EV-FuseMODNet: Moving object detection using the fusion of event camera and frame camera

Haixin Sun, Songming Chen, Minh-Quan Dao, Vincent Frémont

EV-FuseMODNet: Moving object detection using the fusion of event camera and frame camera

Haixin SUN*, Songming CHEN§ Minh-Quan DAO†, Vincent FREMONT‡
Nantes Université, Ecole Centrale de Nantes, CNRS LS2N
Nantes, France
Email: *haixin.sun@ec-nantes.fr, §Songming.chen@ec-nantes.fr, †minh-quan.dao@ec-nantes.fr, ‡vincent.fremont@ec-nantes.fr

***Abstract*—Moving object detection is a crucial task for autonomous vehicles. Indeed, dynamic objects represent higher collision risk than static ones, so the trajectories of the vehicles must be planned according to the motion forecasting of the moving participants of the scene. For the traditional frame-based cameras, images can provide accumulated pixel brightness without temporal information between them. The optical flow computation is used as the inter-frame motion information. Interestingly, event-based camera can preserve the motion information by delivering the precise timestamp of each asynchronous event data, which is more suitable for the motion analysis. Also, the event-based cameras' high temporal resolution and high dynamic range allow them to work in fast motion and extreme light scenarios. In this work, we propose a new Deep Neural Network, called EV-FuseMODNet for Moving Object Detection (MOD) that captures motion and appearance information from both event-based and frame-based cameras. The proposed method has been evaluated with the extended KittiMoSeg dataset and the generated dark KITTI sequence. An overall 27.5% relative improvement on the extended KittiMoSeg dataset compared to the state-of-the-art approaches has been achieved. The code is released in https://github.com/adosum/EV-FuseMODNet.**

***Index Terms*—sensor fusion, moving object detection, event-based camera, deep learning**

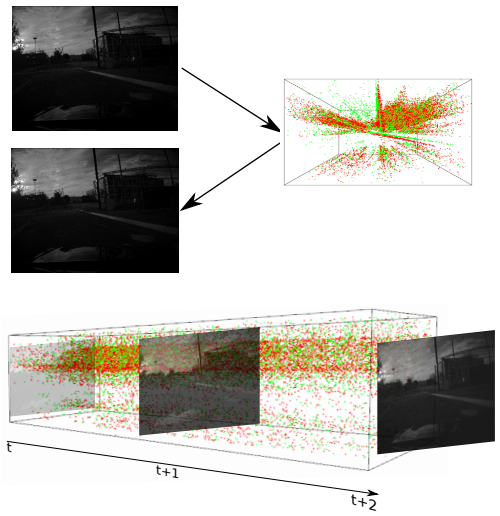


Fig. 1. Visualization of the event data between two grayscale images.

I. INTRODUCTION

An Autonomous Vehicle (AV) needs an accurate perception of its surrounding environment to work reliably and safely. Its perception system should transform raw sensory data such as image pixels into semantic information for scene understanding. For an autonomous vehicle, it is required to fully estimate the motion model of each of the surrounding participants and to plan the ego-trajectories based on their future states to reduce collision risks. There are two main classes of motion in a typical autonomous driving scene: The surrounding moving objects and the scene motion generated by the ego vehicle. Due to the ego vehicle's motion and constraints related to the image formation, it is very challenging to classify the surrounding objects as moving or static because even static objects will be perceived as moving. Motion segmentation implies that the two tasks have to be performed jointly. The first focuses on object segmentation, in which objects of specific interesting classes are highlighted, such as pedestrians or vehicles. The second focuses on motion classification, in which a classifier predicts whether the observed object is moving or static.

Frame-based monocular cameras are one of the most commonly used sensors in the Autonomous Vehicle system. They transmit raw images synchronously, frame by frame, at a fixed rate. This feature leads to a significant drawback of low temporal resolution, redundant information, and low dynamic range. A few years ago, the event-based camera, a bio-inspired silicon retina technology, has been proposed to overcome those limitations and to solve both classical and new computer vision tasks [1]. An event-based camera can have a dynamic range of 130 dB and a minimum of 3 μs latency. Those advantages allow the event-based camera to work in extreme scenarios with low light conditions and fast motions. Typically, event-based cameras are used as sensing modalities on Unmanned aerial vehicle (UAV) [2], mobile robots [3] or wearable electronics [4], where operations are under unrealistic lighting conditions and sensitive to the temporal resolution. The main applications for event-based cameras are object tracking [3], surveillance and monitoring [5], and optical flow estimation [6]. Recently, more and more academic works focus on using event-based cameras for autonomous driving. [7] proposed a method that can predict the vehicle's steering angle according to the event data, and [8] describes a dataset that contains event data along with the vehicle state.

