High-Speed Motion Scene Reconstruction for Spike Camera via Motion Aligned Filtering

Jing Zhao, Ruiqin Xiong, Tiejun Huang
Institute of Digital Media, Peking University, Beijing, China
{jzhaopku, rqxiong, tjhuang}@pku.edu.cn

***Abstract*—A new retina-inspired bionic spike camera has recently shown great potential for capturing high speed movements. Unlike conventional cameras with a fixed low sampling rate, retina-inspired spike camera can well record fast-moving scenes by continuously accumulating luminance intensity and firing spikes. To restore the captured high-speed motion scenes from spike data, several reconstruction methods have been proposed. A typical method utilizes two neighbouring spikes to infer the instantaneous luminance intensity. Although high temporal resolution imaging can be achieved, the signal to noise ratio (SNR) of reconstructions is generally unsatisfactory. For improving the SNR, some methods propose to average the spikes in a big time window. However, the reconstructions may suffer from undesired motion blur, especially when there are objects moving very fast in scenes. To address this issue, we develop a new image reconstruction approach for potential retina-inspired spike camera to recover high-speed motion scenes. Specially, we take the motion of objects into consideration and exploit optical flow to align the scenes of different moments. After motion alignment, a filtering along motion trajectory can be employed to the signals to take the advantage of temporal correlations while not introducing undesired motion blur. Experimental results demonstrate that our proposed method achieves better visual quality than previous reconstruction schemes.**

I. Introduction

With the rapid development of emerging vision applications, such as autonomous driving, robotics and unmanned aerial vehicle, there is a growing demand of high temporal resolution imaging [1]. However, most conventional cameras utilize a fixed low sampling rate and compress the scenes during the exposure time into one frame, which results in serious motion blur in high speed scenarios. Although there are some high speed cameras with a very short exposure time in excess of 1/1000 second, the motion of high-speed objects in microseconds are still missing. Moreover, high speed cameras are generally very expensive [2].

Recently, to mimic the biological retina, many event-based cameras such as dynamic vision sensor (DVS) [3] and ATIS [4] have been proposed to record high speed motion. In contrast to conventional frame-based cameras, event-based cameras only capture variation of light intensity, which yields a very high temporal resolution of a few microseconds [5]. However, event-based cameras can hardly reconstruct textures in scenes as only change information is recorded. Although there are some hybrid sensors combing DVS with conventional image sensors [6] or photo-measurement circuits [7] to reconstruct textures, the mismatch caused by different sampling rate can not be well removed.

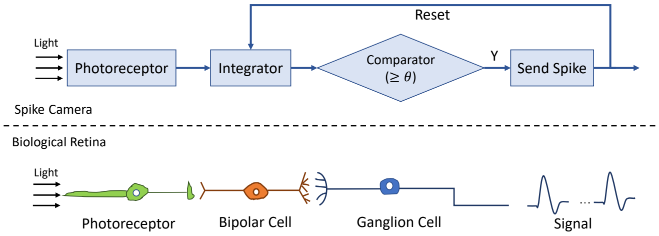


Fig. 1. Illustration of retina-inspired spike camera.

Besides, most of such bio-inspired approaches focus on modeling peripheral vision. However, in human visual system, visual details are acquired by the fovea, which is located in the center of the primate retina and responsible for sharp vision [8], [9]. Inspired by the signal processing of visual neurons, a retina-inspired spike camera with a high temporal resolution of 1/40000 second was proposed for recording high speed moving scenes [10]–[12]. Although some reconstruction methods only taking the advantage of the properties of spikes have been proposed for the spike camera, the performance, such as signal to noise ratio (SNR), is still unsatisfactory.

Addressing this issue, we develop a new texture reconstruction approach for retina-inspired spike cameras, which improves the SNR of reconstructions by utilizing temporal correlations of signals. In order to exploit the relationships between the scenes of different moments, a series of basic reconstructions are firstly estimated using the conventional reconstruction methods. Considering the motion in scenes, we align the basic reconstructions of different moments by taking the advantage of optical flow. Finally, a filtering along temporal direction is applied on the aligned images to reconstruct the ultimate images stably while not introducing motion blur.

