RESEARCH PROPOSAL

November 2021

Two research proposals are attached to this document. Proposal 1 focuses on an unsupervised autonomous agile obstacle avoidance approach using guided depth maps, kinodynamic planning, and model predictive control. Proposal 2 focuses on the 3D reconstruction of indoor environments using a multi-agent collaborative approach.

Proposal 1 - Semantic Guided Monocular Dynamic ObstacleAvoidance for Aerial System

A certain level of autonomy is imperative to the future of aerial vehicles. The research revolves around a robust real-time aerial system framework that can evade obstacles, plan and navigate in indoor or outdoor environments robustly. The process workflow includes three modules: perception, planning, and control. The perception module entails the depth algorithm and semantic segmentation (task-specific). Semantic segmentation pixelates dynamic classes and cross-task guidance is established which optimizes the depth map to effectually detect near-range obstacles from the UAV's Field of View (FOV). The planning module includes the configuration space formation, cost function, and hybrid A* search for a kinodynamic feasible path. B-Spline optimization and iterative time adjustments find the most feasible path. The control module entails the desired navigation expected from the UAV to evade obstacles and reach the setpoint. Stabilization is achieved using position and altitude PID controllers. The system is planned to test with various operating speeds, and compare with state-of-the-art avoidance algorithms.

Proposal 2 - DCPAR: Dynamic Collaborative 3D Reconstruction by maintaining privacy and safety implications inactive construction sites

Inspection and monitoring systems in an active construction site are a rather expensive and time-consuming process that requires expensive sensors and types of equipment. The system designed should not affect the privacy of the stewards working on the site. Also, an active construction site tends to be most dynamic in terms of the infrastructure and working stewards and moving of components or tools frequently. Considering these issues, a Dynamic Collaborative Perception Aware Reconstruction (DCPAR) technique is proposed which ultimately solves all the common issues in reconstruction and inspection in an effective manner. Current multi-agent SLAM techniques fail to address the real-time scene dynamics. Thus, a distributed, scalable multi-agent system is designed with minimal human intervention leveraging monocular vision and inertial sensors capable of working efficiently in dynamic environments constantly updating the global map keeping in mind the safety and privacy of the stewards.

Semantic Guided Monocular Dynamic Obstacle Avoidance for Aerial System

Jerrin Bright
Researcher, Department of Aerospace Engineering,
Indian Institute of Science, Bangalore, India.

October 2021

Abstract

A certain level of autonomy is imperative to the future of aerial vehicles. The research revolves around a robust real-time aerial system framework that can evade obstacles, plan and navigate in indoor or outdoor environments robustly. The process workflow includes three modules: perception, planning and control. The perception module entails the depth algorithm and semantic segmentation (task specific). Semantic segmentation pixelates dynamic classes and cross-task guidance is established which optimizes the depth map to effectually detect near-range obstacles from the UAV's Field of View (FOV). The planning module includes the configuration space formation, cost function and hybrid A* search for a kinodynamic feasible path. B-Spline optimization and iterative time adjustments find the most feasible path. The control module entails the desired navigation expected from the UAV to evade obstacles and reach the setpoint. Stabilization is achieved using position and altitude PID controllers. The system is planned to test with various operating speeds, and compare with state-of-the-art avoidance algorithms.

Objective

This research was aimed to bring in a computationally efficient and robust aerial navigation system which is capable of running in any environment. Some of the major factors aimed to achieve via this research include Eradicating the need for static world assumptions by pixelating the dynamic classes and the usage of spatial-temporal correspondence to perceive the scene alike humans, Small object avoidance by extrapolating the depth map using semantic guided features, Solving lack of ego-motion resulting in insignificant structural change, Simultaneous kinematics and dynamics constraints resulting in agile motion planning.

