Proceedings of the 2023 International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/28/20230140

Applications of visual perception techniques using neural networks in autonomous driving

Hu Kangye

School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai, China

asuna_kirito@sjtu.edu.cn

Abstract. The perception system and the decision system are important components of a complete autonomous driving vehicle. The perception system can help the decision system to obtain the necessary information of external environment and vehicle status. The traditional perception system mainly relies on the on-board radar. But in recent years, vision-based perception techniques have become a hot research topic. Meanwhile, thanks to the excellent performance of neural networks in processing image data, the processing algorithms for visual perception images have also made great progress. Visual perception techniques can not only acquire more information, but also is more cost effective and easier to install. This paper provides an overview of the more mature and promising visual perception techniques, including their principles and data processing algorithms, in terms of acquiring 2D image data and 3D depth information. For acquiring 2D image data, this paper introduces the principle of event camera and reviews the current progress on the event camera. Regarding the acquisition of 3D depth information, three techniques are introduced, namely binocular stereo-vision, time of flight (TOF), and structured light. Their performance when combined with neural networks for autonomous driving applications is also reviewed. Finally, this paper lists the current dilemmas faced by the above 2D and 3D imaging techniques and the possible solutions.

Keywords: autonomous driving, visual perception, neural networks, event camera.

1. Introduction

Since the 1980s, autonomous driving cars have received increasing attention, and many automobile companies, universities, and research institutes have been vigorously promoting research on autonomous driving technology [1]. Thanks to the continuous advances in sensor technology and computer processor, many algorithms that were once limited by hardware capability are now emerging. Although many of them have not shown good performance at this stage, they have certainly provided new ideas for research in autonomous driving cars.

A perception system and a decision system are necessary for an autonomous driving system. The perception system is responsible for information acquisition using vehicle sensors, including information about the external environment as well as the vehicle status information. Due to the recent excellent performance of deep convolutional neural networks (DCNN) in processing image data, there has been an increasing interest in using in-vehicle cameras instead of light detection and ranging (LiDAR) for 3D visual perception. Even autonomous driving systems with vision-only input are somewhat feasible. In addition, for car navigation, localization is extremely important. Using visual

DOI: 10.54254/2755-2721/28/20230140

perception not only allows more information to be gathered, but is also very cheap and lightweight [2]. The camera sensors currently used for visual perception include monocular RGB cameras, event cameras, binocular RGB cameras, TOF cameras, and structured light cameras. After the camera sensors have acquired 2D image data and depth information around the vehicle, the visual odometer performs data processing and estimates the motion of vehicle to enable localization.

The decision system needs to process the information collected by the perception system and control the vehicle based on this information to accomplish the given driving task. Specifically, the decision system can be divided into four parts: the first part is responsible for global route planning, the second part is responsible for determining the vehicle's motion specification based on destination information and traffic rules, the third part is responsible for motion planning, and the fourth part is responsible for correcting errors based on feedback information to achieve closed-loop control of the vehicles [3].

Focusing on visual-based perception systems for autonomous driving cars, this paper provides an overview of several novel and promising approaches for visual perception and how they use neural networks to process image data. The rest of the paper will separately: analyze the differences between using traditional RGB cameras and using event cameras to acquire 2D image data; outline several mature methods for acquiring depth information and provide a horizontal comparison of these methods; and analyze the advantages and challenges of neural-network-based visual perception techniques applied to autonomous driving systems in the context of the previous sections of the paper.

2. Visual perception

In an autonomous driving system, the visual perception system needs to acquire both environmental information and vehicle status information. Among them, the tasks for acquiring environmental information include static obstacle identification and distance measurement, dynamic obstacle identification and motion trend prediction, identification of traffic signals (traffic lights, traffic signs), vehicle positioning, etc. The commonly used sensors include RGB-D cameras, LiDAR, GPS, etc. Collecting vehicle status information includes detecting driving speed, acceleration, steering angle, etc. Commonly used sensors include odometers, inertial measurement units, etc.

2.1. Acquisition of 2D image data

2D image data can assist perception systems in identifying important information, including obstacles, pedestrians, traffic signs, etc. The conventional method of obtaining 2D image data is through RGB cameras, which have a wide range of applications, mature data processing algorithms, and new potential with the help of DCNN. In spite of this, there are still many shortcomings in the field of autonomous driving.

2.1.1. RGB camera. RGB cameras capture the average grey value of each pixel point during the exposure time, but this method is limited by the exposure time and frame rate of the camera, resulting in low temporal resolution. This can make it challenging to obtain stable and clear images when the vehicle is moving at high speeds. The use of high-speed cameras could potentially solve this issue, but the resulting large amount of data would put a significant computing burden on the local computer. Additionally, RGB cameras have poor dynamic range, which results in poor quality 2D image data in low light and strong light environments. This can be fatal in certain driving scenarios, such as driving at night or facing the sun. Autonomous driving systems only require relevant information from specific pixel points, while information from other pixels adds an unnecessary computational burden to the local processor.