This paper is organized as follows. Section II presents the retina-inspired spike camera and introduces the reconstruction methods for moving scenarios and stationary scenarios, respectively. Section III describes the proposed texture reconstruction scheme. Section IV reports the experimental results and Section V concludes the paper.

This work was supported by the National Natural Science Foundation of China (61425025, 61772041).

978-1-7281-3320-1/20/$31.00 ©2020 IEEE

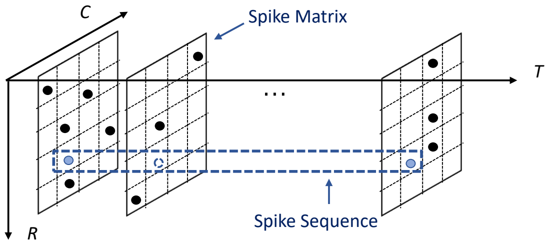


Fig. 2. Spike array generated by retina-inspired spike camera.

II. Retina-inspired Spike Camera

We first briefly reviews the workflow of retina-inspired spike camera. As shown in Fig. 1, the spike camera consists of three major components, "photoreceptor-integrator-comparator", which simulates the one-to-one connections of "photoreceptor-bipolar cell-ganglion cell" in fovea centralis. By continuously capturing the luminance intensity, the integrator accumulates electric charges from the photoreceptor with the aid of photoelectric conversion. Once the total amount of electric charges reaches a threshold, a spike is fired and the model enters a new "integrate-and-fire" cycle. For recording high speed motion, the spike camera adopts a 40000HZ polling, and the raw data can be represented with a $R \times C \times T$ spike array composed of several spike matrixes as shown in Fig. 2. Here solid point denotes that there is a spike fired, $R \times C$ is spatial resolution and T represents the number of total polling times.

To restore the captured scene and bridge the gap between the asynchronous bionic spike data and conventional frame-based vision, several visual texture reconstruction schemes have been proposed. For high-speed motion applications, the luminance of each pixel is changing rapidly. Thus, only two spikes are utilized to estimate a pixel value, as formulated by

$$P_{t_i}(r,c) = \frac{\theta}{\mathrm{ISI}_{t_i}(r,c)}, \tag{1}$$

where $P_{t_i}(r,c)$ is the pixel value at the moment of t_i in pixel (r,c), θ denotes the maximum dynamic range of reconstructions and $\mathrm{ISI}_{t_i}(r,c)$ represents the inter-spike interval at position (r,c), which is equal to the time between two neighbouring spikes of moment t_i.

For stationary scenes, the spike firing characteristics rarely change and more spikes can be used to reduce the effect of noises. To this end, a big time window is adopted to play back the spikes in a specific period. By counting the spikes in the same spatial position, we have

$$P_{t_i}(r,c) = \theta \cdot \frac{\mathcal{N}_w(r,c)}{w}, \tag{2}$$

where $P_{t_i}(r,c)$ is the pixel value at the moment of t_i, θ denotes the maximum dynamic range, w is the size of time window and $\mathcal{N}_w(r,c)$ represents the total number of spikes collected in the time window at position (r,c).

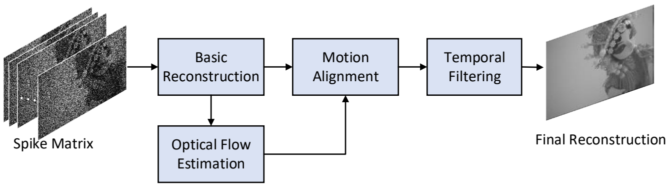


Fig. 3. The framework of proposed reconstruction scheme.