Event-based cameras are asynchronous sensors that monitor

changes in log brightness intensity. When the variation of the brightness of a pixel reaches the threshold, the camera will emit an event. The event signal is usually in the format of a tuple, $e = (x, y, t, p)^T$, where $(x, y)^T$ is the pixel's position, t is the precise timestamp of the event which is accurate up to microseconds, and the polarity p of the change that indicates whether the pixel became brighter or darker. Fig. 1 shows the visualization of the event data between the consecutive frame-based gray-scale images. The positive events are shown in red, and the negative events are in blue. Between two consecutive images, there is a quasi-continuous stream of events that represents all the brightness changes between the two images. The event-based camera's asynchronous nature and tracking in the log image space offer several advantages over traditional frame-based cameras, including extremely low latency for detecting high-speed objects, a very high dynamic range for poor light conditions, and significantly lower power consumption. The cameras' unique output, on the other hand, presents new challenges in algorithm development. Indeed, the events are transmitted asynchronously and lack the pixel's absolute value and spatial neighborhood. Algorithms for traditional frame-based optical flow or object detection, are no longer valid. As a result, a significant research effort has been made to develop new algorithms for event-based cameras to solve these traditional vision problems.

The main contribution of this paper is to propose a novel Deep Neural Network architecture for moving objects detection. As mentioned previously, traditional frame-based cameras cannot provide temporal information, and the event-based cameras lack appearance information and spatial consistency. The two features are both important for moving object detection. We address this problem by proposing a fusion network model that can use the information from the two sensors simultaneously and achieve better performances. We train and evaluate the proposed EV-FuseMODNet network using the extended KittiMoSeg dataset [9]. The results show that the proposed approach outperforms current state-of-the-art methods, by achieving 27.5% improvement compared to the FuseMODNet [9], and 36.7% compared to the MODNet [10].

The paper is structured as follows: In Section II, the related works are discussed. In section III, the methodology is presented, covering the EV-FuseMODNet architecture and the encoding method for the event data. Section IV describes the experimental results, including training details and the comparison results with state-of-the-art approaches.

II. RELATED WORKS

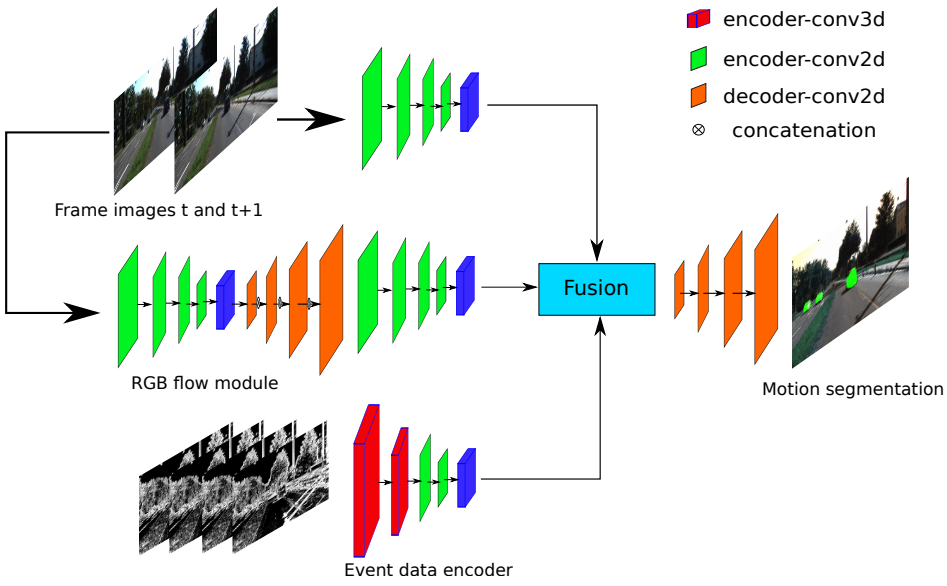
Motion Segmentation using Event-based Cameras: Several approaches have been proposed for motion segmentation with the event-based camera. In [11], a method is described for detecting and tracking a circle with event clutter. It is based on the Hough transform using optical flow information extracted from spatial-temporal windows of events. Segmentation of the moving objects has been also addressed in [12]. It considered more generic object detection than [11] by using event corners as primitives. It adopted a learning technique to separate events caused by camera motion from those by moving objects. However, this method required extra knowledge of the robot joints that control the camera. In [13], [14], segmentation has been addressed using motion-compensated event images [15]. [13] detected moving objects in clutter by fitting a motion-compensation model to the dominant events (the global motion) and detecting inconsistencies concerning that motion (i.e., the motion of moving objects). Those objects were then extracted via morphological operations on the warped image. [14] proposed to jointly estimate the moving object segmentation and the motion parameters of the objects by maximization of an objective function, which depends on the results of the motion-compensated event images.

