cGAN을 이용한 이벤트 카메라기반 High Dynamic Range 이미지 및 비디오 생성방법에 대한 연구

S. Mohammad Mostafavi I.[2]*, Lin Wang[1]*[O], Yo-Sung Ho[2], Kuk-Jin Yoon[1]

[1] 한국과학기술원 기계공학과

[2] 광주과학기술원 전기전자컴퓨터공학부

mostafavi@gist.ac.kr, wanglin@kaist.ac.kr, hoyo@gist.ac.kr, kjyoon@kaist.ac.kr

요 약

Event cameras have a lot of advantages over traditional cameras, such as low latency, high temporal resolution, and high dynamic range. However, since the outputs of event cameras are the sequences of asynchronous events over time rather than actual intensity images, existing algorithms could not be directly applied. Therefore, it is demanding to generate intensity images from events for other tasks. In this paper, we unlock the potential of event camera-based conditional generative adversarial networks to create images/videos from an adjustable portion of the event data stream. The stacks of space-time coordinates of events are used as inputs and the network is trained to reproduce images based on the spatio-temporal intensity changes. The usefulness of event cameras to generate high dynamic range (HDR) images even in extreme illumination conditions and also non blurred images under rapid motion is also shown. The usefulness of event cameras to generate high dynamic range (HDR) images even in extreme illumination conditions and also non-blurred images under rapid motion is also shown. In addition, the possibility of generating very high frame rate videos is demonstrated, theoretically up to 1 million frames per second (FPS) since the temporal resolution of event cameras are about 1 millisecond. Proposed methods are evaluated by comparing the results with the intensity images captured on the same pixel grid-line of events using online available real datasets and synthetic datasets produced by the event camera simulator.

1. Introduction

Event cameras are bio-inspired vision sensors that mimic the human eye in receiving the visual information. While traditional cameras transmit intensity frames at a fixed rate, event cameras transmit the changes of intensity at the time of the changes, in the form of asynchronous events that deliver space-time coordinates of the intensity changes. They have lots of advantages over traditional cameras, *e.g*. low latency in the order of microseconds, high temporal resolution (around 1 μs) and high dynamic range. However, since the outputs of events cameras are the sequences of asynchronous events over time rather than actual intensity images, most existing algorithms cannot be directly applied. Thus, although it has been recently shown that event cam- eras are sufficient to perform some tasks such as 6-DoF pose estimation and 3D reconstruction, it will be a great help if we can generate intensity images from events for other tasks such as object detection, tracking and SLAM.

In this paper, we first propose the event-based domain translation framework that generates better quality images from events compared with active pixel sensor (APS) frames and other previous methods. For this framework, two novel and initiative event stacking methods (see Figure.1) are also proposed based on shifting over the event stream, stacking based on time (SBT) and stacking based on the number of events (SBE), such that we can reach high frame rate and HDR representation with no motion blur, which is, in contrast, impossible for the normal cameras. It turns out that it is possible to generate a video with up to 1 million FPS using these stacking methods.

2. Proposed methods

2.1 Event stacking

In an event camera, each event e is represented as a tuple (u, v, t, p), where u and v are the pixel coordinates and t is the timestamp of the event, and $p = \pm 1$ is the polarity of the event, which is the sign of the brightness change ($p = 0$ for no event). These events are shown as a stream on the left of Fig. 1. Based on the frame rate of intensity camera, we have synchronized APS images and asynchronous events in between two consecutive APS frames. To feed event data input to the network, new representations of event data are required. When denoting the temporal resolution of an event camera by δ_t and the time duration by t_d, the size of the 3D volume is (w, h, n), where w and h represent the spatial resolution of an event camera and $n = t_d / \delta_t$. This is equivalent to have the n-channel image input for the network. This representation preserves all the information about events. However, the problem is that the number of channels is very huge. For this reason, we construct the 3D event volume with small n by forming each channel via merging and stacking the events within a small time

* These two authors contributed equally

interval.

2.2 Network architecture

To reconstruct HDR and high temporal resolution images and videos from events, we exploit currently available deep learning models, such as cGANs, as potential solutions for event vision. cGANs [1] are generative models that learn a mapping from observed image x and random noise vector *z* to the output image *y*, G: {x, z}→y (see Figure.2).

