Proceedings of CONF-FMCE 2026 Symposium: Smart City and Infrastructure Engineering
DOI: 10.54254/2755-2721/2026.CH38284

Vision-Based Autonomous Navigation: Sensors, Algorithms, Challenges and Future Perspectives

Wanyu Li

Qingdao Academy, Qingdao, China
amyli20100212@outlook.com

Abstract. With the rapid development of autonomous transportation and robotics, autonomous navigation has become an important research area. To achieve reliable navigation, advanced sensors and efficient perception algorithms capable of interpreting complex environmental data are required. This paper reviews the development of vision-based autonomous navigation systems, focusing on the sensors, data processing perception systems and current challenges. The characteristics of mainstream sensing approaches, including RGB cameras, infrared cameras and LiDAR, are analyzed through comparisons in their advantages and limitations. Furthermore, major perception algorithms, including visual simultaneous localization and mapping (V-SLAM), point-cloud processing and bird's-eye-view (BEV) perception, are reviewed, putting focus on their processing principles, advantages and computational requirements. Challenges remain in system performance and accuracy, and future development is expected to shift towards improving computational efficiency and enhancing sensing technologies. This paper provides an overview of current progress in vision-based autonomous navigation and insightful directions for achieving more reliable and efficient autonomous navigation systems.

Keywords: Vision-Based Autonomous Navigation, Perception Algorithms, Visual Sensors

1. Introduction

Autonomous navigation has been studied for some time now as an important part of technology for enabling the automated navigation in cars or robots. Their development started in the 15th century when the concepts of systems were only preplanned paths and had no ability to sense the environment. This continued up until the 1960s, when artificial intelligence and advanced sensors were introduced [1], changing the basis of how autonomous systems worked. Methods to improve the reliability of autonomous navigation have been a crucial area of research for different industries due to the shared requirement of a safe and reliable autonomous navigation system. Within autonomous systems, Vision-based sensing plays a significant role in autonomous navigation by providing visual information for tasks such as object recognition and environmental understanding [2]. Thus, the training of artificial intelligence based on visual data has started to show its role in improving reliability.

In recent years, the rising development of artificial intelligence has led to demand for high-quality data in autonomous navigation. Early vision-based navigation started in the early 2000s,

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.CH38284

where algorithms such as early Visual Simultaneous Localization and Mapping (V-SLAM) approaches and visual odometry (VO) were used. These methods obtain a sequence of images from a monocular camera, then process and extract the main features to map the environment or define the position [3].

Since the 2010s, V-SLAM algorithms have evolved rapidly and shifted from feature-based methods to direct methods, using a whole image directly for tracking and mapping to deal with images without features. RGB-D-based approaches grew in popularity with the introduction of low-priced RGB-D cameras. This approach allowed the system to directly procure the 3D structure of the environment as well as its texture information and the scale of the coordinate system [4]. Progressing into the 2020s, data fusion developed, combining sensors like Light Detection and Ranging (LiDAR), cameras, sonars, stereo vision, etc., to utilize the advantages of each. This method when at high level, could also track moving objects and provide a better understanding of the surroundings [2]. During past years, systems have been incorporated with artificial intelligence. The implementation of deep learning and reinforcement learning has addressed the limitations of traditional methods and has been the centre of research due to its robust performance, though the computational costs and requirements need consideration.

Modern autonomous navigation systems rely on a system that integrates sensing hardware and processing algorithms. As shown in Figure 1, environmental information is first collected by sensors and then processed by perception algorithms to estimate vehicle states and understand surrounding objects.