2.1.2. Event-based camera. The concept of event cameras was proposed as a solution to the many problems faced by RGB cameras and has recently become a popular research direction. Although event cameras were first developed in laboratories in the 1990s, the first commercially available time cameras were not introduced until 2008. An "event" in this context refers to a change in the luminance value of a single pixel that exceeds a certain threshold. Event cameras use a new information acquisition

DOI: 10.54254/2755-2721/28/20230140

paradigm, where they record the light intensity change of each pixel at each moment. When the change in pixel light intensity exceeds the threshold, the camera records and emits the event asynchronously through a quaternion array $e_m = (x_m, y_m, t_m, p_m)$, where x, y, t, and p respectively record the location (x, y), time t, and polarity p (+1 for luminance increase, -1 for luminance decrease) of the event. An example of the output of the event camera is shown in figure 1. The event camera is a bio-inspired technology, inspired by the biological visual pathways that are more sensitive to dynamic visual information.

(a) (b)

Figure 1. Examples of the output of the event camera. (a) is the original scene picture and (b) is the semi-dense depth map [4].

Event cameras have the following advantages: (1) High temporal resolution: Because it is extremely fast to just record the luminance change value, event cameras can achieve microsecond level resolution; (2) Low latency: Event cameras record and send out event information asynchronously and independently, as soon as the luminance change of each pixel point is detected. Unlike RGB cameras, event cameras do not have to wait for all pixel points to finish recording synchronously. This results in sub-millisecond latency in practical applications; (3) Low power consumption: Since event cameras only transmit the pixel point where the event occurs, rather than transmitting large amounts of redundant pixel points simultaneously, they typically consume only 10mW of power; (4) High dynamic range: Event cameras take a logarithmic view of luminance changes, so they can adapt to both strong and low light environments [5].

Although event cameras have many advantages mentioned above, their huge volume of event data and novel data formats cannot be processed by traditional CV algorithms. To cope with this challenge, a relatively novel approach is to process the event camera data by spiking neural networks (SNN). SNN combined with reinforcement learning (RL) is also a promising solution for robot obstacle avoidance: Luca et al performed simulation tests of UAV obstacle avoidance on UE4, converting RGB data to dynamic vision sensor (DVS) data as input through v2e toolbox [6]. They use SNN to process the data and then train DDQN(double deep q-learning network) to control one UAV for obstacle avoidance [7]. This scheme has higher accuracy and much lower energy consumption compared to using CNN to train RL. Using SNN to process event camera data can also be used to perform image recognition: Viale et al used the SNN model to finish the task of recognizing cars with the help of the Intel Loihi Neuromorphic Research Chip [8]. The model has an accuracy of 83% with only 0.72ms latency and 310mW power consumption.

Another drawback of event cameras is their vulnerability to noise. To ensure that the event camera does not miss any event, the threshold of the brightness change that triggers an event is often set low. This results in that even slight photon perturbations can be captured by the event camera, thus affecting the accuracy of the event information. Some scholars have tried to solve this problem and have made notable progress [9,10].

In summary, the use of event cameras and the corresponding methods for processing event-based data are becoming more sophisticated today. And the exploration of applying event cameras to the field

DOI: 10.54254/2755-2721/28/20230140

of autonomous driving is increasing: Gehrig et al provided datasets collected using event cameras in driving scenarios with various lighting conditions [11]. Viale et al used SNN to process event-based data and implemented vehicle recognition [8]. Zhu et al improved the event-to-frame conversion method and feature extraction network, aiming at improving the speed and accuracy of pedestrian recognition [12]. Numerous studies have demonstrated the feasibility and great potential of using event cameras instead of RGB cameras to acquire 2D images in autonomous driving.

2.2. Acquisition of depth information

In addition to 2D images, the perception system of autonomous driving cars also needs to acquire depth information and construct 3D scenes. Currently, the more mature depth cameras that can acquire depth information are mainly binocular stereo-vision cameras, TOF cameras, and structured light cameras.

2.2.1. Binocular stereo-vision camera. To acquire depth information, the binocular stereo-vision camera system needs one RGB cameras on the left and another one on the right. By analyzing the frames and combining the known parameters of two cameras, the binocular stereo-vision camera completes stereo matching and obtains the disparity map. Then the depth map is calculated according to trigonometric parallax. Specifically, the whole process consists of the following four steps: (1) camera calibration: internal calibration of a single camera (focal length, distortion coefficient, etc.) and external calibration of the binocular camera (to obtain the conversion relationship of two camera coordinate systems); (2) stereo correction: according to the results of camera calibration, the two original images obtained from two cameras are corrected so that the two images are in the same plane and parallel to each other; (3) stereo matching: according to the corrected image, the position correspondence between each pixel point on one image and the corresponding pixel point on the other image is obtained, and the disparity map is obtained; (4) depth calculation: according to the disparity map, the depth map is calculated. The mechanism of the binocular stereo-vision is shown in figure 2.

