Retinomorphic Sensing: A Novel Paradigm for Future Multimedia Computing

Zhaodong Kang[1]†, Jianing Li[2]†, Lin Zhu[2], Yonghong Tian[2,3]*
{kzd,lijianing,linzhu,yhtian}@pku.edu.cn
[1]School of Electronic and Computer Engineering, Peking University, Beijing, China
[2]Department of Computer Science, Peking University, Beijing, China
[3]Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Conventional frame-based cameras for multimedia computing have encountered important challenges in high-speed and extreme light scenarios. However, how to design a novel paradigm for visual perception that overcomes the disadvantages of conventional cameras still remains an open issue. In this paper, we propose a novel solution, namely retinomorphic sensing, which integrates fovea-like and peripheral-like sampling mechanisms to generate asynchronous visual streams using a unified representation as the retina does. Technically, our encoder incorporates an interaction controller to switch flexibly between dynamic and static sensing. Then, the decoder effectively extracts dynamic events for machine vision and reconstructs visual textures for human vision. The results show that our strategy enables it to sense dynamic events and visual textures meanwhile reduce data redundancy. We further build a prototype hybrid camera system to verify this strategy on vision tasks such as image reconstruction and object detection. We believe that this novel paradigm will provide insight into future multimedia computing. The code can be available at https://github.com/acmmm2021-bni-retinomorphic/retinomorphic-sensing.

KEYWORDS

Retinomorphic sensing, multimedia computing, silicon retina, neuromorphic vision

ACM Reference Format:
Zhaodong Kang[1]†, Jianing Li[2]†, Lin Zhu[2], Yonghong Tian[2,3]. 2021. Retinomorphic Sensing: A Novel Paradigm for Future Multimedia Computing. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3479237

1 INTRODUCTION

What will take place in the real-world within the 33 milliseconds after pressing the shutter? A running car at 140 kph on the expressway can move over 1 meter, a flying bullet at 1800 kph can go through 10 meters, and a rushing rocket at 7200 kph can pass up to 20 floors. However, a conventional RGB camera at 30 FPS will take a photo with a blurred picture. Take a car for example, a horrible accident may occur within the short period between two adjacent frames. Meanwhile, is the existing multimedia computing paradigm [4, 16] that "Computer Vision = Conventional frames + Computer + Algorithms" suitable for every scenario? However, few people doubt whether it can be applied to future multimedia computing. In fact, conventional frames have some limitations in challenging scenarios (e.g., high-speed, low-light, and over-exposure), resulting in low-quality imaging for human vision and performance degradation for machine vision [12, 32]. Therefore, we aim to design a novel visual perception paradigm for future multimedia to overcome the limitation of conventional frame-based cameras.

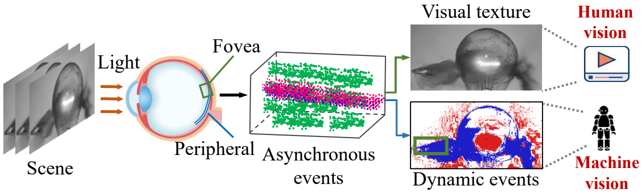


Figure 1: *Retinomorphic sensing system*, integrating fovea-like and peripheral-like visual sampling mechanisms as the retina does, generates asynchronous events using a unified representation, which can effectively reconstruct static information (i.e., visual texture) for human vision and extract effectively dynamic events for machine vision, such as high-speed motion bullet.

Recently, neuromorphic vision sensors [5, 9, 17, 20], namely silicon retinas, imitating biological visual systems, have been gaining more and more attention in computer vision [26, 31] owning to the advantages over conventional cameras: high temporal resolution, high dynamic range, and low power consumption. According to sampling mechanisms, neuromorphic cameras can be broadly classified into two types (i.e., event cameras [24] and spike cameras [8]).

Event cameras (e.g., DVS [15], DAVIS [2], and ATIS [23]) in the former type, mimicking the periphery of the retina, work radically in a different way from frame-based cameras. Each pixel independently responds to intensity changes (i.e., dynamic information) with a stream of asynchronous events using address event representation (AER) [3]. Since asynchronous events are generated only when light changes in scenarios, event cameras are natural motion detectors [12] and perfect for motion sensing [14, 22]. However, only processing DVS events may be hard to reconstruct high-quality visual texture and obtain high-precision recognition.

*Corresponding author: Yonghong Tian (Email: yhtian@pku.edu.cn).
†The first two authors contributed equally to this research.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3479237

Spike cameras (e.g., octopus retina [6], and Vidar [8, 38]) in the latter type, namely time-based vision sensors, adopt the fovea-like sampling model. Each pixel independently generates spikes when the accumulation of photos reaches a threshold. In other words, this brings the ability to reconstruct visual textures using spike frequency or inter-spike interval [41]. An example is Vidar, each pixel is reset asynchronously, and then spikes at the same sampling timestamp are readout synchronous using spike plane. It has a high temporal sampling frequency of 40,000 Hz and is suitable for high-speed vision tasks. However, this integrating manner is hard to provide such high temporal dynamic range as the DVS does. What's more, another drawback of Vidar is high data redundancy due to the spike firing with high frequency for static scenes.

Generally speaking, both two types of neuromorphic cameras have difficulties and limitations in encoding light intensity with both dynamic and static information output. Actually, processing in peripheral and foveal vision is not independent, but is more directly connected than previously thought [29, 30, 33]. In other words, the retina is combined with the fovea and the periphery to sense real-world scenarios. It motivates us to ask: *Can we design a neuromorphic visual sensing system that not only follows the functions of the periphery and fovea of the retina to obtain effectively dynamic and texture information, but also uses a unified representation for two streams towards machine vision and human vision?*

To this end, we put forward a novel paradigm for future multimedia, namely retinomorphic sensing, which integrates fovea-like and peripheral-like visual sampling mechanisms to generate asynchronous streams using a unified representation (i.e., AER), as illustrated in Fig. 1. This visual sensing system can reconstruct high-quality textures for human vision and extract dynamic information for machine vision, which includes retinomorphic encoder, dynamic interacting controller, and retinomorphic decoder. In fact, the goal of this work is not to design a simulator that models the existing realistic neuromorphic cameras (e.g., DVS or Vidar). On the contrary, we aim at overcoming the following challenges: (i) How to design a unified representation to satisfy the machine vision and human vision; (ii) How to dynamically control and switch flexibly between the two modes (i.e., fovea-like and peripheral-like sampling).

In summary, the main contributions are summarized as follows:

- We propose a novel concept of retinomorphic sensing, which overcomes the limitations of conventional frame-based sampling, brings a new paradigm for future multimedia and takes both machine vision and human vision into account.
- We design a dynamical interacting controller using recurrent neural networks, which integrates fovea-like and peripheral-like sampling mechanisms to generate asynchronous streams using a unified representation as the retina does.
- We build a prototype hybrid camera system to verify this strategy on tasks (e.g, image reconstruction and object detection), aiming to solve the disadvantages of conventional cameras, especially in extreme challenging scenarios.