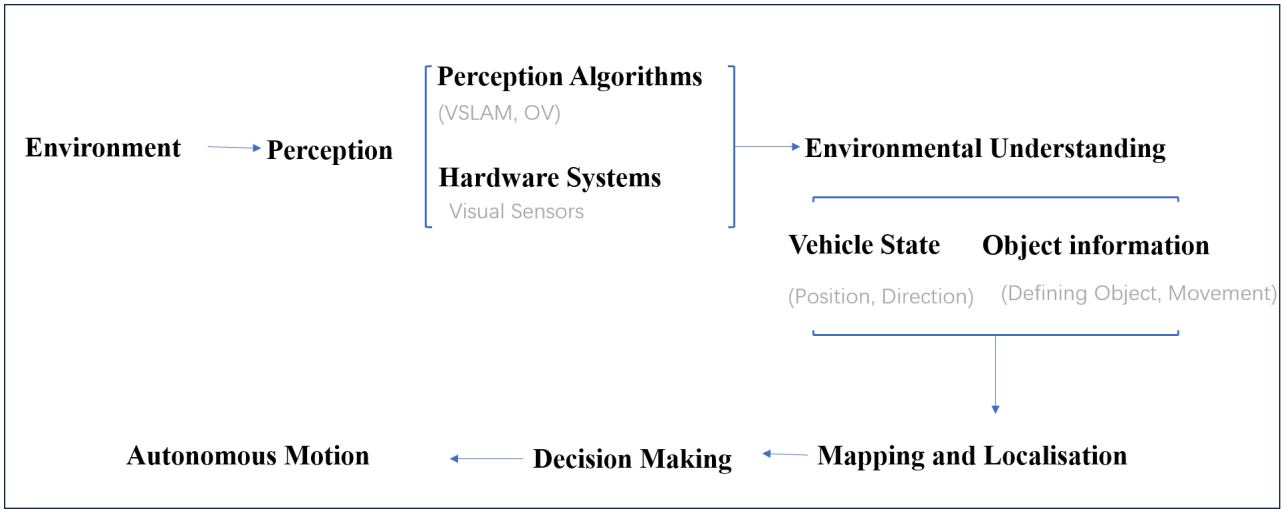

Figure 1. Overview of the framework in autonomous navigation, including perception and environment understanding (picture credit: original)

Although vision-based autonomous system technology has progressed and minimized certain problems, challenges remain in environmental adaptability, sensor reliability, computational efficiency and real-time processing. Hence, this paper will review vision based autonomous system technology, including its sensors and algorithm systems. The paper will then analyse the current challenges and discuss potential future directions, providing insights into the development of more reliable and efficient intelligent navigation systems. The subsequent chapters will discuss visual sensing in autonomous navigation, such as the RGB camera, and complementary sensors such as LiDAR, that are used to overcome the limitations of systems using only vision, the perception algorithms used, then turn to analyse the limitations and prospects, and end with a conclusion.

DOI: 10.54254/2755-2721/2026.CH38284

2. Visual sensing for autonomous navigation

The rapid development of intelligent navigation has meant that visual sensors have become a vital source of information on the surroundings. Current research has focused mainly on improving the performance through the perception capabilities of visual sensors under different conditions, including low-light and adverse weather, and significant progress has been made.

2.1. RGB cameras

To follow the research progress of visual sensors, an understanding of the working principles of basic RGB cameras is needed. An RGB camera is composed of a lens, a colour filter, an image sensor and a signal processor. Visible light is captured through the lens, which focuses it onto a CMOS image sensor (CIS). The image sensor is formed by millions of photosites containing photodiodes, which convert light intensities into electrical signals, then ultimately into digits. Each photosite gives the image a pixel. But before reaching the sensor, the light passes through an infrared cut-off filter (IRCF) and a colour filter array (CFA). The CFA is usually a Bayer filter, with a repeating RGGB pattern. This makes different pixels respond primarily to red, green or blue wavelengths, giving us colour information and not just greyscale values. The image signal processor (ISP) algorithm then uses demosaicing to assign each pixel its colour and reconstruct the RGB information [5].

These RGB cameras have clear advantages: they are low-cost and work similarly to human vision, providing rich and dense contextual information on the environment [5]. It is generally simpler compared to infrared cameras and LiDARs, which will be introduced later, and its early appearance has led to its technical maturity.

Having this in mind, there are also several limitations of the RGB camera. For instance, it cannot measure information on depth so lacks the ability to provide enough 3D image information. It is also especially reliant on lighting conditions, meaning its performance may be poor in the dark or in cloudy or foggy conditions. These bottlenecks led to the exploration of different methods of image capturing such as binocular stereo vision, which uses two cameras to estimate depth from the differences between corresponding images.

2.2. Infrared cameras

Infrared (IR) cameras, as the name suggests, can detect wavelengths that are longer than approximately 700nm and are in the infrared region. Working in basically the same way as RGB cameras, it provides a solution to the limitations of RGB cameras, as they can work in adverse conditions and at night. Certain infrared-based systems can additionally provide depth information used to avoid collisions and map paths. There are three types of commonly used infrared spectra: Near-infrared (NIR), Short-wavelength infrared (SWIR) and Long-wavelength infrared (LWIR), and the three types have their own characteristics.

