Research Proposal on Multi-modal Sensors Fusion

Hejun Wang
Chu Kochen Honor College, Zhejiang University

Summary of the Proposal

Multi-sensor fusion perception algorithms have become a subject of extensive research in recent times. These algorithms hold significant potential in unmanned systems such as autonomous driving technology and robotic systems, as they can effectively compensate for the limitations of a single sensor. The present study aims to address various critical questions in this domain and presents a comprehensive research program. Through this research, we anticipate achieving outcomes that will greatly improve unmanned systems by imparting them with elevated robustness, enabling them to operate effectively across diverse environmental conditions and in instances of individual sensor failure.

Introduction

The primary objective of perception in unmanned systems is to achieve the seamless integration of diverse sensors, ensuring a coherent interpretation of the surrounding environment in a manner that is effective, efficient, and economically viable. Existing research efforts have primarily focused on leveraging sensor technologies such as LiDAR, radar, RGB camera, and IMU to generate high-resolution maps, perform precise semantic segmentation, and enable simultaneous localization. However, when operating under extreme conditions such as adverse weather, unstable information transmission, or unreliable high-level instructions, these systems are susceptible to making errors, thereby limiting their applicability in real-time scenarios that involve human intervention.

To address the aforementioned limitations, researchers have made notable contributions in the field. Researchers like Cai et al. [5], Ma et al. [17], Filos et al. [13], and Wang et al. [22] have proposed approaches that integrate navigation, drivers' intention, and temporal sensing information to make informed decisions. For example, Ma et al. [17] have introduced a sophisticated learning framework that integrates high-level navigation instructions with first-person images to synthesize driving intention. Building upon this framework, Wang et al. [22] have further developed a resilient hierarchical driving model. Nevertheless, it is crucial to acknowledge that these learning frameworks are vulnerable to corruption when confronted with unreliable navigation instructions or blurred images captured by the front-mounted camera. In such instances, the driving intention may veer towards the roadside or buildings, increasing the risk of potential traffic accidents.

The objective of this proposal is to introduce a multi-modal sensor fusion technique aimed at enhancing unmanned systems by providing them with a detailed description of environmental information. This fusion approach is expected to significantly enhance the robustness, reliability, and compatibility of these systems. The proposed methodology involves integrating diverse sensors and developing innovative mathematical representations to accurately interpret the surrounding environment. By achieving this, the proposed work has the potential to greatly advance automated driving technology, human-machine cooperative technology, and other related research areas that are currently in high demand.

The incorporation of additional sensors presents a significant challenge in the field. Surveys conducted in [4, 23, 1, 25] have demonstrated that multi-source and heterogeneous information fusion (MSHIF) effectively addresses perceptual limitations and mitigates random corruption, thereby achieving highly robust, reliable, and adaptive perception. For instance, Bijelic et al. [4] propose a deep multi-modal fusion network that surpasses the limitations of proposal-level fusion and enables enhanced visibility through foggy conditions. However, it is important to note that their method lacks a solid mathematical explanation, which is a common limitation in current deep learning techniques. Additionally, the computational overhead incurred by such fusion approaches should be carefully considered in practical applications.

Furthermore, the development of an innovative mathematical representation for a detailed map of the surrounding environment is of crucial importance. Only by accurately and comprehensively interpreting the environment can the MSHIF framework provide practically valuable results for downstream tasks. The quality and clarity of the environmental interpretation directly impact the effectiveness and applicability of the fusion approach in various real-world applications. Hence, emphasizing the advancement of the mathematical representation is imperative to ensure the practical value of MSHIF in a range of applications.

To summerize, the main interests in this proposal include the followings:

- The proposal focuses on the multi-modal perception under adverse scenarios and the fusion of heterogeneous sensors in real-time unmanned systems such as autonomous driving and robotics.
- There is an emphasis on the development of innovative mathematical techniques that can yield practically valuable results in the domain.
- Solid mathematical explanations and rigorous experimental validations are considered essential for the practical application of the proposed approach in real-world scenarios.

Research Question and Related Works

This proposal concentrates on joint perception using multi-modal sensors that acquire data from various sources. The problem at hand can be further broken down into three distinct subproblems: data representation, data association and simultaneous map construction.