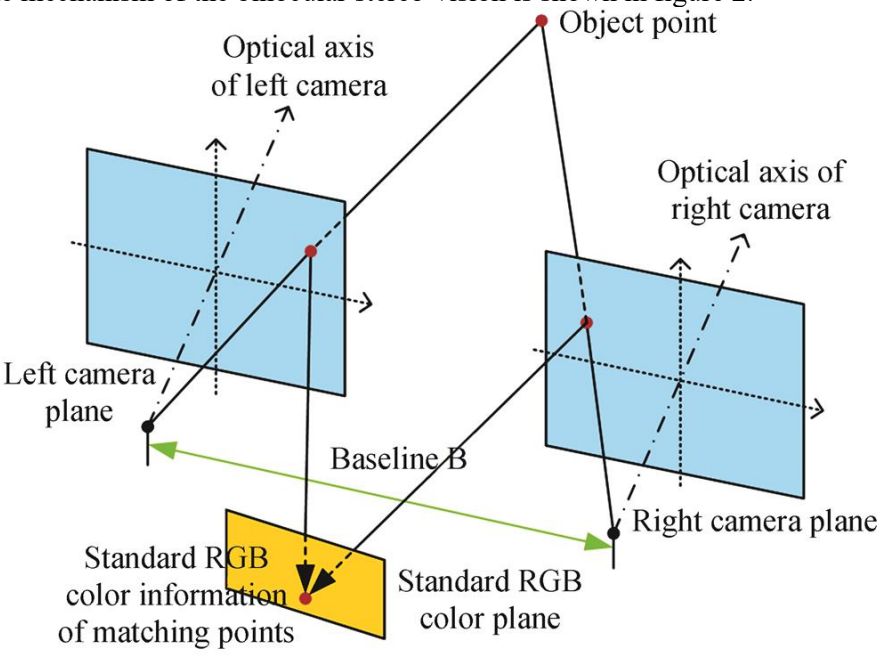


Figure 2. Mechanism of the binocular stereo-vision camera [13].

For a binocular stereo vision camera, the stereo matching algorithm can be the most important. A traditional stereo matching algorithm can be divided into global constraint-based methods and local constraint-based methods. The Traditional stereo matching algorithm has four steps: matching cost calculation, cost aggregation, disparity calculation, and disparity optimization. This paper will focus on the stereo matching algorithms based on the end-to-end neural networks.

DOI: 10.54254/2755-2721/28/20230140

Using a direct end-to-end model is the characteristic of early end-to-end stereo matching algorithms. The model takes the raw images from two cameras as input, and output the disparity maps directly. These models ran faster because they did not require an exact feature matching module. However, these models require a large dataset. It is extremely expensive and cumbersome to acquire such a dataset and calibrate the dataset with accurate depth measurements. Recent end-to-end matching algorithms use different modules to implement different steps of traditional stereo matching algorithms. These models are end-to-end models while reducing the requirements for the dataset [14].

Numerous end-to-end deep learning models can be classified into 2D convolution-based networks and 3D convolution-based networks. A reliable method to evaluate the performance of these models is to compare their performance in the KITTI 2015 dataset [15], with the evaluation metrics D1-bg, D1-fg, D1-all, and runtime given in the KITTI leaderboard. In this paper, distinctive models are selected, whose corresponding papers have been published [16]. The performance of each model is shown in Table 1.

Table 1. Comparison of different models.

Model	Type	D1-bg	D1-fg	D1-all	Runtime
HD^3-Stereo	2D	1.70	3.63	2.02	0.14
DispNet-CSS	2D	1.92	3.32	2.16	0.25
DispNet-C	2D	4.32	4.41	4.34	0.06
CSPN	3D	1.51	2.88	1.74	1.0
GWC-Net	3D	1.74	3.93	2.11	0.32
HSM	3D	1.80	3.85	2.14	0.14
StereoNet	3D	4.30	7.45	4.83	0.02

DispNet-C and StereoNet models run extremely fast and are suitable for applications requiring high real-time performance, but at the cost of their low D1-all scores; CSPN model is one of the most accurate models among all published models; HD^3 and HSM models achieve a good balance of D1-all scores and running time, with good overall performance.

