Event camera aided video frame interpolation

- a report for CS396 Natural and Artificial Vision 2023 Spring

Daizong Tian[a]

This manuscript was compiled on July 14, 2023

Conventional camera-based video frame interpolation (temporal video super-resolution, or VFI in this report) has been a very popular topic in the Computer Vision area. For optical flow-based VFI methods, one of the major challenges is to compensate for inaccurate flow generated from optical flow estimation modules (mainly neural networks). Optical flow estimation is an ill-posed problem since it's only estimated based on pixel values, and a pre-assumption needs to be made on object motions. An event camera can help VFI since it can record a much continuous motion in terms of high frame-rate events, thus could provide helpful information for VFI and bringing in little extra costs.

Event camera | Video frame interpolation | Super resolution | Optical flow

Trade-offs between video resolution (spatial quality) and frame rate (temporal quality) have always been an issue for producers and consumers. The trade-off can come from the limitation of sensor read-out speed for video recorders, the computational power from gaming consoles, the bandwidth for online video streamers, etc... To solve this problem, people usually capture/generate a high spatial but low temporal resolution video first and then generate more frames when displaying it from existing frames. This is also similar to how inter-frame video compression works: select some frames as high-quality I frame. The rest frames (B and P) only contain information that changed compared to the reference I frames. In VFI, it's called MEMC: motion estimation and motion compensation, which is very popular in TVs. In recent years NN-based VFI has become more and more popular. NVIDIA provides Super-resolution (spatial and temporal) support for its video cards.

However, conventional camera frame-based video frame interpolation has drawbacks, as described in the abstract. Luckily event camera perfectly provides the information that conventional camera miss but is critical to VFI: the motion information. Nowadays, event-camera and conventional camera combined VFI methods can achieve far better results than conventional camera-only VFI sota methods. (1–3)

Event camera, a bio-inspired motion extractor

Event-camera is human-vision-system (HVS) inspired. Event cameras are also called silicon retinas sometimes and it actually mimics some part of our HVS. Our human retina is sensitive to light intensity changes instead of absolute light intensity, which can be side proved by the fact that there are blood vessels and neurons in front of our retina but we cannot see them: Since the tissues relative position to the retina does not change, it does not provide any light intensity changes so our retinas ignore it.

We are all familiar with conventional cameras, which captures photo-generated electron within a certain amount of time to generate voltage, then convert it into digital pixel values. In other words, get the absolute light intensity within a certain amount of time. The event camera, however, does not capture the absolution intensity, it only captures if the light intensity becomes bright or darker to a certain step, and the outcome with on or off events for every pixel location (and no event if light intensity not changing). This brings two benefits, 1: it does not need to pre-set a shutter speed so events generate very fast, since a moving object usually continuously brings light intensity changes. And 2: A large part of the scene usually does not change, so the event camera will not and cannot capture them, which saves data bandwidth and computational power. Those two special features make the event camera a very high frame rate (or event rate) while maintaining low power consumption (usually 1/1000 of conventional camera sensors) (4).

One can imagine an event camera is a high-passed version of a conventional high-speed camera with an even higher speed of capturing motion. Although reconstructing a scene directly from the event camera is also a hot topic in Computer vision and computational imaging, currently combining conventional camera and event camera together usually brings better results since we can use the good pixel information from the conventional camera and good motion information from event camera.

Overview of frame-based, optical-flow-based video frame interpolation

Modern optical flow-based conventional camera video frame interpolation can mostly divide into two parts: optical flow estimation (motion estimation) and frame interpolation. FlowNet (5), pwc-net (6), and RAFT (7) are some popular optical flow estimation networks. Most networks estimate optical flow by two input frames based on pixel values or first or second-order statistics, which is an ill-posed problem since the time gap between two frames is large so the possible motion field is very large, not to mention different objects may share same statistics and the same object may have different statistics in different locations. Thus conventional camera-based optical flow estimation gives inaccurate flows, especially on smooth areas like backgrounds.

To overcome this problem, video frame interpolation methods based on optical flow estimation added additional modules in the interpolation part to correct and compensate for inaccurate optical flows. As an example, Super-Slomo (8) not only have an optical residual module to correct inaccurate optical flows generated from FlowNet (5) (and the error introduced during flow-reversal for backward warping), two of the four loss components are used to regulate flow estimation.