To the best of our knowledge, this is the first work to explore such a visual sensing system to integrate fovea-like and peripheral-like sampling as the retina does. We believe this prototype will provide insight into developing the next-generation neuromorphic vision sensor for future multimedia computing.

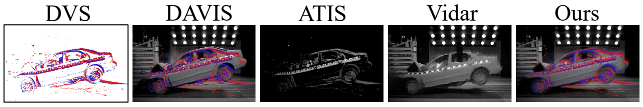


Figure 2: Visualization representation for the existing event cameras (i.e., DVS [15], DAVIS [2], ATIS [23], and Vidar [8]) and our visual sensing system. Note that, our work can effectively sense both dynamic events and visual textures using a unified representation.

Table 1: Attribution analysis for neuromorphic cameras.

Types	DVS [15]	Vidar [8]	ours
Dynamic events	✓		✓
Visual textures		✓	✓
Data (Mbps)	1046.9	5310.9	3497.4

2 RELATED WORKS

This section reviews the related works, which mainly focus on the limitations of frame-based sampling and the existing neuromorphic visual sampling.

Frame-Based Sampling. Conventional cameras have achieved great success for computational photography and computer vision under some conditions [19, 21, 34, 37], but this sampling manner, capturing visual scenes at the fixed frame-rate, has presented several limitations in challenging scenarios (e.g., high-speed, low-light, and over-exposure), resulting in unusable frames and performance degradation for vision tasks. Although some works have designed new types of high-speed cameras [7, 39], the large data size of which brings the difficulty for real-time transmission and storage [28]. Besides, too many frame-based enhanced algorithms have attempted to improve the imaging quality (e.g., de-blurring [10, 13] and HDR imaging [11, 40]) for human vision, but the posting-processing enhancement is hard to solve fundamentally the bottlenecks of conventional frame-based sampling paradigm.

Neuromorphic Visual Sampling. Recently, two types of neuromorphic vision sensors (i.e., event cameras and spike cameras) have emerged. Event cameras (e.g., DVS [15], DAVIS [2], and ATIS [23]) independently respond to intensity changes (i.e., dynamic information) with asynchronous events, thus they may be hard to reconstruct high-quality visual texture and obtain high-precision recognition. To address the shortcoming, several cameras have been developed to output dynamic and static information. For example, DAVIS integrates a traditional active pixel sensor (APS) in the same pixel with DVS. Although APS frames can be provided at a pre-fixed rate, its readout is limited to dynamic range and temporal resolution like a standard camera, and two independent signals may not match for fast-moving objects. Besides, ATIS combines DVS and a conditional exposure measurement circuit triggered by a light change detection. The intensity for one pixel needs to transmit two events coding the temporal interval. Nevertheless, this way brings the disadvantage that only dynamic pixels provide their new intensity values rather than global information. Also, the temporal interval between two events may be long in low light scenes and the intensity measurement process can be interrupted by a new event. Spike cameras (e.g., octopus retina [6], and Vidar [8, 38]) encode the light intensity into spikes when the accumulation of

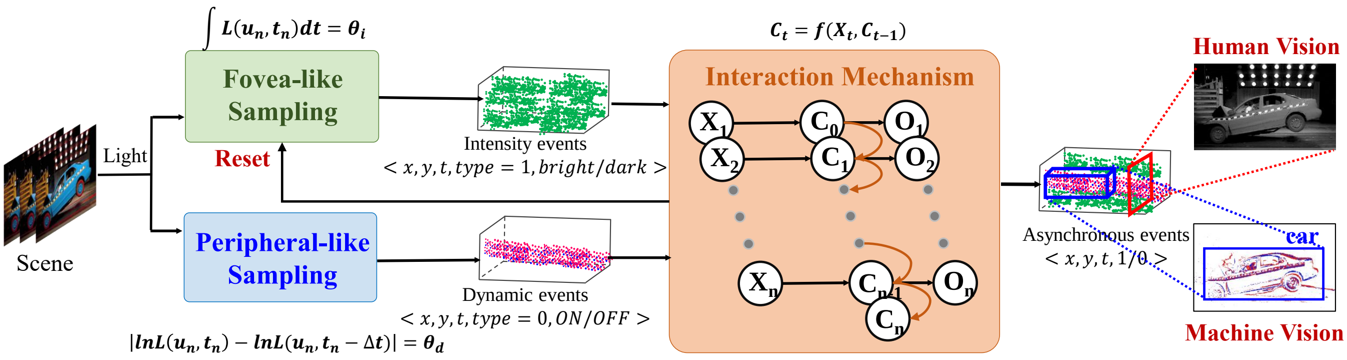


Figure 3: The pipeline of the proposed retinomorhic sensing framework. Initially, we design two sensing modules using peripheral-like sampling and fovea-like sampling mechanisms, respectively. Then, the interaction mechanism, using the recurrent neural network, can switch flexibly between two sensing modules to output both intensity events for human vision and dynamic events for machine vision.

photos reaches the presetting threshold. Although this integrating manner brings the ability to reconstruct visual textures, it is hard to provide high temporal resolution dynamic events for machine vision as the DVS does.

Therefore, this work will design a novel visual sampling manner to output both dynamic events for machine vision and texture information for human vision, which aims at overcoming the limitations of conventional frame-based sampling.

3 PROBLEM STATEMENT

In this section, we will summarize some shortages of the existing neuromorphic visual sampling manners by analyzing visualization results and data statistics.

Visualization Representation. As shown in Fig. 2, we compare our visual sensing system with other sampling manners in the existing neuromorphic cameras (i.e., DVS [15], DAVIS [2], ATIS [23], and Vidar [8]) from a visualization representation perspective. Apparently, DVS produces only dynamic events without static information (i.e., visual textures). Although DAVIS outputs visual texture via incorporating the APS, it exists two shortages that APS frames result in motion blurring and two independent signals fail to match for high-speed scenarios. Unfortunately, ATIS only provides local visual textures in high dynamic areas rather than global information. Vidar can reconstruct high-quality texture without providing high-speed dynamic events as the DVS does. On the contrary, our visual sensing system can sense both dynamic events for machine vision and texture information for human vision.

Data Statistics. We convert a short video into asynchronous streams using the mechanisms of DVS, Vidar [8] and our simulator, and we further make attribution analysis for our visual sensing system and two types of neuromorphic cameras in Table 1. We can see that Vidar provides visual textures without sensing high-speed dynamic events as the DVS does. Besides, another drawback of Vidar is high data redundancy due to the spike firing with high temporal sampling frequency for static scenes. Notably, our simulator can generate not only dynamic events for machine vision as DVS does, but also obtain better visual texture than Vidar (details in Section 5.2) for human vision within low data redundancy.

In this work, we propose a new solution, integrating fovea-like and peripheral-like visual sampling, which can sense both dynamic events and visual textures using a unified representation meanwhile reduce data redundancy.