Motion Segmentation using Frame-based Cameras: Classical approaches have been proposed for moving object detection based on the geometrical understanding of the scene, such as [16], which was used to estimate objects' motion masks. [17] introduced assumptions about the motion model for the background motion in terms of homographies. This approach cannot be used in autonomous driving applications due to the errors arising from camera translations. Classical methods provide poor performances compared to deep learning methods with high complexity. For instance, [16] running time is 50 minutes per frame, making it impossible to use in a real-time application such as autonomous driving. Deep learning models are becoming successful beyond object detection [18] for applications like visual SLAM [19] and semantic segmentation [20]. However, it still needs to be explored for moving object detection tasks. [21] proposed a method to exploit optical flow for generic foreground segmentation. This work is designed for generic object segmentation and does not focus on classifications of objects as moving or static. [10], [22] explored motion segmentation using deep network architectures; however, these networks rely only on the frame images, which is prone to failure in extreme illumination conditions.

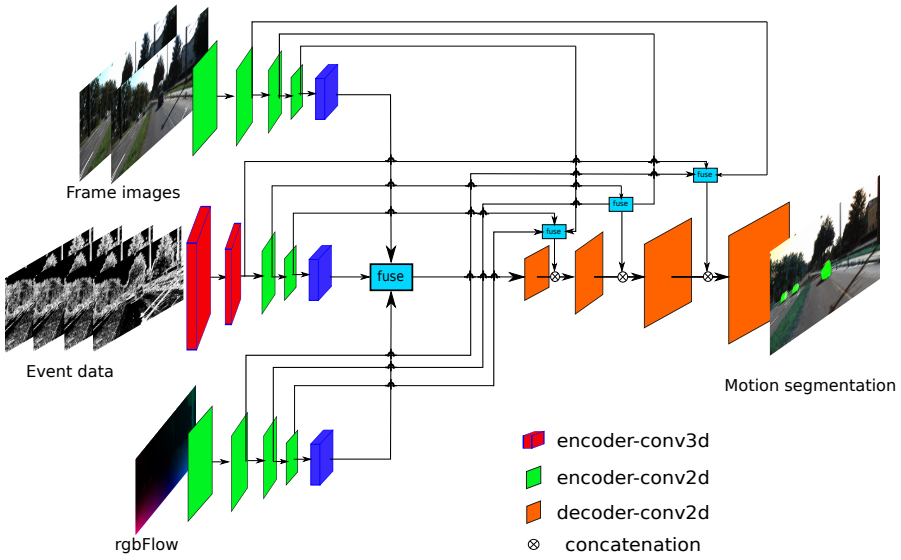
Events and Images Fusion: Several works have been proposed for the fusion between the LiDAR and frame-based camera. [23] proposed an algorithm for 3D semantic segmentation using LiDAR and camera. [24] proposed a semantic segmentation algorithm using fusion between images and optical flow. In [9], a CNN architecture is proposed for Moving Object Detection under low-light conditions by capturing motion information from both camera and LiDAR sensors. It also presented the KittiMoSeg dataset, which provides the moving object mask for the KITTI sequence. Modern vehicles are equipped with various sensors to fully perceive the surrounding environment so the system can be more robust. Data fusion provides improved performances in various tasks such as [10], [21], [25], [26], so it is worth developing the fusion model between different sensors.

III. PROPOSED APPROACH

In this section, we first explain the proposed network structure, including three feature encoders, the fusion structure,



(a) The overall structure of the model



(b) The fusion architecture of the model

Fig. 2. Network structure of the ev-FuseMODNet

and the decoder. We will also discuss our encoding method for the event data because they are asynchronous and cannot be directly fed into the neural network.