Another problem is that due to the huge gap between frames, the actual motion can never be re-constructed but only guessed based on assumptions. Super-Slomo (8) assumes linear motion so the intermediate frame will generate an object in between of locations of the previous and following frames. Quadratic Frame Interpolation (QVI) (9) assumes quadratic motion with the fact that a free-fall object will receive gravity so will have a quadratic motion, and thus needs 4 frames to estimate this

Author affiliations: [a]Northwestern Univerisity

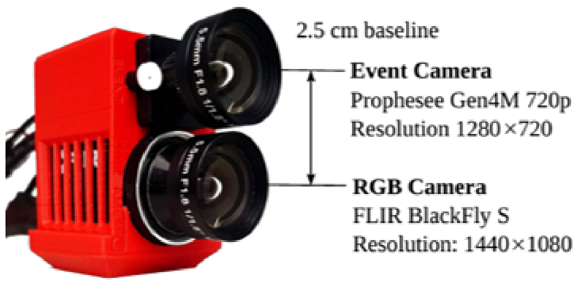
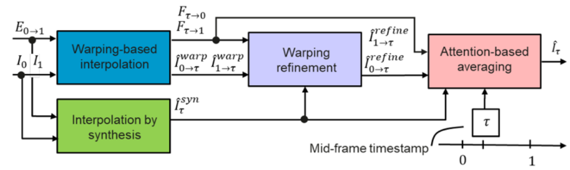


Fig. 1. Structure of TimeLens (1) proposed camera system, constructed with a conventional RGB camera and an event camera



(a) Overview of the proposed method.

Fig. 2. Structure of TimeLens (1), every colored box is a module with a separate Neural Network

quadratic motion with the assumption that the object motion among the 4 frames is continuous. However, those assumptions will not always agree with the ground-truth motion and it cannot possibly be truly reconstructed unless additional motion information is given.

Event camera aided Video Frame Interpolation

The event camera can fill in the gap that conventional camera-only video frame interpolation lacks. The event camera is a high-pass filtered camera that contains and only contains motion information, which is exactly the optical-flow-based VFI needed. In addition, the events already contain the pixel change information, thus extracting motion features (optical flow) by a NN and manually warping pixels might be unnecessary, synthesis-based video frame interpolation can simply take pixel info (frames) and pixel change info (events) to generate intermediate frames.

Optical-flow-based interpolation, and synthesis-based interpolation, both have their pros and cons.

Optical flow-based interpolation, as we mentioned earlier, generates a dense representation of motions and then warps pixel values from the original to destinations. It assumes pixel values of the same object not changing, thus it is struggling to handle illumination change and object reshaping. However, since it's based on dense motion maps and strict warping, the output quality could be ensured if the optical flow map is accurate, just might with wrong illumination.

Synthesis-based interpolation, on the other hand, does not make any assumption. Thus it can handle illumination and shape change if considering an ideal neural network. However, in the real world, due to lacking constraints compared to optical-flow-based methods, it sometimes generates undesired artifacts.

Fig. 3 shows a comparison between the two methods and the ground truth image. TimeLens (1) combines the good from both methods and proposed a conventional camera and event camera bounded video frame interpolation techniques. So I will use it as an example of how event cameras can help VFI. The hardware structure is shown in Fig. 1.

And it uses several Neural Networks with designed structures to combine the benefits from both conventional and event cameras, both optical-flow-based interpolation and synthesis-based interpolation together. The overall design is shown in Fig. 2 The backbone of all NNs are modified hourglass networks.

The warping-based module will use an hourglass network to generate optical flow using event-camera-generated data, warping the previous and following frame to the desired time spot to generate two separate frames. The synthesis-based network will use another hourglass network to receive both event-camera data and conventional camera RGB frames to generate a frame

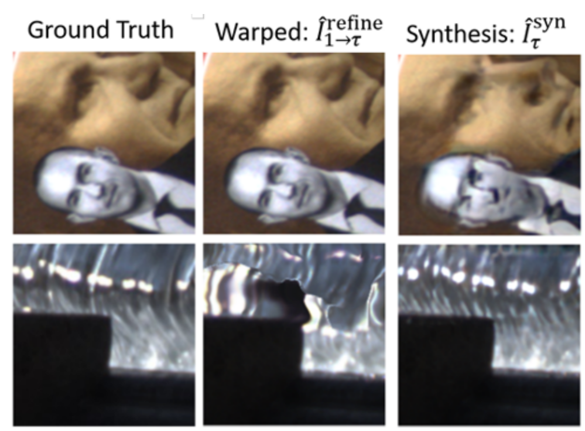


Fig. 3. Ground Truth V.S. Warping-based method V.S. Synthesis-based method. (1)