4 METHODOLOGY

This section will first start with a brief overview of our framework. Then, we elaborate on the details of how to encode dynamic events and intensity events using a unified representation. Finally, we present the decoding strategy to obtain dynamic information and reconstruct visual texture from asynchronous events.

4.1 Framework Overview

Our goal is to convert continuous light intensity $L(\boldsymbol{u_n}, t_n)$ into asynchronous events using a unified representation (i.e., AER). Generally, an event e_n can be represented as a tuple $\langle x_n, y_n, t_n, p_n \rangle$, which is generated from one pixel $\boldsymbol{u}_n = [x_n, y_n]$ at the sampling timestamp t_n. As shown in Fig. 3, our retinomorphic sensing framework includes three parts: *fovea-like sampling*, *peripheral-like sampling*, and *interaction mechanism*. Specifically, our encoder senses the lightness and generates events using peripheral-like sampling and fovea-like sampling mechanisms, respectively. Then, the interaction mechanism using the recurrent neural networks, can switch flexibly between two sensing modules to output asynchronous events. After that, the corresponding decoding framework can obtain dynamic events for machine vision and reconstruct intensity information (i.e., visual texture) for human vision.

4.2 Retinomorphic Encoding Framework

For *fovea-like sampling* part, each pixel outputs an intensity event $\langle x_n, y_n, t_n, q_n \rangle$ once the accumulate of the light intensity $L(\boldsymbol{u_n}, t_n)$ reaches a presetting threshold θ_i. Intuitively, the brighter the illuminance, the higher frequency the event generating, and it can be depicted as follows:

$$\int L(\boldsymbol{u_n}, t_n)dt = \theta_i, \tag{1}$$

where the small integrating window dt (i.e., Δt) may result in an ultra high sampling frequency. At each sampling timestamp, the attribute $p_n \in \{1, 0\}$ of an event denotes bright or dark scenario, respectively. There are two integrators of each pixel and the other integrates the value of $L_{max} - L(\boldsymbol{u_n}, t_n)$, and it will first reach the threshold under dark scenario.

For *peripheral-like sampling* part, each pixel independently responds to changes in the illuminance $L(\boldsymbol{u_n}, t_n)$. A dynamic event e_n can be represented as a tuple $\langle x_n, y_n, t_n, p_n \rangle$, which is generated from one pixel $\boldsymbol{u_n} = [x_n, y_n]$ at the time t_n when the intensity change reaches the threshold θ_d, and it can be described as:

$$\Delta lnL \doteq |lnL(\boldsymbol{u_n}, t_n) - lnL(\boldsymbol{u_n}, t_n - \Delta t_n)| = \theta_d, \tag{2}$$

where the polarity $p_n \in \{1, 0\}$ refers to ON or OFF event respectively, which represents the increasing or decreasing change in the brightness, and Δt_n is the time since the last event at a pixel $\boldsymbol{u_n}$.

For *interaction mechanism*, each pixel maintains a cell unit of the recurrent neural network for the record of the past events. If a new event (dynamic or intensity) with time t_n reaches the interaction controller, the recurrent neural network will take a vector (i.e., its time and type) as input X_n and update the state of the cell by:

$$\hat{C}_n = tanh(W_c[X_n, C_{n-1}]), \tag{3}$$

where W_c is the preset parameters, and C_{n-1} is the last status that the network maintains.

The update gate control G_u for this event can be represented by:

$$G_u = sigmoid(W_u[X_n, C_{n-1}]), \tag{4}$$

where W_u is the pre-setting parameters.

Thus, we can compute new cell status of the recurrent neural network by:

$$C_n = G_u * \hat{C}_n + (1 - G_u) * C_{n-1}, \tag{5}$$

where C_n is the new cell status and used for the next event. We let the cell unit also represent the hidden state. The neural network outputs the current event and resets the accumulators of fovea-like sampling part if G_u is no less than 0.5.

For the spatio-temporal windows Γ_s, the output of asynchronous events $S = \{\langle \boldsymbol{u_n}, t_n, p_n \rangle \mid \boldsymbol{u_n}, t_n \in \Gamma_s, n = 1, ..., N\}$ generated from our retinomorphic encoder can be formulated as:

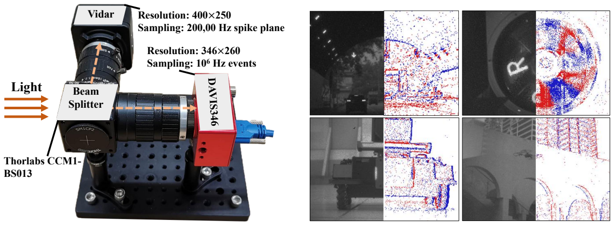
$$S = \bigcup_{n=1}^{N} \{p_n \delta (x - x_n, y - y_n, t - t_n)\}, \tag{6}$$

where N is the number of events in the spatio-temporal windows Γ_s, and $\delta(\cdot)$ refers to the Dirac delta function.

4.3 Retinomorphic Decoding Framework

Retinomorphic decoding framework is to obtain dynamic events S_d and visual textures (i.e., light intensities L) from asynchronous events after the encoder. Specifically, event sequence for one pixel is firstly regarded as the basic processing unit, and we decode each pixel asynchronously in parallel. Then, each event in a sequence can be distinguished as a dynamic event or intensity event using their symbols. Finally, dynamic events are natural moving flows to serve for machine vision, and visual texture can be reconstructed towards high-quality imaging for human vision.

For *intensity events*, visual texture can be reconstructed using the inter-event interval between two events because light intensity $L(\boldsymbol{u_n}, t_n)$ is converted into event sequence under fovea-like sampling. Thus, light intensity for one pixel can be estimated as:



(a) Experimental setup (b) Calibrated views

Figure 4: A hybrid camera system combines DVS and Vidar. (a) A beam splitter is placed in front of two cameras with 50% splitting. (b) Examples of the shared view, the left in each image is the visual texture for Vidar, and the right is event image.

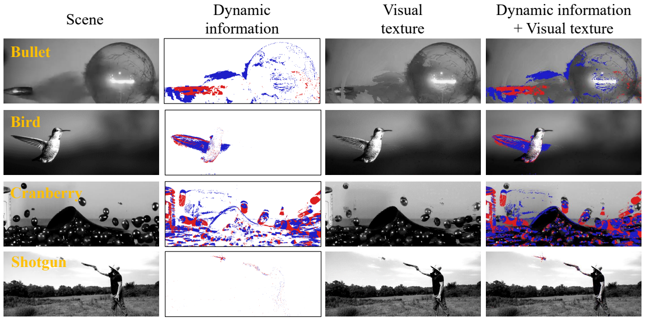


Figure 5: Representative results in high-speed moving scenarios (i.e., flying *bullet*, *bird* flapping wings quickly, flying cranberry, and a *shotgun* taking aim at the moving object). Our framework converts videos into asynchronous events including dynamic and static information. For better visualization, we map dynamic events into event images in the second column, and intensity events are reconstructed visual textures in the third column. Besides, dynamic events and visual textures are merged into frames in the last column.

$$L(\boldsymbol{u_n}, t_n) = \begin{cases} \frac{\theta_i}{t_n - t_{n-1}}, & p_n = 0 \\ L_{max} - \frac{\theta_i}{t_n - t_{n-1}}, & p_n = 1 \end{cases}, \tag{7}$$

where t_n and t_{n-1} are the timestamps of two adjacent intensity events, and the $p_n \in \{1, 0\}$ corresponds to the intensity estimation under bright or dark light scenario. Specially, this equation also works when the first event is a dynamic event.

