \tag{2}$$

When we are looking at the motion between two consecutive images we can get the real metric displacement of the camera by setting the origin to the first camera center ($\mathbf{T}_1 = (0, 0, 0)^T$), defining the coordinate axis being aligned with the camera orientation ($\mathbf{R}_1 = \mathbf{I}$), where $\mathbf{I}$ is the identity matrix) and using $\mathbf{T}_2 = -\mathbf{R}_2\mathbf{C}_2$, where $\mathbf{C}_2$ is the world coordinates of the camera optical centre at the time of taking the second image. This way we are able to get the real metric visual odometry solution, which we will turn into user heading change and speed measurements in our collaborative navigation setup.

We have ensured that the image point matching is providing correct measurements by using RANdom SAmple Consensus (RANSAC) [21], which is a robust estimator discarding erroneous matches. However, it is impossible to avoid measurement errors in computer vision applications and the distance measurements provided by the depth camera are also inconsistent. Therefore, we are using a rule based on rigid transformations in Euclidean space demonstrating that the relative distances between different object points observed from first and second images remain the same [22]. When the object points observed from the first image are defined

as $\mathbf{X}_{1i}, i, k = 1, ...n$ and second image $\mathbf{X}_{2i}, i, k = 1, ...n$ the constraint is

$$\|\mathbf{X}_{1i} - \mathbf{X}_{1k}\| = \|\mathbf{X}_{2i} - \mathbf{X}_{2k}\| \quad (3)$$

By computing the object points using the matched image points and (1) and comparing combinations of object point pairs using (3), we form a subset that most likely contains only inliers and use those for computing the translation parameter $\mathbf{T}$.

A. Improved feature detection in challenging environments

Scale Invariant Feature Transform (SIFT) [23] is an approach based on transforming an image into local feature vectors; SIFT descriptors, describing the intensities around image points that are found as maxima or minima of a difference-of-Gaussian function. Each vector is invariant to image translation, scaling, and rotation and partially invariant to illumination changes and affine or 3D projections. Due to these invariances, SIFT provides good performance for detecting distinctive features that may be matched across images and therefore its improved variants have been developed, such as real-time detector FAST [24] and SURF [25] that provide improved detection accuracy. However, these detectors use Gaussian derivatives as smoothing kernels in the detection process, which smooths relevant detail such as object boundaries from the images, while removing noise. When navigating in an initially feature-poor environment it is essential to detect all features, even the weaker ones and therefore even SURF fails frequently. Kaze [26] is a feature detector and descriptor based on nonlinear diffusion filtering. Kaze provides multiscale features that have high repeatability and distinctiveness. However, creating Kaze detectors and descriptors takes approximately 2.5 times more computation time than SURF and in the safety-critical applications requiring real-time processing Kaze cannot be the only solution. Therefore, in our research we detect first SURF features and then if that fails, look for Kaze features.

Figure 1 shows Kaze feature matching a detection result in a feature poor corridor environment where not a single SURF feature was found.

B. Detecting collaborators

Computation of a camera egomotion relies on tracking features found from static objects. If the tracked features are detected from dynamic objects, camera motion is accidentally fused with the object motion, resulting in erroneous solution. However, the collaborative navigation setting by nature contains multiple moving pedestrians in the nearby area. Therefore, it is essential to detect the pedestrians found in images and remove their features from the tracked ones. Figure 2 shows an example of the common problem; our visual navigation method has detected SURF features and matched them across two images for computing the camera egomotion at the time interval between taking the images. In this case, most of the features have been detected from a dynamic pedestrian resulting in an erroneous visual odometry

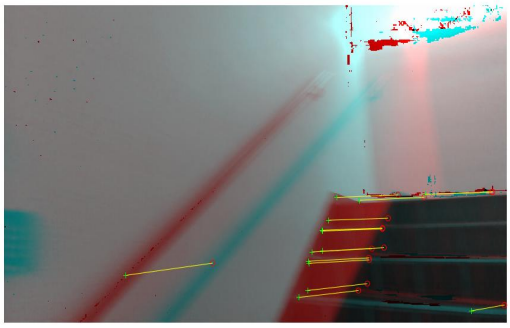

Fig. 1: Kaze features detected and matched (green crosses in image 1 and red circles in image 2) in a feature poor indoor staircase, where no SURF features were found.

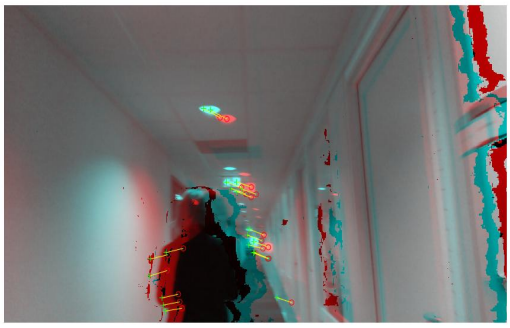

Fig. 2: SURF features detected and matched (green crosses in image 1 and red circles in image 2) in an image from an indoor corridor. Majority of the features are detected from a dynamic pedestrian.

solution. In recent years low-level feature detection models based on Convolutional Neural Networks (CNN) [18] have achieved good performance [27]. Development of methods for detecting high-level features, has been accelerated by the research focusing on autonomous traffic, where the detection of for example pedestrians is crucial [28].

The appearance of humans in the images of a collaborative navigation setting is complicated by many of the challenges found in pedestrian detection, most likely even more than in the largely addressed topic of detecting pedestrians from vehicle cameras. These challenges arise from differences of scale of humans in the images, their attitude, angle of view, illumination changes and occlusion caused by other humans. When in transport applications humans are usually detected outdoors and with moderate distance between the camera and the pedestrian, in indoor collaborative navigation the lighting is challenging and close to each other and thereby the camera. Occlusions arise, resulting in situations where specific body parts are not visible, and a human should be detected from non-human looking figures. Figure 3 shows a typical view from a collaborative navigation situation.