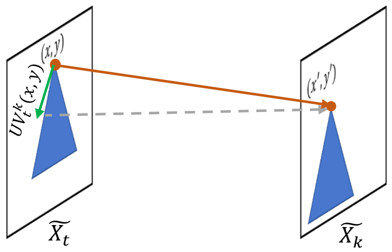


Fig. 4. Motion alignment via optical flow.

III. Proposed Reconstruction via Motion Aligned filtering for Spike Camera

We note that both of the existed reconstruction schemes have some limitations. For the first scheme, which calculates pixel values by averaging the luminance intensity in one inter-spike interval (ISI), the reconstructions are generally with lots of noises as a ISI is too short to make signals stable enough. Although the second scheme selects a big time window to restore signals more stably, it ignores the influence of motions. When there are objects moving very quickly, the luminance intensity accumulated to the same spatial position may be completely different. Thus, simply enlarging time window and averaging the spikes in the same spatial position, as introduced in the second scheme, may cause undesired motion blur.

To address these issues, this paper develops a new reconstruction approach for retina-inspired cameras to exploit the correlations of spikes in a big time-window, while considering the motions in scenes. Fig. 3 shows the framework of the proposed method.

A. Motion Alignment

Suppose X_t is the image of the scene at moment t to reconstruct, while $S \in \mathbb{R}^{R \times C \times T}$ is the array of raw spikes in the time window around t. Before applying a temporal filtering to exploit the correlations of different scenes, we are supposed to align the scenes first. To this end, a series of basic images $\{\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_T\}$, which can roughly represent the instantaneous scenes of different moments, are first reconstructed using Eq. (1). Then, the local relative motion between key image $\tilde{X}_t$ and other image $\tilde{X}_k$ can be estimated by optical flow methods. In this paper, we adopt the classic optical flow estimation algorithm by Horn et al. [13]. It assumes constancy of some image property with a spatial term that models how

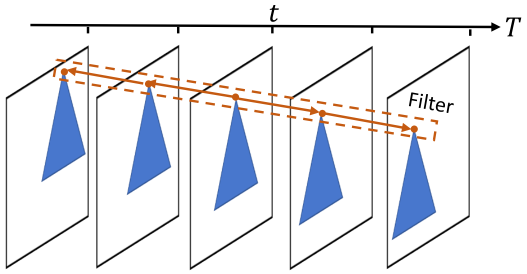


Fig. 5. An example of temporal filtering when time-window is equal to 5.

```
Algorithm 1 Image reconstruction via motion aligned filtering
Input: Array $S$ of raw spikes in the time window around $t$
Output: Ultimate reconstruction $X_t$
Reconstruct basic estimate $\tilde{X}_t$ of moment $t$ using Eq. (1)
for $k = 1 \to T$ do
    if $k \neq t$ then
        Reconstruct basic estimate $\tilde{X}_i$ via Eq. (1)
        Estimate the optical flow from $\tilde{X}_t$ to $\tilde{X}_k$ via Eq. (3)
        Align $\tilde{X}_k$ to $\tilde{X}_t$ based on optical flow via Eq. (4)
    end if
end for
Apply temporal Gaussian filtering on the aligned basic
images using Eq. (5) for final reconstruction $X_t$
```

the flow is expected to vary across the image. We represent the relative motion as

$$\begin{bmatrix} U_t^k \\ V_t^k \end{bmatrix} = HS(\tilde{X}_t, \tilde{X}_k). \tag{3}$$

As shown in Fig. 4, with the aid of relative motion, the spatial relationship between the scene of moment t and neighboring moments can be modeled. Given a pixel (x, y) on $\tilde{X}_t$, its corresponding position (x', y') on $\tilde{X}_k$ can be calculated as follow.

$$\begin{aligned} x' &= x + U_t^k(x, y) \\ y' &= y + V_t^k(x, y) \end{aligned} \tag{4}$$

We denote the mapping as $\mathcal{P}_{t\to k} : (x, y) \to (x', y')$. Similarly, the corresponding positions on other scenes can also be calculated.