Some of the challenges faced by binocular vision cameras include (1) difficulty in camera parameter calibration in outdoor settings [14]；(2) binocular stereo-vision cameras are essentially RGB cameras, which are very sensitive to ambient lighting and perform poorly in bright and low light environments; (3) large data processing operations and high latency, especially for stereo matching algorithms based on end-to-end deep learning; (4) poor spatial resolution and depth resolution; (5) poor recognition of objects lacking texture because of the difficulties of obtaining image features for matching; (6) The measurement accuracy is related to baseline length between two cameras. The accuracy is proportional to the baseline length and inversely proportional to the measurement distance. Yang et al proposed a hierarchical technique to generate on-demand disparity maps by setting an upper resolution limit on the intermediate results [17]. This method achieves a balance between high spatial resolution and high depth resolution with low latency when generating disparity maps for high-resolution image inputs

2.2.2. TOF camera. Time-of-flight (TOF) method acquires depth information by continuously sending light pulses from an infrared emitter to a target object. The light reflected from the target object is focused by an optical lens and imaged on the sensor. Then the distance of the target object is derived according to the round-trip time of the light. Compared with binocular stereo vision cameras that use passive light detection, TOF uses active light detection; compared with 3D laser sensors that scan point by point, TOF can get depth information of the whole image at the same time. The mechanism of TOF method is shown in figure 3.

DOI: 10.54254/2755-2721/28/20230140

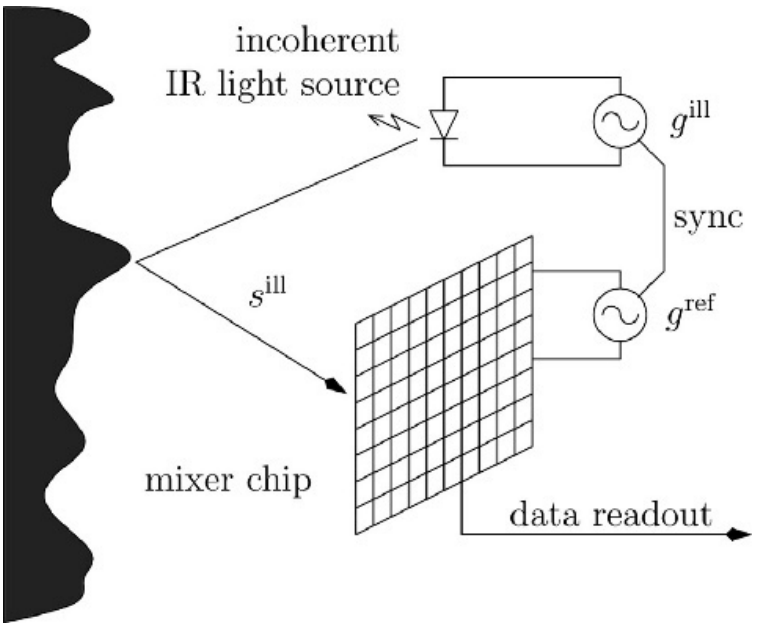


Figure 3. TOF Schematic diagram [18].

Time-of-flight methods can be direct(dTOF) or indirect(iTOF). dTOF measures the round-trip time of the light pulses, which are directly emitted to the target. iTOF emits a continuous modulated wave to the target and calculates the phase difference between emitted and received light. Then the distance to the object is derived according to the phase difference.

The measurement principle of dTOF is much simpler, but its physical hardware requirements are extremely high. On one hand, high temporal measurement accuracy is required because of the extremely short round-trip time of the light pulse. Although this makes it more difficult for using dTOF in close-range measurements (because the short optical pulse round-trip time severely affects the measurement accuracy), dTOF performs well for long-range measurements because its accuracy does not degrade as the measurement distance increases. This feature reflects the advantages of dTOF applications in autonomous driving. On the other hand, dTOF requires high-frequency and high-intensity pulses to be generated at the transmitter, which requires the use of a single-electron avalanche diode (SPAD) as an optical pulse detector [19]. Although the high manufacturing difficulty of SPAD leads to fewer mature commercial dTOF cameras, the irradiance power of light pulses is much higher than that of ambient light, thus improving the immunity of dTOF to ambient light and enabling dTOF to be used outdoors. In the field of autonomous driving, Niskanen et al tried to use a 2D TOF profiler to generate 3D point cloud maps for vehicles and achieved high accuracy [20]. However, this method has a low frame rate and can only be used for vehicles traveling at low speeds. This method is also susceptible to weather conditions.