Preliminary Literature Review

The sole of the perception module is an extension of SGDepth [8], where they proposed a guided depth segmentation approach derived from the research progression of [3]. Some of the benchmarked depth algorithms evaluated includes monodepth [5] that estimates depth maps from left-right consistency based reconstruction, monodepth2 [6] that works on the downside of [5] thereby reducing visual artifacts and re-projection error, densedepth [1] deals with high resolution depth maps using transfer learning and augmentation, and PyD-net [9] capable to run with CPU in real-time. There has been some interesting research works by [2], [7] and [10] that uses monocular vision to do obstacle avoidance using depth maps and HSV, depth map and collision networks, depth and confidence based pCNN respectively. An kinodynamic path planning [4] version proposed by [11], leverages hybrid A*, B-Spline and iterative time adjustment methods.

Methodology

The proposed system has three major modules - perception, planning and control. In figure 1, the overall architecture of our proposed system is shown.

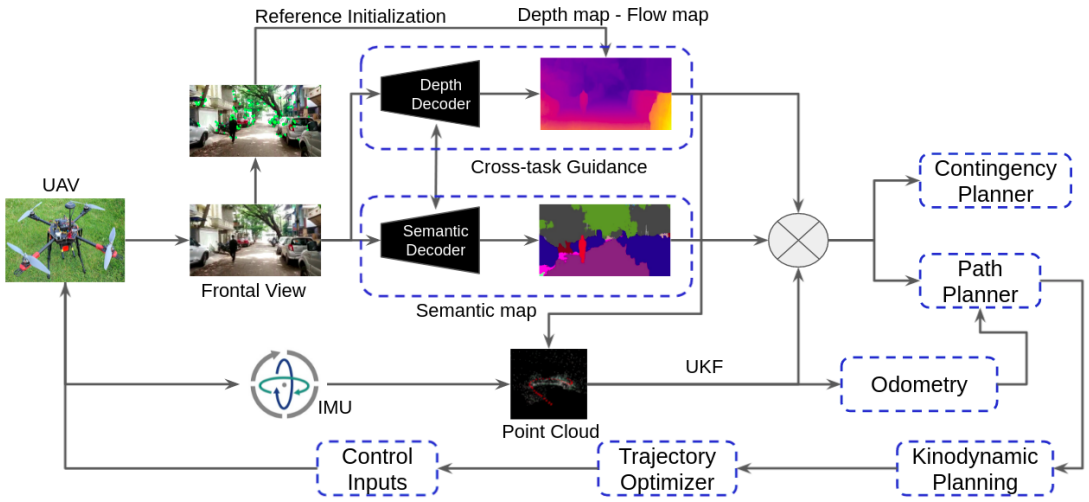

Figure 1: Proposed Architecture

The perception module is split into two sub-modules- depth and semantic map. Real-time Semantic Segmentation is done over the video frames in a supervised manner, which is sent as semantic guidance to the depth encoders, which processes the pixel-wise information, resulting in a depth map (self-supervised). The generated depth maps are processed through spatio-temporal correspondence using matching and optical flow approach to obtain a flow map which are combination of depth information isolated for dynamic classes enabling perception similar to a human.