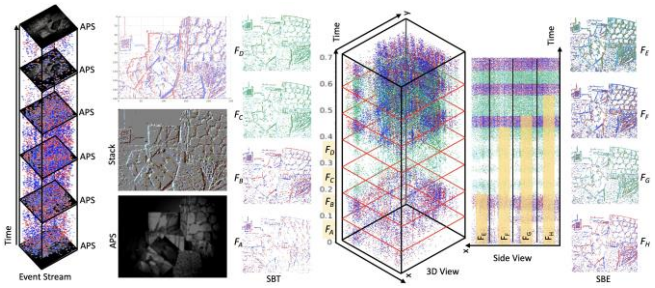

Figure 1. The event stream and construction of stacks by SBT and SBE. Two main color tuples of (Red(+), Blue(-)) and (Green(+), Cyan(-)) express the event polarity (plus, minus) throughout this paper.

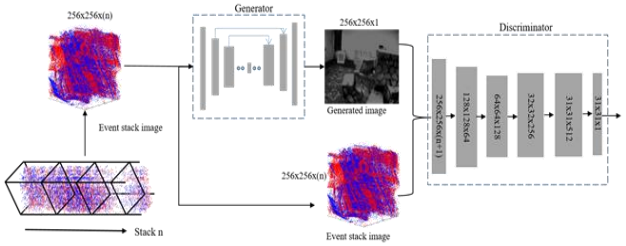

Figure 2. The proposed framework with the generator and discriminator networks.

3. Experiment and evaluation

To explore the capability of our method, we conduct intensive experiments on the both real-world and simulated datasets, and also use another open-source dataset with three real sequences (Face, jumping, and ball) [2] for comparison.

Figure 3. From left to right, input events, active pixel sensor (APS) images from the DAVIS camera, and our results. Our methods construct HDR images with more details that normal cameras could not reproduce as in APS frames.

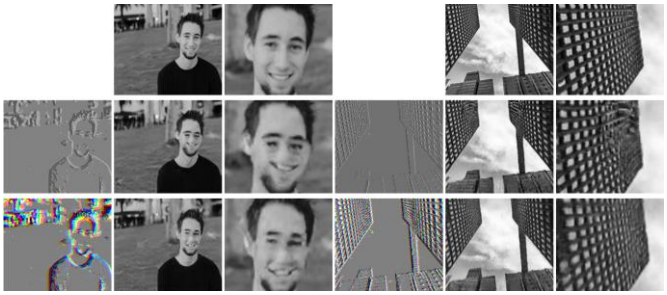

Figure 4. Reconstructed outputs from the inputs generated by simulator.

Table.1 Comparison of BRQUE scores with [2] and [3].

Sequence	Face	Jumping	Ball
Bardow [2]	22.27±8.81	29.39±7.27	29.37±9.61
Munda [21]	27.29±7.27	48.18±6.70	34.98±9.31
Ours(n = 3)	**48.26**±3.14	**48.34**±2.18	**39.18**±3.49

Figure.5 Comparison to the methods of Bardow *et al*. [2] and Munda *et al.* [3]

4. Conclusion

We demonstrated how our cGANs-based approach can benefit from the properties of event cameras to accurately reconstruct HDR non-blurred intensity images and high frame rate videos from pure events. We first proposed two initiative event stacking methods (SBT and SBE) for both image and video reconstruction from events using the network. We then showed the advantages of using event cameras to generate high dynamic range images and high frame rate videos through experiments based on our datasets made of online available real-world sequences and simulator.

감사의 글

This work was supported by National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (NRF-2018R1A2B3008640).

참고문헌

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image- to-image translation with conditional adversarial networks. *arXiv preprint*, 2017. 3, 4, 6.

[2] C. Reinbacher, G. Graber, and T. Pock. Real-time intensity-image reconstruction for event cameras using manifold regularisation. *arXiv preprint arXiv:1607.06283*, 2016.

[3] arXiv preprint, 2017. 3, 4, 6P. Bardow, A. J. Davison, and S. Leutenegger. Simultaneous optical flow and intensity estimation from an event camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 884–892, 2016. 1, 2, 5, 7, 8.