The emitter of iTOF is usually an LED or light-emitting diode [19]. Compared with dTOF, iTOF is susceptible to the interference of infrared light in ambient light and thus is basically used in indoor scenes. A relatively mature commercial iTOF camera is the Kinect camera, and Michal et al tested the performance of Microsoft's Azure Kinect camera released in 2019 and concluded that this kind of camera is almost impossible to use outdoors [21]. This shows that iTOF performs poorly outdoors. The main reasons for this are the high ambient light brightness that causes iTOF pixels to saturate and the noise in the ambient light that affects the performance of the iTOF camera. To solve this problem, the following methods can be adapted: increase the peak power of the emitter to reduce integration time; use a global shutter to prevent the effect of ambient light during array readout; divide the field of view into multiple integration regions and then integrate region by region to shorten the integration time [22]. Miller et al provided a method for tuning and calibrating the iTOF camera by adjusting the integration time, modulation frequency, and offset parameters for the Sentis 3D-M420 ToF camera [23].

DOI: 10.54254/2755-2721/28/20230140

A problem faced by both dTOF and iTOF cameras is multipath mitigation: the light pulses or continuously modulated waves received by the photodetector come not only from the reflection of the target object (direct path), but also from the reflection of other objects (indirect path). For dTOF, the photodetector first receives the signal from the direct path, and the challenge is to distinguish the signal of the direct path from the signal of the indirect path that arrives after a short time interval; for iTOF, the modulated waves from the direct and indirect paths are superimposed, affecting the judgment of the true phase difference. Cyrus et al compiled several feasible solutions for the multipath mitigation problem [22].

Finally, Table 2 summarizes and compares the features of dTOF and iTOF.

Table 2. Comparison of dTOF and iTOF.

Metric	dTOF	iTOF
Imaging Principle	Measure time difference	Judge phase difference
Dynamic range	Good	poor
Imaging Accuracy	not decrease with increasing distance	decrease with increasing distance
Resolution	Low	High
Anti-interference	Strong	Weak
Imaging frame rate	High	Low
Application	Outdoor	Indoor

2.2.3. Structured light camera. Structured light 3D imaging technology can be seen as an improved form of stereo vision method. The difference is that: binocular stereo vision cameras perform passive detection through two RGB cameras, which are strongly influenced by ambient light; while structured light cameras are active detection, using near-infrared lasers to project images with certain structural features onto the surface of the target object. Then the images are distorted by the stereoscopic shape of the object and returned. According to the parameters and positions of the laser and the camera, the computational unit converts the structural changes of the returned image into depth information. The mechanism of the structured light camera is shown in figure 4.

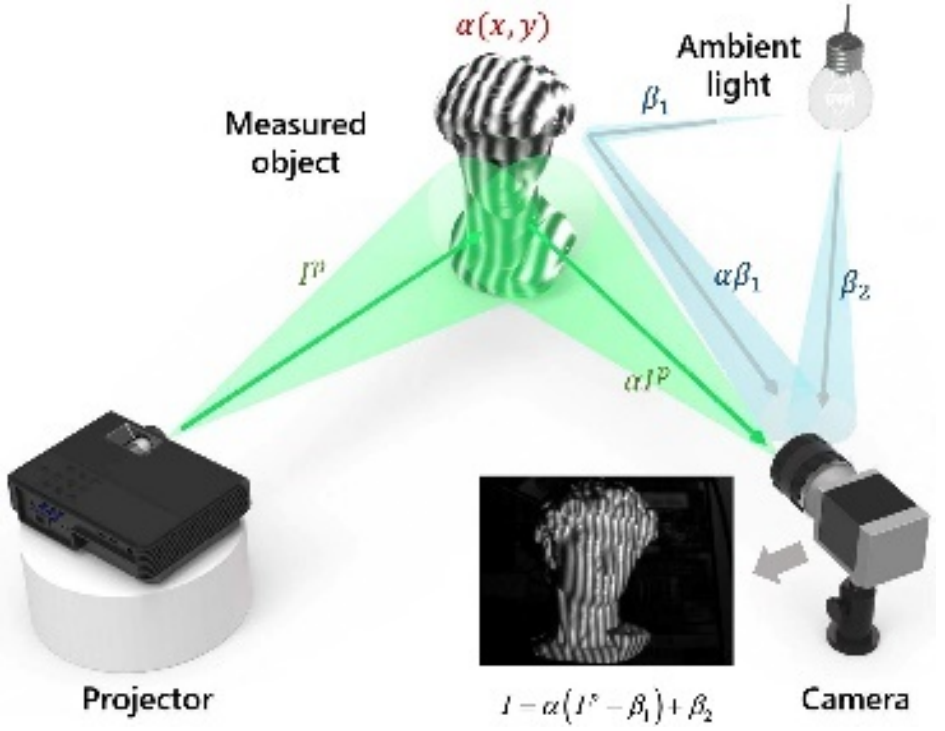
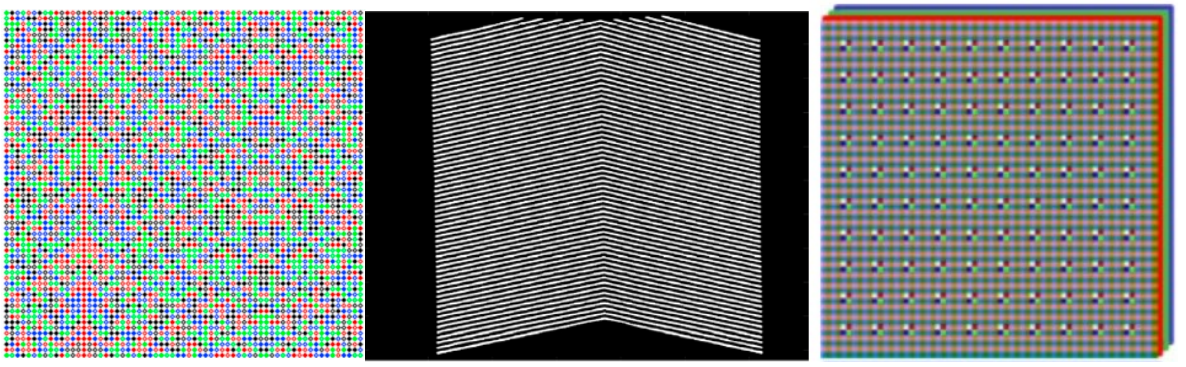