NIR imaging can enhance visual perception in low-light environments. NIR cameras can also be converted into an active night vision system when combined with external NIR illuminators. SWIR can deal with adverse conditions like fog, as its long wavelengths experience less scattering when propagating through a medium [5]. In contrast, LWIR cameras detect thermal radiation emitted by objects and can therefore provide information related to temperature. This makes them particularly useful for detecting pedestrians at night when visible light is limited [6].

DOI: 10.54254/2755-2721/2026.CH38284

Despite these advantages, infrared cameras also have limitations. Compared with RGB cameras, some infrared cameras have lower resolution, thus providing less detailed texture and colour information. They are also more complex and require specialized, expensive hardware.

2.3. LiDAR

LiDAR is an active sensing technology that obtains information about the target by emitting laser pulses toward the environment. It measures the time taken for reflected light to return to calculate the distance and can also procure information like the object's shape or size. The repeated measurements generate an HD map by capturing dense 3-D point cloud data. However, raw LiDAR measurements cannot be directly used for decision-making and require further processing, including object detection, tracking and semantic understanding [7].

Due to its accurate depth perception, it is complementary to the RGB camera. LiDAR has low resolution and provides less colour information, while the RGB camera cannot provide depth information. Hence on many occasions, the two are combined. The LiDAR has clear benefits like its long detection range and the ability to provide 3-D information [7].

Current LiDAR systems can be broadly classified into time-of-flight (ToF) and frequency-modulated continuous-wave (FMCW) approaches. ToF LiDAR determines distance from the round-trip travel time of emitted laser pulses, while FMCW LiDAR uses the frequency difference between transmitted and reflected signals to determine target distance and velocity [8]. Although ToF remains the main approach in current LiDAR systems due to its relatively simple implementation and lower cost, FMCW LiDAR has attracted increasing research attention as an emerging alternative.

Like other sensors, the LiDAR also has its flaws. It is expensive compared with other traditional visual sensors and produces large amounts of point-cloud data, resulting in high computational requirements and high cost. Like the RGB camera, its performance can be affected by adverse conditions like rain, fog or snow.

Modern research has a growing concern on LiDAR point-cloud processing, the LiDAR perception system, Deep Learning methods that use neural networks to extract features and LiDAR-camera fusion, rather than simply the LiDAR alone.

3. Visual data processing and perception algorithms

Apart from the hardware needed, autonomous navigation systems also require high- performing algorithms to process the raw data obtained in the previous process, therefore, the development of perception algorithms is also essential for mapping and decision-making. With the increasing complexity of sensor data, algorithms have gradually progressed from conventional data processing towards those that combine deep learning and multi-sensors. Modern research focuses on not only improving accuracy but also reducing computational costs. Therefore, this section will review the development of V-SLAM, point-cloud processing and Bird's Eye View (BEV) perception, with attention to their processing approaches, technical advantages, limitations and computational efficiency.

3.1. V-SLAM

Simultaneous Localization and Mapping (SLAM) is composed of two parts: estimating the sensor pose (localization) and building a map of the environment (mapping). V-SLAM specifically refers to SLAM systems that use cameras as data input sources. A traditional V-SLAM system consists of a

DOI: 10.54254/2755-2721/2026.CH38284

front end and a back-end. The front-end extracts and matches visual features between successive frames to estimate the camera motion, while the back-end optimizes the estimated trajectory and map. The estimated camera motion is then used to update the map, and loop closure is employed to correct accumulated drift. In modern systems, deep learning has been introduced into VO for feature extraction, matching and loop closure detection [9].

A representative feature-based V-SLAM system is ORB-SLAM, proposed by Mur-Artal et al., which integrated tracking, local mapping, loop closure detection and optimization framework. By using ORB features, the system achieved accurate localization and enabled real-time performance [10].

3.2. Point-cloud processing

Point-cloud processing is a processing method based on LiDAR sensors. A major task in this process is 3D object detection, for which several methods have been developed. A point cloud is a collection of 3-D points representing the spatial structure of the surrounding environment. In order to generate the point cloud, LiDAR sensors measure the distance and angular information of nearby objects, which can be represented in spherical coordinates and converted into Cartesian coordinates. There are then various different types of approaches for 3D object detection based on how the point-cloud data are represented and processed. In point-based 3D object detection, the object is directly detected using raw points. This process is comprised of point cloud sampling and feature learning. Its limitations are clear: due to its sequential characteristic, the method requires more time and has limited ability in real-time detection. These limitations led to the development of a more manageable grid-based representation.