Early CNN based object detectors processed images in two phases, first they detected where the objects in the image

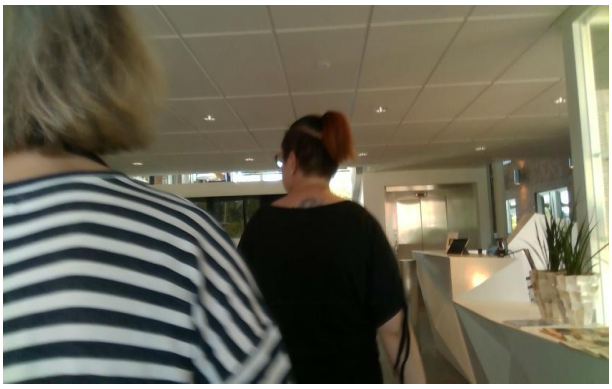

Fig. 3: Typical camera view in a collaborative navigation setting.

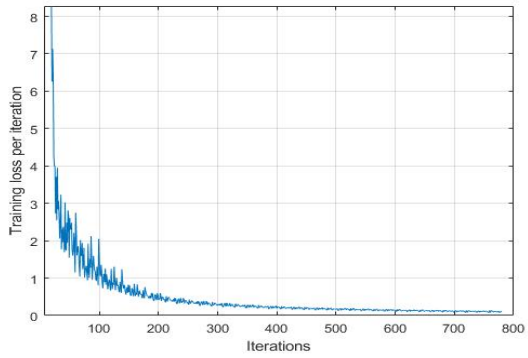

Fig. 4: Training loss per iteration.

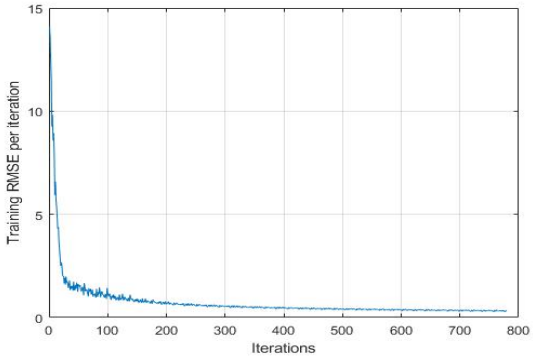

Fig. 5: Training error per iteration.

were and then identified them. First CNN based object detector achieving significant improvement over the conventional object detectors was R-CNN [29]. However, it was slow to train and required a lot of disk space and therefore more efficient variants emerged [30] [31]. Although the processing time decreased on each R-CNN variant, it is not yet suitable for most time critical applications. The invention of You Only Look Once (YOLO) model family; YOLO [32], YOLO9000 also known as YOLOv2 [33] and the recent version YOLOv3 [34] which performed detection in one phase, revolutionized the time demand. As the first YOLO implementation lost clearly to R-CNNs in accuracy, the situation has improved for newer versions. At present, YOLOs are the cutting edge of the object detection. They usually use a pre-trained CNN as basis for feature extraction. YOLOs split the input image into a grid of cells and for each cell directly predicts a bounding box locating an object if there is one inside the box and simultaneously provide its classification.

Here we have used a method based on YOLOv2 model for detecting pedestrians in images in the collaborative navigation setup. Deficiency in YOLOv2 is that it is not able to utilize effectively low-level features, and therefore detect pedestrians that are at a long distance from the camera appearing small in the images. However, this inability does not degrade our navigation solution. Objects with a larger distance than 10 meters from the camera are discarded due to the RealSense depth detection limits and therefore, the presence of distant pedestrians is not disturbing the computation. YOLOv2 object detection network is composed of two sub-networks. A feature extraction network followed by a detection network. Our YOLOv2 based implementation uses a pre-trained ResNet-50 CNN as a backbone for feature extraction and Activation40ReLU for detection trained using images collected during various cooperative navigation experiments. Before training the object detector with 1000 selected images, we augmented the data to avoid over-fitting the model to the data [35]. In this case we are not increasing the number of images in the data set, but warping them to increase variability. Out of the 1000 manually labelled images we used 600 for training and the remaining 400 images for evaluation of the method. For optimization we used Stochastic Gradient Descent with Momentum (SGDM) with mini-batch size 16 and learning rate 0.001. We trained the model using 20 epochs. When the batch size is 16 for 600 images, one epoch takes 38 iterations to complete, and thereby the number of iterations for 20 epochs is 760.

Figure 4 shows the training loss for each iteration, Figure 5 the training root mean square error (RMSE) and Figure 6 precision (ratio of the detected pedestrians, ie. true positives, to all instances that were inferred to be pedestrians) and recall (ratio of detected pedestrians to all pedestrians in the data set) for the training and learning processes. Average precision was 70%, which is in line with the accuracy of conventional pedestrian detectors in challenging indoor lighting environments [36].

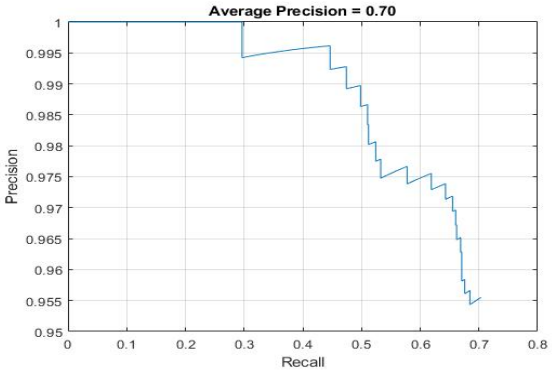

Fig. 6: Precision and recall for pedestrian detection.

C. Our visual odometry solution

The final goal of our research is to first fuse tightly inertial sensor measurements with visual odometry and then use cooperative navigation for correcting the resulting individual navigation solutions. Due to the challenging navigation environment and setup, the challenges degrading the visual odometry solution have to be understood and mitigated. Therefore, we have computed a stand-alone visual odometry solution using a RealSense D435 depth camera attached rigidly to the user's torso, and incorporated the improvements presented in this section.

Image points were detected and matched from the images using SURF features providing efficient computation, but when there were too few points, Kaze features were detected. Wrong matches were discraded using the RANSAC algorithm. The presence of pedestrians in the image was detected using the YOLOv2 detector and the points inside the boxes were discarded. Also, points inside a 30 pixel wide area from the image left side were discarded as non-reliable due to the RealSense depth computation approach [14]. RealSense D435 provides the depth solution reliably only for objects closer than 10 meters from the camera. Unfortunately, the resulting depth measurements are not consistent over an image in a dynamic scenario and for objects further away from the camera. These challenges affected mainly the speed solution of the visual odometry and the attitude with lesser extent. The resulting speed solution had random oscillation, which was filtered by using a very basic Kalman filter.