B. Temporal Filtering

We assume that there is brightness constancy between corresponding positions in neighboring scenes. As shown in Fig. 5, after motion alignment, a temporal filtering can be applied to the basic images to exploit temporal correlations so that the ultimate reconstruction becomes more stable. It should be noted that (x, y)'s corresponding position (x', y') may be a sub-pixel as the derived relative motion may be decimal. In this paper, we adopt the bilinear interpolation, which takes the advantage of the four nearest neighboring pixels, to calculate the luminance of (x', y'). Then, the ultimate pixel value of pixel (x, y) can be calculated by applying a filtering along temporal direction as

TABLE I
DETAIL INFORMATION OF SPIKE SEQUENCE

Name	Size	Sampling rate	Description
Car	400×250×1600	40000 HZ	run in 100km/h
Doll	400×250×400	40000 HZ	fall freely

$$X_t(x, y) = \frac{1}{C} \sum_{k=1}^{T} w_t^k \cdot \tilde{X}_k\left(\mathcal{P}_{t\to k}(x, y)\right) \tag{5}$$

where $\mathcal{P}_t^k(x, y)$ is the corresponding position of (x, y) on $\tilde{X}_k$, w_t^k denotes the weight of the scene at moment k on reconstructing the scene at moment t and $C = \sum_{k=1}^{T} w_t^k$ is the normalization factor. In this paper, we adopt Gaussian filtering, in which farther scenes are awarded smaller weights and the weight is defined as

$$w_t^k = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{\frac{-(t-k)^2}{2\sigma^2}}, \tag{6}$$

where σ is standard deviation of distribution. In this paper, it is determined by time window.

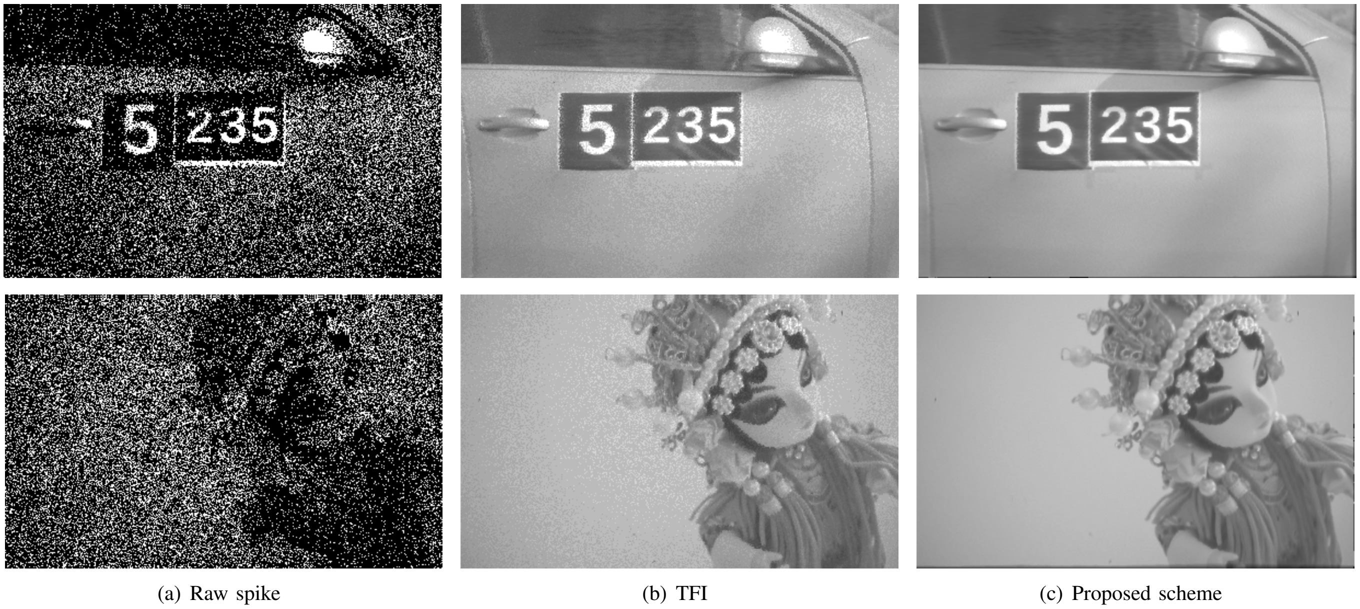
In summary, motion alignment and temporal filtering are two key processes of our proposed reconstruction method and a global overview is outlined in Algorithm 1.