For *dynamic events*, we directly collect each dynamic event into spatio-temporal point sets for machine vision. Besides, we also push each dynamic event into a stack $\{\langle t_1, p_1 \rangle, \langle t_2, p_2 \rangle, ..., \langle t_{n-1}, p_{n-1} \rangle\}$ and wait for a new intensity event to compute light intensity. The first dynamic event at the top of the stack can use the light intensity $L(\boldsymbol{u_n}, t_n)$ from the adjacent intensity event. After that, the corresponding dynamic events at middle or bottom of the stack can utilize the polarity p_n and the threshold θ_i to calculate the light intensity. If we have known $L(\boldsymbol{u_n}, t_{n-1})$ of the top of stack S, the $L(\boldsymbol{u_n}, t_{n-2})$ of the second top element can be computed as follows:

$$L(\boldsymbol{u_n}, t_{n-2}) = \begin{cases} L(\boldsymbol{u_n}, t_{n-1}) + \theta_d, & p_{n-1} = 0 \\ L(\boldsymbol{u_n}, t_{n-1}) - \theta_d, & p_{n-1} = 1 \end{cases}, \tag{8}$$

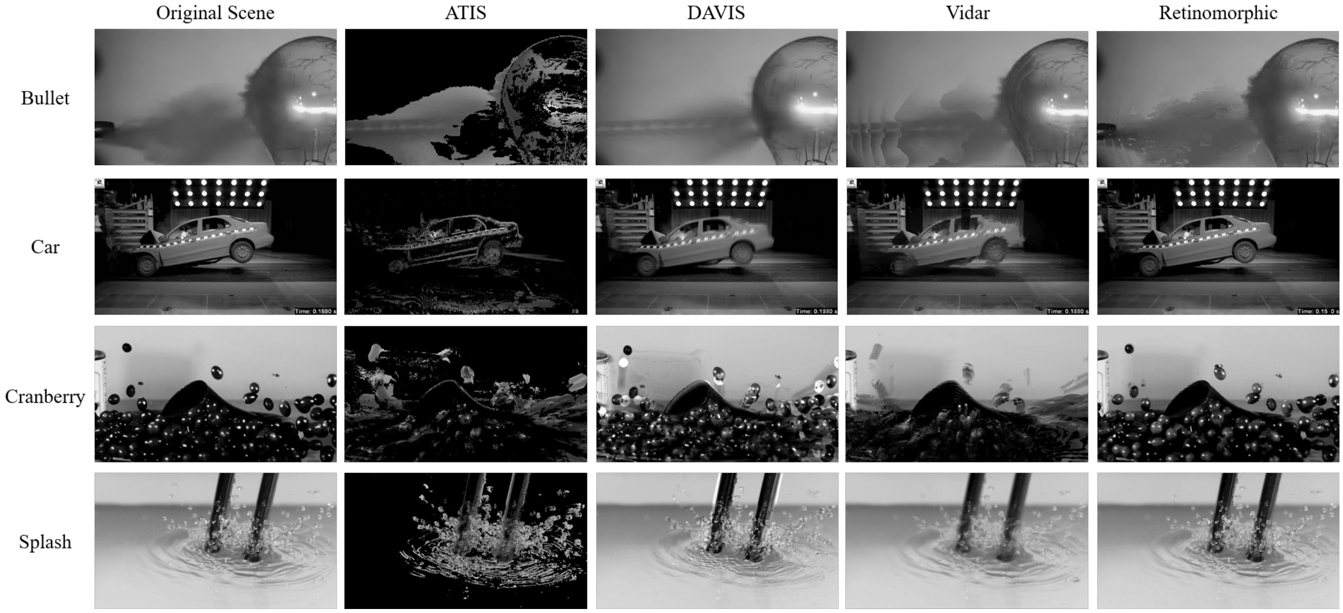


Figure 6: Representative image reconstruction results in four high-speed moving scenarios. ATIS of the second column can restore intensity information only for dynamic areas rather than global textures. DAVIS of the third column usually mismatches two streams and brings motion blur for conventional frames. Vidar of the fourth column may result in low-quality imaging in the low sampling frequency. On the contrary, our framework of the last column has better performance on reconstructing the scene and restoring detailed textures.

where $L(\boldsymbol{u_n}, t_{n-1})$ and $L(\boldsymbol{u_n}, t_{n-2})$ are the light intensities of the top and the second element of the stack, respectively.

To restore the light intensity at any time between t_i and t_n. we adopt a linear interpolation strategy as:

$$L\left(\boldsymbol{u_n}, t\right) = aL\left(\boldsymbol{u_n}, t_{i+1}\right) + (1-a)L\left(\boldsymbol{u_n}, t_i\right), \tag{9}$$

where $a = (t - t_i)/(t_{i+1} - t_i)$, and $t_i < t < t_{i+1}$. For example, we record the light intensity in the last intensity event as $L(\boldsymbol{u_n}, t')$, then we can restore the light intensity at time from t' to t_1 using the mentioned linear interpolation Eq. (9).

5 EXPERIMENTS

In this section, detailed experimental settings, representative results, parameter discussions, and test prototype on our hybrid camera system can be found as follows.

5.1 Experimental Settings

Simulated Settings. To verify the effectiveness of our retinomorphic sensing, we collect a high-speed moving dataset involving 6 typical scenarios (named bullet, car, cranberry, bird, shotgun, and splash) recorded by ultra high-speed cameras. Our collected dataset includes various illumination conditions, spatial resolutions, frame numbers, and dynamic ratios.