Visual odometry (VO) requires at least five successfully matched eligible image features when a calibrated camera is used for computing the relative camera pose and four points to get the pose into the orientation and translation with respect to the world coordinates. Our VO solution uses 3D to 2D Perspective-n-Point (PnP) algorithm for computing the camera pose [15].Bucketing is generally used in computer vision research for computing the VO solution [16]. The benefits of using bucketing are reducing the number of features for decreasing computational complexity of the algorithm as well as guaranteeing good distribution of the image points through the image. However, in our use case the challenge is the scarcity of feature points, so neither of the benefits would be achieved and therefore bucketing is not used. The VO solution would be improved using a method called keyframing [17]. Keyframing means that feature correspondences are not only computed to two consecutive images, but to a number of previous images, keyframes making the pose estimation more robust. The goal of this paper was to address the challenges due to monocular depth ambiguity, moving objects and scarcity of features, therefore keyframing will be left for future research.

Our navigation setup includes multiple areas, where less image points than required by VO are detected in total, such as indoor staircases, or areas where all detected features are farther than five meters from the user and thereby must be discarded due to an erroneous depth solution. In these situations the computation is evaluated to have failed and such epochs are excluded from navigation. Figure 7 shows

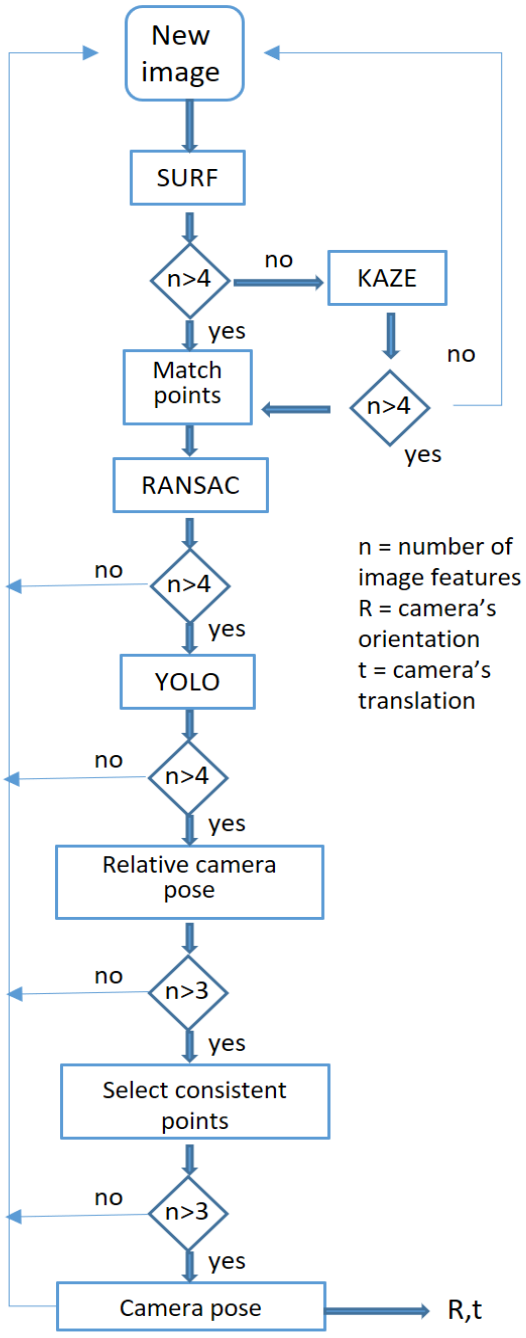

Fig. 7: Flowchart of the visual odometry process.

the flow-chart of the visual odometry processing. As the VO solution is lost frequently due to these challenges, using conventional frame-to-frame methods, such as keyframing, would not have been robust enough. Therefore, we have used the basic Kalman filter mentioned before. The Kalman filter state $\mathbf{x}$ consisted of only the horizontal speed s $\mathbf{x} = s$ filtering out only large changes in the speed solution due to erroneous VO observations. This setup resulted in very simple Kalman filtering implementation, namely state-transition matrix $\mathbf{A} = 1$, covariance of the process noise $\mathbf{Q} = 0.001$. The covariance of the observation noise ($\mathbf{R}$) was adapted to the error detection in the observations described earlier. When the VO solution

was evaluated accurate $\mathbf{R} = 0.05$, and when unreliable $\mathbf{R}$ was set to $\mathbf{R} = 0.5$.

III. Collaborative Navigation

Collaborative Navigation (CN, also known as cooperative or peer-to-peer navigation) is an approach where the navigating units within the same area share their location and possibly other information. CN is based on two central assumptions. First, that each individual user is able to independently estimate their navigation state and state uncertainty, and second that they have the ability to measure or estimate range and range uncertainty to other cooperating users then communicate the result to them. By exchanging in real time estimates of state, relative range and state uncertainty each cooperating user reduces the rate at which the error of the group of cooperating users accumulates [37]. In our study, we are computing the navigation solution for one user (target user U) by using the visual odometry, IMU and barometer measurements and then improving the solution by using collaborative navigation via measuring range between the user and another user in the group (initiating user S). The range is measured by transmitting signals using Ultra-Wide Band (UWB) sensors and computing the Round Trip Time (RTT) of transmitted and received signal as shown in Figure 8 and (4).

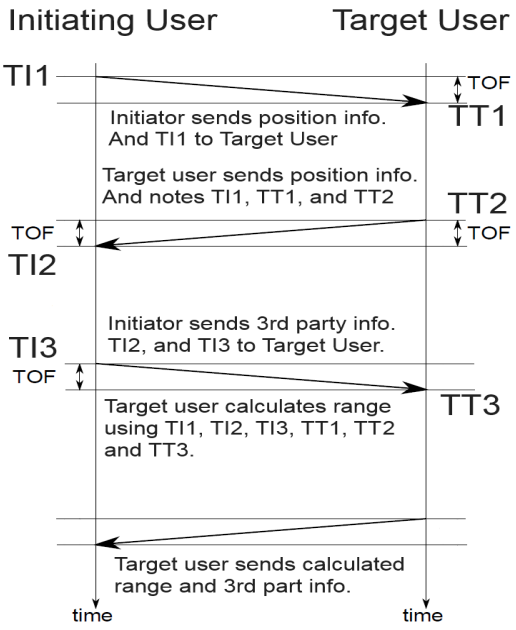

Fig. 8: Computing the Round Trip Time solution.