Figure 4. Structured light camera schematic [24].

DOI: 10.54254/2755-2721/28/20230140

The core of structured light cameras is encoding structured light. Depending on the nature of the projected structured light, structured light 3D imaging technology can be divided into point, line, and surface structured light method. In the point structured light method, the laser projects a point laser beam, which can be specifically divided into single-point and multi-point (projecting laser dot matrix). The point structured light method scans the target object point by point and obtains a high-density point cloud with the highest accuracy. But at the same time, the efficiency is very low and the real-time performance is also poor. Therefore, the point structured light method is not suitable for automatic driving situations. The line structured light consists of single line structure (projecting stripe light spot) and multi-line structure (projecting multiple stripe light spots). Line structured light method only requires one-dimensional scanning compared to the point structured light method, thus slightly improving efficiency but still does not meet the needs of autonomous driving. For multi-visual line structured light cameras, Ge et al proposed a global calibration method using an auxiliary camera, which improves the efficiency of line structured light cameras to some extent [25].

A more common method is using surface structured light, which projects the encoded image onto the surface of the target object and simultaneously captures the encoded image through the camera. Two encoding methods are spatial encoding and temporal encoding. Spatial encoding will only project one image onto the surface of the target object, and the local encoding of this image should be unique from the global encoding. The spatial encoding can be point, line, or cross-line [26]. Figure 5 shows some examples of the spatial encoding method. Although spatial encoding is less accurate and more difficult to decode, it is better in real-time and suitable for applications in dynamic measurements such as autonomous driving. Because the full depth information can be obtained by only one projection and the computational effort is small. Temporal coding is to project a series of simple coded images in time sequence. The coding is realized by the change of bright and dark pixels over time. Time encoding includes binary encoding, gray code, phase shift code, etc. Time encoding is characterized by high accuracy, simple decoding, and strong anti-interference ability. But it requires multiple projections for one encoding, so the real-time performance is poor and only suitable for static measurement.



(a) (b) (c)

Figure 5. These are the examples of the spatial encoding method**.** (a) is point spatial encoding [27]. (b) is line spatial encoding [28]. (c) is cross-line spatial encoding [29].

Since convolutional neural networks (CNNs) prove to be an excellent way of extracting features of images, many scholars have attempted to decode images from structured light cameras by using CNN. Sam et al and Feng et al both designed a CNN for analyzing single fringe pattern input and outputting 3D height maps, achieving high accuracy [30,31]. Analyzing road conditions is needed for autonomous driving systems: Jia et al. [32] proposed a depth measurement convolutional neural network (DMCNN) based on structured light cameras. DMCNN first extracts the multiscale fusion features of images and then classifies and regresses them by parallel branches. After that, the network uses the four-step phase shift algorithm to successfully construct a high-precision ground height map. As the autonomous driving system needs to face complex road conditions, the perception system needs to be robust and highly resistant to interference. Wang et al designed an end-to-end deep neural network based on U-net for

DOI: 10.54254/2755-2721/28/20230140

extracting image features of shape-coded structured light [33]. Their network has high robustness, high detection accuracy, and high coding density for structured light images.

2.2.4. Horizontal comparison. Table 3 compares the three 3D imaging methods according to several metrics.

Table 3. Comparison of 3D imaging methods.

