Multimodal perception of vulnerable road users for autonomous driving in urban environments

Vincent BREBION
Directeurs de thèse : Franck DAVOINE, Julien MOREAU
Équipe : SyRI

***Abstract*—Urban areas are complex driving environments where vehicles and personal mobility devices (pedestrians, cyclists, ...) often operate close to each other. Autonomous driving in these environments represents a challenge, due to the highly unpredictable behaviour of these vulnerable users. The objective of the thesis is to develop multimodal perception systems to improve their detection and avoid any collision by means of computer vision, multi-sensor fusion, and machine learning.**

***Index Terms*—Computer vision; Multi-sensor fusion; Machine learning; Neuromorphic cameras**

I. INTRODUCTION

Autonomous driving in open environments calls for deep understanding abilities from the self-driving vehicle to make it able to navigate safely. This understanding process requires the detection and recognition of potential obstacles. The use of perception sensors allows for such capabilities, especially powered by the rise of machine-learning-based methods.

Most of the results from the literature, however, were achieved in favorable conditions (adequate lighting and weather, clearly visible objects), which only represent a fraction of the real-life situations a driver is confronted with. Recent studies have particularly shown the limits of these approaches in more complex conditions (at dawn/dusk, when the vulnerable user is partially occluded or very close to the ego-vehicle, ...), raising multiple safety questions [1].

In parallel, the navigation in urban environments has been deeply changing over the past few years, with the rise of soft mobility solutions (bicycles, scooters, skateboards, rollerblades, hoverboards, ...). While they allow for more flexible movements in urban areas, they especially put their user at risk in case of a collision with a traditional vehicle. This risk is further amplified by the erratic behaviour these users may have: slaloming between cars, alternating between the use of the road and the sidewalks, not respecting the road markings, navigating close to the other vehicles, etc.

As an answer to these issues, the objective of the thesis is to reinforce the detection of these vulnerable users in difficult visual conditions typical in urban environments. To reach it, two complementary approaches are being employed:

- the fusion of data from proven and novel sensors (RGB and event-based cameras, LiDARs), allowing for redundancy and for perception even under complex lighting and/or weather conditions;
- the use of machine learning techniques, able to provide state-of-the-art results for perception tasks even with noisy data from multiple asynchronous sensors.

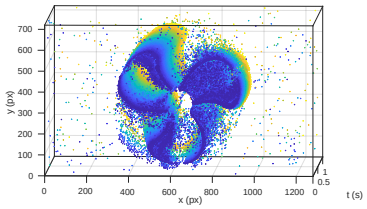

Fig. 1. Events from a simple "rotating fan" sequence. Events are plotted in 3D along the x, y, and t axes, each dot representing an event (colored based on their timestamp for a better visualization). Please note how the events correspond to the edges of the blades of the fan, i.e. the pixels for which light intensity changes over time.

II. WORK CONDUCTED

A. A short introduction to event-based cameras

In opposition to "conventional" frame-based cameras, which accumulate light during short periods of time to create dense images, event cameras produce only small pulses of information (called events), which report asynchronously and independently per pixel the changes of light in the observed visual scene (see Fig. 1).

Thanks to their ability to perceive movement asynchronously and with very low reaction times, these event-based cameras are particularly suited for perception in complex scenes with fast dynamics, e.g. driving scenes.

B. Event-based optical flow

Optical flow is a central problem in the field of computer vision, where the goal is to estimate the pixelwise motion of a visual scene. It is a key enabler for other applications, including especially the object detection and recognition issues.

As part of the first year of the thesis, we developed a novel pipeline-based framework, allowing for the computation of optical flow with both low- and high-resolution event-based sensors in real-time. Our method consists in accumulating events over short temporal windows to create binary images. These images are then denoised, densified, and finally fed to an image-based optical flow library to recover the final optical flow for each event.

If the reader is interested, more details about this work are available in last year's report, or in the published articles (see Section III).

C. Depth estimation for events

While events by themselves are an interesting component for the detection of vulnerable road users, they lack the

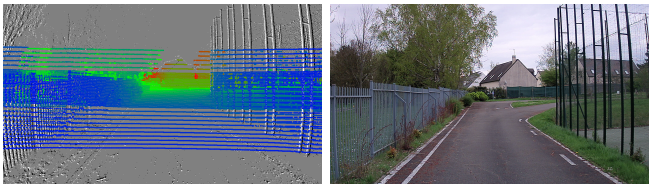

Fig. 2. Left: Events (in black and white) and the corresponding projected LiDAR point cloud (colored dots, based on their depth), recorded while driving on the SEVILLE track. Notice how sparse the LiDAR scans are for the uppermost and lowermost parts of the image. Right: Reference RGB image of the scene.