Grid-based approaches divide the 3D space into regular cells and assign each raw point to a corresponding cell. There are three main types of grid-based representations. The first is voxels, where a voxel is represented by each cell. Raw point cloud data is usually sparse, so most of the voxels end up empty. The non-empty voxels are then chosen for use in feature extraction and multi-view and multi-scale voxels are used to enhance object detection. Another type of representation is pillars, which are basically voxels with unlimited size vertically. The last type is BEV. The different grid representations balance accuracy and computational efficiency. Voxel-based methods preserve more 3D information but require greater computational and memory resources. BEV-based methods are more efficient but lose some 3D information, while pillar-based methods provide a compromise between information preservation and processing efficiency [11].

Introduced in 2018, VoxelNet first proposed a voxel-based feature encoding layer, which transforms sparse point clouds into structured volumetric representations. The voxel feature encoding layer enabled end-to-end learning for 3D object detection and became important foundation for later voxel-based LiDAR perception methods [12].

Despite the improvements in processing efficiency offered by the algorithms, real-time deployment remains challenging as 3D perception requires advanced computational and memory resources. As a result, more recent studies have concentrated on hardware optimization of the algorithms. This is demonstrated in Gao et al.'s research, in which progressing onboard for a quadrotor is proposed and validated for real-time use [13]. They used specialized hardware architectures, enabling real-time 3D perception on platforms with limited resources. This approach enhances the operational efficiency of autonomous systems while also paving the way for broader applications in dynamic environments.

In order to overcome the computational bottlenecks present in challenging 3D perception tasks and ensure reliable performance without the reduction of speed or accuracy, it becomes crucial to

DOI: 10.54254/2755-2721/2026.CH38284

combine algorithmic optimization with hardware acceleration.

3.3. BEV

BEV 3D perception transforms sensor features obtained from perspective-view images or point clouds into a top-view image. BEV has evolved from geometric transformation to deep learning transformation, which can be divided into three approaches: depth-based, Multilayer Perceptron (MLP) -based and transformer-based. For 3D objects, these methods are required to reduce artifacts [14].

The depth-based method estimates depth and lifts 2D image features into 3D space before obtaining BEV and MLP-based learns the transformation using multilayer perceptrons. Cross-attention has been introduced by transformer-based techniques to identify input-dependent correspondences between image features and BEV queries. Compared with MLP-based methods, this allows more expressive feature interactions and has become an important direction in modern BEV perception [14].

BEV is widely adopted in autonomous driving because of its high performance in localization of surrounding objects. For instance, Tesla uses multi-camera BEV networks to map road layouts and objects.

To convert multi-camera image features into BEV representations, transformer-based BEV techniques such as BEVFormer integrated temporal and spatial attention mechanisms. By understanding the connections between grid-shaped BEV queries and image features, BEVFormer was able to achieve high performance in 3D object detection [15].

4. Limitations and prospects

Although vision-based autonomous navigation has undergone substantial development in recent years, parts of it still lack real-world application reliability. The increasing complexity of sensing systems and perception algorithms introduces challenges in processing efficiency, adaptability and system reliability, which can limit their practical application. Thus, further development requires advancements in both sensors and data-processing algorithms, as well as closer integration between hardware and algorithms.

The current challenges of vision-based autonomous navigation are mainly related to real-time performance and insufficient reliability in adverse environments. This is due to the large amount of sensor data algorithms have to handle, increasing the computational demands of advanced perception algorithms and visual sensors. This section will look into the challenges that current systems are facing as well as highlight potential future research directions.

4.1. Processing speed

For applications of autonomous navigation, it is next to always that the systems are required to be able to sense its surroundings in real-time, however, high-resolution cameras and LiDAR sensors generate large amounts of computational data, alongside the fact that algorithms like point-cloud processing and BEV increase the computational complexity. This results in a needed trade-off between accuracy and speed.

As research continues on how to solve this limitation, future systems will likely keep focus on reducing the computational force needed while maintaining accuracy to achieve higher efficiency. While real-time onboard implementation is still a crucial path for autonomous systems, as shown by

DOI: 10.54254/2755-2721/2026.CH38284

point-cloud processing navigation systems installed directly on quadrotors [13], event-based cameras and faster LiDAR technologies like FMCW LiDAR may also be able to offer higher temporal information [5, 8].