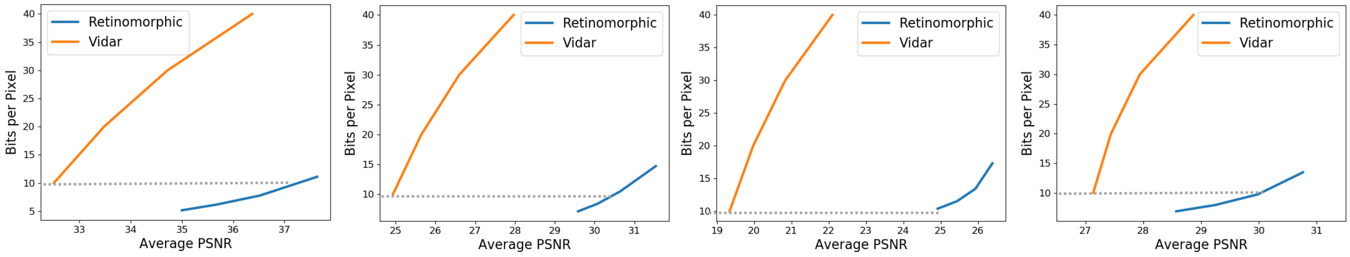
In implementation details, we implement the simulator of ATIS, DAVIS, Vidar, and our retinomorphic sensing. DVS is not implemented because DAVIS totally includes the feature of DVS. All mentioned simulators take frame series from high-speed videos as the input and output asynchronous events in the format of $\langle x_n, y_n, t_n, p_n \rangle$. Differently, Vidar simulator outputs spike arrays and DAVIS simulator outputs additional conventional frames. The two evaluation metrics [35] (i.e., PSNR and SSIM) are adopted to report the performance scores in reconstructing visual textures.

Hybrid Camera System Setting. As depicted in Fig. 4(a), we collect an event camera (i.e., DAVIS346, resolution of 346×260) and a spike camera (i.e., Vidar, resolution of 400×250). The input light is equally divided into two cameras by a beam splitter (i.e., Thorlabs CCM1-BS013) [36]. For temporal calibration, we write a synchronized script to start two cameras simultaneously. For spatial calibration, we map DVS events into event images and reconstruct visual textures from Vidar with the shared view, and we consider the homography between two views. After spatiotemporal calibration, we present some representative examples in Fig. 4(b).

5.2 Effective Test

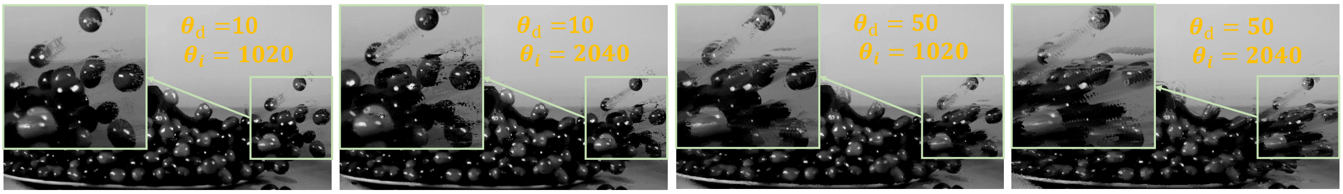
We will explore several experiments to see why and how our framework works well from two perspectives as follows.

Visualization Evaluation. Some representative visualization results on high-speed moving scenarios are illustrated in Fig. 5. Our framework first uses a unified representation (i.e., AER) to encode each video into asynchronous events, which includes dynamic and texture information. Then, our decoding framework can effectively split dynamic events and intensity events. Dynamic events are directly mapped into 2D image-like representations (i.e., event images) in the second column of Fig. 5. Visual textures are reconstructed via the inter-event interval between two events in the third column of Fig. 5. Note that, our framework not only obtains dynamic information for machine vision and reconstructs high-quality visual textures for human vision.



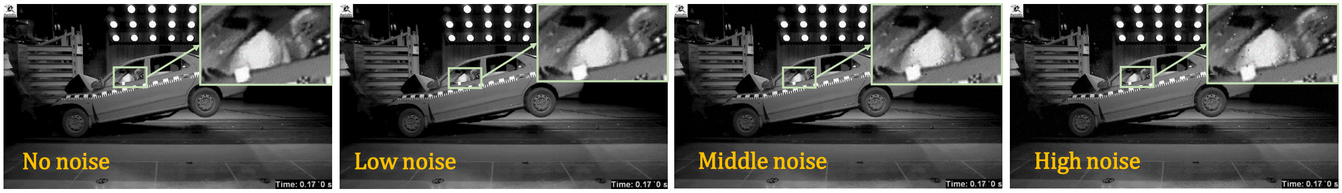
Bullet Car Cranberry Splash

Figure 7: Quantization results on average PSNR of visual reconstruction with different methods. Four sub-graphs depict the performance of our retinomorphic sensing and Vidar under four typical scenarios (i.e., bullet, car, cranberry, and splash). The horizontal axis represents the average PSNR, and the vertical axis is calculated by the bits used for each pixel per frame on average. Notably, the horizontal gray lines show that our retinomorphic sensing system has better performance in imaging quality than Vidar with the same data size.



PSNR=26.640, SSIM=0.977 PSNR=25.496, SSIM=0.975 PSNR=25.355, SSIM=0.974 PSNR=24.309, SSIM=0.972

Figure 8: Average PSNR of visual reconstruction with different thresholds. Notably, the thresholds θ_i and θ_d of fovea-like sensing and peripheral-like sensing significantly influence the quality of image reconstruction. We can see that the threshold smaller, the image reconstruction performance better.



PSNR=30.644, SSIM=0.984 PSNR=29.970, SSIM=0.982 PSNR=28.705, SSIM=0.978 PSNR=27.361, SSIM=0.971

Figure 9: Average PSNR of visual reconstruction with different noise intensities. Specifically, random noises, obeying the normal distribution, are attached to the timestamp of each event. Obviously, the PSNR slightly decreases with the creasing of the noise intensity. It indicates that our framework is robust against random noise.

Besides, we further compare our framework with the existing sampling manners using neuromorphic cameras (i.e., ATIS, DAVIS, and Vidar) in Fig. 6. It demonstrates that our approach performs better than other sampling methods especially on reconstructing high-speed objects (e.g., the details of the wheels for the crashing car). Specifically, in contrast to ATIS, our framework, using the global integrating sensing, can reconstruct the visual texture of the whole pixel array instead of only dynamic area. Meanwhile, our approach can overcome the limitations (e.g., motion blur) of conventional frames in DAVIS. Compared to Vidar, we effectively extract high-speed dynamic events for machine vision as DVS does.

Quantization Results. To conduct a quantitative evaluation of reconstruction images, we compare our framework with the best competitor (i.e., Vidar), which utilizes the integrating sampling manner for each pixel and outputs spike arrays to encode the light intensity. As shown in Fig. 7, we report the average PSNR between visual textures from two reconstruction methods and original frames from each video. Apparently, our retinomorphic sensing system consistently achieves better performance than Vidar with the same data size on each scenario, with an average increase of 5 in PSNR when the bits per pixel is set to 10. This is because that the dynamic events can be used to improve significantly the quality of visual reconstruction.

5.3 Scalability Test

Beyond the effective test, we next conduct several ablation tests to take a deep look at the impact of each parameter choice of our retinomorphic sensing framework, and more details are demonstrated as follows.

Threshold Parameters for Visual Texture. As shown in Table 2, the quality of image reconstruction is improved with the

Table 2: Quantitative evaluation on visual reconstruction with different thresholds. θ_i and θ_d are the preset threshold for fovea-like sensing and periphery-like sensing. It depicts that the smaller thresholds, the better performance of image reconstruction.