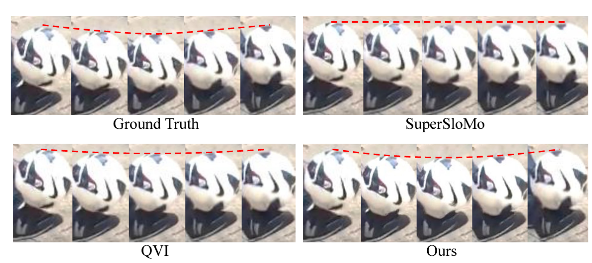


Fig. 4. Time Replayer (2) V.S. conventional camera VFI

at desired time spot. Then the synthesized result and warped result and optical flow will be fed into another hourglass network to compensate for possible optical flow errors like misalignment (notice that there is 2.5cm baseline in Fig.2 between two cameras). In the following attention module, there's a fourth hourglass network that takes both warped frames, optical flow, and synthesized frames to generate weights to average the 3 generated frames to get a final interpolated output. With 4 networks and information from both cameras, it achieves around 4-9 dB increment in terms of PSNR on popular datasets like GoPro(10), Vimeo90k(11). Notice that since those dataset does not provide event data, the events are actually simulated using an event simulator (12) with skipped frames.

TimeReplayer (2), another event and conventional camera combined VFI, proves that the combined method can record and thus interpolate the actual object movements. Fig.4 clearly shows the "Ours" (which means TimeReplayer's method(2)) recovers the football motion much better than linear-assumption SuperSloMo(8) and quadratic-assumption QVI(9).

Conclusion

The event camera, with its benefits of high temporal resolution and low data rate, perfectly fills the need for conventional camera video frame interpolations. I truly believe it will make VFI much easier with less power consumption and higher quality, and can come into a real-world application like how ToF lens becomes standard in today's smartphones.

I have slides also publicly viewable on https://makiseasuka.github.io/

Ethics

As an image-based neural network, this area has some potential ethics concerns: Firstly, it raises privacy issues as the system may inadvertently capture and process images that individuals may not wish to be shared. Moreover, inherent biases within the data used to train these neural networks can lead to biased outputs, potentially causing inequities or unfair treatment, especially considering event cameras might be more sensitive to some colors since it only records intensity changes and different colors may have different scale factor in terms of colors.

Therefore strict privacy regulations should be implemented regarding data collection and usage. Although there are public datasets available, only verified datasets should be used, also, to address bias, it's important to ensure that training datasets are diverse and representative of the populations. After training, detailed tests should be conducted for all different situations.

1 S Tulyakov, et al., Time lens: Event-based video frame interpolation in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* pp. 16155–16164 (2021).
2 W He, et al., Timereplayer: Unlocking the potential of event cameras for video interpolation in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* pp. 17804–17813 (2022).
3 J Dong, K Ota, M Dong, Video frame interpolation: A comprehensive survey. *ACM Transactions on Multimed. Comput. Commun. Appl.* (2022).
4 G Gallego, et al., Event-based vision: A survey. *IEEE transactions on pattern analysis machine intelligence* **44**, 154–180 (2020).
5 E Ilg, et al., Flownet 2.0: Evolution of optical flow estimation with deep networks in *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 2462–2470 (2017).
6 D Sun, X Yang, MY Liu, J Kautz, Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume in *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 8934–8943 (2018).
7 Z Teed, J Deng, Raft: Recurrent all-pairs field transforms for optical flow in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16.* (Springer), pp. 402–419 (2020).
8 H Jiang, et al., Super slomo: High quality estimation of multiple intermediate frames for video interpolation in *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 9000–9008 (2018).
9 X Xu, L Siyao, W Sun, Q Yin, MH Yang, Quadratic video interpolation. *Adv. Neural Inf. Process. Syst.* **32** (2019).
10 S Nah, T Hyun Kim, K Mu Lee, Deep multi-scale convolutional neural network for dynamic scene deblurring in *Proceedings of the IEEE conference on computer vision and pattern recognition.* pp. 3883–3891 (2017).
11 T Xue, B Chen, J Wu, D Wei, WT Freeman, Video enhancement with task-oriented flow. *Int. J. Comput. Vis.* **127**, 1106–1125 (2019).
12 D Gehrig, M Gehrig, J Hidalgo-Carrió, D Scaramuzza, Video to events: Recycling video datasets for event cameras in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* pp. 3586–3595 (2020).