A. Network structure

The overall architecture of the EV-FuseMODNet is shown in Fig. 2 (a). It contains three main parts: frame-based optical flow, frame-based image processing, and event data processing.

For the frame-based optical flow, we directly adopt the RAFT [27] model since it achieves the state of the art performance and high efficiency in inference time. A 2D-encoder is used after the optical flow to extract features for the fusion step, shown as the RGB flow module in Fig. 2 (a). This module is designed to provide precise motion information to the model, and only the 2D-encoder will be trained during the training process.

For the frame image processing, we also use a 2D-encoder structure to extract the features as shown at the top of Fig. 2 (a). Two consecutive images will be sent into the encoder so it can provide the appearance and general motion features. Note that this encoder is pretrained with the cityscape segmentation

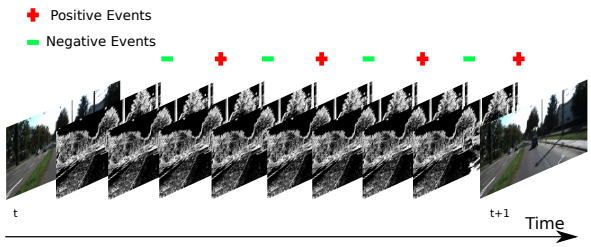


Fig. 3. Visualization of our event encoding representation.

dataset.

We adopt the 3D-encoder [28] for the event data processing. This choice is linked to the fact that it can better preserve the spatial-temporal information of the event data compared to the 2D-encoder. Since the event data is asynchronous and cannot be used in the network directly, an encoding method is also required here. Given a set of N input events $E_N = (x_i, y_i, t_i, p_i), i \in [1, N]$, and a time depth D to discretize the time dimension of event data, we accumulate each group of event data into images as follows:

$$\begin{aligned} t_{norm} &= (t - t_0)/(t_N - t_1) * (D - 1) \\ I(x, y, t, p) &= \sum_i \delta(p - p_i) k_b(x - x_i) k_b(y - y_i) k_b(t - t_{norm}) \\ k_b(a) &= max(0, 1 - |a|) \end{aligned} \tag{1}$$

Here, (x, y) denotes the position of the event, p is the polarity of the event, and δ is the Kronecker delta operator. $k_b(\cdot)$ denotes bi-linear sampling kernel. The generated event image I is a $(2, D, H, W)$ tensor, where the number 2 represents the positive and negative polarity, D is the discretized time depth, and (H, W) are respectively the height and width of the image. Fig. 3 is an example of the event data representation where $D = 4$.

B. The fusion architecture

The fusion architecture of the EV-FuseMODNet is shown in Fig. 2 (b). A mid fusion strategy is adopted in our model. Mid-Fusion represents feature-level-fusion where features are extracted from each input separately using an encoder that is exclusive to each input. The fusion is done by concatenating feature maps generated from each stream before feeding them into the decoder. There is a skip connection from each encoder to the corresponding decoder. For the skip connection between 2D-encoder and the decoder, the activation of the encoder is directly concatenated with the activation of the decoder. For the skip connection of the 3D-encoder for the event data processing, the 3D activation $(C \times D \times W \times H)$ is flattened into a 2D tensor $((C \cdot D) \times W \times H)$ first, then it is concatenated with the activation of the decoder.

IV. Experiments

A. Dataset and Implementation Details

We trained all proposed models end-to-end with weighted binary cross-entropy loss for 100 epochs and batch size of 8.

Methods	Moving IoU
KITTI	
RGB-only	32.7
RGB+rgbFlow [10]	49.36
RGB+LidarFlow	41.64
RGB+rgbFlow+LidarFlow [9]	51.46
Ours-RGB+rgbFlow+Events	**63.16**
Dark-KITTI	
RGB-only	26.5
RGB+rgbFlow [10]	39.5
RGB+LidarFlow	38.5
RGB+rgbFlow+LidarFlow [9]	43.5
Ours-RGB+rgbFlow+Events	**57.51**