$$\begin{aligned}
\mathbf{T}_{RoundTripA} &= TI2 - TI1 \\
\mathbf{T}_{RoundTripB} &= TT3 - TT2 \\
\mathbf{T}_{ReplyA} &= TI3 - TI2 \\
\mathbf{T}_{ReplyB} &= TT2 - TT1 \\
\mathbf{RTT} &= 4 \cdot TOF = \\
\mathbf{T}_{ReplyB}(\mathbf{T}_{RoundTripA} &- \mathbf{T}_{ReplyA}) + \\
(\mathbf{T}_{RoundTripB} &- \mathbf{T}_{ReplyB})
\end{aligned} \tag{4}$$

In this study, an Extended Kalman filter (EKF) was used for estimating the state vector $\mathbf{x}_k$ including 16 states, namely the position (x, y, z), velocity (v_x, v_y, v_z), attitude (pitch, roll, yaw), gyroscope bias (g_x, g_y, g_z), accelerometer bias (a_x, a_y, a_z) and barometer bias (b_z). The state is predicted using the visual odometry, IMU, and barometer measurements. As is implicit in the state selection, integration of aiding sensors with the IMU is accomplished via loose coupling, typically via the position and or velocity states depending on the specific source.The noise model adopted assumes white gaussian noise dominates from the INS accelerometers and gyroscopes,while the accelerometer and gyroscope bias states are modelled as first order Markov processes.

The measurement model relating the RTT measurements with the state is

$$z_k = H_k \cdot x_k, \tag{5}$$

where $H_k = [dxdp, dydp, dzdp, 0, \ldots, 0]$. The $dxdp$ term is calculated from the predicted position X-coordinate of the initiating user (S) and target user (U) as

$$dxdp = \frac{X_s - X_u}{R_{pred}} \tag{6}$$

and R_{pred} is

$$R_{pred} = \sqrt{(X_s - X_u)^2 + (Y_s - Y_u)^2 + (Z_s - Z_u)^2}. \tag{7}$$

dydp and dzdp are computed similarly using the predicted Y and Z coordinates, respectively.

Additional performance is achieved through sharing of specific state flags which reflect whether for example a user is presently static (ZUPT condition) or whether a user has an independent ability to measure its position such as a GNSS fix. These conditions frequently arise when some but not all members of a team enter a building, denying GNSS to those inside but often still allowing the higher power RTT and communications signals to pass through the building. In this study, decentralized collaborative navigation [4] was pursued in order to allow the cooperating group of users to dynamically fragment and reform without disruption as communication through building material allowed at any given time. While the benefit of decentralized processing is the implicit tolerance of the ensemble to communication dropouts, a primary drawback is the need to mitigate the impact of the circulation of stale data which in this study was addressed through injection of stabilizing noise to the position state of users operating without an external reference. In order to reduce the impact of information re-circulation within the network, an additional noise term is added to the position states of the Q matrix of each user with density equivalent to a 5 metre uncertainty in each axis after 5 minutes of operation, but is only applied to dynamic users without GNSS reference. Users under ZUPT conditions or with an external position solution do not receive this Q matrix modification. Situational awareness is maintained through the forwarding of 3rd party state information such that even a

single point of contact shared between two groups of users will allow knowledge of the others state even if range measurement between two given users is at that point not possible due to building material obstruction. Since the measurement model makes direct use of the estimated position states of both the initiating and responding user, the formed vector is both scaled and rotated by position state errors in either user. For constant position uncertainty in the initiating and target users, the error in the direction of the vector formed in equation 6 will tend to grow larger with decreasing predicted range. To account for this, the measurement covariance matrix for inter-user ranging measurements is de-weighted when Rpred values fall below 5 metres.

IV. Experiments and Results

We collected data for evaluating the performance of the improved computer vision based navigation method. The test was done in December 2019 in Trondheim, Norway, with snowy conditions. We started navigating outdoors, entered an office building where we walked around for five minutes, and ended the data collection outdoors. Figure 10 shows the test conditions outdoors. The whole experiment lasted approximately 13 minutes and resulted in 6740 images, the whole route, computed with the reference solution, is shown in Figure 11. RealSense depth camera provides RGB and depth images separately, and our system aligned them at the time of capture. Image rate varied between 1Hz and 15Hz, 8Hz on average, however this was accommodated using time stamps from image processing.

Testing was carried out using a team of collaborating users each carrying a combined navigation system based on an ADIS16488A tactical grade Micro-Electro-Mechanical (MEMS) IMU, a uBlox M8T multi-constellation GNSS receiver (GPS, Glonass, Galileo, Beidou), barometer (integrated within the ADIS16488A), and two forms of RTT measurement and communication radios. The first radio operates on 802.15.4a Chirp Spread Spectrum (CSS) in the 2.4 GHz band, while the Second based on the decawave DW1000 IC uses a 500 MHz UWB pulse implementation. The RealSense D435 camera camera was carried by a single user designated user 1 rigidly mounted to a carrying frame which also carried the reference navigation system as shown in Figure 12.

The reference system comprised a Novatel SPAN system [38] with professional GNSS receiver and ISA100C IMU, which was initialized outdoors using GNSS Real-time Kinematic (RTK) corrections [39]. Figure 9 shows a block-diagram of the system setup and communication between the modules. More details about the system may be found from [4]. Solution was post-processed to form a reference trajectory accurate and consistent to the decimeter level through the course of the trajectory including the more than five minutes of GNSS signal absence when operating within the building. Using this trajectory it is possible to calculate the navigation state errors of the collaborative navigation system, as well as the fixed Realsense camera solution. While the reference system only directly measures the navigation errors of the primary user directly, the test trajectory was configured such that secondary

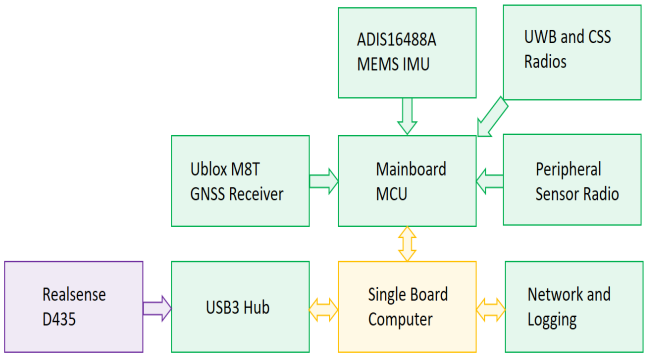

Fig. 9: Block-diagram of the system setup and communication between the modules. MCU stands for Microcontroller Unit and CSS Chirp Spread Spectrum. More details in [4].

users would follow the same trajectory until stopping at staggered locations while the primary user proceeded with the remaining group. At the end of the indoor trajectory segment, the primary user exited on to the roof, for 30 seconds before doubling back and retracing the trajectory in reverse. This approach allows approximation of the errors of the secondary users even absent direct measurement though with increased uncertainty. For the remainder of the paper we will focus on the error distribution of the primary user with the camera.