Data from multi-modal sensors can be represented in three main data structures: cubic tensors, point clouds, and sequences of points. Vision data, captured from different perspectives such as RGB cameras and depth cameras, is often represented as a 3-dimensional cubic tensor. On the other hand, radar and LiDAR sensors provide point cloud data. GPS and IMU sensors, on the other hand, generate sequences of points that are discrete in time. However, these structurally and physically different data representations pose challenges for data fusion processes. Researchers in [19, 20, 7, 6] have proposed methods to convert point clouds into sparse high-dimensional tensors and extract features using sparse convolutional networks. By treating radar or LiDAR signals as special types of images, they can be better integrated with camera-acquired images. However, using sparse tensors as a representation for point clouds may not be ideal because it increases space complexity, especially when high resolution is required. In contrast, researchers in [3, 24] have explored the use of transformer mechanisms to process point clouds. This approach takes advantage of the inherent characteristics of point cloud data but also poses new challenges in network design. To solve this problem, it is necessary to develop novel mathematical representations for both point clouds and images and design networks that can effectively extract features from both types of data. This would enable more efficient and accurate fusion of multi-modal sensor data in various applications.

Joint perception of the environment using different sensors has the potential to significantly improve perception robustness[23], although it also presents several challenges. Firstly, the callback frequencies of different sensors may not be equal, which can complicate the synchronization and integration of sensor data. Secondly, reconciling data points perceived through different pathways poses a challenge. Each sensor may have its own biases, noise characteristics, and limitations, necessitating the development of robust algorithms to reconcile and combine the information. Thirdly, it is important to investigate how different sensors can compensate for each other's blind spots. By leveraging the strengths of each sensor modality, it is possible to create a more comprehensive and accurate perception of the environment. Deep learning techniques, along with algorithms proposed in works such as [6, 14, 16, 21, 26], have achieved state-of-the-art performance in object detection, localization, and semantic segmentation. For instance, H. Song et al.[21] propose a robust vision-based relative-localization approach that integrates LiDAR and RGB-Depth camera data, showcasing improved performance. Similarly, X. Zhang et al.[26] successfully identify lane lines clearly from noisy road environments by leveraging the integration of LiDAR and camera data. In this proposal, the applicant aims to utilize a learning-based approach to correlate different types of sensors in real-time, thereby achieving accurate and robust perception of the environment. By leveraging the power of deep learning and sensor fusion techniques, the proposal seeks to advance the state-of-the-art in multi-modal perception and address the challenges associated with integrating diverse sensor data.

Simultaneous location and map construction (SLAM) remains a significant problem in robot autonomous navigation. Efforts have been made to address this challenge using multi-modal sensor systems. For example, M. Bijelic et al. [4] have developed an algorithm that leverages multi-modal sensors to see through dense fog, enabling the identification of objects even in challenging atmospheric conditions. However, while their algorithm successfully detects objects, it lacks the ability to provide further interpretation of the surrounding environment. In this proposal, the applicants aim to use multi-sensor synergistic sensing methods to construct accurate maps that can serve various downstream tasks. The focus of this proposal extends beyond object detection alone and includes the broader goal of map construction. By leveraging the combined information from multiple sensors, the applicants aim to create detailed and precise maps that can be utilized for navigation and other applications. Additionally, to ensure real-time operation, the proposal emphasizes the importance of minimizing computational consumption. Efficient algorithms and optimization strategies will be explored to handle the computational requirements and enable real-time processing.

By addressing the challenges of SLAM and leveraging the synergy of multi-modal sensors, the proposal aims to advance the state-of-the-art in autonomous robot navigation and pave the way for efficient and reliable robotic systems in various domains.

As mentioned above, this proposal highlights the following issues to be resolved:

- Well-designed mathematical representation of differing sensing data format: The proposal recognizes the need for appropriate mathematical representations of the different data formats from various sensors. Developing suitable data structures and representations that can effectively capture the characteristics of each sensor modality is crucial for accurate fusion and analysis.

- Elaborate and accurate point-wise detection or semantic segmentation with a combination of image and point cloud: The proposal acknowledges the challenge of integrating image and point cloud data for tasks such as object detection or semantic segmentation. Developing algorithms that can leverage both modalities and extract detailed information from each source is essential for achieving accurate and reliable results.