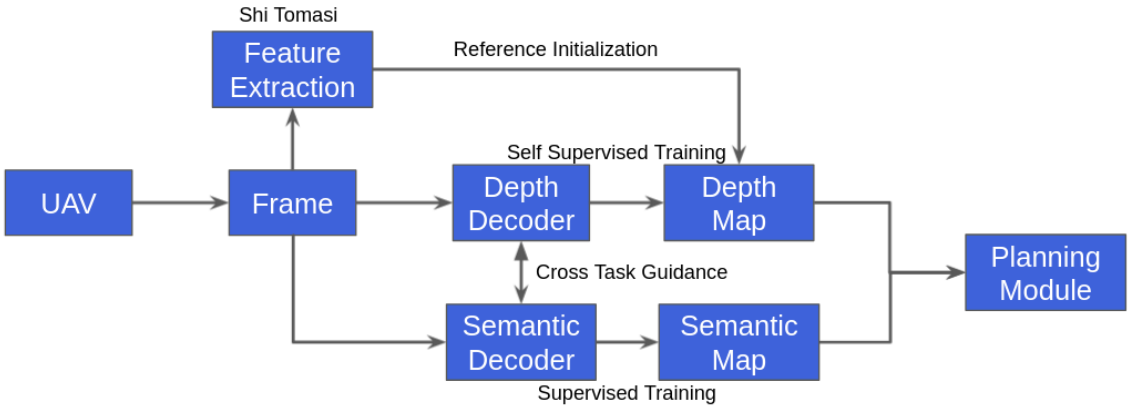

Figure 2: Perception Module

ORB-SLAM based sparse mapping is used to localize the UAV and Unscented Kalman Filters (UKF) is used to fuse the localization data obtained with IMU. ORB features are also used for reference initialisation to determine the absolute scale values of the depth maps obtained in the perception module. Then, point clouds are extrapolated and are stored in a Vector Field Histogram (VFH) on which RRT* path planning algorithm is used to explore the histogram. This in turn finds the most effectual path and way points are generated which is then fed to the trajectory planner to sketch the trajectory for navigation.

Figure 3: Frame - Segmentation - Guided Depth Map

Figure 3 summarizes a clear visualization of the perception module. When obstacles are very close to the drone, the potential of all algorithms deviated catastrophically, but this algorithm uses information from semantic map and guides the depth map, resulting in no deviations.

Figure 4: Depth Map of thin objects (wires)

Figure 4 depicts depth map of thin obstacles like cable wires and branch side which wasn't detected in any of the previous algorithms included [6]. Thus, the semantic guided approach enables detection of thin objects and can be enhanced line detectors algorithms like canny-edge.

Figure 5: Point Cloud from depth map

Scale dependent point clouds were estimated from the semantically guided depth maps obtained in the perception module. The point clouds are assigned voxels in an occupancy grid based on probability using Bayes Rule, thus creating the configuration space of the environment that the UAV perceives. ORB SLAM's localization data is used for global frame initialization of the voxels inside the occupancy grid. Now the map generated will be used by the planning module to find the control points for efficient navigation.

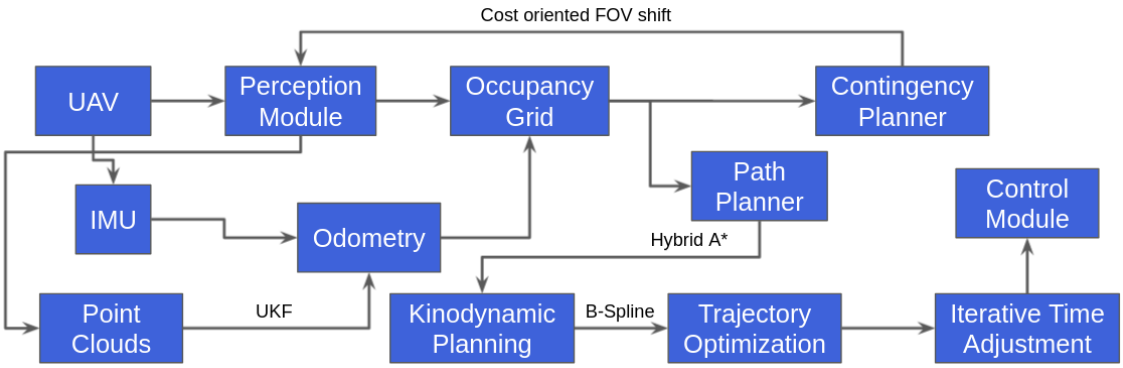

Figure 6: Planning Module

Kinodynamic path planning approach is adapted to navigate from the information perceived from the perception model. B-spline optimization is done to improve the smoothness and clearance of the estimated trajectory. Hybrid State A* algorithm is adapted to search a feasible path between the current pose and desired goal pose as a graph based problem over the generated configuration space. Cost is assigned for each generated trajectory and heuristics enforce faster search.