A. Results

First, we computed the visual odometry stand-alone solution presented in Section II. From 6740 images, the solution was evaluated to be error free for 4269 images. Pedestrians were detected from 2448 images using the YOLOv2 detector. Kaze features were successfully detected from 350 images when detection of SURF points failed, but no features were detected from 192 images.

To evaluate the accuracy of our visual odometry solution, we computed user's speed from the translation from visual odometry and respective image rate. Speed was compared with the reference speed solution. Figure 13 shows the speed profile of visual odometry using conventional visual odometry and Figure 14 for our method. Both Figures show the speed solution using red points when it was evaluated to be correct, based on the error detection explained in Section II, and blue for the reference. When the speed solution is evaluated to be erroneous, it will not be used for cooperative navigation, and those points are omitted from the Figures.

The visual odometry speed profile follows the ground truth speed profile quite well except for two sections, roughly for images 300-350 and 3800-3850. The reason for the failure for those parts was the snowy outdoor environment. The only features closer than 10 meters from the camera, and thereby usable for the depth computation, were found from one side of the camera only. This resulted in a degenerated feature configuration and erroneous visual odometry solution [40]. The speed mean errors were 0.41 m/s for conventional visual odometry and 0.29 m/s for our solution and standard deviations 0.60 m/s and 0.43 m/s, respectively. Thereby, the

Fig. 10: Collaborative navigation team and equipment in a test campaign.

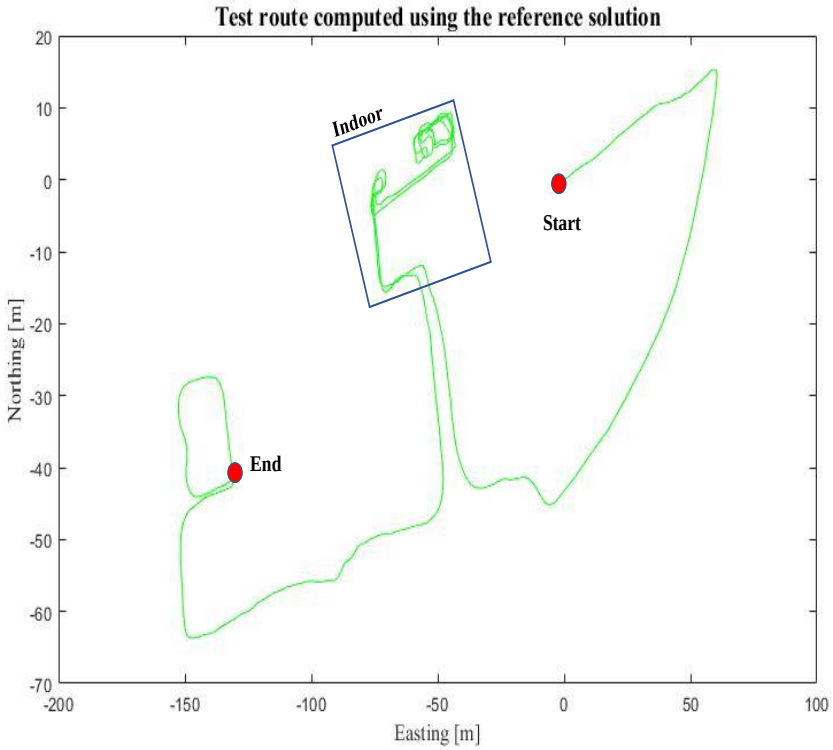

Fig. 11: Test campaign route computed using the reference solution. Start and end positions are shown with red dots and the indoor area with a blue square.

improvement of the visual odometry solution was significant in this dataset.

To compare our VO solution with the state-of-the-art methods we have computed an Relative Pose Error (RPE) translation measure that is mainly used for benchmarking VO and SLAM research. RPE first aligns the two trajectories and then evaluates directly the absolute pose differences [41]. As our paper concentrated on solving the challenges arising from the challenging indoor navigation situations, we computed RPE for the 288 meters long route indoors. When the VO solution was lost for longer time for the reasons discussed in section II we re-initialized the heading using the SPAN system. The average RPE was 8.8 *m* and 8.6 *m* at the end point, resulting in 3% error. We computed also the Distance Root Mean Square (DSRM) measure used in the navigation domain, and it was 4 *m*.

Our VO solution performs provides good performance when compared to conventional VO solutions using RGB-D cameras. Using KITTI dataset [42] achieved 5.8% average translation error. When the environment is more challenging for VO and the motion of the camera unconstrained as in our case, the RPE value increases. In such cases, the error percentages varied between 7% and even 400% for short paths depending on the motion and the environment in [43].

Secondly, we computed a loosely coupled visual and inertial measurement fusion based cooperative navigation solution

Fig. 12: System setup in our collaborative navigation test campaign.

using the Extended Kalman filter discussed in Section III and compared that to the reference trajectory. Figure 15 shows the positioning error for all users in the cooperative setup, User1 being the main user computing the visual odometry solution and getting corrections from the cooperative solution.

On further analysis of the error states of the collaborating users, it is found that the dominant position errors accumulate during the portion of the trajectory related to initial building entry and mirrored in the final building exit. The two factors that make this section of the trajectory particularly challenging are that it is simultaneously the longest portion of the trajectory without the opportunity for the user carried navigation systems to enter a ZUPT state, but also occurs when the users are operating in close proximity in a confined and feature poor stairwell. The consequence of this long ZUPT free trajectory with a near total lack of useful visual odometry information providing the large majority of solution error is that the combined solution is only slightly improved relative to the baseline without the additional visual information in terms of position error.