- Joint learning of diverse perception to simultaneously construct a detailed high-resolution map: Simultaneously constructing a detailed high-resolution map requires the integration and fusion of information from multiple perception modalities. The proposal emphasizes the need for joint learning approaches that can leverage the strengths of each sensor to construct a comprehensive and accurate map representation.

- Achieving a balance between computational cost and performance: Real-time operation is crucial in many robotic applications, and it is essential to strike a balance between computational cost and performance. The proposal recognizes the need to develop efficient algorithms and optimization techniques to achieve real-time processing capabilities while maintaining high performance and accuracy.

Previous Work of the Applicant

The applicant's research background includes participation in the Zhejiang New Talent Program under the supervision of Prof. Rong Xiong. The applicant's research focused on generating continuous trajectories for driverless vehicles using an imitation learning strategy, following the work of Y. Wang et al.[17][22]. This research provided the applicant with a deep understanding of the challenges and approaches in driverless technology, as well as extensive hands-on experience with deep learning techniques.

The research acknowledged that the hierarchy for trajectory generation introduced by Y. Wang et al.[17][22] has limitations in robustness and adaptability to the environment. The architecture utilizes a generative module to synthesize driving intentions from first-person images and high-level instructions. However, in certain extreme situations, such as when the vehicle can only turn left but is given an instruction to turn right, the synthesized driving intention can lead to incorrect actions and potential traffic accidents.

To address this limitation, the applicant worked on augmenting the capability of the autonomous driving system to conduct semantic reasoning while avoiding excessive computational

overhead or complex integration operations. The research drew inspiration from works such as [10, 2, 15, 11] that introduced additional modules for semantic analysis, at the cost of increased computational requirements.

In the applicant's work, a semantic supervisor was developed based on techniques found in [9, 8, 18]. This supervisor provided knowledge distillation during the training process, enabling human-like semantic reasoning. The effectiveness of this approach was confirmed through applications in CARLA simulation[12]. The research paper, titled *Knowledge Distillation on Driving Intention Generator: Learn Human-like Semantic Reasoning*, was accepted by *Real-time Computing and Robotics 2023* in June.

Additionally, in the summer of 2023, the applicant conducted research on weather-shift effects on LiDAR sensors. The aim was to simulate the impact of adverse weather conditions, such as fog, rain, and snowfall, on LiDAR performance. Unlike traditional approaches that simulate weather based on physical models, the applicant explored data-driven techniques to generate synthetic LiDAR data under adverse weather conditions. This research seeks to provide insights into addressing the challenges posed by adverse weather in LDAR-based autonomous systems.

Schedule and Expectation

The applicant's schedule and expectations are as follows:

- Fall 2024 - Spring 2026: The applicant plans to compare and contrast current high-performing sensor fusion technologies. They aim to provide a mathematically rigorous explanation of each approach and publish a review paper in a major journal or conference.
- Spring 2026 - Winter 2026: Building upon the previous research, the applicant intends to focus on object detection in extreme weather environments. They aim to develop methods and algorithms that can effectively detect objects in challenging weather conditions and publish one top journal or conference paper based on the findings.
- Spring 2027 - Summer 2028: The applicant aims to realize real-time mapping tasks under multi-sensor fusion. By leveraging the research results from previous stages, they will work towards constructing accurate and detailed maps using data from multiple sensors. The expectation is to publish 1-2 top journal or conference papers based on the outcomes of this research.

By following this timeline, the applicant aims to contribute to the field of sensor fusion and autonomous navigation, advancing the understanding and development of technology that can effectively handle extreme weather conditions, improve object detection, and enable real-time mapping capabilities.

References

[1] Ahmed Alkhateeb, Gouranga Charan, Tawfik Osman, Andrew Hredzak, Joao Morais, Umut Demirhan, and Nikhil Srinivas. Deepsense 6g: A large-scale real-world multi-modal sensing and communication dataset. *IEEE Communications Magazine*, pages 1–7, 2023.

[2] Gerrit Bagschik, Till Menzel, and Markus Maurer. Ontology based scene creation for the development of automated vehicles. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1813–1820, 2018.

[3] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1099, June 2022.

[4] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather, 2020.

[5] Peide Cai, Yuxiang Sun, Hengli Wang, and Ming Liu. Vtgnet: A vision-based trajectory generation network for autonomous vehicles in urban environments, 2020.