IV. Experimental Results

To evaluate the performance of our proposed reconstruction method on recording high-speed moving scenes, we compare our method with two conventional texture reconstruction methods, i.e., “Texture from ISI (TFI)” and “Texture from Playback (TFP)” [10]. We conduct our experiments on spike sequences captured by the spike camera. Table I shows the detail of each spike sequence.

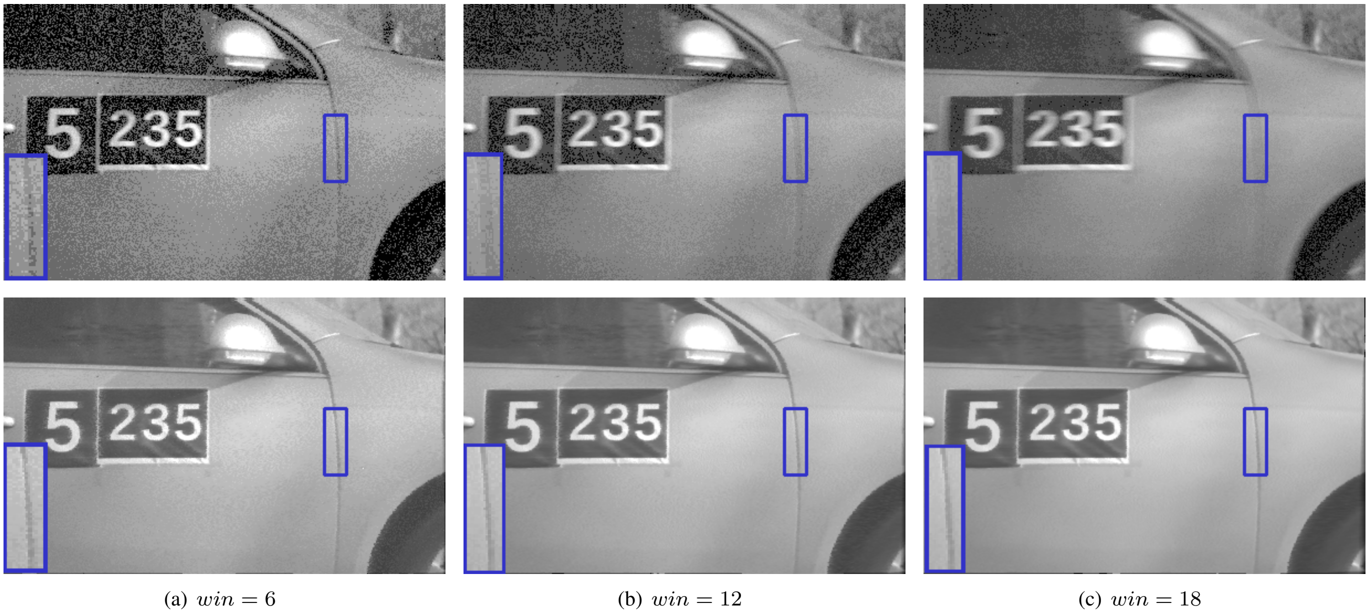
Fig. 6 shows the visual results compared with TFI. We note that TFI can well reconstruct the outline of fast moving objects. However, as the number of spikes used for reconstruction is limited, the signal to noise ratio (SNR) is unsatisfactory. For example, there are many shining points on the body of car. In contrast, our proposed method provides a stable reconstruction, while restoring clear textures and details.