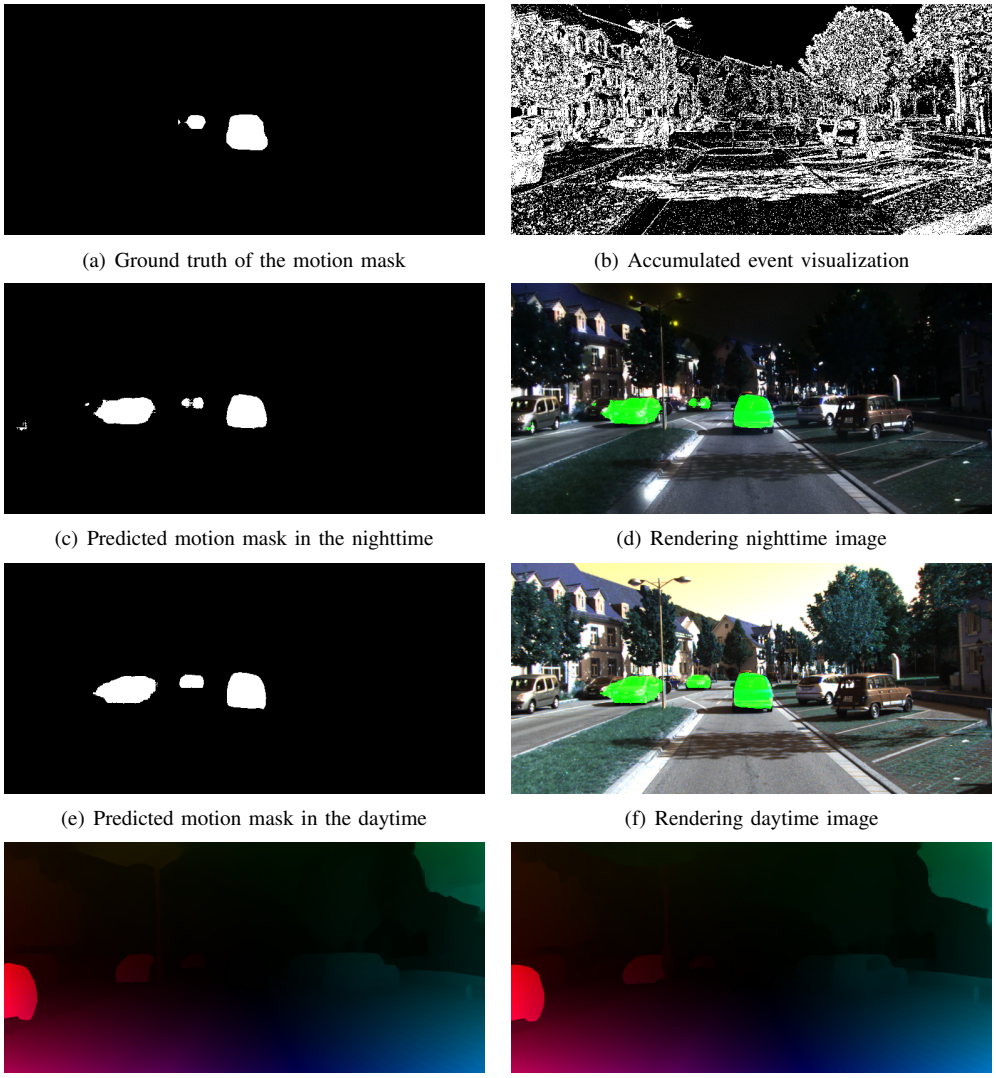
TABLE I
QUANTITATIVE ASSESSMENT OF OUR APPROACH COMPARED TO THE STATE-OF-THE-ART METHODS

The Adam optimizer is used with a learning rate of 1e-5 and a weight decay rate of 5e-4. We evaluate our methods for both day and night images with the extended KittiMoSeg dataset and the generated dark-Kitti sequence [29]. The extended KittiMoSeg provides the motion mask for the raw KITTI sequence and contains 12919 images. Since KITTI do not provide event data, we generate the event data with the event camera simulator [30].

B. Results

Table IV-B shows the results of the Intersection-over-Union (IoU) for the moving objects in comparison to previous moving object detection approaches [9], [10]. The results show that our approach shows a 22.7% improvement in the daytime KITTI sequence and a 31.2% improvement in the generated nighttime KITTI sequence. We attribute the improvement to the fusion with event data. Compared to the LiDAR, the event-based camera is a better complementary to the frame-based camera for the moving object detection task. The frame-based camera transmits images at a fixed rate. So it provides the appearance information but no motion knowledge. On the opposite, the event-based camera monitors the brightness change of each pixel and provides the precise timestamp value of each event data. So it lacks appearance information but can give more precise motion features. Also, the improvement in the Dark-KITTI sequence also proves that the event-based camera can improve the robustness of the model in a bad light environment.