Metric	Binocular stereo-vision	TOF	Structured light
Imaging Method	Passive	Active	Active
Imaging Accuracy	Millimeter at close range	Up to centimeter	0.01-1mm at close range
Resolution	Very high	Very low	Medium
Imaging frame rate	High and low	High(100+fps)	Low(30fps)
Measurement range	Very close	Far	Close
Response time	Medium	Fast	Slow
Power consumption	Low	High	Medium
Software Complexity	Complex	Simple	Medium
Dark light recognition	Weak	Strong	Strong

3. Conclusion

In the field of autonomous driving, visual-based perception systems have become a hot research topic in recent years. Thanks to the great progress in image processing technology represented by CNNs, visual-based perception systems now successfully serve as an alternative to the traditional LiDAR-based perception systems. This paper provides an overview of the current promising and mature visual-based perception technologies from the perspectives of acquiring 2D image data and acquiring depth information. This paper also analyzes how these technologies use neural networks to improve their performance, and reviews the recent performance of visual-based perception technologies for autonomous driving applications.

In acquiring 2D images, event cameras show many advantages that RGB cameras do not have. Although the algorithms for processing event camera data streams are not mature, a growing number of studies in this field have provided many feasible solutions for this problem. Meanwhile, algorithms for acquiring depth information using event cameras have also made progress in recent years.

In terms of acquiring depth information, commercial TOF cameras represented by Kinect2 and commercial structured light cameras represented by RealSense have shown excellent performance. Meanwhile, the improvement in stereo matching algorithms with the help of neural networks also results in better performance of binocular stereo vision cameras. All three of these depth camera technologies have great potential in the future to assist autonomous driving systems to build 3D scene maps more accurately and efficiently

References

[1] Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., . . . De Souza, A. F. 2021 Self-driving cars: A survey *Expert Systems with Applications* 165 113816.

[2] Cheng, J., Zhang, L., Chen, Q., Hu, X., and Cai, J. 2022 A review of visual SLAM methods for autonomous driving vehicles *Engineering Applications of Artificial Intelligence* 114 104992

[3] Paden, B., Cap, M., Yong, S. Z., Yershov, D., and Frazzoli, E. 2016 A Survey of Motion Planning and Control Techniques for Self-Driving Urban Vehicles *IEEE Transactions on Intelligent Vehicles* 1(1) 33-55.

[4] Rebecq, H., Gallego, G., Mueggler, E., and Scaramuzza, D. 2018 EMVS: Event-Based Multi-View Stereo3D Reconstruction with an Event Camera in Real-Time *International Journal of Computer Vision* 126(12) 1394-1414.

[5] Gallego, G., Delbrück, T., Orchard, G., Bartolozzi, C., Taba, B., Censi, A., . . . Scaramuzza, D. 2022 Event-Based Vision: A Survey I*EEE Transactions on Pattern Analysis and Machine Intelligence* 44(1) 154-180.

DOI: 10.54254/2755-2721/28/20230140

[6] Zanatta, L., Di Mauro, A., Barchi, F., Bartolini, A., Benini, L., and Acquaviva, A. (2022). Directly-Trained Spiking Neural Networks for Deep Reinforcement Learning: Energy Efficient Implementation of Event-Based Obstacle Avoidance on a Neuromorphic Accelerator. *Available at SSRN* 4272378.

[7] Hu, Y. H., Liu, S. C., Delbruck, T., and Soc, I. C. (2021, Jun 19-25). v2e: From Video Frames to Realistic DVS Events. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Electr Network.

[8] Viale, A., Marchisio, A., Martina, M., Masera, G., Shafique, M., and Ieee. (2021, Jul 18-22). CarSNN: An Efficient Spiking Neural Network for Event-Based Autonomous Cars on the Loihi Neuromorphic Research Processor. *International Joint Conference on Neural Networks (IJCNN)*, Electr Network.

[9] Deng, Y. J., Chen, H., and Li, Y. F. 2022 MVF-Net: A Multi-View Fusion Network for Event-Based Object Classification *IEEE Transactions on Circuits and Systems for Video Technology* 32(12) 8275-8284.

[10] Duan, P. Q., Wang, Z. H. W., Shi, B. X., Cossairt, O., Huang, T. J., and Katsaggelos, A. K. 2022 Guided Event Filtering: Synergy Between Intensity Images and Neuromorphic Events for High Performance Imaging *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(11) 8261-8275.

[11] Gehrig, M., Aarents, W., Gehrig, D., and Scaramuzza, D. 2021 DSEC: A Stereo Event Camera Dataset for Driving Scenarios *IEEE Robotics and Automation Letters* 6(3) 4947-4954.

[12] Wan, J. X., Xia, M., Huang, Z. K., Tian, L., Zheng, X. Y., Chang, V., . . . Wang, H. 2021 Event-Based Pedestrian Detection Using Dynamic Vision *Sensors Electronics* 10(8).