Threshold	Scene	θ_d =10		θ_d =15		θ_d =20		θ_d =25		θ_d =30		θ_d =40		θ_d =50	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
θ_i =1020	bullet	**38.125**	**0.981**	37.862	0.980	37.638	0.979	37.515	0.979	37.458	0.979	37.414	0.978	37.397	0.978
	car	**31.636**	**0.985**	31.682	0.985	31.540	0.985	31.351	0.985	30.985	0.984	30.771	0.984	30.491	0.983
	cranberry	**26.640**	**0.977**	26.528	0.977	26.387	0.977	26.206	0.976	25.871	0.976	25.536	0.975	25.355	0.974
	splash	**31.716**	**0.985**	31.198	0.984	30.760	0.983	30.442	0.982	30.121	0.981	29.944	0.981	29.722	0.980
θ_i =1530	bullet	37.291	0.979	36.826	0.977	36.520	0.976	36.346	0.975	36.179	0.974	36.025	0.974	35.991	0.973
	car	30.652	0.984	30.685	0.984	30.644	0.984	30.629	0.984	30.287	0.984	30.080	0.983	29.826	0.982
	cranberry	26.126	0.977	26.082	0.977	25.931	0.977	25.749	0.977	25.158	0.975	24.858	0.974	24.781	0.973
	splash	30.398	0.979	30.355	0.982	29.908	0.981	29.627	0.980	29.219	0.979	28.916	0.977	28.713	0.976
θ_i =2040	bullet	36.717	0.978	36.098	0.975	35.667	0.972	35.349	0.970	35.128	0.968	34.983	0.967	34.849	0.967
	car	30.066	0.982	30.153	0.983	30.073	0.983	29.955	0.983	29.592	0.982	29.326	0.981	29.150	0.981
	cranberry	25.496	0.975	25.541	0.976	25.436	0.976	25.256	0.975	24.976	0.973	24.512	0.972	24.309	0.972
	splash	27.846	0.962	29.243	0.976	29.250	0.978	29.011	0.978	28.874	0.975	28.423	0.974	28.033	0.974
θ_i =2550	bullet	36.209	0.977	35.333	0.970	35.000	0.969	34.594	0.965	34.303	0.962	34.129	0.961	33.971	0.961
	car	29.443	0.979	29.634	0.981	29.586	0.981	29.450	0.981	29.107	0.980	28.826	0.979	28.691	0.979
	cranberry	24.462	0.969	24.881	0.972	24.919	0.973	24.789	0.973	24.232	0.972	24.002	0.971	23.868	0.970
	splash	24.447	0.933	27.765	0.965	28.570	0.973	28.569	0.975	28.166	0.973	27.862	0.972	27.795	0.972

Table 3: Data size (i.e., events per pixel) with different thresholds. We can see that the thresholds θ_i and θ_d in our framework can control the data size of asynchronous events.

Threshold	Scene	θ_d =10	θ_d =15	θ_d =20	θ_d =25
θ_i =1020	bullet	0.205	0.183	0.174	0.169
	car	0.271	0.245	0.230	0.221
	cranberry	0.395	0.311	0.271	0.248
	splash	0.275	0.230	0.210	0.200
θ_i =1530	bullet	0.155	0.132	0.121	0.116
	car	0.213	0.180	0.164	0.155
	cranberry	0.338	0.252	0.210	0.187
	splash	0.220	0.173	0.152	0.141
θ_i =2040	bullet	0.131	0.106	0.096	0.090
	car	0.182	0.147	0.131	0.122
	cranberry	0.310	0.223	0.180	0.156
	splash	0.195	0.146	0.125	0.113
θ_i =2550	bullet	0.117	0.091	0.080	**0.075**
	car	0.162	0.128	0.112	**0.103**
	cranberry	0.294	0.206	0.163	**0.138**
	splash	0.179	0.130	0.108	**0.096**

decreasing of the preset parameters θ_d and θ_i in our retinomorphic sensing system. This is because the smaller threshold of θ_d brings higher temporal sampling resolution of dynamic areas, and the smaller threshold of θ_i decreases the influence of slow change in scenarios. In addition, we also present the corresponding visualization reconstruction results with different thresholds in Fig. 8. This indicates the thresholds θ_i and θ_d of fovea-like sensing and peripheral-like sensing significantly influence the quality of image reconstruction.

Threshold Parameters for Data Size. We further analysis the data size of asynchronous events with different thresholds θ_d and θ_i in Table 3. It is obvious that smaller thresholds θ_d and θ_i are easier to trigger an event, which brings larger data size in our framework. However, the data size is a very important indicator for high-speed cameras. In other words, the thresholds θ_i and θ_d can flexibly control the data size of asynchronous events to facilitate data transmission and storage.

According to Table 2 and Table 3, it is surprise for us to find that the preset thresholds θ_d and θ_i will affect the final results involving imaging quality and data size. In fact, we attempt to decrease the thresholds to improve imaging quality meanwhile brings increasing of data size. We believe that an optimized trade-off between imaging quality and data size can be useful in some scenarios that require more high-quality imaging meanwhile reducing the data redundancy. In the future, we also explore an effective optimization scheme to adaptively adjust the thresholds in our framework.

Robustness to Noise. Generally speaking, neuromorphic vision sensors are always with random noises in sampling circuits and data transmission. In this work, we further explore the effects of random noises on our retinomorphic sensing system. Specifically, random noises with three intensities, obeying the normal distribution, are attached to the timestamp of each event in our framework. As shown in Fig. 9, the reconstruction results display that the PSNR slightly decreases with the increasing of the noise intensity. In short, we can draw the conclusion that our framework is robust against random noise. Actually, the noise always appears at dynamic areas because the interval between dynamic events is usually short, and timestamp increments bring larger relative change to dynamic events than intensity events.

5.4 Test Prototype on Hybrid Camera System

To verify our visual sensing systems integrating fovea-like and peripheral-like sampling, we build a prototype hybrid camera system combining two types of neuromorphic cameras (see Fig. 4). We evaluate the effectiveness of two tasks involving image reconstruction for human vision and object detection for machine vision.