Fig. 4 is the qualitative result of our approach. Results show the benefit of fusion, where the network was able to segment the moving objects in both daytime and nighttime scenes. Note that the ground truth mask is imperfect because it misses the left vehicle. According to the rendering image, our approach performs better since it recognizes the left-moving vehicle. However, the performance of our model in the nighttime is downgraded; Fig. 4 (c) shows that there is noise around the left vehicles, and the shape of the middle vehicle is not satisfactory. This is because the quality of the frame images downgrades under the low-illumination environment.



(a) Ground truth of the motion mask

(b) Accumulated event visualization

(c) Predicted motion mask in the nighttime

(d) Rendering nighttime image

(e) Predicted motion mask in the daytime

(f) Rendering daytime image

(g) rgbFlow of the daytime image

(h) rgbFlow of the nighttime image

Fig. 4. Qualitative result of the ev-FuseMODNet

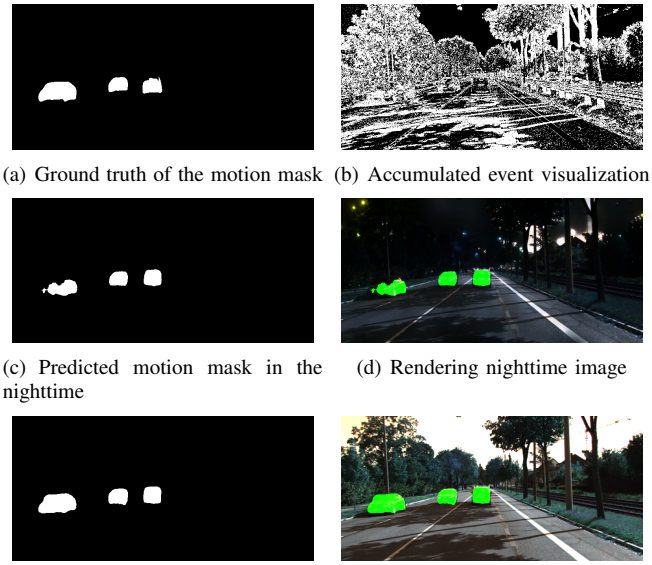
Fig. 5 shows an example of failure of the ev-FuseMODNet. In this sample, the ego vehicle is moving forward, and the target moving objects are two opposite-direction cars and one same-direction car. In the daytime, our approach can provide an accurate estimation of moving objects because the frame camera provides enough appearance features to the model. However, during the nighttime, the prediction for the left car is not satisfying. Their two main reasons: the left car is the most dark area of the RGB image, which makes the frame camera completely fails. Also, the left car is moving beside the trees and grass. This creates a textured background and downgrades the performance of the event camera. We can verify this from the accumulated event visualization, the left car is almost invisible because of the noise background, but the contour of the rest car is still clear.

V. CONCLUSIONS

In this work, we propose EV-FuseMODNet, a deep fusion neural network for moving object detection using a fusion of the frame-based camera and event-based camera data. This architecture fuses appearance features and motion information that is captured from both frame-based cameras and event-based cameras. The results show that our approach can generate more accurate (27.5%-36.7%) moving object segmentation due to the fusion of the event-based camera. In the future, adding more sensors (e.g. LiDAR) to the fusion model is a perspective direction; better fusion architecture is also required since more sensors will increase the computation time.

VI. ACKNOWLEDGMENT

This work was carried out in the framework of the NExT Senior Talent Chair DeepCoSLAM, which were funded by the French Government, through the program Investments for the



(a) Ground truth of the motion mask (b) Accumulated event visualization

(c) Predicted motion mask in the nighttime (d) Rendering nighttime image

(e) Predicted motion mask in the daytime (f) Rendering daytime image

Fig. 5. Failure case of the ev-FuseMODNet

Future managed by the National Agency for Research (ANR-16-IDEX-0007), and with the support of Région Pays de la Loire and Nantes Métropole. This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012128R1 made by GENCI.

REFERENCES

[1] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240 × 180 130 db 3 µs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[2] N. Sanket, C. Parameshwara, C. Singh, A. V. Kuruttukulam, C. Fermüller, D. Scaramuzza, and Y. Aloimonos, "Evdodgenet: Deep dynamic obstacle dodging with event cameras," in *IEEE International Conference on Robotics and Automation (ICRA)*, 05 2020, pp. 10 651–10 657.