The root-mean-square (RMS) and maximum errors for the cooperative solution with (VO) and without visual odometry (No VO) for all four users are shown in Table I for the whole navigation path. What's important to note with these data sets is that only user 1 has VO, yet all four of the closely cooperating users see improvements in their peak, and RMS error distributions when the VO is activated on only user 1. This shows the ability of one user's augmentation information to propagate through the network in an a direct way.

TABLE I: RMSE and maximum errors for all users in the cooperative navigation setting with (VO) and without (No VO) visual odometry

	VO	No VO
RMS1	7.2	7.4
RMS2	8.1	8.2
RMS3	8.5	8.8
RMS4	9.3	10.0
Max1	13.4	13.5
Max2	13.6	13.8
Max3	14.2	14.6
Max4	14.3	14.3

V. Conclusions

This paper discussed the development of an improved visual odometry solution, where a small and low-cost RGB-D camera was used for solving the scale issue, a problem arising when using a monocular camera. Because the low-cost camera had inconsistencies in the depth detection, a methods for making the detection more robust were used. The goal of our research was to obtain an accurate infrastructure-free navigation solution for a group of people interacting indoors, therefore our method detected pedestrians from the images for avoiding to track their feature points. Accuracy of the resulting visual odometry solution was significantly improved over the conventional visual odometry solution for the real-world collaborative navigation dataset. However, the main issue in the indoor visual odometry, the lack of useful features, remained at harmful level despite our solution. This led to loosing the visual odometry solution completely at some challenging areas, mostly at staircases where the camera was able to perceive only featureless white wall. Therefore, the fusion of the visual odometry into the collaborative navigation setup improved the positioning accuracy only incrementally.

Collaborative navigation remains a powerful tool for teams when entering environments which are unknown and cannot be prepared for navigation in advance. Therefore, our future research includes development of deep learning methods for improving the feature detection in the challenging and feature poor indoor environment. Then, after solving the remaining vision based issues we will continue in fusing tightly the camera and inertial sensors for the collaborative setting. The tight fusion of the camera and inertial sensors means that the feature detection and motion computation phase incorporate information from the inertial measurement processing and vice versa, resulting in improved complementary error detection and mitigation.

References

[1] L. Ruotsalainen, M. Kirkko-Jaakkola, J. Rantanen, M. Mäkelä, Error modelling for multi-sensor measurements in infrastructure-free indoor navigation, Sensors, 18 (2), 590. 2018.

[2] Ali, A., El-Sheimy, N. Low-Cost MEMS-Based Pedestrian Navigation Technique for GPS-Denied Areas. Journal of Sensors. 2013. DOI:10.1155/2013/197090.

[3] Pires, I. M., Garcia, N. M., Pombo, N., Flrez-Revuelta, F. From Data Acquisition to Data Fusion: A Comprehensive Review and a Roadmap for the Identification of Activities of Daily Living Using Mobile Devices. Sensors (Basel, Switzerland), 16(2), 184. 2016. https://doi.org/10.3390/s16020184

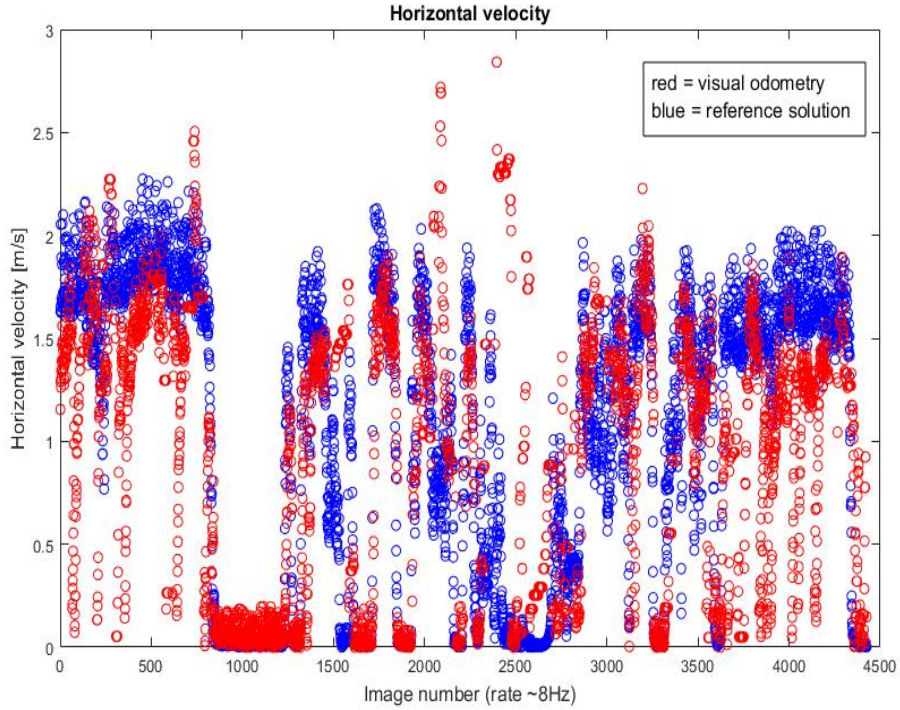

Fig. 13: Speed computed from the conventional visual odometry solution (red points). Speed shown with blue points is obtained from the reference system.