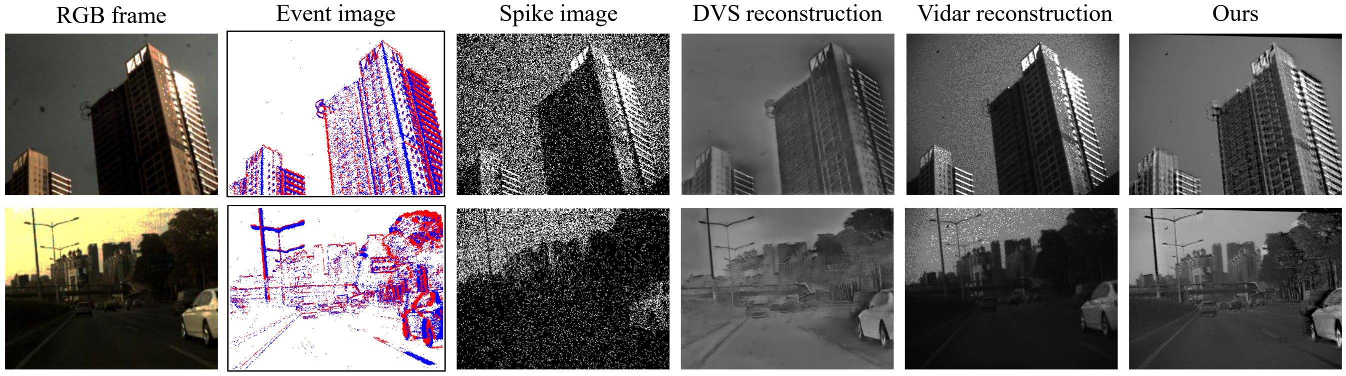


Figure 10: Representative reconstruction images using our hybrid camera system (i.e., DAVIS346 and Vidar). Note that, our strategy, combining the HDR property from Vidar and high-quality textures from Vidar, can obtain better visual results than other methods using the single-modality (e.g., DVS reconstruction utilizing the E2VID [25] and Vidar reconstruction by computing the interval between two spikes).

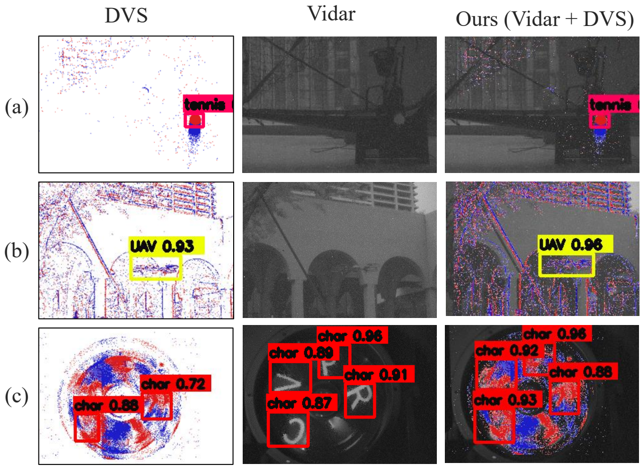


Figure 11: Representative object detection results using our hybrid camera system (i.e., DAVIS346 and Vidar). (a) high-speed moving tennis in low-light. (b) UAV in low-light. (c) High-speed rotation characters with 2600r/min. We can see that our detector, inheriting the HDR property from DVS and high-speed visual textures from Vidar, obtains better performance than single-modality in challenging scenarios (i.e., high-speed and low-light).

Image Reconstruction. Inspired by the complementary filter [27], we use our decoder to reconstruct high-speed images using DVS events and Vidar spikes. As shown in Fig. 10, our reconstruction approach, integrating Vidar and DVS, achieves better performance for human vision than other method using the single-modality (i.e., RGB frame from DAIVS346, event image by mapping DVS events, spike image via projecting Vidar spikes, Vidar reconstruction image via computing the interval between two spikes, and DVS reconstruction image utilizing the E2VID [25]). This is because that only DVS is hard to reconstruct clear textures, and Vidar fails to capture the light intensity in extreme low-light scenarios. Apparently, our hybrid camera system, as the prototype of the proposed retinomorphic sensing system, can provide a simulated platform to take advantages of the HDR property of DVS and high-speed textures of Vidar for human vision.

Object Detection. As depicted in Fig. 11, we exhibit representative results on three typical scenarios (i.e., high-speed moving tennis in low-light, UAV in low-light, high-speed rotation characters) using our newly built hybrid neuromorphic camera system. We directly map each event stream into event images and input the reconstructed Vidar images for the detector (i.e., SSD [18]) respectively, then we further adopt the post-processing fusion strategy (i.e., non-maximum suppression [1]) to combine the bounding boxes of two streams into the final results. For example, we can find that DVS, taking the advantage of its HDR, has brought a new perspective to overcome the shortage of Vidar in low-light scenarios. Meanwhile, Vidar can provide high-speed visual textures to serve for high-quality object detection in high-speed rotation characters. In other words, the joint detection, using the retinomorhic sampling manner, integrates DVS and Vidar to overcome the limitation of conventional frames in challenging scenarios (i.e., high-speed and low-light).

6 CONCLUSION

In this paper, we present a novel paradigm for future multimedia computing, namely retinomorphic sensing, which can effectively sense both dynamic information and visual textures as the retina does. To the best of our knowledge, this is the first work to integrate fovea-like and peripheral-like visual sampling mechanisms to generate asynchronous events using a unified representation (i.e., AER). The results show that our retinomorphic visual sensing system can extract dynamic events for machine vision and reconstruct visual textures from intensity events for human vision. We further build a prototype hybrid camera system to verify this strategy on tasks such as image reconstruction and object detection. We believe this work will aim at addressing the shortages of conventional cameras towards the next-generation neuromorphic vision sensor. We also believe that this novel paradigm will provide insight into future multimedia computing.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Contract 62027804, Contract 61825101 and Contract 62088102.

REFERENCES

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-NMS–improving object detection with one line of code. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 5561–5569.

[2] Christian Brandli, Raphael Berner, Minhao Yang, Shih-Chii Liu, and Tobi Delbruck. 2014. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor. *IEEE journal of Solid-State Circuits* 49, 10 (2014), 2333–2341.

[3] Vincent Chan, Shih-Chii Liu, and Andr van Schaik. 2007. AER EAR: A matched silicon cochlea pair with address event representation interface. *IEEE Transactions on Circuits and Systems I: Regular Papers* 54, 1 (2007), 48–59.

[4] Chang Wen Chen. 2020. Internet of Video Things: Next-Generation IoT With Visual Sensors. *IEEE Internet of Things Journal* 7, 8 (2020), 6676–6685.

[5] Denis Guangyin Chen, Daniel Matolin, Amine Bermak, and Christoph Posch. 2011. Pulse-modulation imaging—Review and performance analysis. *IEEE Transactions on Biomedical Circuits and Systems* 5, 1 (2011), 64–82.

[6] Eugenio Culurciello, Ralph Etienne-Cummings, and Kwabena A Boahen. 2003. A biomorphic digital image sensor. *IEEE journal of Solid-State Circuits* 38, 2 (2003), 281–294.

[7] Chao Deng, Yuanlong Zhang, Yifeng Mao, Jingtao Fan, Jinli Suo, Zhili Zhang, and Qionghai Dai. 2021. Sinusoidal Sampling Enhanced Compressive Camera for High Speed Imaging. *IEEE Transactions Pattern Analysis Machine Intelligence* 43, 4 (2021), 1380–1393.

[8] Siwei Dong, Tiejun Huang, and Yonghong Tian. 2017. Spike camera and its coding methods. In *Proceedings of the IEEE Data Compression Conference (DCC)*. 437–437.