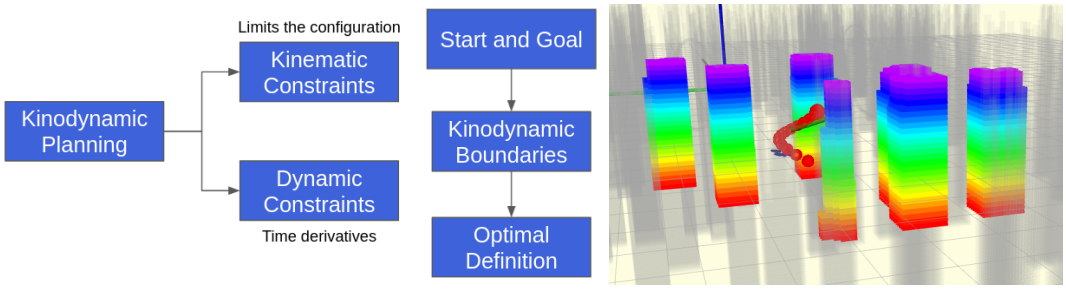

Figure 7: Kinodynamic Path Planning

Deriving the mathematical model of the UAV includes understanding the reference frames, deriving kinematic and dynamic equations of the UAV. The inertial frame, body frame and the vehicle frame upon transformation can be shifted to a common reference frame. Here formulation of equation of motion is done on the body frame considering inertial measurements are in this frame (symmetry of UAV tends to simplify mathematical equations). Euler angles are derived in the vehicular frames and angular rate from body frame.

The control module majorly consist of Model Predictive Controller (MPC) for navigating effectually from the information and commands understood from the perception and the planning modules. Altitude and Position are studied as outputs by the PID controller. Genetic algorithm (optimization and search technique) based tuning is done for the PID controller coefficients.

Preliminary Results

Literature review of stable avoidance algorithms varying from supervised to self-supervised methods using various sensor types are studied. A completely teleoperable quadcopter is made ready (Tarot 680, Jetson Xavier NX, Pixhawk, Logitech C930e). Perception module successfully completed which encompasses depth and semantic map establishing a guided approach. Compared with benchmarked algorithms including [5], [6], [1], [9], etc. and got significantly better results. Currently working on defining mathematical equations for kinodynamic path planning.

Future Works

Some of the most interesting add-ons to the proposed system are:

- Adding optical flow information to the guided depth map thereby making the optical flow map which can not just gives depth information but also gives moving obstacle information.
- The pre-processing module that is the guided depth estimation is planned to be tested with neuromorphic vision sensor (Dynamic and Active-pixel Vision Sensor (DAVIS) Camera). It processes frames at 4000 fps which is much advanced than traditional vision sensors.
- The hybrid A* method is planned to be replaced using a control barrier function which is computationally cheap and helps in reducing latency.
- Multi-trajectory formation using monte-carlo sampling techniques like Metropolis-Hastings Sampling, after which based on cost for control points, navigation trajectory can be chosen.

Conclusion

A robust and agile aerial system framework that can sense and avoid obstacles in a robust manner in terms of computation and accuracy. Texture independent system, lack of ego motion, small object detection and avoidance, fast and agile navigation are some imperative results of this robust system proposed via this research. The algorithm leverages guided segmentation approach to derive steadier depth maps and further extended to detect small objects including thin wires and poles. These small objects are marked as high potential regions and the UAV moves away from its direction (considering the depth information for small objects is not very reliable). As per my knowledge, this extension of segmentation map and the use of guided perception module is the first to ever be implemented in an avoidance system (aerial and ground). Hybrid A* search for the most feasible kinodynamic path enhanced by B-Spline and iterative time adjustment optimizations in the devised configuration space with an adjusted cost map entails the planning module. The system can be efficiently used in a lot of applications including search operations in disaster prone constructions, forests, delivery purposes in high risk environments, etc. to name a few.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *CoRR*, abs/1812.11941, 2018.

[2] Kumar Bipin, Vishakh Duggal, and Madhava Krishna. Autonomous navigation of generic monocular quadcopter in natural environment. *IEEE International Conference on Robotics and Automation*, 2015:1063–1070, 06 2015.