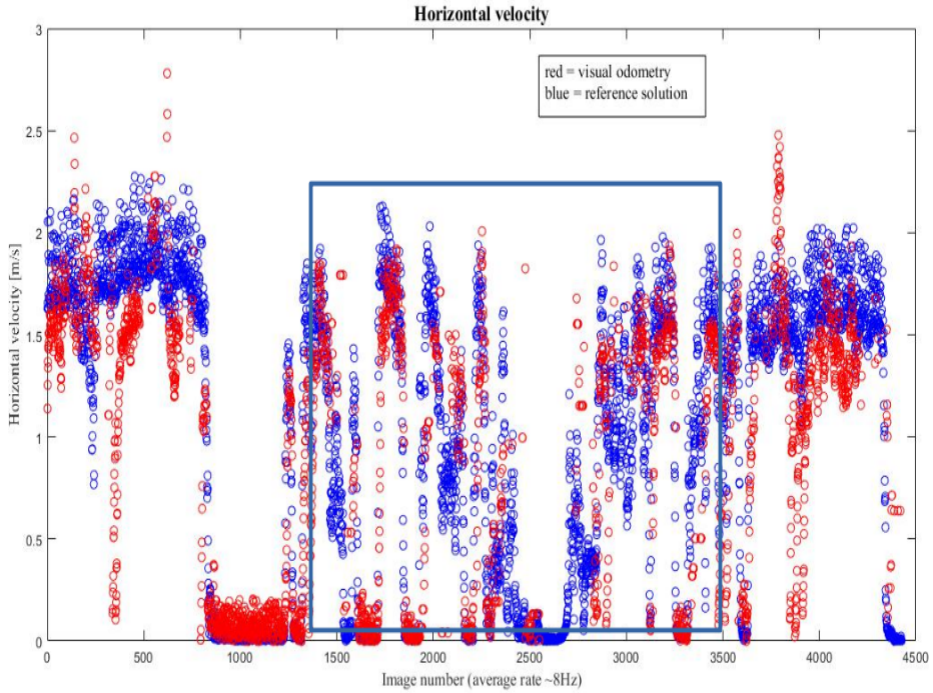

Fig. 14: Speed computed from the visual odometry solution using our method (red points). Speed shown with blue points is obtained from the reference system. Blue rectangle encloses the indoor area.

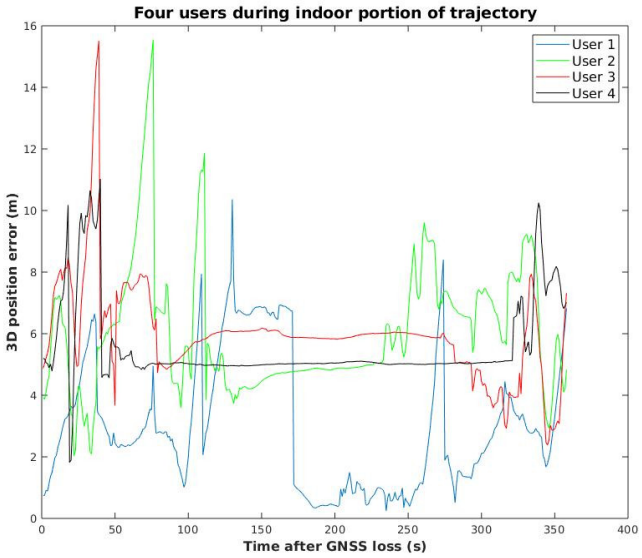

Fig. 15: Positioning errors for all users in our cooperative navigation setup. User number 1 is the user whose position solution is corrected with the cooperative solution.

[4] A. Morrison, L. Ruotsalainen, M. Mäkelä, J. Rantanen, N. Sokolova, Combining visual, pedestrian, and collaborative navigation techniques for team based infrastructure free indoor navigation, Proceedings of the ION GNSS. 2019.

[5] Nilsson, J.-O., Rantakokko, J., Hndel, P., Skog, I., Olsson, M., Hari, K.V.S., "Accurate Indoor Positioning of Firefighters Using Dual Foot-mounted Inertial Sensors and Inter-agent Ranging," Proceedings of IEEE/ION PLANS 2014, Monterey, CA, May 2014, pp. 631-636.

[6] Abdulrahim, K., Moore, T., Hide, C., Hill, C., "Understanding the performance of zero velocity updates in MEMS-based pedestrian navigation", International Journal of Advancements in Technology, 2014

[7] J. Zhang, M. Ren, P. Wang, J. Meng, and Y. Mu, Indoor Localization Based on VIO System and Three-Dimensional Map Matching, Sensors, 20, 2790 (2020)

[8] Bonin-Font, F., Ortiz, A., Oliver, G. Visual Navigation for Mobile Robots: A Survey. Journal of Intelligent and Robotic Systems. 2008. 53(263). DOI:10.1007/s10846-008-9235-4.

[9] N.W. Campbell, M.R. Pout, M.D.J. Priestly, E.L. Dagless, B.T. Thomas, Autonomous road vehicle navigation, Engineering Applications of Artificial Intelligence, Volume 7, Issue 2, 1994, Pages 177-190, https://doi.org/10.1016/0952-1976(94)90022-1.

[10] N. Kronenwett and G. F. Trommer, "Multi Sensor Pedestrian Navigation System for Indoor and Outdoor Environments," 2019 DGON Inertial Sensors and Systems (ISS), Braunschweig, Germany, 2019, pp. 1-21, doi: 10.1109/ISS46986.2019.8943692.

[11] Al-Kaff, A., Martn, D., Garca, F., de la Escalera, A., Armingol, J.M. Survey of computer vision algorithms and applications for unmanned aerial vehicles. Expert Systems with Applications. Volume 92, 2018, Pages 447-463, https://doi.org/10.1016/j.eswa.2017.09.033.

[12] J. Xu, H. Cao, D. Li, K. Huang, C. Qian, L. Shangguan, Z. Yang, Edge Assisted Mobile Semantic Visual SLAM, IEEE Conference on Computer Communications, Toronto, ON, Canada, pp. 1828-1837, 2020

[13] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. Speeded-Up Robust Features (SURF). Computer Vision and Image Understanding, Volume 110, Issue 3, 2008, Pages 346-359, https://doi.org/10.1016/j.cviu.2007.09.014.

[14] Intel, RealSense depth camera, https://www.intelrealsense.com/depth-camera-d435/, last accessed 28.10.2020.

[15] Dawei Leng and W. Sun, "Finding all the solutions of PnP problem," 2009 IEEE International Workshop on Imaging Systems and Techniques, 2009, pp. 348-352, doi: 10.1109/IST.2009.5071663.

[16] Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong, A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry, Artifical Intelligence, vol. 78, no. 1 2, pp. 87 119, 1995.

[17] Leutenegger S, Lynen S, Bosse M, Siegwart R, Furgale P. Keyframe-based visualinertial odometry using nonlinear optimization. The International Journal of Robotics Research. 2015;34(3):314-334. doi:10.1177/0278364914554813

[18] Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, http://www.deeplearningbook.org,2016.

[19] Zhang L., Lin L., Liang X., He K. (2016) Is Faster RCNN Doing Well for Pedestrian Detection?. In: Leibe B., Matas J., Sebe N., Welling M. (eds) Computer Vision ECCV 2016. ECCV 2016. Lecture Notes in Computer Science, vol 9906. Springer, Cham.