[13] Zhuang, S., Ji, Y., Tu, D and Zhang, X. 2022 Underwater RGB-D Camera Based on Binocular Stereo Vision RGB-D *Guangzi Xuebao/Acta Photonica Sinica* 51(4) 161-175.

[14] Laga, H., Jospin, L. V., Boussaid, F., and Bennamoun, M. 2022 A Survey on Deep Learning Techniques for Stereo-Based Depth Estimation *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(4) 1738-1764.

[15] Menze, M., Geiger, A., and Ieee. (2015, Jun 07-12). Object Scene Flow for Autonomous Vehicles. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA.

[16] Poggi, M., Tosi, F., Batsos, K., Mordohai, P., and Mattoccia, S. 2022 On the Synergies Between Machine Learning and Binocular Stereo for Depth Estimation From Images: A Survey IEEE *Transactions on Pattern Analysis and Machine Intelligence* 44(9) 5314-5334.

[17] Yang, G. S., Manela, J., Happold, M., Ramanan, D., and Soc, I. C. (2019, Jun 16-20). Hierarchical Deep Stereo Matching on High-resolution Images. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA.

[18] Sarbolandi, H., Lefloch, D., and Kolb, A. 2015 Kinect range sensing: Structured-light versus Time-of-Flight Kinect *Computer Vision and Image Understanding* 139 1-20.

[19] Horaud, R., Hansard, M., Evangelidis, G., and Menier, C. 2016 An overview of depth cameras and range scanners based on time-of-flight technologies *Machine Vision and Applications* 27(7) 1005-1020.

[20] Niskanen, I., Immonen, M., Hallman, L., Yamamuchi, G., Mikkonen, M., Hashimoto, T., Heikkilä, R. 2021 Time-of-flight sensor for getting shape model of automobiles toward digital 3D imaging approach of autonomous driving *Automation in Construction* 121 103429.

[21] Tölgyessy, M., Dekan, M., Chovanec, Ľ., and Hubinský, P. 2021 Evaluation of the Azure Kinect and Its Comparison to Kinect V1 and Kinect V2 *Sensors*, 21(2) 413.

[22] Bamji, C., Godbaz, J., Oh, M., Mehta, S., Payne, A., Ortiz, S., . . . Thompson, B. 2022 A Review of Indirect Time-of-Flight Technologies *IEEE Transactions on Electron Devices* 69(6) 2779-2793.

[23] Miller, L., García, A. R., Lorente, P. J. N., Andrés, C. F., and Morón, R. B. (2020, 2020//). Time of Flight Camera Calibration and Obstacle Detection Application for an Autonomous Vehicle. *Information Systems Architecture and Technology: Proceedings of 40th Anniversary*

DOI: 10.54254/2755-2721/28/20230140

International Conference on Information Systems Architecture and Technology – ISAT 2019, Cham.

[24] Zuo, C., Feng, S., Huang, L., Tao, T., Yin, W., and Chen, Q. 2018 Phase shifting algorithms for fringe projection profilometry: A review *Optics and Lasers in Engineering* 109 23-59.

[25] GE, Q., YAN, S., CHEN, W., WU, L., and XU, X. 2022 Research on the global calibration of the multi-vision line structured light measurement system based on auxiliary camera *SCIENTIA SINICA Technologica* 52(8) 1274-1284.

[26] Wang, Z. 2020 Review of real-time three-dimensional shape measurement techniques *Measurement* 156 107624.

[27] Lin, H., Nie, L., and Song, Z. 2016 A single-shot structured light means by encoding both color and geometrical features *Pattern Recognition* 54 178-189.

[28] Wang, Z., and Yang, Y. 2018 Single-shot three-dimensional reconstruction based on structured light line *Pattern Optics and Lasers in Engineering* 106 10-16.

[29] Di Martino, M., Flores, J., and Ferrari, J. A. 2018 One-shot 3D scanning by combining sparse landmarks with dense gradient information *Optics and Lasers in Engineering* 105 188-197.

[30] Van der Jeught, S., and Dirckx, J. J. J. 2019 Deep neural networks for single shot structured light profilometry *Optics Express* 27(12) 17091-17101.

[31] Feng, S., Chen, Q., Gu, G., et al. 2019 Fringe pattern analysis using deep learning *Advanced Photonics*, 1(2), 025001-025001.

[32] Jia, T., Liu, Y., Yuan, X., Li, W., Chen, D., and Zhang, Y. 2022 Depth measurement based on a convolutional neural network and structured light *Measurement Science and Technology* 33(2) 025202.

[33] Wang, S., Song, Z., Du, H., and Gu, F. (2022, 17-22 July 2022). Highly-Robust Feature Detection Method in Shape-Coded Structured Light Based on End-to-End Deep Neural Network. *2022 IEEE International Conference on Real-time Computing and Robotics (RCAR)*.