[3] Po-Yi Chen and Yu-Chiang Frank Liu, Alexander H.and Wang. Towards scene understanding: Unsupervised monocular depth estimation with semantic-aware representation. June 2019.

[4] Bruce Donald, Patrick Xavier, John Canny, and John Reif. Kinodynamic motion planning. *J. ACM*, 40(5):1048–1066, November 1993.

[5] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016.

[6] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018.

[7] Kyle Hatch, John Mern, and Mykel J. Kochenderfer. Obstacle avoidance using a monocular camera. *CoRR*, abs/2012.01608, 2020.

[8] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. *CoRR*, abs/2007.06936, 2020.

[9] Matteo Poggi, Filippo Aleotti, Fabio Tosi, and Stefano Mattoccia. Towards real-time unsupervised monocular depth estimation on CPU. *CoRR*, abs/1806.11430, 2018.

[10] Xin Yang, Jingyu Chen, Yuanjie Dang, Hongcheng Luo, and Kwang-Ting. Fast depth prediction and obstacle avoidance on a monocular drone using probabilistic convolutional neural network. *IEEE Transactions on Intelligent Transportation Systems*, 22(1):156–167, 2021.

[11] Boyu Zhou, Chuhao, and Shaojie Shen. Robust and efficient quadrotor trajectory generation for fast autonomous flight. *IEEE Robotics and Automation Letters*, 4(4):3529–3536, 2019.

DCPAR: Dynamic Collaborative 3D Reconstruction by maintaining privacy and safety implications in active construction sites

Jerrin Bright
Researcher, Department of Aerospace Engineering,
Indian Institute of Science, Bangalore, India.

November 2021

Abstract

Inspection and monitoring systems in an active construction site are a rather expensive and time-consuming process that requires expensive sensors and types of equipment. The system designed should not affect the privacy of the stewards working on the site. Also, an active construction site tends to be most dynamic in terms of the infrastructure and working stewards and moving of components or tools frequently. Considering these issues, a Dynamic Collaborative Perception Aware Reconstruction (DCPAR) technique is proposed which ultimately solves all the common issues in reconstruction and inspection in an effective manner. Current multi-agent SLAM techniques fail to address the real-time scene dynamics. Thus, a distributed, scalable multi-agent system is designed with minimal human intervention leveraging monocular vision and inertial sensors capable of working efficiently in dynamic environments constantly updating the global map keeping in mind the safety and privacy of the stewards.

Objective

Some of the major contributions of the proposed system include:

- Multi-agent dynamic real-time reconstruction technique working in a distributed collaborative manner.
- The system doesn't rely on texture variances considering the use of depth maps.
- Use of virtual loop closure for better odometry corrections by reducing re-projection error using key-frame from other agents.
- Communicating via a map module which is a database storing key-frames aids in saving computational cost significantly.
- Dynamic Class Removal (DCR) for privacy and safety implications which also gives more accurate depth maps.

Preliminary Literature Review

Benchmarked multi-system collaborative scene recreation techniques include [2] (two agents focusing on handheld/ wearable devices), [10] (three to twelve agents for handheld devices) resulting in one global map and techniques like [12] (three agents focusing on UAVs) and [6] (four agents on sequences of EUROC dataset) resulting in multiple local maps. There has been a lot of monocular depth algorithm including [3], [4] and [1] to name a few. The depth technique adapted by the proposed system is inspired from [7]. Image in-painting was done using [8] which is a generative image in-painting technique using adversarial edge learning. Algorithms like [11] and [9] work on low overlap feature matching which could be used to attain more robust global maps.

Methodology

In this section, the architecture of the proposed system is explained in detail. Figure 1 visualizes the architecture workflow for two agents. The architecture can also be extended for multiple agents communicating with the 'map module' which will contain the key-frames estimated by the respective agent. Communication via key-frames also aids in less consumption of computation. Firstly key-frames are selected and sent to the decoders for depth estimation. This solves redundancy issues of frames and thereby saves computational cost significantly.