[3] T. Delbruck and M. Lang, "Robotic goalie with 3 ms reaction time at 4% cpu load using event-based dynamic vision sensor," *Frontiers in Neuroscience*, vol. 7, p. 223, 2013. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnins.2013.00223

[4] T. Delbruck, "Neuromorphic vision sensing and processing," in *2016 46th European Solid-State Device Research Conference (ESSDERC)*, 2016, pp. 7–14.

[5] M. Litzenberger, B. Kohn, A. Belbachir, N. Donath, G. Gritsch, H. Garn, C. Posch, and S. Schraml, "Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor," in *2006 IEEE Intelligent Transportation Systems Conference*, 2006, pp. 653–658.

[6] M. Almatrafi and K. Hirakawa, "Davis camera optical flow," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 396–407, 2020.

[7] A. Maqueda, A. Loquercio, G. Gallego, N. Garcia, and D. Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 06 2018, pp. 5419–5427.

[8] Y. Hu, J. Binas, D. Neil, S.-C. Liu, and Delbruck, "Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction," in *IEEE International Conference on Intelligent Transportation Systems*, 11 2020.

[9] H. Rashed, M. I. Ramzy, V. Vaquero, A. E. Sallab, G. Sistu, and S. K. Yogamani, "Fusemodnet: Real-time camera and lidar based moving object detection for robust low-light autonomous driving," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pp. 2393–2402, 2019.

[10] M. Siam, H. Mahgoub, M. Zahran, S. Yogamani, M. Jagersand, and A. El-Sallab, "Modnet: Motion and appearance based moving object detection network for autonomous driving," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2859–2864.

[11] A. Glover and C. Bartolozzi, "Event-driven ball detection and gaze fixation in clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 2203–2208.

[12] V. Vasco, A. Glover, E. Mueggler, D. Scaramuzza, L. Natale, and C. Bartolozzi, "Independent motion detection with event-driven cameras," 07 2017.

[13] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 1–9.

[14] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, "Event-based motion segmentation by motion compensation," 10 2019, pp. 7243–7252.

[15] T. Stoffregen and L. Kleeman, "Event cameras, contrast maximization and reward functions: An analysis," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 292–12 300.

[16] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3061–3070.

[17] S. Wehrwein and R. Szeliski, "Video segmentation with background motion models," 01 2017.

[18] M. Siam, S. Elkerdawy, M. Jagersand, and S. Yogamani, "Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges," 10 2017, pp. 1–8.

[19] S. Milz, G. Arbeiter, C. Witt, B. Abdallah, and S. Yogamani, "Visual slam for automated driving: Exploring the applications of deep learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 360–36 010.

[20] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1568494618302813

[21] S. Jain, B. Xiong, and K. Grauman, "Fusionseg: Learning to combine motion and appearance for fully automatic segmentation of generic objects in videos," 07 2017, pp. 2117–2126.

[22] M. Siam, S. Eikerdawy, M. Gamal, M. Abdel-Razek, M. Jagersand, and H. Zhang, "Real-time segmentation with appearance, motion and geometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 5793–5800.

[23] K. Madawy, H. Rashed, A. Sallab, O. Nasr, H. Kamel, and S. Yogamani, "Rgb and lidar fusion based 3d semantic segmentation for autonomous driving," 06 2019.

[24] H. Rashed, S. Yogamani, A. El-Sallab, P. Krizek, and M. El-Helw, "Optical flow augmented semantic segmentation networks for automated driving," *arXiv preprint arXiv:1901.07355*, 2019.

[25] H. Rashed, A. El Sallab, S. Yogamani, and M. ElHelw, "Motion and depth augmented semantic segmentation for autonomous navigation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019, pp. 364–370.

[26] S. Chen, H. Sun, and V. Frémont, "Mono-vision based moving object detection using semantic-guided ransac," in *2021 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2021, pp. 1–6.

[27] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *European conference on computer vision*. Springer, 2020, pp. 402–419.

[28] H. Sun, M.-Q. Dao, and V. Fremont, "3d-flownet: Event-based optical flow estimation with 3d representation," in *2022 IEEE Intelligent Vehicles Symposium, IV 2022, Aachen, Germany, June 4-9, 2022*. IEEE, 2022.

[29] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," 03 2017.

[30] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: an open event camera simulator," *Conf. on Robotics Learning (CoRL)*, Oct. 2018.