[6] Luca Caltagirone, Mauro Bellone, Lennart Svensson, and Mattias Wahde. Lidar–camera fusion for road detection using fully convolutional neural networks. *Robotics and Autonomous Systems*, 111:125–131, 2019.

[7] Chi Chen, Ang Jin, Bisheng Yang, Ruiqi Ma, Shangzhe Sun, Zhiye Wang, Zeliang Zong, and Fei Zhang. Dcpld-net: A diffusion coupled convolution neural network for real-time power transmission lines detection from uav-borne lidar data. *International Journal of Applied Earth Observation and Geoinformation*, 112:102960, 2022.

[8] Dian Chen, Brady Zhou, Vladlen Koltun, and Philipp Krähenbühl. Learning by cheating, 2019.

[9] Wuyang Chen, Xinyu Gong, Xianming Liu, Qian Zhang, Yuan Li, and Zhangyang Wang. Fasterseg: Searching for faster real-time semantic segmentation, 2020.

[10] Mark Colley, Benjamin Eder, Jan Ole Rixen, and Enrico Rukzio. Effects of semantic segmentation visualization on trust, situation awareness, and cognitive load in highly automated vehicles. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[11] Deepak Kumar Dewangan and Satya Prakash Sahu. Road detection using semantic segmentation-based convolutional neural network for intelligent vehicle system. In K. Ashoka Reddy, B. Rama Devi, Boby George, and K. Srujan Raju, editors, *Data Engineering and Communication Technology*, pages 629–637, Singapore, 2021. Springer Singapore.

[12] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[13] Angelos Filos, Panagiotis Tigas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. Can autonomous vehicles identify, recover from, and adapt to distribution shifts?, 2020.

[14] Chunrui Han, Jianjian Sun, Zheng Ge, Jinrong Yang, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception, 2023.

[15] Radmila Juric and Olav Madland. Semantic framework for creating an instance of the ioe in urban transport: A study of traffic management with driverless vehicles. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, pages 1–8, 2020.

[16] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V. Le, Alan Yuille, and Mingxing Tan. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17182–17191, June 2022.

[17] Huifang Ma, Yue Wang, Li Tang, Sarath Kodagoda, and Rong Xiong. Towards navigation without precise localization: Weakly supervised learning of goal-directed navigation cost map, 2019.

[18] Seyed-Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant, 2019.

[19] Jaehyun Park, Chansoo Kim, Soyeong Kim, and Kichun Jo. Pcscnet: Fast 3d semantic segmentation of lidar point cloud for autonomous car using point convolution and sparse convolution network. *Expert Systems with Applications*, 212:118815, 2023.

[20] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, 2021.

[21] Haryong Song, Wonsub Choi, and Haedong Kim. Robust vision-based relative-localization approach using an rgb-depth camera and lidar sensor fusion. *IEEE Transactions on Industrial Electronics*, 63(6):3725–3736, 2016.

[22] Yunkai Wang, Dongkun Zhang, Jingke Wang, Zexi Chen, Yuehua Li, Yue Wang, and Rong Xiong. Imitation learning of hierarchical driving model: From continuous intention to continuous trajectory. *IEEE Robotics and Automation Letters*, 6(2):2477–2484, 2021.

[23] Zhangjing Wang, Yu Wu, and Qingqing Niu. Multi-sensor fusion in automated driving: A survey. *IEEE Access*, 8:2847–2868, 2020.

[24] Zhixiang Xue, Xiong Tan, Xuchu Yu, Bing Liu, Anzhu Yu, and Pengqiang Zhang. Deep hierarchical vision transformer for hyperspectral and lidar data classification. *IEEE Transactions on Image Processing*, 31:3095–3110, 2022.

[25] Cheng Zhang, Hai Wang, Yingfeng Cai, Long Chen, Yicheng Li, Miguel Angel Sotelo, and Zhixiong Li. Robust-fusionnet: Deep multimodal sensor fusion for 3-d object detection under severe weather conditions. *IEEE Transactions on Instrumentation and Measurement*, 71:1–13, 2022.

[26] Xinyu Zhang, Zhiwei Li, Xin Gao, Dafeng Jin, and Jun Li. Channel attention in lidar-camera fusion for lane line segmentation. *Pattern Recognition*, 118:108020, 2021.