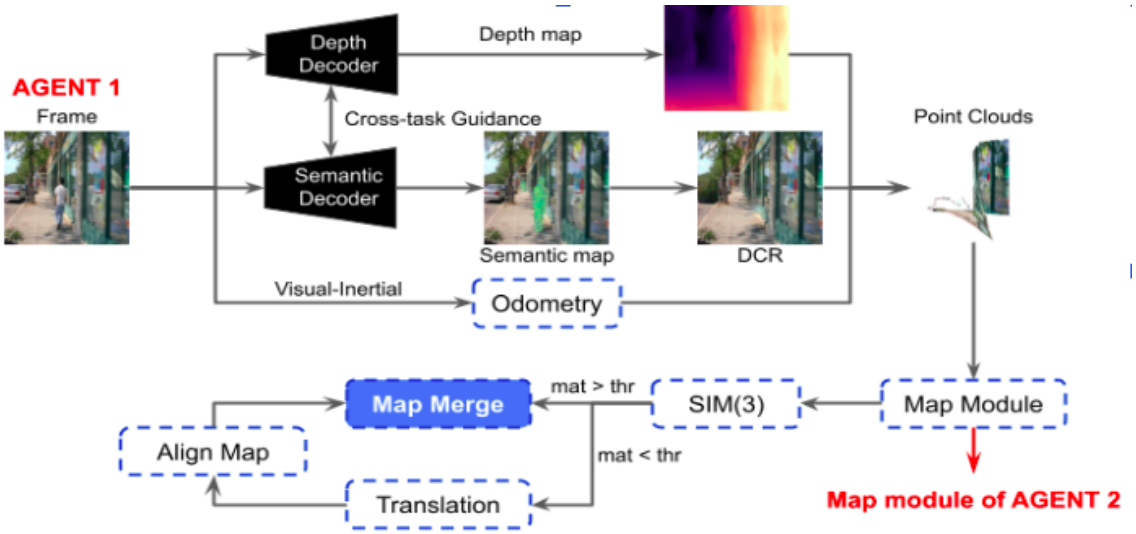

Figure 1: Architecture of the proposed system

Dynamic Class Removal (DCR) is considered to be a segmentation and image in-painting problem. Semantic Segmentation using dense U-Net backbone was leveraged and adversarial edge learning and connecting technique was used for image in-painting. Dilated convolution layers and residual blocks were used for in-painting the 'hole' as a result of semantic identification. In the below figure 2, the DCR algorithm is tested with people to be the only dynamic class to be screened considering the privacy of stewards in construction sites.

Figure 2: Dynamic Class Removal (DCR)

Monocular Depth Estimation is done using the semantic map from the DCR framework which guides and estimates the depth maps in an unsupervised manner by establishing cross-task guidance with semantic segmentation (U-Net). In the below figure 3, the proposed depth algorithm is trained and tested in the frames fed by the DCR framework which was a by-product of the desired screening of dynamic classes depending on the use-case. Once the output key-frame from the DCR is obtained, pre-processing is done. Pre-processing includes depth estimation and conversion of point clouds from the depth image.

Open3D based point clouds are obtained from the depth map. These point clouds are then sent to the map module. Fusion of local maps takes place in respective map modules in a decentralized manner thereby reducing latency. In the map module, virtual loop closure between

the respective agent takes place upon the high matching percentage of key features. This will in turn be used to correct the re-projection error giving more accurate pose estimation resulting in better map fusion.

Figure 3: Pre Processing

The matching probability will be compared with a threshold and if it exceeds, the maps will be merged instantaneously. If the overlap is very less, the virtual loop closure will be triggered and map merge will take place after aligning with odometry information. This way, computation, and efficient mapping are established.