Fig. 7 shows the visual results compared to TFP with different time window sizes. We note that the visual performance of proposed method is always superior to TFP. When the size of time window is small, the reconstruction of TFP is noisy, which is similar to TFI. As the size of time window increases, there is severe motion blur on the reconstructions of TFP, which seriously influences visual performance. Conversely, the visual performance of our proposed method is improved with the size of time window increasing. In particular, we note that when the size of time window is big enough, our proposed method achieves stable reconstruction while not introducing undesired motion blur.



(a) Raw spike (b) TFI (c) Proposed scheme

Fig. 6. Visual results compared with TFI. Please enlarge the figure for better comparison.



(a) $win = 6$ (b) $win = 12$ (c) $win = 18$

Fig. 7. Visual results compared to TFP with different time window sizes. Top row: TFP. Bottom row: Proposed scheme. Please enlarge the figure for better comparison.

V. Conclusion

In this paper, we focus on the inferior performance of conventional reconstruction methods on retina-inspired spike camera, and develop a new reconstruction approach for the spike camera. We take the motions of scenes into consideration and align the scenes of different moments by integrating optical flow. To take the full advantage of temporal correlations between different scenes, we conduct a temporal filtering on the aligned scenes in a relatively big time window in order to improve the performance of reconstructions. Experimental results confirm that the proposed method achieves competitive reconstruction performance, which can not only recover sharp textures and fine details but also produce promising perceptual quality in the smooth regions.

References

[1] MOESLUND, B. Thomas, HILTON, Adrian, KRGER, and Volker, "A survey of advances in vision-based human motion capture and analysis," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[2] M. G. Ballarotti, M. M. F. Saba, and O. P. Jr, "High-speed camera obsevations of negative ground flashes on a millesecond-scale," *Geophysical Research Letters*, vol. 32, no. 23, pp. 338–360, 2005.

[3] P. Lichtsteiner, C. Posch, and T. Delbruck, “A 128$\times$128 120 db 15μs latency asynchronous temporal contrast vision sensor,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[4] C. Posch, D. Matolin, and R. Wohlgenannt, “A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds,” *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, 2010.

[5] M. V. Srinivasan, R. J. D. Moore, S. Thurrowgood, D. Soccol, and D. Bland, *From biology to engineering: Insect vision and applications to robotics*, 2012.

[6] C. Brandli, R. Berner, M. Yang, S. C. Liu, and T. Delbruck, “A 240x180 130db 3us latency global shutter spatiotemporal vision sensor,” *IEEE Journal of Solid-State Circuits*, 2014.

[7] C. Posch, D. Matolin, and R. Wohlgenannt, “An asynchronous time-based image sensor,” in *IEEE International Symposium on Circuits and Systems*, 2008.

[8] S. Gould, J. Arfvidsson, A. Kaehler, B. Sapp, and A. Y. Ng, “Peripheral-foveal vision for real-time object recognition and tracking in video,” *Ijcai*, pp. 2115–2121, 2007.

[9] M. Robles and Rebeca, *Merging Attention and Segmentation: Active Foveal Image Representation*, 2013.

[10] L. Zhu, S. Dong, T. Huang, and Y. Tian, “A retina-inspired sampling method for visual texture reconstruction,” in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, July 2019.

[11] S. Dong, L. Zhu, D. Xu, Y. Tian, and T. Huang, “An efficient coding method for spike camera using inter-spike intervals,” in *Data Compression Conference, DCC 2019, Snowbird, UT, USA, March 26-29, 2019*, 2019, p. 568.

[12] S. Dong, T. Huang, and Y. Tian, “Spike camera and its coding methods,” in *2017 Data Compression Conference, DCC 2017, Snowbird, UT, USA, April 4-7, 2017*, 2017, p. 437.

[13] B. K. P. Horn and B. G. Schunck, “Determining optical flow,” *Artificial Intelligence*, vol. 17, no. 1C3, pp. 185–203, 1980.