[9] Guillermo Gallego, Tobi Delbruck, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jorg Conradt, Kostas Daniilidis, et al. 2020. Event-based Vision: A Survey. *IEEE Transactions Pattern Analysis Machine Intelligence* (2020).

[10] Qing Guo, Wei Feng, Ruijun Gao, Yang Liu, and Song Wang. 2021. Exploring the Effects of Blur and Deblurring to Visual Object Tracking. *IEEE Transactions on Image Processing* 30 (2021), 1812–1824.

[11] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. 2020. Neuromorphic camera guided high dynamic range imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1730–1739.

[12] Yuhang Hu, Delbruck Tobi, and Shih-Chii Liu. 2020. Learning to Exploit Multiple Vision Modalities by Using Grafted Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 85–101.

[13] Tae Hyun Kim, Seungjun Nah, and Kyoung Mu Lee. 2018. Dynamic Video Deblurring Using a Locally Adaptive Blur Model. *IEEE Transactions Pattern Analysis Machine Intelligence* 40, 10 (2018), 2374–2387.

[14] Jianing Li, Siwei Dong, Zhaofei Yu, Yonghong Tian, and Tiejun Huang. 2019. Event-based vision enhanced: A joint detection framework in autonomous driving. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 1396–1401.

[15] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck. 2008. A 128×128 120 dB 15μs latency asynchronous temporal contrast vision sensor. *IEEE journal of Solid-State Circuits* 43, 2 (2008), 566–576.

[16] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. 2020. Deep learning for generic object detection: A survey. *International Journal of Computer Vision* 128, 2 (2020), 261–318.

[17] Shih-Chii Liu, Bodo Rueckauer, Enea Ceolini, Adrian Huber, and Tobi Delbruck. 2019. Event-driven sensing for efficient perception: Vision and audition algorithms. *IEEE Signal Processinge Magazine* 36, 6 (2019), 29–37.

[18] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 2016. SSD: Single shot multibox detector. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 21–37.

[19] Yadan Luo, Zi Huang, Zijian Wang, Zheng Zhang, and Mahsa Baktashmotlagh. 2020. Adversarial Bipartite Graph Learning for Video Domain Adaptation. In *Proceedings of the ACM Multimedia (ACM MM)*. 19–27.

[20] Carver Mead. 2020. How we created neuromorphic engineering. *Nature Electronics* 3, 7 (2020), 434–435.

[21] Weiqing Min, Yonghong Tian, Zi Huang, Wen-Huang Cheng, and Abdulmotaleb El Saddik. 2020. Urban Multimedia Computing: Emerging Methods in Multimedia Computing for Urban Data Analysis and Applications. *Proceedings of the ACM Multimedia (ACM MM)* 27, 3 (2020), 8–11.

[22] Anton Mitrokhin, Zhiyuan Hua, Cornelia Fermuller, and Yiannis Aloimonos. 2020. Learning Visual Motion Segmentation Using Event Surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 14414–14423.

[23] Christoph Posch, Daniel Matolin, and Rainer Wohlgenannt. 2010. 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* 46, 1 (2010), 259–275.

[24] Christoph Posch, Teresa Serrano-Gotarredona, Bernabe Lins-Barranco, and Tobi Delbruck. 2014. Retinomorphic event-based vision sensors: bioinspired cameras with spiking output. *Proc. IEEE* 102, 10 (2014), 1470–1484.

[25] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. 2019. Events-to-video: Bringing modern computer vision to event cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3857–3866.

[26] Kaushik Roy, Akhilesh Jaiswal, and Priyadarshini Panda. 2019. Towards spike-based machine intelligence with neuromorphic computing. *Nature* 575, 7784 (2019), 607–617.

[27] Cedric Scheerlinck, Nick Barnes, and Robert Mahony. 2018. Continuous-time intensity estimation using event cameras. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. Springer, 308–324.

[28] Xiao Shu and Xiaolin Wu. 2018. Real-Time High-Fidelity Compression for Extremely High Frame Rate Video Cameras. *IEEE Transactions on Computational Imaging* 4, 1 (2018), 172–180.

[29] Raunak Sinha, Mrinalini Hoon, Jacob Baudin, Haruhisa Okawa, Rachel OL Wong, and Fred Rieke. 2017. Cellular and circuit mechanisms shaping the perceptual properties of the primate fovea. *Cell* 168, 3 (2017), 413–426.

[30] Emma EM Stewart, Matteo Valsecchi, and Alexander C Schütz. 2020. A review of interactions between peripheral and foveal vision. *Journal of vision* 20, 12 (2020), 1–35.

[31] Dmitri Strukov, Giacomo Indiveri, Julie Grollier, and Stefano Fusi. 2019. Building brain-inspired computing. *Nature Communications* 10, 1 (2019), 1–6.

[32] Keqiang Sun, Wayne Wu, Tinghao Liu, Shuo Yang, Quan Wang, Qiang Zhou, Zuochang Ye, and Chen Qian. 2019. FAB: A Robust Facial Landmark Detection Framework for Motion-Blurred Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 5462–5471.

[33] Panqu Wang and Garrison W Cottrell. 2017. Central and peripheral vision for scene recognition: A neurocomputational modeling exploration. *Journal of vision* 17, 4 (2017), 1–22.

[34] Weiying Wang, Jieting Chen, and Qin Jin. 2020. VideoIC: A Video Interactive Comments Dataset and Multimodal Multitask Learning for Comments Generation. In *Proceedings of the ACM Multimedia (ACM MM)*. 2599–2607.

[35] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions Image Processing* 13, 4 (2004), 600–612.

[36] Zihao W Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. 2020. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1609–1619.

[37] Lifang Wu, Zhou Yang, Jiaoyu He, Meng Jian, Yaowen Xu, Dezhong Xu, and Chang Wen Chen. 2019. Ontology-based global and collective motion patterns for event classification in basketball videos. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 7 (2019), 2178–2190.

[38] Jiangtao Xu, Liang Xu, Zhiyuan Gao, Peng Lin, and Kaiming Nie. 2020. A Denoising Method Based on Pulse Interval Compensation for High-Speed Spike-Based Image Sensor. *IEEE Transactions on Circuits and Systems for Video Technology* (2020).

[39] Michitaka Yoshida, Toshiki Sonoda, Hajime Nagahara, Kenta Endo, Yukinobu Sugiyama, and Rin-Ichiro Taniguchi. 2020. High-Speed Imaging Using CMOS Image Sensor With Quasi Pixel-Wise Exposure. *IEEE Transactions on Computational Imaging* 6 (2020), 463–476.

[40] Chu Zhou, Hang Zhao, Jin Han, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. 2020. UnModNet: Learning to Unwrap a Modulo Image for High Dynamic Range Imaging. In *Proceedings of the Advances in Neural Information Processing Systems (NeuIPS)*. 1–12.

[41] Lin Zhu, Siwei Dong, Jianing Li, Tiejun Huang, and Yonghong Tian. 2020. Retina-Like Visual Image Reconstruction via Spiking Neural Model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1438–1446.