Preliminary Results

Pre-processing including dynamic class in-painting, depth map and point cloud estimation from key-frames has been implemented and tested using real-time data. The key-frames where first sent into the in-painting module to remove dynamic classes thereby ensuring privacy and depth accuracy is maintained. Testing was done using Logitech C930e monocular camera and Jetson Xavier NX development board. The entire pre-processing module runs at 10 key-frames per second when tested in on-board Jetson board. Also, multi-session SLAM techniques including CORB and RTABMAP where exploited and studied extensively to bring forth a robust map merging technique.

Timeline

The research has to start with conducting extensive thorough literature review to identify gaps in knowledge and the scope of the research. Basic experimentation especially with the frame pre-processing has been done. But modifying its performance and reducing computational cost is mandatory requirement. Guided Depth map is a very novel technique and it will approximately take quarter of a year to complete it with good accuracy. Another two to three months to optimize the pre-processing modules with the best algorithms. Then local map has to be made from the attained point cloud and stored in a database followed by map merging. Map merging is a very intricate process which will drastically change the efficiency and computation of the system. There is lot of approaches to do it with different techniques, constraints and assumptions which will take the most number of duration in this research estimated to be about half a year. The proposed architecture can also be enhanced with the future works mentioned in the upcoming sections. Coupling the guided depth map with optical flow information resulting in optical maps capable of comprehending both depth and dynamic scene understanding can be done in a month or two. A total of one and a half years will be perfect for the completion of the proposed system with few additions of the future works.

Conclusion

In this proposal, a novel architecture is proposed for collaborative image reconstruction in continually changing environments (infrastructure) with non-dynamic scene constraints. With very less sensor usage, that is monocular vision sensor and inertial sensor, the proposed system can work effectually in a robust manner. Safety and privacy implications were also taken into account while formulating the architecture. The proposed system was focused on construction sites primarily but

has a wide range of applications especially in the inspection domain and in agile system navigation in unknown cluttered dynamic environments.

Future Works

Some of the immediate future add-ons to this architecture that will ultimately make a significant difference to the performance of this system include:

- Feature matching and registration in low overlap conditions, inspired from the works of [5].
- Unsupervised or Semi-Supervised semantic segmentation approach by coupling in optical flow information like Lucas Kanade Technique from the monocular vision sensor.
- Bring in precise constraints to limit updating the map thereby significantly reducing the computational cost of the architecture.
- Design and implement a more modular system, wherein each module can be interchanged and tested with different algorithms with ease to improve the efficiency of the system.

References

[1] Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. *CoRR*, abs/1812.11941, 2018.

[2] Robert Castle, Georg Klein, and David W. Murray. Video-rate localization in multiple maps for wearable augmented reality. In *2008 12th IEEE International Symposium on Wearable Computers*, pages 15–22, 2008.

[3] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. *CoRR*, abs/1609.03677, 2016.

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018.

[5] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. PREDATOR: registration of 3d point clouds with low overlap. *CoRR*, abs/2011.13005, 2020.

[6] Marco Karrer, Patrik Schmuck, and Margarita Chli. Cvi-slam—collaborative visual-inertial slam. *IEEE Robotics and Automation Letters*, 3(4):2762–2769, 2018.

[7] Marvin Klingner, Jan-Aike Termöhlen, Jonas Mikolajczyk, and Tim Fingscheidt. Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. *CoRR*, abs/2007.06936, 2020.

[8] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z. Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *CoRR*, abs/1901.00212, 2019.

[9] Milos Prokop, Salman Shaikh, and Kyoungsook Kim. Low overlapping point cloud registration using line features detection. *Remote Sensing*, 12:61, 12 2019.

[10] Patrik Schmuck and Margarita Chli. CCM-SLAM: Robust and efficient centralized collaborative monocular simultaneous localization and mapping for robotic teams. In *Journal of Field Robotics (JFR)*, 2018.

[11] John Stechschulte, Nisar Ahmed, and Christoffer Heckman. Robust low-overlap 3-d point cloud registration for outlier rejection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7143–7149, 2019.

[12] Danping Zou and Ping Tan. Coslam: Collaborative visual slam in dynamic environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):354–366, 2013.