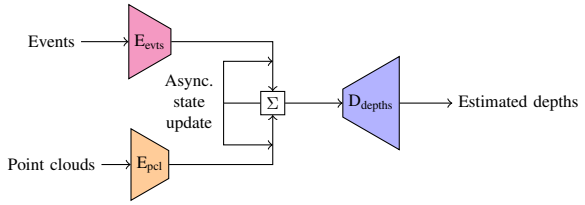

Fig. 3. Our "depth estimation for events" network, currently under construction. It is composed of two separate encoders (one for events, the other one for point clouds), an asynchronous state update module, and a decoder to obtain the final depths output. Figure inspired by [2].

depth and the color/texture information a LiDAR and an RGB camera can bring. Therefore, as part of the thesis, we propose to estimate a depth for each event, leveraging information from a LiDAR sensor. By adding this depth information:

- we give more context to each event, allowing us to distinguish events representing sharp edges of objects from those corresponding to continuous textures, and thus simplifying the tasks of isolating independent objects in a scene;
- each event can be reprojected to the 3D world then back-projected in an RGB camera, allowing for the fusion/superimposition of information from the two sensors.

Estimating the depth of each event however is not an easy task, as LiDARs only provide sparse 3D point clouds, as illustrated by Fig. 2. While some authors have tried to propose heuristic methods for estimating a depth for each event based on the nearest LiDAR points [3], we propose here to leverage the recent advancements in machine learning to carry this task.

Inspired by the work conducted by Gehrig et al. [2], our current work is centered around the task of designing a neural network able, from asynchronous LiDAR and event inputs, to derive a depth estimation for each event. A model of our network currently under construction is displayed in Fig. 3.

D. Simulated and real datasets generation

In order to be able to train and validate the depth estimation method, large scale datasets are needed, including events, LiDAR, and dense ground truth depth measurements. To this day, however, such specific datasets are not available.

Therefore, as part of the thesis, two complementary datasets are currently being constructed:

- a simulated dataset, using CARLA;
- and a real-world dataset, using sensors mounted on top of the Zoe cars of the laboratory.

The creation of a simulated dataset is bound to the fact that acquiring large amount of driving data in varying environments and varying traffic/lighting/weather conditions is complex and time consuming, while it is nearly effortless with a simulator. Furthermore, dense ground truth depth values can easily be recovered in CARLA, while such an operation is not straightforward with real data.

Yet, we plan to also acquire driving sequences in the real-world, as simulation can not reproduce perfectly the noise and dynamics of the sensors and the environment. This real-world data will be used for fine-tuning the network, as well as for validating the method on complex real-world scenarios. In preparation for these recordings, a sensor-synchronization technique as well as a calibration method compatible with event-based cameras have been developed and implemented on the Zoe vehicles of the laboratory.

III. Publications and collaboration

As part of the thesis work, a publication titled "Real-Time Optical Flow for Vehicular Perception with Low- and High-Resolution Event Cameras" [4] was published in IEEE Transactions on Intelligent Transportation Systems in early 2022. Its accompanying dataset is available on Heudiasyc's platform (https://datasets.hds.utc.fr/share/er2aA4R0QMJzMyO). A second publication, titled "Estimation de flot optique basé évènements en temps réel" [5], is also going to be presented as part of the RFIAP conference in Vannes in July 2022.

A formal collaboration with Prophesee[1], a company designing event-based cameras, is also being put in place.

IV. Conclusion and future work

As of today, the focus is set on the depth estimation application. If accurate results are obtained, they will open large opportunities for the final "vulnerable road users detection" application, which will be treated in the third and final year of the thesis.

Acknowledgment

This work is supported by the Hauts-de-France Region and SIVALab (Renault-UTC-CNRS).

References

[1] T. S. Combs, L. S. Sandt, M. P. Clamann, and N. McDonald, "Automated vehicles and pedestrian safety: Exploring the promise and limits of pedestrian detection." *American journal of preventive medicine*, vol. 56 1, pp. 1–7, 2019.

[2] D. Gehrig, M. Rüegg, M. Gehrig, J. Hidalgo-Carrió, and D. Scaramuzza, "Combining events and frames using recurrent asynchronous multimodal networks for monocular depth prediction," *IEEE Robotics and Automation Letters*, vol. 6, pp. 2822–2829, 2021.

[3] B. Li, H. Meng, Y. Zhu, R. Song, M. Cui, G. Chen, and K. Huang, "Enhancing 3-d lidar point clouds with event-based camera," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.

[4] V. Brebion, J. Moreau, and F. Davoine, "Real-time optical flow for vehicular perception with low- and high-resolution event cameras," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[5] ——, "Estimation de flot optique basé évènements en temps réel," in *RFIAP 2022 (Congrès Reconnaissance des Formes, Image, Apprentissage et Perception)*, Vannes, France, Jul. 2022.

[1] https://www.prophesee.ai