4.2. Environment and sensor reliability

The vulnerability of hardware sensors to environmental changes acts as a prominent obstacle to obtaining the raw data information. Sensors such as RGB cameras struggle under low light and adverse weather conditions, while LiDAR sensors are highly likely to be impacted by conditions such as fog or snow. V-SLAM can also fail when visual features are insufficient or when dynamic objects interfere with localization, so no single module can provide reliable perception under all operating conditions.

This implies that in order to reduce this issue, a solution that gathers the advantages and makes the drawbacks negligible is desired, and hence, future research will likely focus on multimodal perception and sensor fusion. Combining modules can enhance the individual advantages and reduce the dependence on a single sensor [5]. Future algorithms may also be able to select or weight sensor information according to environmental conditions, while improved calibration and uncertainty estimation could further enhance system reliability.

5. Conclusion

In conclusion, this paper has reviewed the development of vision-based autonomous navigation systems, focusing on the progression from sensors to data processing and perception algorithms. The capabilities and limitations of major sensing methods, including RGB cameras, infrared cameras and LiDAR were examined, showing that different sensors provide complementary information for the environment. The main perception approaches, such as V-SLAM, point-cloud processing and BEV 3D perception, along with their working principles, advantages and drawbacks are then discussed, highlighting the problem regarding reliable real-time deployment, especially when it comes to computational requirements and efficiency. Future development will likely rely on more efficient algorithms and advanced sensor fusion to minimize the limitations found. This review provides an overview of current progress in vision-based autonomous navigation and identifies key areas for future research necessary to achieve more reliable autonomous systems.

References

[1] Stanchev, P., & Geske, J. (2016). Autonomous cars. History. State of art. Research problems. *Communications in Computer and Information Science, 601*.

[2] Kolar, P., Benavidez, P., & Jamshidi, M. (2020). Survey of datafusion techniques for laser and vision based sensor integration for autonomous navigation. *Sensors, 20*, 2180.

[3] Yasuda, Y. D. V., Martins, L. E. G., & Cappabianco, F. A. M. (2020). Autonomous visual navigation for mobile robots: A systematic literature review. *ACM Computing Surveys, 53*(1), 1-34.

[4] Taketomi, T., Uchiyama, H., & Ikeda, S. (2017). Visual SLAM algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications, 9*, 16.

[5] Li, Y., Moreau, J., & Ibanez-Guzman, J. (2023). Emergent visual sensors for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems, 24*(5), 4716-4737.

[6] Miethig, B., Liu, A., Habibi, S., & Mohrenschildt, M. v. (2019). Leveraging thermal imaging for autonomous driving. *2019 IEEE Transportation Electrification Conference and Expo*, 1-5. Detroit, MI, USA.

[7] Li, Y., & Ibanez-Guzman, J. (2020). Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine, 37*(4), 50-61.

DOI: 10.54254/2755-2721/2026.CH38284

[8] Liang, D., Zhang, C., Zhang, P., et al. (2024). Evolution of laser technology for automotive LiDAR, an industrial viewpoint. *Nature Communications, 15*, 7660.
[9] Kazerouni, I. A., Fitzgerald, L., Dooly, G., & Toal, D. (2022). A survey of state-of-the-art on visual SLAM. *Expert Systems with Applications, 205*, 117734.
[10] Mur-Artal, R., Montiel, J. M. M., & Tardós, J. D. (2015). ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics, 31*(5), 1147-1163.
[11] Mao, J., Shi, S., Wang, X., & Li, H. (2023). 3D object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision, 131*, 1909–1963.
[12] Zhou, Y., & Tuzel, O. (2018). VoxelNet: End-to-end learning for point cloud based 3D object detection. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4490–4499.
[13] Gao, F., Wu, W., Gao, W., & Shen, S. (2019). Flying on point clouds: Online trajectory generation and autonomous navigation for quadrotors in cluttered environments. *Journal of Field Robotics, 36*, 710–733.
[14] Ma, Y., Wang, T., Bai, X., Yang, H., Hou, Y., Wang, Y., Qiao, Y., Yang, R., & Zhu, X. (2024). Vision-centric BEV perception: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 10978–10997.
[15] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., & Dai, J. (2025). BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(3), 2020–2036.