[20] Radu Horaud, Miles Hansard, Georgios Evangelidis, Clment Mnier. An Overview of Depth Cameras and Range Scanners Based on Time-of-Flight Technologies. Machine Vision and Applications, Springer Verlag, 2016, 27 (7), pp.1005-1020.

[21] M. Fischler and R. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Communications of the ACM, vol. 6, no. 24, pp. 381395, 1981.

[22] H. Hirschmuller, P.R. Innocent, J.M.Garibaldi. Fast, Unconstrained Camera Motion Estimation from Stereo without Tracking and Robust Statistics, ICARCV, December 2002, Singapore, 2002.

[23] D. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision, vol. 60, no. 2, pp. 91110, 2004.

[24] Rosten E, Drummond T. Machine Learning for High-Speed Corner Detection. European Conference on Computer Vision. Springer-Verlag, 2006:430-443

[25] H Bay, T Tuytelaars, L Gool. Surf: Speeded up robust features. European Conference on Computer Vision. Springer-Verlag, 2006:404-417.6.

[26] Alcantarilla P.F., Bartoli A., Davison A.J. (2012) KAZE Features. In: Fitzgibbon A., Lazebnik S., Perona P., Sato Y., Schmid C. (eds) Computer Vision - ECCV 2012. ECCV 2012. Lecture Notes in Computer Science, vol 7577. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-33783_316.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S.Reed, C.- Y Fu, and A.C. Berg. Ssd: Single shot multibox detector. In European conference on computer vision, pages 2137. Springer, 2016.

[28] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, Yinan Yu, High-level Semantic Feature Detection: A New Perspective for Pedestrian Detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 5182-5191, doi: 10.1109/CVPR.2019.00533.

[29] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.

[30] Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Neural Information Processing Systems (NIPS) (2015)
[31] Girshick, R.: Fast R-CNN. In: IEEE International Conference on Computer Vision (ICCV) (2015)
[32] J. Redmon, S. Divvala, R. GirshickA. Farhadi.You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. Pp.779-788. 2016
[33] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 65176525. IEEE, 2017.
[34] J Redmon, A Farhadi. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017. 7268, 2017. Yolov3: An incremental improvement.
[35] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2016, pp. 1310-1315.
[36] K. Chen, J. D. Deng and Y. Hwang, "A High-Performance Pedestrian Detector and Its Implementation on Embedded Systems for Hypermarket Environment," 2019 International SoC Design Conference (ISOCC), Jeju, Korea (South), 2019, pp. 154-155, doi: 10.1109/ISOCC47750.2019.9027682.
[37] M. Mäkelä, M Kirkko-Jaakkola, J Rantanen, L Ruotsalainen, Proof of Concept Tests on Cooperative Tactical Pedestrian Indoor Navigation, 21st International Conference on Information Fusion (FUSION), 1369-1376, 2018.
[38] S. Kennedy, J. Hamilton and H. Martell, "Architecture and System Performance of SPAN -NovAtel's GPS/INS Solution," 2006 IEEE/ION Position, Location, And Navigation Symposium, Coronado, CA, USA, 2006, pp. 266-, doi: 10.1109/PLANS.2006.1650612.
[39] E. Kaplan and D. Hegarty, Eds., Understanding GPS Principles and Applications. Norwood, MA, USA: Artech House, 2006.
[40] P. Decker, D. W. R. Paulus, T. Feldmann, Dealing with degeneracy in essential matrix estimation, International Conference on Image Processing (ICIP), DOI: 10.1109/ICIP.2008.4712167, October 12-15, 2008, San Diego, California, USA.
[41] Sturm, J., Engelhard, N., Endres, F., Burgard, W., Cremers, D.: A benchmark for the evaluation of RGB-D SLAM systems. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 573580 (2012).
[42] Aladem, M., Rawashdeh, S. A. (2018). Lightweight Visual Odometry for Autonomous Mobile Robots. Sensors (Basel, Switzerland), 18(9), 2837.
[43] Jaramillo, C., Yang, L., Muoz, J.P. et al. Visual odometry with a single-camera stereo omnidirectional system. Machine Vision and Applications 30, 11451155 (2019).

Laura Ruotsalainen is an Associate Professor of Spatiotemporal Data Analysis for Sustainability Science at the Department of Computer Science at the University of Helsinki, Finland. Her current research interests cover development of sophisticated data analysis means, mainly based on deep learning, for robust navigation and situational awareness especially in urban areas and for GNSS interference detection and mitigation. The goal of the research group lead by her and consisting of nine researchers, is to form and utilize spatiotemporal data for the development of sustainable smart cities, mainly via the development of autonomous systems.She is an Associate Editor of the IEEE Intelligent Transport Systems Magazine.

Aiden Morrison received his PhD degree in 2010 from the University of Calgary, where he worked on ionospheric phase scintillation characterization using multi frequency civil GNSS signals. Currently, he works as a senior research scientist at SINTEF. His main research interests are in the areas of GNSS and multi-user collaborative navigation systems and GNSS RFI monitoring and analysis.

Jesperi Rantanen received his M.Sc. (Tech.) degree in Geomatics from Aalto University School of Engineering, Finland, in 2015 and he is currently pursuing a doctoral degree at the University of Tampere. He works as a researcher at the Finnish Geospatial Research Institute focusing on developing adaptive navigation systems.

Maija Mäkelä is a Research Scientist at the Department of Navigation and Positioning in Finnish Geospatial Research Institute (FGI). She received her M.Sc. (Tech.) degree in 2016 from the Department of Mathematics, Tampere University of Technology, where she also worked as a research assistant studying ionospheric correction methods in GNSS. She joined FGI in 2017 and is also a doctoral student in Tampere University. Her current research interests include infrastructure-free navigation methods, cooperative positioning algorithms and machine learning in situational awareness.

Nadezda Sokolova received her PhD degree in 2011 from Norwegian University of Science and Technology (NTNU), where she worked on weak GNSS signal tracking and use of GNSS for precise velocity and acceleration determination. She is now working as a senior research scientist at SINTEF, and adjunct associate professor at the Engineering Cybernetics Department, NTNU, focusing on GNSS integrity and multi-sensor navigation for autonomous system operations.